19th Computer Vision Winter Workshop
Zuzana Kúkelová and Jan Heller (eds.)
Křtiny, Czech Republic, February 3–5, 2014

The VOT2013 challenge: overview and additional results

M. Kristan[1], R. Pflugfelder[2], A. Leonardis[3], J. Matas[4], F. Porikli[5], L. Čehovin[1], G. Nebehay[2], G. Fernandez[2], and T. Vojir[4]

[1] University of Ljubljana, Slovenia [2] Austrian Institute of Technology, Austria
[3] University of Birmingham, United Kingdom [4] Czech Technical University in Prague, Czech Republic
[5] NICTA and Australian National University, Australia

Abstract *Visual tracking has attracted a significant attention in the last few decades. The recent surge in the number of publications on tracking-related problems have made it almost impossible to follow the developments in the field. One of the reasons is that there is a lack of commonly accepted annotated data-sets and standardized evaluation protocols that would allow objective comparison of different tracking methods. To address this issue, the Visual Object Tracking (VOT) challenge and workshop was organized in conjunction with ICCV2013. Researchers from academia as well as industry were invited to participate in the first VOT2013 challenge which aimed at single-object visual trackers that do not apply pre-learned models of object appearance (model-free). In this paper we provide an overview of the VOT2013 challenge, point out its main results and document the additional previously unpublished experiments and results.*

1 Introduction

Visual tracking is a rapidly evolving field of computer vision that has been increasingly attracting attention of the vision community. One reason is that it offers many challenges as a scientific problem. Second, it is a part of many higher-level problems of computer vision, such as motion analysis, event detection and activity understanding. Furthermore, the steady advance of technology in terms of computational power, form factor and price, opens vast application potential for tracking algorithms. Applications include surveillance systems, transport, sports analytics, medical imaging, mobile robotics, film post-production and human-computer interfaces.

Single-object trackers that do not apply pre-learned models of object appearance (model-free) are of particular interest due to their large application domain. The activity in the field is reflected by the abundance of new tracking algorithms presented and evaluated in journals and at conferences, and summarized in the many survey papers, e.g., [13, 29, 11, 17, 30, 43, 26]. Despite the efforts invested in proposing new trackers, these have not been accompanied with established evaluation methodology.

One of the most influential performance analysis efforts for object tracking is PETS (Performance Evaluation of Tracking and Surveillance) [44]. The first PETS workshop that took place in 2000, aimed at evaluation of visual tracking algorithms for surveillance applications. However, its focus gradually shifted to high-level event interpretation algorithms. Other frameworks and datasets have been presented since, but these focussed on evaluation of surveillance systems and event detection, e.g., CAVIAR[1], i-LIDS [2], ETISEO[3], change detection [15], sports analytics (e.g., CVBASE[4]), or specialized on tracking of specific objects like faces, e.g. FERET [33] and [20]. In general, the evaluation of new tracking algorithms, and their comparison to the state-of-the-art, depends on three essential components: (1) an evaluation system, (2) a dataset, (3) performance evaluation measures.

The Evaluation system. For objective and rigorous evaluation, an evaluation system that performs the same experiment on different trackers using the same dataset is required. Ideally, the system should support multiple OS and programming languages and allow easy integration of new trackers. Furthermore, a certain level of interaction with the tracker is desirable, for instance to allow for detection of tracking failures. Currently, the most notable and general systems are the ODViS [18], VIVID [4] and ViPER [8] toolkits. These, however, do not allow for interaction with the tracker. Recently, Wu et al. [41] have performed a large-scale benchmark of several trackers and developed an evaluation kit that allows integration of third-party trackers as well. However, in our experience, the integration is not straightforward due to a lack of standardization of the input/output communication between the tracker and the evaluation kit.

Dataset A trend has emerged in the single-object model-free tracking community to test newly proposed trackers on larger datasets that include different real-life visual phenomena like occlusion, clutter and illumination change. As a consequence, various authors nowadays compare their trackers on many publicly-available sequences, of which some have became a de-facto standard in evaluation of new

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1
[2] http://www.homeoffice.gov.uk/science-research/hosdb/i-lids
[3] http://www-sop.inria.fr/orion/ETISEO
[4] http://vision.fe.uni-lj.si/cvbase06/

trackers. However, many of these sequences lack a standard ground truth labeling, which makes comparison of proposed algorithms difficult. Furthermore, authors usually do not use datasets with various visual phenomena equally represented. In fact, many popular sequences exhibit the same visual phenomenon, which makes the results biased toward some particular types of the phenomena. To address this issue, Wu et al. [41] annotated each sequence with several visual attributes and report tracker performance with respect to each attribute separately. For example, a sequence is annotated as "occlusion" if the target is occluded anywhere in the sequence, etc. However, visual phenomena like occlusion do not usually last throughout the entire sequence. For example, an occlusion might occur at the end of the sequence, while a tracker might fail due to some other effects occurring at the beginning of the sequence. In this case, the failure would be falsely attributed to occlusion. Thus a per-frame dataset labeling is required to facilitate a more precise analysis.

Performance measures. A wealth of performance measures have been proposed for single-object tracker evaluation. These range from basic measures like center error [34], region overlap [25], tracking length [24] and failure rate [22, 21] to more sophisticated measures, such as CoTPS [31], which combine several measures into a single measure. A nice property of the combined measures is that they provide a single score to rank the trackers. A downside is that they offer little insight into the tracker performance. In this respect the basic measures, or their simple derivatives, are preferred as they usually offer a straight-forward interpretation. While some authors choose several basic measures to compare their trackers, the recent study [37] has shown that many measures are correlated and do not reflect different aspects of tracking performance. In this respect, choosing a large number of measures may in fact again bias results toward some particular aspects of tracking performance.

VOT2013. In order to address the above stated issues, the Visual Object Tracking (VOT2013) challenge was organized. Its aim was to provide an evaluation platform that goes beyond the current state-of-the-art. In particular, the authors of the challenge have compiled a labeled dataset collected from widely used sequences showing a balanced set of various objects and scenes. All the sequences are labeled per-frame with different visual attributes to aid a less biased analysis of the tracking results. An evaluation kit[5] was developed in *Matlab/Octave* that automatically performs experiments on a tracker using the provided dataset. A new tracker performance comparison protocol based on basic performance measures was also proposed. A significant novelty of the proposed evaluation protocol was that it explicitly addresses the statistical significance of the results and addresses the equivalence of trackers. A dedicated VOT2013 homepage[6] has been set up, from which the dataset, the evaluation kit and the results are publicly available. The authors of tracking algorithms have an opportunity to publish their source code at the VOT homepage as well, thus pushing the field of visual tracking towards reproducible research. The results of the challenge have been presented at the VOT2013 workshop in conjuction with the ICCV2013 and documented in the supporting paper [23]. In this paper we provide an overview of the VOT2013 challenge with a particular focus on the evaluation methodology and provide additional results that have not been published in [23].

[5] https://github.com/vicoslab/vot-toolkit
[6] http://www.votchallenge.net/

2 Summary of the tracking experiments

The VOT2013 benchmark is designed for single-object, single-camera, short-term causal trackers. The tracker is initialized at the beginning of a sequence using the ground truth bounding box and is required to predict a single bounding box of the target for each frame of the sequence. Causality requires the tracker to solely process the frames from the initialization up to the current frame without using any information from the future frames. Since we are evaluating short-term tracking, whenever the tracker fails, a complete reinitialization is performed so that any previously learned information (such as appearance and dynamics) is discarded. The challenge consists of three experiments:

- Baseline: Ground truth bounding boxes are used for initialization.
- Noise: Randomly perturbed bounding boxes are used for initialization, where the perturbation is in order of ten percent of the ground truth bounding box size.
- Grayscale: Color information is removed from the sequences.

The evaluation kit runs each tracker 15 times on each experiment to obtain a reliable estimate of the performance.

3 The dataset

We collected a large pool of sequences that have been used by various authors in the tracking community. Many sequences may be visually similar and would not contribute to diversification of the dataset, while significantly prolong the execution of the experiments. We have therefore reduced the set to 16 sequences, while keeping the dataset rich in visual phenomena. We represented each sequence in the pool of sequences as a 6-dimensional vector of global visual features and clustered them into 16 clusters by affinity propagation [10]. From each cluster a single sequence was manually selected.

The six global visual features were defined as follows: The *illumination change* as the maximal difference in average intensities computed from the bounding boxes; the *object size change* as the average of sequential differences in the ground-truth bounding box size; the *object motion* as the average of changes in bounding box center over the frames; *clutter* as the histogram difference within and outside the ground truth bounding box; *camera motion* as the per-pixel average difference outside the bounding box; *blur* was measured by a camera focus measure.

Since the bounding boxes were annotated by various authors, there was no common guideline for the process of manually annotating the sequences. It seemed that most authors followed the strategy of maintaining a high fore-

ground/background ratio within the bounding box (at least $> 60\%$). In most cases, this ratio is quite high since the upright bounding box tightly fits the target. But in some cases, (e.g., the *gymnastics* sequence) where an elongated target is rotating significantly, the bounding box contains a large portion of the background at some frames as well. After inspecting all the bounding box annotations, we have re-annotated those sequences in which the original annotations were out of place.

Additionally, we manually or semi-manually labeled each frame in each selected sequence with five visual attributes that reflect a particular challenge in appearance degradation: occlusion, illumination change, motion change, size change and camera motion. In case a particular frame did not correspond to any of the five degradations, we denoted it as non-degraded.

4 Evaluation methodology

There exists an abundance of performance measures in the field of visual tracking (e.g., [40, 32, 15, 20, 41]). Our approach to choosing the performance measures was the interpretability of the measures while selecting as few measures as possible to provide a clear comparison among trackers. Based on the recent analysis of widely-used performance measures [37] we have chosen two weakly-correlated measures: (i) accuracy and (ii) robustness.

The accuracy measures how well the bounding box predicted by the tracker overlaps with the ground truth bounding box. The tracking accuracy at time-step t is defined as the overlap between the tracker predicted bounding box A_t^T and the ground truth bounding box A_t^G

$$\phi_t = \frac{A_t^G \cap A_t^T}{A_t^G \cup A_t^T}. \tag{1}$$

On the other hand, the robustness was measured by the failure rate measure, which counts the number of times the tracker drifted from the target and had to be reinitialized. A failure is indicated when the overlap measure (Eq. 1) drops to zero.

The reinitialization of trackers might introduce a bias into the performance measures. Typically, if a tracker fails at a particular frame it will likely fail again immediately after re-initialization. To reduce this bias, we re-initialize the tracker five frames after the failure and denote the skipped frames as invalid for accuracy computation. This number was determined experimentally on a separate dataset. A similar bias occurs in the accuracy measure, as the overlap measure in the frames right after the initialization are biased towards higher values for several frames (burn-in period, Figure 1). In a preliminary study we have determined by a large-scale experiment that the burn-in period is approximately ten frames. The burn-in frames are also labeled invalid and are not used in the computation of accuracy.

Let $\Phi_t(i,k)$ denote the accuracy of i-th tracker at frame t at experiment repetition k. The per frame accuracy is obtained by taking the average over these, i.e., $\Phi_t(i) = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} \Phi_t(i,k)$. The average accuracy of the i-th tracker, $\rho_{\text{A}}(i)$, over some set of N_{valid} valid frames is then

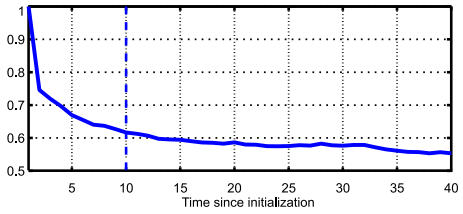

Figure 1: Overlaps after reinitialization averaged over a large number of trackers and many reinitializations.

calculated as the average of per-frame accuracies

$$\rho_{\text{A}}(i) = \frac{1}{N_{\text{valid}}} \sum_{j=1}^{N_{\text{valid}}} \Phi_j(i). \tag{2}$$

In contrast to accuracy measurements, we obtain a single measure of robustness per experiment repetition. Let $F(i,k)$ be the number of times the i-th tracker failed in the experiment repetition k over a set of frames. The average robustness of the i-th tracker is then

$$\rho_R(i) = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} F(i,k). \tag{3}$$

Note that in the dataset some attributes are more frequently presented than the others, which would introduce a bias into the results. To address this, we calculate the accuracy (2) and robustness (3) separately for each attribute. For a particular attribute we calculate the two measures only on the subset of frames in the dataset that contain that attribute (attribute subset). To compare different trackers one might average the accuracy and robustness over all the attribute subset frames. However, these will likely be at a different scale across the attribute sequences in which case direct averaging of performance measures is not appropriate. Instead, we have developed a ranking-based methodology akin to [6, 9, 15]. We start by ranking all the trackers with respect to each measure on each attribute subset separately. Let $r(i,a,m)$ be the rank of the i-th tracker on the attribute subset a using the performance measure m, which can either be accuracy (A) or robustness (R). Now we can calculate the average rank for the i-th tracker by averaging over the attributes $r(i,m) = \frac{1}{N_{\text{att}}} \sum_{a=1}^{N_{\text{att}}} r(i,a,m)$. Giving an equal weight to each performance measure, we average the two corresponding rankings as

$$r(i) = \frac{1}{2} \sum_{m \in \{A,R\}} r(i,m). \tag{4}$$

The averaging over attribute subsets assures that every attribute contributes equally to the final ranking. Since the frequency of the attributes is uneven and some frames contain several attributes, it means that some frames contribute more than the other to the final rank. This is a subtlety that might not be immediately apparent, but has to be kept in mind when interpreting the results.

A group of trackers may perform equally well on a given attribute subset, in which case they should be assigned an equal rank. In particular, after ranking trackers on an attribute set, we calculate for each i-th tracker its corrected

rank as follows. We determine for each tracker, indexed by i, a group of equivalent trackers, which contains the i-th tracker as well as any tracker that performed equally well as the selected tracker. The corrected rank of the i-th tracker is then calculated as the average of the ranks in the group of equivalent trackers. Note that this equality is not transitive. For example, consider trackers T_1, T_2 and T_3. It may happen that a tracker T_2 performs equally well as T_1 and T_3, but this does not necessarily mean that T_1 performs equally well as both, T_2 and T_3 – the equivalence groups should be established for each tracker separately.

To determine for each tracker the group of equivalent trackers, we require an objective measure of equivalence on a given sequence. In case of accuracy measure, a per-frame accuracy is available for each tracker. One way to gauge equivalence in this case is to apply a paired test to determine whether the difference in accuracies is statistically significant. In case the differences are distributed normally, the Student's t-test, which is often used in the aeronautic tracking research [3], is the appropriate choice. However, in a preliminary study we have applied Anderson-Darling tests of normality [1] and have observed that the accuracies in frames are not always distributed normally, which might render the t-test inappropriate. As an alternative, we apply the Wilcoxon Signed-Rank test as in [6]. In case of robustness, we obtain several measurements of the number of times the tracker failed over the entire sequence in different runs. However, these cannot be paired, and we use the Wilcoxon Rank-Sum (also known as Mann-Whitney U-test) [6] instead to test the difference in the average number of failures.

When establishing equivalence, we have to keep in mind that statistical significance of performance differences does not directly imply a practical difference [7]. One would have to define a maximal difference in performance of two trackers at which both trackers are said to perform practically equally well. By varying the practical difference from zero to 0.1 we have not observed significant changes in ranking. However, since we could not find clear means to objectively define this difference, we reserve our methodology only to testing the statistical significance of the differences.

5 Result analysis

In the following section we overview the results of the VOT2013 challenge. We briefly overview the results from [23] and focus on results that were not yet published.

5.1 Submitted trackers

In total 27 trackers were evaluated for the challenge, 19 original submissions and 8 baseline well-known trackers that were contributed by the VOT committee. In interest of space, we cite [23] for all trackers submitted and/or presented at VOT2013. We also refer the reader to the appendix of the VOT supporting paper [23] for short descriptions. The set of trackers was very diverse, ranging from trackers that use background-subtraction (MORP [23], STMT [23]), are based on optical-flow or motion cues (FoT [39], TLD [19], SwATrack [27]), key-points (SCTT [23], Matrioska [28]), use complex generative (IVT [34], MS [5], CCMS [23], DFT [35], ORIA [42], EDFT [23], AIF [23], CACTuS-FL [12], PJS-S [45], SwATrack) or discriminative (MIL[2], STRUCK [16], PLT [23], CT [46], RDET [23], ASAM [23], GSDT [23]) models to trackers that use geometrical constellation of parts (HT [14], LGT [38], LGT++ [23], LT-FLO [23]).

5.2 Conclusions for the challenge experiments

We overview only the major conclusions for experiments 1, 2 and 3 and refer the reader to [23] for further details. For reference, we show the accuracy-robustness rank plots in Figure 2. Averaging ranks across all three basic experiments, the top performing trackers were PLT, FoT, EDFT, LGT++ and LT-FLO. The top performing PLT is a single-scale, detection-based tracker that applies online structural SVM on color and grayscale pixels and grayscale derivatives. In all experiments the PLT achieved best robustness, however in the Baseline and Noise experiment, the part-based LGT++ and the original LGT tightly followed. The three trackers (PLT, LGT++ and LGT) perform well even in noisy initializations, but in accuracy, the top performing tracker on average was FoT, followed by SCTT and a RANSAC-based edge tracker LT-FLO. We have noticed that the top ranked trackers in the averaged ranks remain at the top also with respect to each attribute separately, with two exceptions. When considering the size change, the best robustness is still achieved by PLT, however, the trackers that yield best trade-off between the robustness and accuracy are the LGT++ and the size-adaptive mean shift tracker CCMS. When considering occlusion, the PLT and STRUCK seem to share the first place in the best trade-off. Note that the evaluation kit was also measuring the tracker speed during the experiments. Since the trackers were coded in different programming languages and run on different machines, the measurements are not directly comparable. However, two trackers stood out in performance. These were the PLT and FoT, whose speed was higher than 150fps.

We ranked the individual types of visual degradation according to the tracking difficulty they present to the tested trackers. Our results imply that the subsequences that do not contain any specific degradation present little difficulty for the trackers in general. Most trackers do not fail on such intervals and achieve best average overlap. On the other hand, camera motion is the hardest degradation in this respect. One way to explain this is that most trackers focus primarily on appearance changes of the target and do not explicitly account for changing background. Note that camera motion does not necessarily imply that the object is significantly changing position in the image frame. For accuracy, the hardest degradation is the change of object size. This is plausible as many trackers do not adapt in this respect and sacrifice their accuracy for a more stable visual model that is more accurate in situations where the size of the target does not change.

In summary, the sparse discriminative tracker PLT seems to address the robustness quite well, despite that it does not adapt the target size, which reduces its accuracy when the size of the tracked object is significantly changing. On the other hand, the part-based trackers with a rigid part constel-

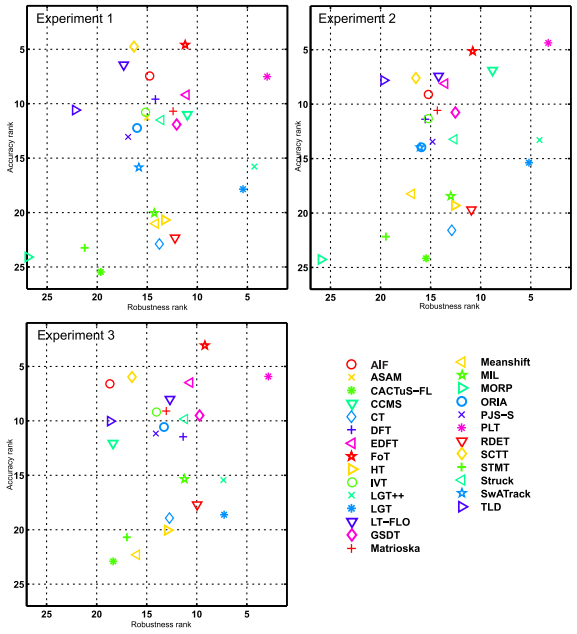

Figure 2: The accuracy-robustness ranking plots with respect to the three experiments. An optimal tracker would reside in the top-right corner of the plot.

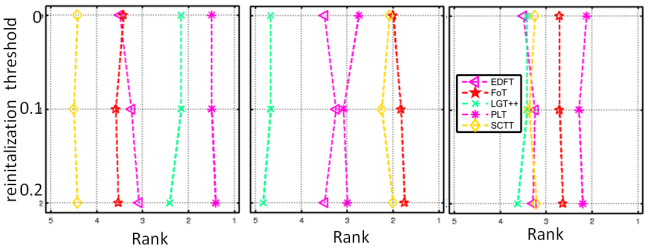

Figure 3: Effects of failure threshold on ranking.

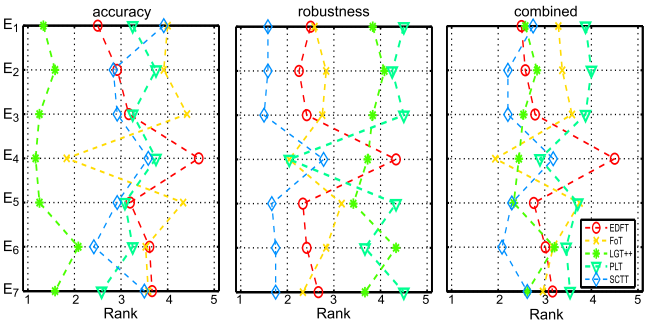

Figure 4: Results of top performing trackers on Baseline (E_1), Noise (E_2), Gray (E_3), Empty frames (E_4), Frame skipping (E_5), Frame Resize (E_6) and Reverse order (E_7) experiments.

lation yield a better accuracy at reduced robustness. The robustness is increased with part-based models that relax the constellation, but this on average comes at a cost of significant drop in accuracy.

5.3 Additional experiments

In addition to the official challenge experiments, the VOT committee has also performed four additional experiments on the top five submitted entries that had been attached a working executable version of the tracker. The aim of the first experiment was to evaluate the effect of the failure threshold on the overall ranking outcome. The remaining four experiments were designed to offer further insights into the tracker performance.

5.3.1 Effects of the failure threshold Recall that the evaluation kit proclaimed a failure if the overlap between the predicted and ground-truth bounding box became zero. To study how increasing the threshold affects the ranking of the trackers, we have repeated the baseline experiment with thresholds 0.1 and 0.2. The results are shown in Figure 3. We have observed that the failure rate increased with the threshold, however, the increase is approximately the same for all five trackers. From Figure 3 we see that the two top performing trackers do not change rank, but there is a slight change in ranking of the last three trackers. This change is due to ranking change in the accuracy rankings, since the practical difference in accuracy is in fact small for these trackers. We can conclude that the applied ranking scheme is sufficiently stable across reasonable values of failure thresholds.

5.3.2 Sequence degradation We have considered four diverse challenging scenarios of sequence degradation:

- **Empty frames**: The adaptability of the employed visual models is tested by replacing every fifth frame in the sequence by a black image.
- **Skipping frames**: To simulate frame-drops that can occur in video transmission, the original sequences were modified by removing every third frame from the sequence.
- **Reduction of image size**: To study how the size of the target affects the tracking, the size of the images is reduced by 60%.
- **Reversed sequence**: To test the importance of the specific sequence of events in the sequence, the order of the frames is reversed.

The overall results for the four additional results above are shown in Figure 4. In all but one experiment the ranking results do not change a lot, meaning that the trackers are equally well adapted to these degradations. In the experiment with black frames, the performance of the FoT and PLT significantly decreased relative to other trackers, while the performance of EDFT relatively increased. Note that the absolute performance decreased for all trackers, but this reduction was greater for FoT and PLT than it was for EDFT. The significant jump in ranking for the FoT can be explained by the way this tracker adapts its visual model. In particular, the FoT performs full adaptation in each frame. Once a black frame occurs the visual model becomes completely corrupted, which leads to failure. In case of PLT the decrease is most likely a result of fixed color model that is initialized at the first frame and is used to determine regions that most likely belong to the object. Once a black frame arrives, the discriminative power of model is rendered useless, which, may lead to unrepairable false adaptations of the visual model. EDFT on the other hand is better suited for this

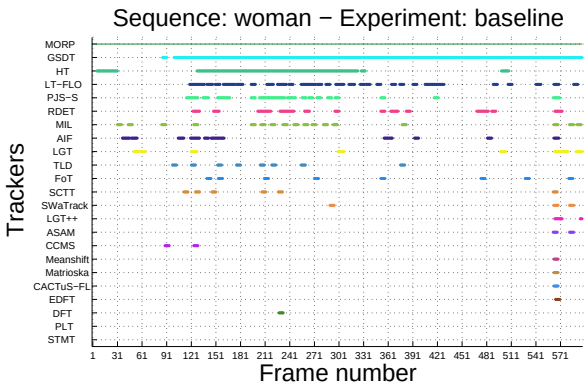

Figure 5: Scatter plot for the *woman* sequence shows the failures for each tracker w.r.t. frame number.

kind of changes, likely because of lazy adaptation of the visual model and a well designed motion model, which help it to survive short-term image degradations.

5.4 Sequence analysis

The second part of our analysis focused on the selected sequences. In particular, we have first analyzed the difficulty of each sequence presented to the trackers. Namely, for each sequence we have recorded if a particular tracker fails at least once at a particular frame. Using this approach we have constructed for each sequence a scatterplot of failures over each tracker (see the example in Figure 5). We visualize the level of difficulty for each sequence by summing the scatterplots vertically. This yields the failure curve (e.g., Figure 6a,b,c,d) which shows how many trackers failed at each frame. From these curves we derive two objective measures of sequence difficulty: *area* and *max*. The *area* is just the sum of frame-wise values from the failure curve normalized by the number of frames, while the *max* is the maximum on this curve, i.e. the maximal number of failed trackers in a frame. While the *area* indicates the average level of difficulty of a sequence, the *max* is localized and focuses on a particular frame that presented most difficult part of the sequence. For example, the *area* in the failure curve for the David sequence (Figure 6a) is smaller than the *area* for the *woman* sequence (Figure 6b), which suggests that *david* sequence is less challenging that the *woman* sequence. Furthermore, a significant peak in the *woman* sequence (frame 565) suggests that this sequence contains a subsequence which is challenging to most of the trackers. Table 1 summarizes the *area* and *max* values for all sequences.

Given the measures of *area*, we intuitively clustered the remaining 15 sequences by hand into three clusters according to their level of difficulty: Hard, Intermediate, Easy. The results are summarized in Table 1. *david*, *face*, *bicycle*, *juice*, *jump*, *car* and *cup* sequences do not present a significant challenge to most of the trackers; on average, less than a single tracker fails. Surprisingly the *david* sequence (Figure 6a) shows a small *area* in this study, although the sequence is usually considered in literature to be challenging. The reason might be that this sequence is so frequently used for tracker evaluation that the authors might have over-

sequence	area	max	frame	difficulty
bolt	4.28	13	242	hard
diving	4.23	9	105	hard
hand	4.22	14	51	hard
gymnastics	3.13	12	98	interm.
woman	2.86	15	565	interm.
sunshade	2.79	11	85	interm.
torus	2.67	8	189	interm.
iceskater	2.38	6	227	interm.
singer	1.68	4	268	interm./easy
david	1.36	4	337	easy
face	1.22	3	140	easy
bicycle	1.22	11	178	easy
juice	1.12	4	242	easy
jump	0.93	4	203	easy
car	0.92	5	253	easy
cup	0.22	2	232	easy

Table 1: The sequence analysis results. The area under the failure curve (area), the maximal number of simultaneously failed trackers (max), the frame number with maximum number of failures (frame), and the difficulty label (difficulty).

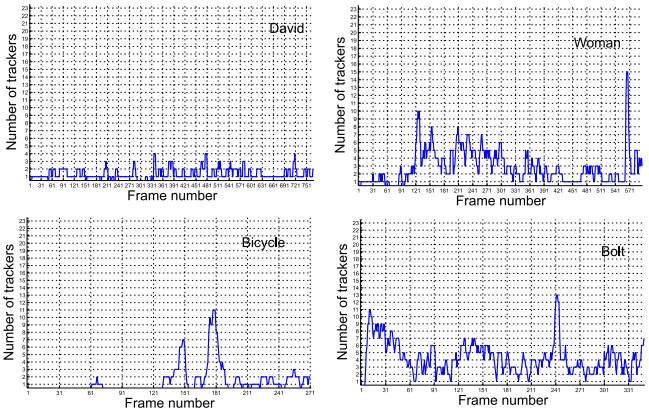

Figure 6: Failure curves for *david*, *woman*, *bicycle* and *bolt* sequences.

fitted to this sequence. The analysis also shows that the *bolt*, *diving* and *hand* sequences are the most challenging ones, followed by sequences of intermediate difficulty, in particular, the *gymnastics*, *woman*, *sunshade*, *torus*, and *iceskater* sequences and the *singer* sequence which seems to be easy-to-intermediate.

Most of the difficulties in hard and intermediate sequences arise from changes in camera and object motion as well as from rapid changes in object size. For example, *bolt* is hard, as all three aforementioned nuisances occur simultaneously in the sequence. The *diving* sequence shows significant changes in object size while the *hand* sequence shows challenging pose variations of the person's hand.

Easy to intermediate sequences might remain valuable for tracker comparison as these sequences still conceal challenges in particular frames. We can identify those sequences by considering *max* in Table 1. For example, the *woman* sequence at frame 565 (Figure 6b) shows camera zooming which lets 15 out of 23 trackers fail. Similarily, the *bicycle* sequence at frame 178 (Figure 6c) shows a peak in the failure curve. Here, an object occlusion happens and immediately after that, the tracked object is partially covered by a

shadow. A large peak is also present in the challenging Bolt sequence (Figure 6d) at frame 242. Almost half of the trackers fail here. A closer look at the frame and its neighboring frames shows significant object motion between frames as a cause of failures.

6 Conclusions and Future Work

This paper reviewed the VOT2013 challenge and its results. The challenge provides an evaluation kit comprising an evaluation protocol and dataset of 16 sequences which allows simple and objective tracker comparison. VOT2013 also provides attributes such as illumination change, occlusion, etc. on frame level for each sequence. First results show that trackers tend to specialize either for robustness or accuracy. None of the trackers consistently outperformed the others by all measures at all sequence attributes. It is currently impossible to conclusively say what kind of tracker design works best in general, however, there is some evidence showing that accuracy tends to be better for the trackers that do not apply global models, but rather split their visual models into parts. On the other hand, robustness is pretty much achieved by discriminative learning where variants of Structured SVM, e.g. PLT, seems very promising. The analysis of our dataset showed that some sequences are challenging on average, other sequences are very challenging at particular frames and some of them were well tackled by all the trackers. While we believe that it is difficult to overfit a tracker to a visually diverse dataset, tuning parameters may very likely contribute to higher ranks. Because of this unavoidable dependence on implementation and parameter tuning, care has to be taken when deciding for or against a new tracker based on performance scores. Rather than waging decision on absolute scores, comparative evaluation should be used to position trackers against baseline implementations and further focus on detailed analysis per visual properties. Our future work will focus on revising and carefully enriching the dataset with new sequences, e.g. including sequences from related datasets like the recent [36] with the aim of significantly increasing diversity while keeping the number of sequences on a useful level. We also intend to improve the evaluation kit, allowing faster execution of more complex experiments. Our work will focus on organizing further VOT challenges and pushing towards a standard for tracker comparison.

Acknowledgement

This work was supported in part by the following research programs and projects: Slovenian research agency research programs P2-0214, P2-0094, Slovenian research agency projects J2-4284, J2-3607, J2-2221 and the EPiCS project from European Union seventh framework programme under grant agreement no 257906. Jiri Matas and Tomas Vojir were supported by CTU Project SGS13/142/OHK3/2T/13 and by the Technology Agency of the Czech Republic project TE01020415 (V3C – Visual Computing Competence Center)

References

[1] T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212, 1952.

[2] B. Babenko, M. H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1619–1632, 2011.

[3] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan. *Estimation with Applications to Tracking and Navigation*, chapter 11, pages 438–440. John Wiley & Sons, Inc., 2001.

[4] R. Collins, X. Zhou, and S. K. Teh. An open source tracking testbed and evaluation web site. In *Perf. Eval. Track. and Surveillance*, 2005.

[5] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(5):564–577, 2003.

[6] J. Demšar. Statistical comparisons of classifiers over multiple datasets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[7] J. Demšar. On the appropriateness of statistical tests in machine learning. In *Workshop on Evaluation Methods for Machine Learning in conjunction with ICML*, 2008.

[8] D. Doermann and D. Mihalcik. Tools and techniques for video performance evaluation. In *Proc. Int. Conf. Pattern Recognition*, page 167170, 2000.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010.

[10] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972976, 2007.

[11] P. Gabriel, J. Verly, J. Piater, and A. Genon. The state of the art in multiple object tracking under occlusion in video sequences. In *Proc. Advanced Concepts for Intelligent Vision Systems*, page 166173, 2003.

[12] A. Gatt, S. Wong, and D. Kearney. Combining online feature selection with adaptive shape estimation. In *Image and Vision Computing New Zealand (IVCNZ), 2010 25th International Conference of*, pages 1–8. IEEE, 2010.

[13] D. M. Gavrila. The visual analysis of human movement: A survey. *Comp. Vis. Image Understanding*, 73(1):82–98, 1999.

[14] M. Godec, P. M. Roth, and H. Bischof. Hough-based tracking of non-rigid objects. *Comp. Vis. Image Understanding*, 117(10):1245–1256, 2013.

[15] N. Goyette, P. M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar. Changedetection.net: A new change detection benchmark dataset. In *CVPR Workshops*, pages 1–8. IEEE, 2012.

[16] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc J. Van Gool, editors, *Int. Conf. Computer Vision*, pages 263–270. IEEE, 2011.

[17] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey

on visual surveillance of object motion and behaviors. *IEEE Trans. Systems, Man and Cybernetics, C*, 34(30):334–352, 2004.

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

[19] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1409–1422, 2012.

[20] R. Kasturi, D. B. Goldgof, P. Soundararajan, V. Manohar, J. S. Garofolo, R. Bowers, M. Boonstra, V. N. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):319–336, 2009.

[21] M. Kristan, S. Kovacic, A. Leonardis, and J. Perš. A two-stage dynamic model for visual tracking. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 40(6):1505–1520, 2010.

[22] M. Kristan, J. Perš, M. Perše, M. Bon, and S. Kovačič. Multiple interacting targets tracking with application to team sports. In *International Symposium on Image and Signal Processing and Analysis*, pages 322–327, September 2005.

[23] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Cehovin, Georg Nebehay, Fernandez G., T. Vojir, and et al. The visual object tracking vot2013 challenge results. In *ICCV2013 Workshops, Workshop on visual object tracking challenge*, pages 98 –111, 2013.

[24] J. 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 *Comp. Vis. Patt. Recognition*, pages 1208–1215. IEEE, 2009.

[25] H. Li, C. Shen, and Q. Shi. Real-time visual tracking using compressive sensing. In *Comp. Vis. Patt. Recognition*, pages 1305–1312. IEEE, 2011.

[26] X. Li, W. Hu, C. Shen, Z. Zhang, A. R. Dick, and A. Van den Hengel. A survey of appearance models in visual object tracking. *arXiv:1303.4803 [cs.CV]*, 2013.

[27] M. Lim, C. Chan, D. Monekosso, and P. Remagnino. Swatrack: A swarm intelligence-based abrupt motion tracker. In *In proceedings of IAPR MVA*, page pages 3740, 2013.

[28] M. E. Maresca and A. Petrosino. Matrioska: A multi-level approach to fast tracking by learning. In *Proc. Int. Conf. Image Analysis and Processing*, pages 419–428, 2013.

[29] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Comp. Vis. Image Understanding*, 81(3):231–268, March 2001.

[30] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Comp. Vis. Image Understanding*, 103(2-3):90–126, November 2006.

[31] T. Nawaz and A. Cavallaro. A protocol for evaluating video trackers under real-world conditions. *IEEE Trans. Image Proc.*, 22(4):1354–1361, 2013.

[32] Y. Pang and H. Ling. Finding the best from the second bests – inhibiting subjective bias in evaluation of visual tracking algorithms. In *Int. Conf. Computer Vision*, 2013.

[33] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104, 2000.

[34] D. A. Ross, J. Lim, R. S. Lin, and M. H. Yang. Incremental learning for robust visual tracking. *Int. J. Comput. Vision*, 77(1-3):125–141, 2008.

[35] L. Sevilla-Lara and E. G. Learned-Miller. Distribution fields for tracking. In *Comp. Vis. Patt. Recognition*, pages 1910–1917. IEEE, 2012.

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

[37] L. Čehovin, M. Kristan, and A. Leonardis. Is my new tracker really better than yours? Technical Report 10, ViCoS Lab, University of Ljubljana, Oct 2013.

[38] L. Čehovin, M. Kristan, and A. Leonardis. Robust visual tracking using an adaptive coupled-layer visual model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(4):941–953, 2013.

[39] T. Vojir and J. Matas. Robustifying the flock of trackers. In *Comp. Vis. Winter Workshop*, pages 91–97. IEEE, 2011.

[40] H. Wu, A. C. Sankaranarayanan, and R. Chellappa. Online empirical evaluation of tracking algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(8):1443–1458, 2010.

[41] Y. Wu, J. Lim, and M. H. Yang. Online object tracking: A benchmark. In *Comp. Vis. Patt. Recognition*, 2013.

[42] Y. Wu, B. Shen, and H. Ling. Online robust image alignment via iterative convex optimization. In *Comp. Vis. Patt. Recognition*, pages 1808–1814. IEEE, 2012.

[43] A. Yilmaz and M. Shah. Object tracking: A survey. *Journal ACM Computing Surveys*, 38(4), 2006.

[44] D. P. Young and J. M. Ferryman. Pets metrics: On-line performance evaluation service. In *ICCCN '05 Proceedings of the 14th International Conference on Computer Communications and Networks*, pages 317–324, 2005.

[45] A. Zarezade, H. R. Rabiee, A. Soltani-Frani, and A. Khajenezhad. Patchwise joint sparse tracker with occlusion detection using adaptive markov model. *preprint in arXiv*, 2013.

[46] K. Zhang, L. Zhang, and M. H. Yang. Real-time compressive tracking. In *Proc. European Conf. Computer Vision*, Lecture Notes in Computer Science, pages 864–877. Springer, 2012.