The Thermal Infrared Visual Object Tracking VOT-TIR2015 Challenge Results

Michael Felsberg[4], Amanda Berg[4,9], Gustav Häger[4], Jörgen Ahlberg[4,9], Matej Kristan[1], Jiri Matas[2], Aleš Leonardis[3], Luka Čehovin[1], Gustavo Fernández[5], Tomáš Vojíř[2], Georg Nebehay[5], Roman Pflugfelder[5], Alan Lukežič[1], Alvaro Garcia-Martin[8], Amir Saffari[10], Ang Li[11], Andrés Solís Montero[13], Baojun Zhao[16], Cordelia Schmid[24], Dapeng Chen[11], Dawei Du[26,27], Fahad Shahbaz Khan[4], Fatih Porikli[19,20], Gao Zhu[19], Guibo Zhu[22], Hanqing Lu[22], Hilke Kieritz[17], Hongdong Li[19,21], Honggang Qi[26,27], Jae-chan Jeong[15], Jae-il Cho[15], Jae-Yeong Lee[15], Jianke Zhu[12], Jiatong Li[25,16], Jiayi Feng[14], Jinqiao Wang[22], Ji-Wan Kim[15], Jochen Lang[13], Jose M. Martinez[8], Kai Xue[23], Karteek Alahari[24], Liang Ma[23], Lipeng Ke[26,27], Longyin Wen[26], Luca Bertinetto[6], Martin Danelljan[4], Michael Arens[17], Ming Tang[14], Ming-Ching Chang[26], Ondrej Miksik[6], Philip H S Torr[6], Rafael Martin-Nieto[8], Robert Laganière[13], Sam Hare[7], Siwei Lyu[26], Song-Chun Zhu[18], Stefan Becker[17], Stephen L Hicks[6], Stuart Golodetz[6], Sunglok Choi[15], Tianfu Wu[18], Wolfgang Hübner[17], Xu Zhao[14], Yang Hua[24], Yang Li[12], Yang Lu[18], Yuezun Li[26], Zejian Yuan[11], and Zhibin Hong[25]

[1]University of Ljubljana, Slovenia
[2]Czech Technical University, Czech Republic
[3]University of Birmingham, United Kingdom
[4]Linköping University, Sweden
[5]Austrian Institute of Technology, Austria
[6]Oxford University, United Kingdom
[7]Obvious Engineering, United Kingdom
[8]Universidad Autónoma de Madrid, Spain
[9]Termisk Systemteknik AB, Sweden
[10]Affectv, United Kingdom
[11]Xi'an Jiaotong University
[12]Zhejiang University, China
[13]University of Ottawa, Canada
[14]Institute of Automation, Chinese Academy of Sciences, China
[15]Electronics and Telecommunications Research Institute, Korea
[16]Beijing Institute of Technology, China
[17]Fraunhofer IOSB, Germany
[18]University of California, USA
[19]Australian National University, Australia
[20]NICTA, Australia
[21]ARC Centre of Excellence for Robotic Vision, Australia
[22]NLPR, Chinese Academy of Sciences, China
[23]Harbin Engineering University, China
[24]INRIA Grenoble Rhône-Alpes, France
[25]University of Technology, Australia
[26]University at Albany, USA
[27]SCCE, Chinese Academy of Sciences, China

Abstract

The Thermal Infrared Visual Object Tracking challenge 2015, VOT-TIR2015, aims at comparing short-term single-object visual trackers that work on thermal infrared (TIR) sequences and do not apply pre-learned models of object appearance. VOT-TIR2015 is the first benchmark on short-term tracking in TIR sequences. Results of 24 trackers are presented. For each participating tracker, a short description is provided in the appendix. The VOT-TIR2015 challenge is based on the VOT2013 challenge, but introduces the following novelties: (i) the newly collected LTIR (Linköping TIR) dataset is used, (ii) the VOT2013 attributes are adapted to TIR data, (iii) the evaluation is performed using insights gained during VOT2013 and VOT2014 and is similar to VOT2015.

1. Introduction

Visual tracking is a challenging task that has attracted significant attention in the past two decades, e.g. [16, 29, 32]. The number of accepted motion or tracking papers in high profile conferences, such as ICCV, ECCV, and CVPR, has been consistently high (∼40 papers annually), summing up to a few hundred relevant papers in the field. However, the lack of established performance evaluation methodology combined with this large number of publications makes it difficult to assess and understand the advancements made in the field. Several initiatives have attempted to establish a common ground in tracking performance evaluation, starting with PETS [43] and more recently with the Visual Object Tracking (VOT) challenges [26, 27, 23] and the Object Tracking Benchmark [42, 41].

In recent years, thermal cameras improved in image quality and resolution while decreased in both price and size. This development has opened up new application areas [15]. Historically, thermal cameras have delivered noisy images with low resolution, used mainly for tracking small objects (point targets) against colder backgrounds and have mainly been of interest for military purposes. Today, they are commonly used in various applications, e.g., cars and surveillance systems. Increasing image quality allows exploration of new application areas, often requiring methods for tracking of extended dynamic objects. Further, for some applications, the methods cannot be restricted to stationary platforms. The main advantages of thermal cameras are their ability to see in total darkness, their robustness to illumination changes and shadow effects, and reduced privacy intrusion.

This paper describes the first thermal infrared (TIR), short-term tracking challenge, the Visual Object Tracking TIR (VOT-TIR2015) challenge, and the results obtained. Like the VOT challenge, the VOT-TIR challenge considers single-camera, single-target, model-free, causal trackers, applied to short-term tracking. It has been featured as a sub-challenge to VOT2015, organized in conjunction with ICCV2015. The challenge enabled participants not only to evaluate their results on visual data, but also to benchmark their trackers on thermal infrared sequences.

Available datasets for evaluation of tracking in thermal infrared have become outdated [3]. This causes researchers to evaluate their methods on proprietary datasets, which makes it difficult to get an overview of advancement made in the field. Inconsistent performance measures across different papers contributes to this difficulty. The Visual Object Tracking challenge, provides an established evaluation methodology for data in the visible spectrum. The main idea of VOT-TIR2015 is to carry these ideas to the area of TIR data, based on a recently collected dataset [3].

1.1. Related work

A large number of benchmarks exist in the area of visual tracking, but far fewer for TIR tracking. Among visual spectrum (RGB) tracking, the most closely related investigations to the approach presented here is the VOT2015 challenge [23], as well as those of previous years [26, 27]. The online tracking benchmark (OTB) by Wu et al. [42, 41] contains 100 sequences and is a widely used tracking benchmark. In the OTB, trackers are compared using a precision score and a success score, without restarting a failed tracker. The precision score is the percentage of frames where the estimated bounding box is within some fixed distance to the ground truth, while the success score measures the area under the curve of number of frames where the overlap is greater than some fixed percentage. This area has been shown to be equivalent to the average overlap [37, 38]. For further discussion on OTB we refer to [42, 41] and for comparisons with the VOT evaluation to [25, 24].

The series of workshops on Performance Evaluation of Tracking and Surveillance (PETS) [43] have organized thermal infrared challenges on two occasions. The first has taken place in 2005 and the second in 2015, where the challenge was detection, multi-camera/long-term tracking and behavior (threat) analysis. In contrast to VOT-TIR, the challenges concerned multiple research areas while VOT-TIR focuses on the problem of short-term tracking only. The lack of further related work within the area of thermal infrared tracking challenges motivates the VOT-TIR initiative.

1.2. The VOT-TIR2015 challenge

The VOT-TIR2015 challenge targets a specific set of trackers. All participating trackers are required to be: (i) Causal – sequence frames have to be processed in sequential order; (ii) Short-term – trackers are not required to handle reinitialization; (iii) Model-free – pre-built models of object appearances are not allowed.

Performance of participating trackers is automatically measured using the VOT2014 evaluation kit [27]. The toolkit performs a standardized experiment and stores resulting bounding boxes. If the tracker fails, it is reinitialized. Participants are required to integrate their trackers into the toolkit. Tracking results have been analyzed using the VOT2015 evaluation methodology [23].

Participants were expected to submit a single set of results per tracker as well as binaries for result verification. A different set of parameters does not constitute a new tracker. Tracker parameters set by the participant is required to be equal for all test sequences. Detection (by the tracker) of a specific test sequence in order to set hand-tuned parameters is not permitted. However, the tracker itself is allowed to internally change parameters using, e.g., the bounding box size. Further details regarding participation rules are available from the challenge homepage[1].

Differences from the visual spectrum challenge

Compared to the visual equivalent, VOT2015 [23], there are some differences in annotation as well as acquisition and evaluation procedure. The annotated bounding boxes are not allowed to rotate. Further, due to the limited amount of freely available thermal infrared datasets and sequences, sequence selection could not be done as in VOT2015. A new dataset, LTIR (the Linköping Thermal IR dataset) [3], was created for this purpose. Seven different sources were asked to contribute with data and the provided data that contained sufficiently challenging tracking events were included in the dataset. A more detailed description can be found in Section 2.

The VOT-TIR2015 challenge applies the same evaluation methodology as VOT2015 [23], except for the practical difference evaluation. This evaluation requires multiple annotations, which are not (yet) available for the LTIR dataset.

1.3. Outline

The dataset used in the VOT-TIR2015 challenge is described in Section 2. Section 3 briefly summarizes the performance measures and evaluation methodology used in the challenge. Analysis and results are presented in Section 4 and, finally, conclusions are drawn in Section 5. In addition, short descriptions of all participating trackers can be found in Appendix A.

2. The VOT-TIR2015 dataset

The dataset used in VOT-TIR2015 is LTIR, the Linköping Thermal IR dataset [3]. Sequences included in the dataset were collected from seven different sources using eight different types of sensors. The included sequences originate from industry, universities, a research institute and an EU FP7 project. Resolutions range from 320×240 to 1920×480 pixels and the average sequence length is 563 frames. Some sequences in the LTIR dataset are available with both 8- and 16-bit pixel values, however, for this challenge, only 8-bit sequences were used. The main reason for this restriction is that several of the submitted methods cannot deal with 16-bit data. There are sequences from indoor and outdoor environments, and the outdoor sequences were recorded in different weather conditions. Example frames from four sequences are shown in Fig. 1.

All benchmark annotations are in accordance with the VOT2013 annotation process [26] and have been done manually. One object within each sequence is annotated in each frame with a bounding box that encloses the object throughout the sequence. The bounding box is allowed to vary in size but not to rotate. In addition to the bounding box annotations, global attributes are per-sequence annotated and local attributes per-frame annotated.

Global attributes The per-sequence global attributes from VOT have to be adapted to the properties of TIR in order to be useful. Below, the global attributes have been arranged according to similarity to VOT-attributes.

- Attributes different from VOT: *Dynamics change* and *temperature change* have been introduced instead of *illumination change* and *object color change*. Not all cameras provide the full 16-bit range, instead, an adaptively changing 8-bit dynamics are sometimes used. *Dynamics change* indicates whether the dynamics is fixed during the sequence or not. *Temperature change* refers to changes in the thermal signature of the object during the sequence
- Attributes similar to VOT: In TIR, *Blur* indicates blur due to motion, high humidity, rain or water on the lens.
- Attributes equal to VOT: *Camera motion*, *object motion*, *background clutter*, *size change*, *aspect ratio change*, *object deformation*, and *scene complexity*.

Local attributes The local, per-frame annotated attributes are: *motion change*, *camera motion*, *dynamics change*, *occlusion*, and *size change*. The attributes are used in the evaluation process to weigh tracking results. They can also be used to evaluate the performance of the method on frames with specific attributes.

3. Performance measures and evaluation methodology

The performance measures as well as evaluation methodology for VOT-TIR2015 are equal to the ones for VOT2015, except for the practical difference evaluation. Therefore, only a brief summary is given here, further details can be found in [23].

[1] http://www.votchallenge.net/vot2015/participation.html

(a) running rhino (b) quadrocopter (c) crowd (d) street

Figure 1: Snapshots from four sequences included in the LTIR dataset. The annotated bounding box is marked in yellow.

Similar to the VOT2015 challenge, the two weakly correlated performance measures, accuracy and robustness, are used due to their high level of interpretability [37, 38]. The accuracy measurement measures the overlap between the predicted bounding box and the ground truth while the robustness measurement measures how many times the tracker fails. If a tracker is considered to have failed, it is re-initialized five frames later. Overlap calculations, re-initialization, definition of a failure and the rank-based evaluation methodology is further explained in [23].

4. Analysis and results

4.1. The VOT2015 experiments

In our evaluation, and in contrast to VOT2014 [27], we considered the baseline experiment only. We did not consider the region noise experiment for three reasons: First, the results of previous experiments hardly differed [27]. Second, the experiments need significantly more time. Finally, the reproducibility of results would have required to store the seed, which has not been foreseen in the evaluation kit.

4.2. Submitted trackers

In total, 24 trackers were included in the VOT-TIR2015 challenge. Among them, 20 trackers were submitted and 4 trackers were added by the VOT Committee (3 novel and 1 baseline trackers). The committee have used the accompanying binaries/source code for result verification. For the baseline trackers, the default parameters were selected, or, when not available, were set to reasonable values. All entries are briefly described below and references to the original papers are given in the Appendix A where available.

Twenty trackers participated in both the VOT2015- and VOT-TIR2015 challenge while 4 trackers were only entered in the VOT-TIR2015 challenge.[2]

One tracker, EBT (A.11), uses object proposals [48] for object position generation or scoring. Several trackers are based on Mean Shift tracker extensions [8], ASMS (A.21), PKLTF (A.4), SumShift (A.14), and its derivative DTracker (A.19). CMIL is based on online boosting (A.18) and sPST (A.20) is based on tracking-by-detection learning. A number of trackers can be classified as part-based trackers. These were LDP (A.16), G2T (A.8), AOGTracker (A.7), MCCT (A.3), and FoT (A.22). A number of trackers come from a class of holistic models that apply regression-based learning for target localization. Out of these, one is based on structured SVM learning, Struck[3] (A.5). Several regression-based trackers use correlation filters [5, 20] as visual models. Some correlation filer based trackers maintain a single model for tracking, i.e., NSAMF (A.10), OACF (A.6), SRDCFir (A.15), sKCF (A.2), STC (A.23), MKCF+ (A.12), CCFP (A.13), and several trackers apply multiple templates to model appearance variation, i.e., SME (A.9), and KCFv2 (A.1). One tracker, ABCD (A.17), applies a global, generative model exploiting channel representations. Finally, the VOT Committee added a baseline tracker, the HotSpot tracker, to the set of submitted trackers. Tracking by detecting hot areas is still state-of-the-art in many TIR applications, e.g. pedestrian detection [22]. The HotSpot tracker detects objects by pixel intensity thresholding and tracks detections using a Kalman filter with a Global Nearest Neighbor approach to the association problem.

4.3. Results

The results are summarized in sequence pooled and attribute normalized AR rank and AR raw plots in Figure 2. The sequence pooled AR rank plot is obtained by concatenating the results from all sequences and creating a single rank list, while the attribute normalized AR rank plot is created by ranking the trackers over each attribute and averaging the rank lists. Similarly the AR raw plots were constructed. The raw values for the sequence pooled results are also given in Table 1.

[2]Here, we consider SRDCF and SRDCFir being the same, despite the fact that SRDCFir uses a slightly different feature vector, see Appendix A.15.

[3]The implementation used here is a recent improvement of [18].

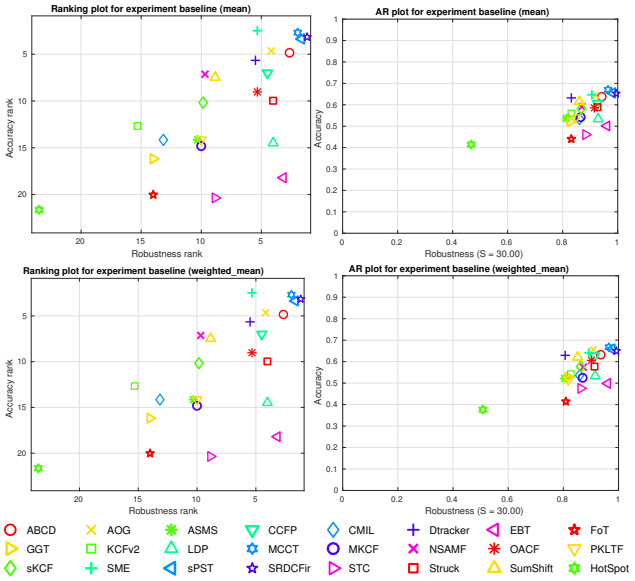

Figure 2: The AR rank plots and AR raw plots generated by sequence pooling (upper) and by attribute normalization (below).

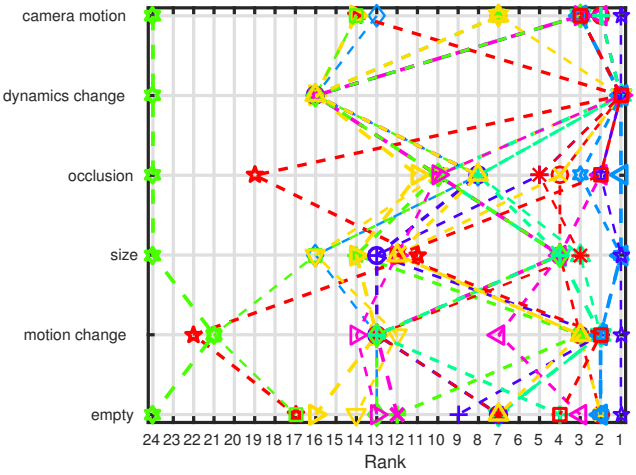

Figure 3: Robustness plots with respect to the visual attributes. See Figure 2 for legend.

The following trackers appear either very accurate or very robust among the top performing trackers (closest to the upper right corner of rank plots): SME (A.9), MCCT (A.3), sPST(A.20), SRDCFir (A.15), ABCD (A.17), and AOG (A.7). In contrast to VOT2014, where methods based on correlation filters were largely dominating [27], top performers in VOT-TIR2015 belong to several different classes.

The robustness ranks with respect to the visual attributes are shown in Figure 3. The top three trackers with respect to the different visual attributes are mostly SRDCFir, sPST, and MCCT. A significant exception is camera motion, where SME and EBT (A.11) come second and third.

Tracker	A	R	$\hat{\Phi}$	Speed	Impl.
SRDCFir	0.65	**0.58**	**0.70**	3.17	M C
sPST	*0.66*	*2.18*	*0.64*	0.61	M C
MCCT	**0.67**	3.34	0.55	15.05	M C
EBT	0.50	3.50	0.43	1.08	M C
CCFP	0.63	8.55	0.36	1.03	M C
ABCD	0.63	5.81	0.34	6.88	M
Struck	0.58	8.48	0.30	2.90	C
SME	0.64	9.97	0.30	6.67	M C
LDP	0.53	8.33	0.29	6.96	M C
NSAMF	0.57	12.63	0.28	10.69	M
OACF	0.61	9.57	0.28	3.22	M C
AOG	0.65	8.76	0.27	1.27	*binary*
sKCF	0.58	13.90	0.27	**255.13**	C
CMIL	0.54	14.04	0.25	5.31	C
MKCF+	0.52	12.61	0.24	1.60	M C
KCFv2	0.54	17.81	0.23	14.78	M
STC	0.48	13.85	0.23	29.92	M
SumShift	0.62	15.67	0.19	19.78	C
G2T	0.53	18.59	0.18	0.39	M C
FoT	0.41	19.40	0.17	131.57	C
PKLTF	0.52	19.30	0.16	23.65	C
Dtracker	0.63	19.69	0.16	11.55	C
ASMS	0.52	20.03	0.14	*163.42*	C
HotSpot	0.38	62.27	0.04	5.98	M

Table 1: The table shows raw accuracy and the average number of failures, expected average overlap, tracking speed (in EFO), and implementation details (M is Matlab, C is C or C++).

The latter turns also out to rank well in the overall criterion *expected average overlap*, see Figure 4. The expected average overlap curve is given by the average bounding-box-overlap averaged over a set of sequences of certain length, plotted over the sequence length N_s [23]. These curves confirm previous statements on the three top performing methods MCCT, sPST, and SRDCFir, where the latter gives the best overall performance. The fact that EBT is ranked fourth underpins the importance of robustness for the expected average overlap.

Apart from tracking accuracy, robustness, and expected average overlap at N_s frames, the tracking speed is also crucial in many realistic tracking applications. We therefore visualize the expected overlap score with respect to the tracking speed measured in EFO units in Figure 5. To put EFO units into perspective, a C++ implementation of a NCC tracker provided in the toolkit runs with average 140 frames per second on a laptop with an Intel Core i5-2557M processor, which equals to approximately 160 EFO units.

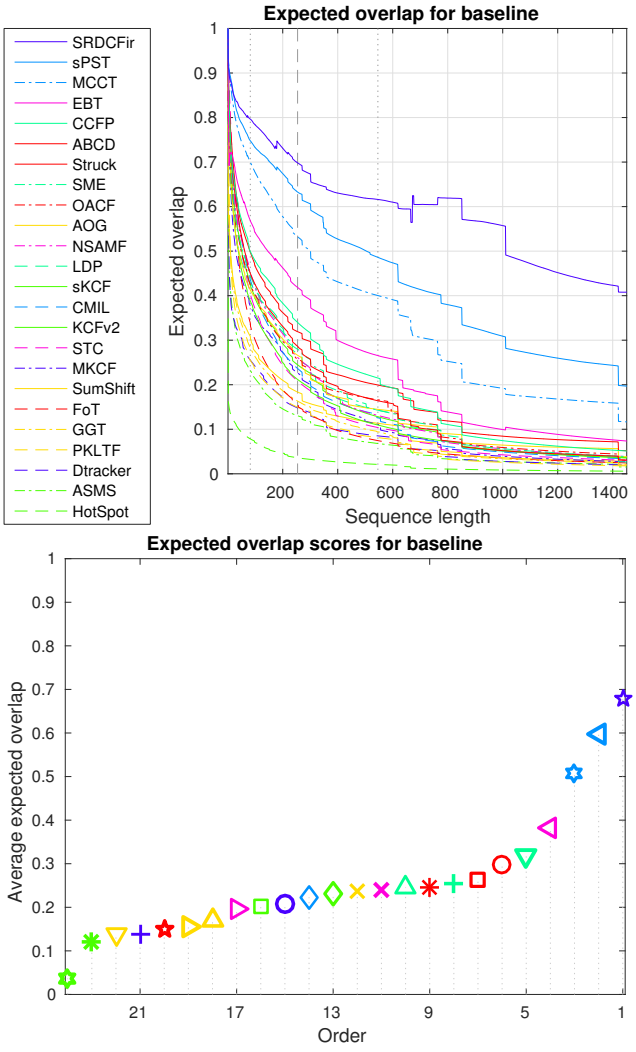

Figure 4: Expected average overlap curve (above) and expected average overlap graph (below) with trackers ranked from right to left. The right-most tracker is the top-performing according to the VOT2015 expected average overlap values. See Figure 2 for legend. The vertical lines in the upper plot show the range of typical sequence lengths.

The vertical dashed line in Figure 5 indicates the real-time speed (equivalent to approximately 20fps). Among the three top-performing trackers, MCCT comes closest to real-time performance. The top-performing tracker in terms of expected overlap among the trackers that exceed the real-time threshold is at the same time the overall fastest tracker, sKCF (A.2).

4.4. TIR-specific analysis and results

A particular interesting question in context of VOT-TIR is the effect of the differences between RGB sequences and TIR sequences on the ranking of the trackers. For this purpose, the joint ranking for VOT and VOT-TIR of the 20 common trackers[2] is shown in Figure 6. The only VOT-TIR trackers that have not been run on VOT are MCCT, CCFP, ABCD, and the HotSpot detector.

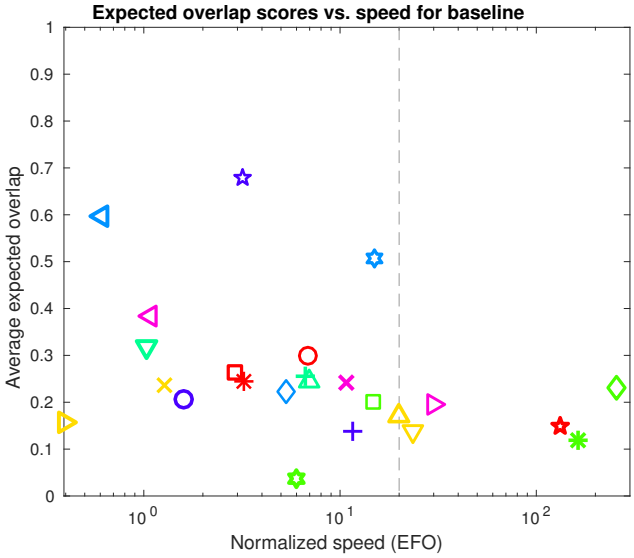

Figure 5: Expected average overlap scores w.r.t. the tracking speed in EFO units. The dashed vertical line denotes the estimated real-time performance threshold of 20 EFO units. See Figure 2 for legend.

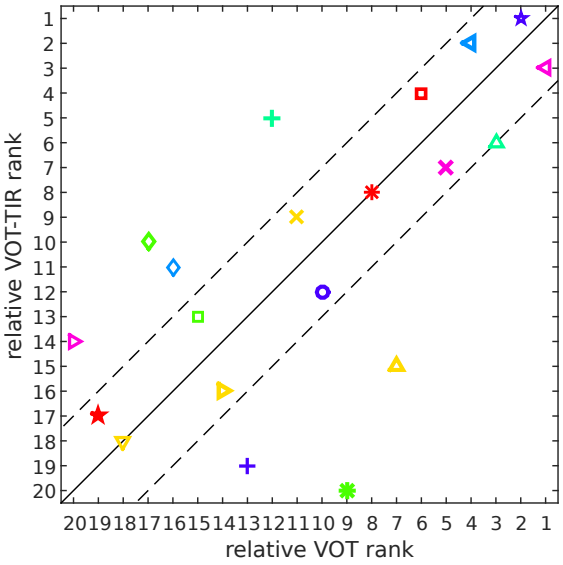

Figure 6: Comparison of relative ranking of 20 trackers in VOT and VOT-TIR. See Figure 2 for legend.

The dashed lines are the margin of a rank-change by more than three positions. Any change of rank within this margin is considered insignificant and only 7 trackers change their rank by more than three positions. The most

dramatic change occurs for ASMS, which ranks 23 in VOT-TIR, but 20 (out of more than 60) in VOT, corresponding to rank 9 within the set of 20 trackers. Other trackers that perform significantly worse are SumShift, and DTracker.

On the other hand, SME, sKCF, STC, and CMIL perform significantly better on VOT-TIR than on VOT according to the relative ranking. Similar as for the overall performance, it is difficult to identify a systematic correlation between improvement and type of tracking methods. Tracking methods that do not use color are likely to perform better on TIR sequences than color-based methods, such as ASMS, SumShift, and DTracker. Also the size of targets differ between VOT (larger) and VOT-TIR (smaller). It is also believed that the tuning of input features is more important to maintain good performance on VOT-TIR, e.g. SRDCFir introduces additional features beyond HOG (see Appendix A.15) and works better on TIR sequences than SRDCF with features as used in VOT2015.

5. Conclusions

The VOT-TIR challenge received 20 submissions and compared in total 24 trackers, which we consider a good success and the results presumably give a good guidance to future research within TIR tracking. Best overall performance has been achieved by SRDCFir, closely followed by sPST and MCCT. However, further analysis of the results will be required in order to draw deeper conclusions.

For future challenges, the dataset needs to be extended to become larger and more challenging. Annotation and evaluation need to be adapted to the current VOT standard: multiple annotations and rotating bounding boxes. Also challenges with mixed sequences (RGB and TIR) might be interesting to perform.

Acknowledgments

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 European Union 7th Framework Programme under grant agreement 257906. J. Matas and T. 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). M. Felsberg and G. Häger were supported by the Swedish Foundation for Strategic Research through the project CUAS and the Swedish Research Council trough the project EMC^2. J. Ahlberg and A. Berg were supported by the European Union 7th Framework Programme under grant agreement 312784 (P5) and the Swedish Research Council through the contract D0570301. Some experiments where run on GPUs donated by NVIDIA.

A. Submitted trackers - VOT TIR

In this appendix we provide a short summary of all trackers that were considered in the VOT-TIR2015 challenge.

A.1. Restore Point guided Kernelized Correlation Filters (KCFv2)

Liang Ma, Kai Xue
mllx01161110@hotmail.com, xuekai@hrbeu.edu.cn

The Kernelized Correlation Filters [20] have been shown effective for target tracking in VOT2014 challenges. Its success lies in the fast online Support Vector Machine learning process in Fourier domain. Due to the fact that there is only one positive sample and the negative samples are generated virtually by circulant matrices at each frame, the KCF tracker would learn a biased model during tracking and the bias would definitely increase over time. The original KCF tracker adopts a linear interpolation method in the newly trained model to alleviate this bias. However, the linear interpolation method cannot handle target appearance change caused by camera motion, occlusion or target deformation at a moderate level. Our approach, the RP-KCF tracker, enhances its robustness by examine the similarity between each candidate patch generated by the KCF tracker and the Restore Point patch. A restore point patch is a base patch that can characterize target appearance in a short time period. In short-term target tracking, the restore point patch can be directly set to be the ground truth patch at first frame; whereas, in long-term tracking, the restore point patch should be updated over time. We measure the similarity likelihood of top k candidate positions produced by the KCF tracker at neighboring scales, and the likelihood function involves the histogram of gray-level and gradient.

A.2. Scalable Kernel Correlation Filter with Sparse Feature Integration (sKCF)

Andrés Solís Montero, Jochen Lang, Robert Laganière
asolismon@uottawa.ca,
{jlang,laganiereg}@eecs.uottawa.ca

Fast scalable solution based on the Kernelized Correlation Filter (KCF) framework. We introduce an adjustable Gaussian window function and keypoint-based model for scale estimation to deal with the fixed size limitation in the Kernelized Correlation Filter. Furthermore, we integrate the fast HoG descriptors and Intels Complex Conjugate Symmetric (CCS) packed format to boost achievable frame rates.

A.3. Motion-aware Complex Cell Tracker (MCCT)

Dapeng Chen, Ang Li, Zejian Yuan
dapengchenxjtu@foxmail.com

The proposed tracker is a novel variant of CCT proposed in [6]. CCT utilizes intensity histogram and oriented gradient histogram as cell descriptors, which is not sufficient for

tracking in VOT-TIR 2015. This is because the thermal infrared images contain no color information and less texture information. We observed that many of the sequences in VOT-TIR 2015 are captured by a fixed surveillance camera. This justifies the utilization of frame difference, as in this situation the frame difference encodes the contour of moving the object. We compute the absolute values of frame difference for the region surround the object, then generate a binary image by a small threshold value, and finally compute the oriented gradient histogram of the binary image to describe the motion contour. Now, each cell is described by the histogram of intensity, the oriented gradient, and the motion contour, but the three visual cues can not always be effective due to dynamically changing environment. A score normalization strategy, which is similar to the fusion method of the complex cells as introduced in [6], is adopted to weight different visual cues. The other components are same with CCT, including using the obtained cell descriptors to describe complex cells, using score normalization to mediate different visual cues and different types of complex cells, and inferring the occlusion and stability situation for each complex cells.

A.4. Point-based Kanade Lukas Tomasi color-Filter (PKLTF)

Rafael Martin-Nieto, Alvaro Garcia-Martin, Jose M. Martinez
{*rafael.martinn, alvaro.garcia, josem.martinez*}*@uam.es*

PKLTF [17] is a single-object long-term tracker that supports high appearance changes in the target, occlusions, and is also capable of recovering a target lost during the tracking process. It was originally designed for long term tracking but it has been adapted to the VOT short term sequences.

A two stages algorithm has been designed for this single-target object tracker. The first stage is based on the Kanade Lukas Tomasi approach (KLT) [34] to choose the object features (using color and motion coherence) in order to track relatively large object displacements. The second stage is based on mean shift gradient descent [7] to place the bounding box into the exact position of the object. Besides the color model is updated adding weight to the pixels which are present in the original histogram.

The object model is based on the RGB color and the luminance gradient. The model consists of a histogram including the quantized values of the color components, and an edge binary flag. The histogram is generated with all the pixels of this first frame located inside the object image patch. All pixels in this patch contribute with the same weight to the histogram, regardless of their position/location in the bounding box. After that, using the CBWH method [33], the histogram is corrected reducing the effect caused by the background pixels which are present in the initial patch.

A.5. Struck

Stuart Golodetz, Sam Hare, Amir Saffari, Stephen L Hicks, Philip H S Torr
sgolodetz@gxstudios.net, sam@samhare.net,
amir@ymer.org, stephen.hicks@ndcn.ox.ac.uk,
philip.torr@eng.ox.ac.uk

Struck [18] is a framework for adaptive visual object tracking based on structured output prediction. By explicitly allowing the output space to express the needs of the tracker, the need for an intermediate classification step is avoided. The method uses a kernelized structured output support vector machine (SVM), which is learned online to provide adaptive tracking. The version of Struck submitted to VOT 2015 uses multi-kernel learning (MKL) and larger feature vectors than were used in the past. In particular, we combine a Gaussian kernel on 192D Haar features with an intersection kernel on 480D histogram features. This significantly improves the tracking performance, but at a cost in speed. The reader is referred to [18] for details.

A.6. Object-Aware Correlation Filter Tracker (OACF)

Luca Bertinetto, Ondrej Miksik, Stuart Golodetz, Philip H.S. Torr
{*luca.bertinetto, ondrej.miksik*}*@eng.ox.ac.uk,*
stuart.golodetz@ndcn.ox.ac.uk, philip.torr@eng.ox.ac.uk

Correlation trackers have achieved excellent performance in single-target model-free tracking. Several versions spurred from the original MOSSE [5], incorporating multi-channel features (like HOG), kernels [19] and scale adaptation [9]. A common trait is that they all train a new filter at each frame by imposing a Gaussian *desired response* (which acts as a *soft label*) in correspondence of the center of the currently estimated bounding box. A global filter is then updated with a (slow) running average. This approach is doomed to fail when the object quickly changes its appearance for two reasons. (a) The global filter cannot handle fast changes because of its slow update rule, that is however necessary to have a robust representation. (b) In general, HOG features do not cope well with changes of shape, and sometimes they are simply not adequate to discriminate between target object and background. To tackle this problem, we build on the scale adaptive DSST [9] and we compute a per-pixel likelihood map of the target (implemented with grayscale histograms) [4]. In this way we can estimate, for each pixel $\mathbf{x}$, the probability that it belongs to the object to track $\mathcal{O}$, i.e. $L = P(\mathbf{x} \in \mathcal{O}|O, B)$, where O, B are the areas delimiting foreground and background. With this information, we can refine the estimation of the correlation filter and also make sure that the new learned filter is centered on the target, simply by shifting the peak of the Gaussian desired response in correspondence of the center of mass of the likelihood map.

A.7. AOGTracker

Tianfu Wu, Yang Lu and Song-Chun Zhu
{*tfwu, yanglv*}*@ucla.edu, sczhu@stat.ucla.edu*

This method consists of a framework for simultaneously tracking, learning and parsing objects in video sequences with a hierarchical and compositional And-Or graph (AOG) representation. We call our tracker AOGTracker. The AOG explores latent discriminative part configurations to represent objects. It is discriminatively learned online to account for the appearance (e.g., lighting and partial occlusion) and structural (e.g., different poses and viewpoints) variations of the object, as well as the distractors (e.g., similar objects) in the scene background. The AOGTracker is formulated under the Bayesian framework and a spatial-temporal dynamic programming (DP) algorithm is derived to infer the state of the object (i.e., bounding box) on the fly in tracking. During online learning, the AOG is updated iteratively with two steps in the latent structural SVM framework: (i) Identifying the false positives and false negatives of the current AOG in a new frame by exploiting the spatial and temporal constraints observed in the trajectory; (ii) Updating the structure of the AOG based on the intrackability of the current AOG, and re-estimating the parameters based on the augmented training dataset. In experiments, the proposed method is tested on both VOT2015 and VOT-TIR2015 with the same parameter setting (except for the appearance features).

A.8. Geometric Structure Hyper-Graph based Tracker (G2T)

Yuezun Li, Dawei Du, Longyin Wen, Lipeng Ke, Ming-Ching Chang, Honggang Qi, Siwei Lyu
{*liyuezun, cvdaviddo, wly880815, lipengke1, mingching, honggangqi.cas, heizi.lyu*}*@gmail.com*

G2T tracker is especially designed for tracking deformable objects. G2T represents the target object by a geometric structure hyper-graph, which integrates the local appearance of the target with higher order geometric structure correlations among target parts. In each video frame, tracking is formulated as a hyper-graph matching between the target geometric structure hyper-graph and a candidate hyper-graph. Multiple candidate associations between the nodes of both hyper-graphs are built. The weight of the nodes indicate the reliability of the candidate associations based on the appearance similarity between the corresponding parts of each hyper-graph. A matching between the target and a candidate is solved by applying the extended pairwise updating algorithm of [31].

A.9. Scale-adaptive Multi-Expert Tracker (SME)

Jiatong Li, Zhibin Hong, Richard Yi Da Xu, Baojun Zhao
{*Jiatong.Li-3@student., Zhibin.Hong@student., yida.xu@*}*uts.edu.au, zbj@bit.edu.cn*

SME is a multi-expert based scale adaptive tracker. Inspired by [44], SME adopts the current tracker as well as the historical trained tracker snapshots to constitute the expert ensemble. At each frame, each expert decide the target state independently. If a disagreement among the experts is reported, the best expert is selected by their accumulated score. Unlike [44], SME proposes a trajectory consistency based score function as the expert selection criteria. Furthermore, an effective scale adaptive scheme is introduced to handle scale changes on-the-fly. Multi-channel based correlation filter tracker [19] is adopted as the base tracker, where HOG and image illumination features are concatenated to enhance the performance.

A.10. NSAMF

Yang Li, Jianke Zhu
{*liyang89, jkzhu*}*@zju.edu.cn*

As the correlation filter-based trackers [19, 5] have achieved the competitive results both on accuracy and robustness in VOT2014 challenge, we present a tracker based on the correlation filter framework. The proposed tracker is an improved version of our previous method, SAMF [30]. The main difference is that NSAMF employs color probability rather than color name. In addition, the final response map is a fusion of multi-models based on the different features. The extensive empirical evaluation on the VOT 2015 dataset demonstrates that the proposed tracker is very promising for the various challenging scenarios.

A.11. Edge Box Tracker (EBT)

Gao Zhu, Fatih Porikli, Hongdong Li
{*gao.zhu, fatih.porikli, hongdong.li*}*@anu.edu.au*

Human visual system is adept at tracking shapes without any texture. Motivated by this, we incorporated an object proposal mechanism that uses sparse yet informative contours to score proposals based on the number of contours they wholly enclose into a detection-by-tracking process for visual tracking. Our method is able to execute search in the entire image quickly and focus only on those high-quality candidates to test and update our discriminative classifier. Using high-quality candidates to chose better positive and negative samples, we reduce the spurious false positives and improve the tracking accuracy. Since our tracker employs only a few candidates to search the object, it has potential to use higher-dimensional features if needed. More importantly, our method can track randomly and very fast moving objects. It is robust to full occlusions as it is able to rediscover the object after occlusion. More details can be found in [46]. The reader is referred to [46] for details.

A.12. Multi-kernelized Correlation Filter plus (MKCF+)

Ming Tang, Jiayi Feng, and Xu Zhao
{*tangm, jiayi.feng, xu.zhao*}*@nlpr.ia.ac.cn*

Our tracker is implemented based on the multi-kernelized correlation filter tracker (MKCF) [36] and background modeling algorithm ViBe [2]. MKCF, as its name suggest, combines the multiple kernel learning and correlation filter techniques. Compared to traditional correlation filter trackers, MKCF explores diverse features (gray and HOG in this experiment) simultaneously to improve tracking performance. In addition, an optimal search technique and PSR (peak to sidelobe ratio) are also utilized in MKCF to estimate object scales. PSR is supposed to reach maximum when the bounding box fits target scale properly. Although MKCF performs well on challenging sequences, it can not prevent itself from model drift problem. Therefore, ViBe is adapted to our MKCF+ to alarm its locating failures. ViBe is launched only on frames with stable scenes. And in such case, it is probable for ViBe to find out the possible locations of the target in searching area. The candidate locations are then tested by MKCF to determine which one should be the target.

A.13. Clustering Correlation Tracking with Foreground Proposals (CCFP)

Guibo Zhu, Jinqiao Wang, Hanqing Lu
{*gbzhu, jqwang, luhq*}*@nlpr.ia.ac.cn*

CCFP tracker is mainly based on the idea of collaborative correlation tracking [47]. Some confident candidate proposals are generated through online detection or background modeling, and used to improve the overall tracking capability of the correlation filter-based tracker. To be specific, it relies on an incremental appearance clustering algorithm for evaluation, discriminative scale space tracker [9] and background modeling [2]. The collaborative combination of three parts constructs the CCFP tracker which is robust to heavily occlusion and fast motion.

A.14. SumShift

Jae-Yeong Lee, Sunglok Choi, Jae-chan Jeong, Ji-Wan Kim, Jae-il Cho
{*jylee, sunglok, channij80, giraffe, jicho*}*@etri.re.kr*

SumShift tracker is an implementation of the histogram-based tracker suggested in [28]. SumShift improves conventional histogram-based trackers (e.g., meanshift tracker) in two ways. First it uses a partition-based object model represented by multiple patch histograms to preserve geometric structure of the color distribution of the object. Secondly the object likelihood is computed by the sum of the patch probabilities which are computed from each corresponding patch histograms, which enables more robust and accurate tracking. The reader is referred to [28] for details.

A.15. Spatially Regularized Discriminative Correlation Filter Tracker for IR (SRDCFir)

Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, Michael Felsberg
{*martin.danelljan, gustav.hager, fahad.khan, michael.felsberg*}*@liu.se*

SRDCFir adapts the SRDCF approach proposed in [10] to thermal infrared data. Standard Discriminative Correlation Filter (DCF) based trackers such as [9, 11, 20] suffer from the inherent periodic assumption when using circular correlation. The resulting periodic boundary effects leads to inaccurate training samples and a restricted search region. The SRDCF mitigates these problems by introducing a spatial regularization function that penalizes filter coefficients residing outside the target region. This allows the size of the training and detection samples to be increased without affecting the effective filter size. By selecting the spatial regularization function to have a sparse Discrete Fourier Spectrum, the filter is efficiently optimized directly in the Fourier domain. Instead of solving for an approximate filter, as in previous DCF based trackers (e.g. [9, 11, 20]), the SRDCF employs an iterative optimization based on Gauss-Seidel that converges to the exact filter. The detection step employs a sub-grid location estimation. In addition to the HOG features used in [10], SRDCFir also employs channel coded intensity features. SRDCFir also employs a motion feature channel, computed by thresholding the difference between the current and previous frame. The result is a binary image that indicates if a pixel has changed its value compared to the previous frame. The intensity and motion features are averaged over the 4×4 HOG cells and then concatenated, giving a 43 dimensional feature vector at each cell.

A.16. Layered Deformable Parts tracker (LDP)

A. Lukežič, L. Čehovin, Matej Kristan
alan.lukezic@gmail.com

LDP is a part-based correlation filter composed of a coarse and mid-level target representations. Coarse representation is responsible for approximate target localization and uses HoG as well as color features. The mid-level representation is a deformable parts correlation filter with fully-connected parts topology and applies a novel formulation that threats geometric and visual properties within a single convex optimization function. The mid-level as well as coarse level representations are based on the kernelized correlation filter from [20].

A.17. Adaptive object region and Background weighted scaled Channel coded Distribution field tracker (ABCD)

Amanda Berg, Jörgen Ahlberg, Michael Felsberg
{*amanda.,jorgen.ahl,michael.fels*}*berg@liu.se*

The ABCD tracker is based on the Enhanced Distribution Field tracker [14]. In order to avoid background contamination of the object template, the ABCD tracker exploits background information for the online template update and it adaptively selects the object region used for tracking. Moreover, background information is also used to estimate object scale change.

A.18. Multi-Channel Multiple-Instance-Learning Tracker (CMIL)

Hilke Kieritz, Stefan Becker, Wolfgang Hubner, Michael Arens
{*hilke.kieritz, stefan.becker, wolfgang.huebner, michael.arens*}*@iosb.fraunhofer.de*

The Multi-Channel Multiple-Instance-Learning Tracker is a deterministic version of the MIL-Tracker ('Online Multiple Instance Learning Visual Tracker', [1]). Their work is extended by the use of multiple feature channels in compliance with the ICF person detector ('Integral Channel Features', [13]). Similar to the MIL-Tracker the appearance of the target object is learned via online multiple instance boosting and updated in each frame. This tracker uses a tracking-by-detection approach, where the classifier output is used to update the position. Different to the MIL-Tracker the Multi-Channel MIL-Tracker uses multiple features channels and only the sum of one region per feature. Developed to work in combination with a person detector similar to [13, 12] the Multi-Channel MIL-Tracker uses the same feature channels as the person detector: LUV-color channels, six per gradient direction quantized gradient magnitude channels and the gradient magnitude channel. To track the object over scale changes the feature responses are scaled using a scaling factor depended on the feature channel [12].

A.19. DTracker

Jae-Yeong Lee, Jae-chan Jeong, Sunglok Choi, Ji-Wan Kim, Jae-il Cho
{*jylee, channij80, sunglok, giraffe, jicho*}*@etri.re.kr*

DTracker extends the sumshift tracker [28] with an optical flow tracker and the NCC tracker. The color distribution of an object is modeled by kernel density estimation (KDE) to provide continuous measure of color similarity. Similarity evaluation of the KDE color model and the NCC template matching acts as global localizer to bound possible drift of the tracker and the optical flow tracker has a role of adopting frame to frame variation.

A.20. simplified Proposal Selection Tracker (sPST)

Yang Hua, Karteek Alahari, Cordelia Schmid
firstname.lastname@inria.fr

The simplified Proposal Selection Tracker (sPST) is based on our ICCV2015 paper [21]. sPST operates in two phases. Firstly, we propose a set of candidate object locations computed by tracking-by-detection framework [35], where we use the frame as is and rotate them according to the ground truth annotation in the initial frame if applicable. Secondly, we determine the best candidate as the tracking result by two cues: detection confidence score and an objectness measure computed with edges [48]. Note that the full version of our tracker uses additional proposals and motion boundaries calculated with optical flow. But it is not included in this submission due to the computational cost of the optical flow method. The reader is referred to [21] for details.

A.21. ASMS

Submitted by VOT Committee

The mean-shift tracker optimize the Hellinger distance between template histogram and target candidate in the image. This optimization is done by a gradient descend. The ASMS [40] method address the problem of scale adaptation and present a novel theoretically justified scale estimation mechanism which relies solely on the mean-shift procedure for the Hellinger distance. The ASMS also introduces two improvements of the mean-shift tracker that make the scale estimation more robust in the presence of background clutter – a novel histogram color weighting and a forward-backward consistency check.

A.22. Flock of Trackers (FoT)

Submitted by VOT Committee

The Flock of Trackers (FoT) [39] is a tracking framework where the object motion is estimated from the displacements or, more generally, transformation estimates of a number of local trackers covering the object. Each local tracker is attached to a certain area specified in the object coordinate frame. The local trackers are not robust and assume that the tracked area is visible in all images and that it undergoes a simple motion, e.g. translation. The Flock of Trackers object motion estimate is robust if it is from local tracker motions by a combination which is insensitive to failures.

A.23. Spatio-temporal context tracker (STC)

Submitted by VOT Committee

The STC [45] is a correlation filter based tracker, which uses image intensity features. It formulates the spatio temporal relationships between the object of interest and its locally dense contexts in a Bayesian framework, which models the statistical correlation between features from the target and its surrounding regions. For fast learning and detection the Fast Fourier Transform (FFT) is adopted.

References

[1] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 983–990, 2009.

[2] O. Barnich and M. V. Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):1709–1724, 2011.

[3] A. Berg, J. Ahlberg, and M. Felsberg. A thermal object tracking benchmark. In *12th IEEE International Conference on Advanced Video- and Signal-based Surveillance, Karlsruhe, Germany, August 25-28 2015*. IEEE, 2015.

[4] L. Bertinetto, M. O., J. Valmadre, G. S., and P. Torr. The importance of estimating object extent when tracking with correlation filters. *Preprint*, 2015.

[5] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[6] D. Chen, Z. Yuan, Y. Wu, G. Zhang, and N. Zheng. Constructing adaptive complex cells for robust visual tracking. In *Int. Conf. Computer Vision*, 2013.

[7] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Comp. Vis. Patt. Recognition*, volume 2, pages 142–149, 2000.

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

[9] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *Proc. British Machine Vision Conference*, 2014.

[10] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Learning spatially regularized correlation filters for visual tracking. In *Int. Conf. Computer Vision*, 2015.

[11] M. Danelljan, F. S. Khan, M. Felsberg, and J. Van de Weijer. Adaptive color attributes for real-time visual tracking. In *Comp. Vis. Patt. Recognition*, 2014.

[12] P. Dollar, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *Proc. British Machine Vision Conference*, volume 2, page 7, 2010.

[13] P. Dollar, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *Proc. British Machine Vision Conference*, volume 2, page 5, 2009.

[14] M. Felsberg. Enhanced distribution field tracking using channel representations. In *Vis. Obj. Track. Challenge VOT2013, In conjunction with ICCV2013*, 2013.

[15] R. Gade and T. B. Moeslund. Thermal cameras and applications: A survey. *Machine Vision & Applications*, 25(1), 2014.

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

[17] A. González, R. Martín-Nieto, J. Bescós, and J. M. Martínez. Single object long-term tracker for smart control of a PTZ camera. In *International Conference on Distributed Smart Cameras*, pages 121–126, 2014.

[18] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In D. N. Metaxas, L. Quan, A. Sanfeliu, and L. J. V. Gool, editors, *Int. Conf. Computer Vision*, pages 263–270. IEEE, 2011.

[19] J. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(3):125–141, 2014.

[20] J. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2015.

[21] Y. Hua, K. Alahari, and C. Schmid. Online object tracking with proposal selection. In *Int. Conf. Computer Vision*, 2015.

[22] J.-E. Kllhammer, D. Eriksson, G. Granlund, M. Felsberg, A. Moe, B. Johansson, J. Wiklund, and P.-E. Forssén. Near Zone Pedestrian Detection using a Low-Resolution FIR Sensor. In *Intelligent Vehicles Symposium, 2007 IEEE*, Intelligent Vehicles Symposium, Istanbul, Turkey, 2007. IEEE.

[23] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Čehovin, G. Fernández, T. Vojíř, G. Nebehay, R. Pflugfelder, and G. Hger. The visual object tracking vot2015 challenge results. In *ICCV workshop on VOT2015 Visual Object Tracking Challenge*, 2015.

[24] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin. A novel performance evaluation methodology for single-target trackers. *arXiv:1503.01313*, 2015.

[25] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Cehovin, G. Nebehay, G. Fernandez, and T. Vojir. The vot2013 challenge: overview and additional results. In *Computer Vision Winter Workshop*, 2014.

[26] 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 Workshops*, pages 98–111, 2013.

[27] M. Kristan, R. P. Pflugfelder, A. Leonardis, J. Matas, L. Cehovin, G. Nebehay, T. Vojir, G. Fernandez, A. Lukezi, A. Dimitriev, A. Petrosino, A. Saffari, B. Li, B. Han, C. Heng, C. Garcia, D. Pangersic, G. Hger, F. S. Khan, F. Oven, H. Possegger, H. Bischof, H. Nam, J. Zhu, J. Li, J. Y. Choi, J.-W. Choi, J. F. Henriques, J. van de Weijer, J. Batista, K. Lebeda, K. Ofjall, 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. YueYing 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 *Computer Vision - ECCV 2014 Workshops*, volume 8926 of *Lecture Notes in Computer Science*, pages 191–217. Springer, 2014.

[28] J.-Y. Lee and W. Yu. Visual tracking by partition-based histogram backprojection and maximum support criteria. In *Proceedings of the IEEE International Conference on Robotics and Biomimetic (ROBIO)*, 2011.

[29] 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.

[30] Y. Li and J. Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *Proceedings of the ECCV Workshop*, pages 254–265, 2014.

[31] H. Liu, X. Yang, L. J. Latecki, and S. Yan. Dense neighborhoods on affinity graph. *Int. J. Comput. Vision*, 98(1):65–82, 2012.

[32] 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.

[33] J. Ning, L. Zhang, D. Zhang, and C. Wu. Robust meanshift tracking with corrected background-weighted histogram. *IET Computer Vision*, 6(1):62–69, 2012.

[34] J. Shi and C. Tomasi. Good features to track. In *Comp. Vis. Patt. Recognition*, pages 593 – 600, June 1994.

[35] J. S. Supancic and D. Ramanan. Self-paced learning for long-term tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[36] M. Tang and J. Feng. Multi-kernel correlation filter for visual tracking. In *Int. Conf. Computer Vision*, 2015.

[37] L. Čehovin, M. Kristan, and A. Leonardis. Is my new tracker really better than yours? *WACV 2014: IEEE Winter Conference on Applications of Computer Vision*, 2014.

[38] L. Čehovin, A. Leonardis, and M. Kristan. Visual object tracking performance measures revisited. arXiv:1502.05803 [cs.CV], 2013.

[39] T. Vojir and J. Matas. The enhanced flock of trackers. In R. Cipolla, S. Battiato, and G. M. Farinella, editors, *Registration and Recognition in Images and Videos*, volume 532 of *Studies in Computational Intelligence*, pages 113–136. Springer Berlin Heidelberg, Springer Berlin Heidelberg, January 2014.

[40] T. Vojir, J. Noskova, and J. Matas. Robust scale-adaptive mean-shift for tracking. *Pattern Recognition Letters*, 49(0):250 – 258, 2014.

[41] Y. Wu, J. Lim, and M. Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2014.

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

[43] 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.

[44] J. Zhang, S. Ma, and S. Sclaroff. Meem: Robust tracking via multiple experts using entropy minimization. In *Comp. Vis. Patt. Recognition*, 2014.

[45] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang. Fast visual tracking via dense spatio-temporal context learning. In *Proc. European Conf. Computer Vision*, pages 127–141, 2014.

[46] G. Zhu, F. Porikli, and H. Li. Tracking randomly moving objects on edge box proposals. In *CoRR*, 2015.

[47] G. Zhu, J. Wang, Y. Wu, and H. Lu. Collaborative correlation tracking. In *Proc. British Machine Vision Conference*, 2015.

[48] C. L. Zitnick and P. Dollar. Edge boxes: Locating object proposals from edges. In *Proc. European Conf. Computer Vision*, pages 391–405, 2014.