A Thermal Object Tracking Benchmark

Amanda Berg, Jörgen Ahlberg and Michael Felsberg

Amanda Berg, Jörgen Ahlberg and Michael Felsberg, A Thermal Object Tracking Benchmark, 2015, 12th IEEE International Conference on Advanced Video- and Signal-based Surveillance.

A Thermal Object Tracking Benchmark

Amanda Berg[1,2], Jörgen Ahlberg[1,2], Michael Felsberg[2]
[1]Termisk Systemteknik AB, Diskettgatan 11 B, 583 35 Linköping, Sweden
[2]Computer Vision Laboratory, Dept. EE, Linköping University, 581 83 Linköping, Sweden
{amanda.,jorgen.ahl}berg@termisk.se, {amanda.,jorgen.ahl,michael.fels}berg@liu.se

Abstract

Short-term single-object (STSO) tracking in thermal images is a challenging problem relevant in a growing number of applications. In order to evaluate STSO tracking algorithms on visual imagery, there are de facto standard benchmarks. However, we argue that tracking in thermal imagery is different than in visual imagery, and that a separate benchmark is needed. The available thermal infrared datasets are few and the existing ones are not challenging for modern tracking algorithms. Therefore, we hereby propose a thermal infrared benchmark according to the Visual Visual Object Tracking (VOT) protocol for evaluation of STSO tracking methods. The benchmark includes the new LTIR dataset containing 20 thermal image sequences which have been collected from multiple sources and annotated in the format used in the VOT Challenge. In addition, we show that the ranking of different tracking principles differ between the visual and thermal benchmarks, confirming the need for the new benchmark.

1. Introduction

Tracking of objects in video is a problem that has been subject to extensive research [18]. Indicators of the popularity of the topic are challenges/contest like the recurring Visual Object Tracking (VOT) challenge [12, 13] the Online Object Tracking (OTB) benchmark [18], and the series of workshops on Performance Evaluation of Tracking and Surveillance (PETS) [20].

Thermal infrared tracking has historically been of interest mainly for military purposes. Thermal cameras have delivered noisy images with low resolution, useful mainly for tracking small objects (point targets) against colder backgrounds. In recent years, thermal cameras have decreased in both price and size while image quality and resolution has improved, which has opened up new application areas [7]. Thermal cameras are now commonly used, *e.g.*, in cars and in surveillance systems. The main advantages of thermal cameras are their ability to see in total darkness, their robustness to illumination changes and shadow effects, and less intrusion on privacy.

In spite of the popularity of tracking in visual video and availability of a range of benchmark datasets, the currently available thermal infrared datasets are either outdated or address other problems. As a consequence, many papers describing new thermal tracking methods perform the evaluation on proprietary sequences. This makes it hard to get an overview of the current status and advances within the field. Furthermore, tracking in thermal infrared video poses different challenges compared to tracking in visual video [22], hence, a separate benchmark is needed. For these reasons, we have prepared and will make publicly available a new thermal infrared benchmark for short-term single-object (STSO) tracking methods.

Contribution Our contribution is a publicly available benchmark, including a dataset of annotated thermal infrared image sequences to be used for comparing tracking methods. The benchmark contains previously available sequences as well as several newly recorded and annotated sequences for this specific purpose. The benchmark has been integrated in the VOT2015 Challenge.

2. Background and motivation

The infrared wavelength band is usually divided into different bands according to their different properties: near infrared (NIR, wavelengths 0.7–1 μm), shortwave infrared (SWIR, 1–3 μm), midwave infrared (MWIR, 3–5 μm), and longwave infrared (LWIR, 7.5–12 μm). Other definitions exist as well. These bands are separated by regions where the atmospheric transmission is very low (*i.e.*, the air is opaque) or where sensor technologies have their limits. LWIR, and sometimes MWIR, is commonly referred to as thermal infrared (TIR). TIR cameras should not be confused with NIR cameras that are dependent on illumination and in general behave in a similar way as visual cameras.

In thermal infrared, most of the captured radiation is *emitted* from the observed objects, in contrast to visual and

978-1-4673-7632-7/15/$31.00 ©2015 IEEE

near infrared, where most of the radiation is *reflected*. Thus, knowing or assuming material and environmental properties, temperatures can be measured using a thermal camera (*i.e.*, the camera is said to be *radiometric*).

Thermal cameras are either cooled or uncooled. High-end cooled cameras can deliver hundreds of HD resolution frames per second and have a temperature sensitivity of 20 mK. Images are typically stored as 16 bits per pixel to allow a large dynamic range, for example 0–382.2K with a precision of 10 mK. Uncooled cameras usually have bolometer detectors and operate in LWIR. They give noisier images at a lower framerate, but are smaller, silent, and less expensive. Some uncooled cameras provide access to the raw 16-bit (radiometric) intensity values, while others convert the images to 8 bits and compress them *e.g.* using MPEG. In the latter case, the dynamic range is adaptively changed in order provide an image that looks good to the eye, and the temperature information is lost.

2.1. Why is TIR tracking different?

There are two common beliefs regarding object tracking in TIR. One is that it is all about *hotspot tracking*, that is, tracking warm (bright) objects againts a cold (dark) background. In certain military applications, such as missile warning, this assumption is valid, but for most other applications the situation is more complex. The other is that TIR tracking is identical to tracking in grayscale visual imagery, and, as a consequence, that a tracker that is good for visual tracking is good for TIR tracking. However, there are differences between the two types of imagery that indicate that this is not the case.

First, there are no shadows in TIR. A tracker that is optimized to handle shadows might thus be suboptimal for TIR.

Second, the noise characteristics are different. Compared to a visual camera, a TIR camera typically has more blooming, lower resolution and a larger percentage of dead pixels. As a consequence, a tracker that depends heavily on (high resolution) spatial structure is presumably suboptimal for TIR imagery.

Third, visual color patterns are discernible in TIR only if they correspond to variations in material or temperature. Again, the consequence is that trackers relying on (high resolution) spatial patterns might be suboptimal for TIR imagery, and, moreover, re-identification and resolving occlusions might need to be done differently. For example, two persons with differently patterned or colored clothes might look similar in TIR.

Fourth, in most applications, the emitted radiation change much slower than the reflected radiation. That is, an object moving from a dark room into the sunlight will not immediately change its appearance (as it would in visual imagery). Thus, trackers that exploit the absolute levels (for example, distribution field trackers) should have an advantage in TIR. This is especially relevant for radiometric 16-bit cameras, since they have a dynamic range large enough to accommodate relevant temperature intervals without adapting the dynamic range to each frame.

In conclusion, we argue that:

1. Hotspot tracking is applicable for specific applications only; see Fig. 1c for an example where it is not.

2. Benchmarking trackers on visual and TIR imagery will give different results. We intend to show that by the evaluation described in Sec. 4.

3. TIR tracking should exploit different principles and image features compared to visual tracking. Trackers exploiting absolute values rather than spatial patterns will have an advantage in TIR. We intend to show that as well in Sec. 4.

2.2. Related work

A summary of currently available civilian TIR datasets for benchmarking of tracking methods is provided in Table 1. The most common datasets for evaluation of TIR tracking methods are the OTCBVS datasets [3, 4, 14]. They were published in 2005 and are characterized by low resolution, warm objects against cold backgrounds (*i.e.*, easily tracked objects) and few challenging events. Since then, both cameras and tracking techniques have advanced and the OCTBVS datasets have become outdated.

Another dataset that also mainly contains warm objects moving against cold backgrounds without any occlusions is the LITIV dataset [17]. The included sequences are heavily compressed, resulting in severe compression artifacts. Furthermore, there is no groundtruth for tracking.

The ASL-TID [15] dataset provides sequences simulating a thermal camera mounted on a UAV. This is the only publicly available dataset including sequences with a moving camera. The included sequences are of varying difficulty, high/low object resolution, cluttered backgrounds and occlusions. The dataset is primarily designed for object detection, not tracking.

Furthermore, only one of the existing datasets, the BU-TIV dataset [19], provides high-resolution 16-bit sequences captured with a cooled sensor. The purpose of the dataset is various visual analysis tasks, *i.e.*, it is not specifically designed for tracking.

3. Description of the benchmark

As mentioned, existing publicly available datasets for thermal infrared tracking have become outdated or do not address the specific task of short-term, single-object (STSO) tracking given an initial bounding box. Because there are too few available sequences suitable for this task,

Table 1: Properties of the available civilian datasets for benchmarking of TIR-tracking methods. Our proposed dataset (LTIR) is included in the comparison as well. Bpp is the number of bits per pixel, Stat/Mov if the camera is static or moving, and Vis if there are recordings in the visual domain of the same scenario.

Name	Purpose	Resolution	Bpp	Stat/Mov	Vis
OSU Pedestrian [3]	Pedestrian detection and tracking.	360×240	8	Y/N	N
OSU Color-Thermal [4]	Pedestrian detection, tracking and thermal/visual fusion.	360×240	8	Y/N	Y
Terravic Motion [14]	Detection and tracking	320×240	8	Y/N	N
LITIV [17]	Visible-infrared registration.	320×240	8	Y/N	Y
ASL-TID [15]	Object detection and tracking.	324×256	8/16	N/Y	N
BU-TIV [19]	Various visual analysis tasks. Single-object, multiple-object and multiple sensor tracking as well as motion patterns.	Up to 1024×1024	16	Y/N	N
Ours: LTIR	Short-term single-object tracking of different objects with varying challenging events.	Up to 1920×480	8/16	Y/Y	N

we hereby propose a thermal object tracking benchmark. The benchmark consists of a new thermal infrared dataset LTIR (Linköping Thermal InfraRed), object annotations, local and global attribute annotations, and an evaluation metric, as described below.

3.1. Dataset design criteria

The aim when designing the dataset was to fulfil a number of predefined criteria. An explicit purpose was, for example, to collect a dataset from various sources recorded with different sensors. Thus, several other owners or producers of thermal image sequences have been contacted and asked if they would like to contribute to the dataset. Further, the dataset should contain representative sequences of the presently most common application areas. Different environments, natural as well as man-made backgrounds, indoors as well as outdoors, should all be represented. Other aims were to include sequences recorded from different platforms (static, hand-held, moving, flying) as well as having various natures of the objects to track (humans, animals, non-deformable objects, objects on ground, objects that fly). Finally, the sequences should span the space of local and global attributes described below.

3.2. Data collection

Sequences to be included in the dataset have been collected from seven different sources using eight different types of sensors, see Table 2. 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 some of the sequences are available with both 8- and 16-bit pixel values. There are sequences from indoor and outdoor environments, and the sequences outdoors have been recorded in different weather conditions. The average sequence length is 563 frames. All included sequences are listed in Table 2 and described further below. The format of the included sequences and annotations have been standardized. The image data is stored as 8- or 16-bit PNG files, and the annotations are stored in ground truth text files which contain the corner coordinates of the bounding boxes, one row per frame. This format is in accordance with the VOT sequence- and annotation-format [13].

3.3. Included sequences

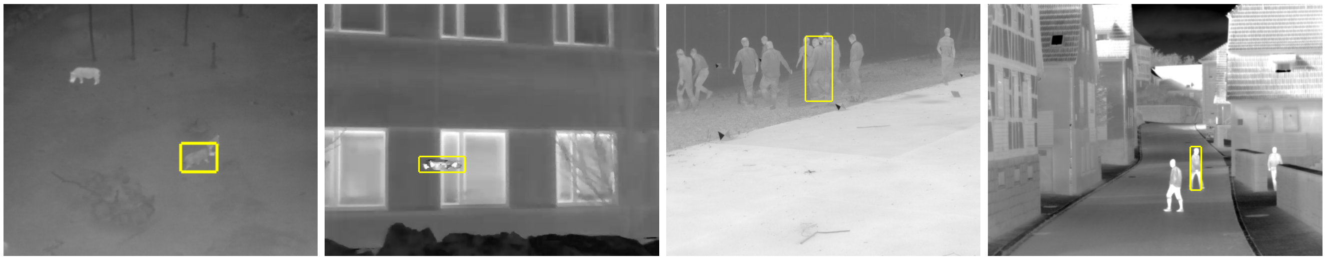
All included sequences found in Table 2 are further described below. In addition, example frames from four sequences are shown in Fig. 1.

1–2 The first two sequences, *rhino behind tree* and *running rhino*, contain natural background and are recorded from a UAV. The objects to track are two rhinoceros at the Kolmården Zoo. The sequences were provided by the Smart Savannah project at Linköping University.

3–4 *garden* and *horse* are both collected using a moving, hand-held camera. They contain natural background (with a few man-made elements) and the objects to track are a human and a horse respectively. The sequences originate from the ASL-TID [15] dataset but have been cut and annotations converted in accordance with the format.

5–6 *hiding* and *mixed distractors* were originally recorded by the School of Mechanical Engineering at University of Birmingham [16] for the purpose of thermal/visual fusion tracking. The sequences are recorded indoors and the object to track is a human.

7–8 *saturated* and *street* are recorded in the German test village of Bonnland by the Fraunhofer IOSB institute [11].



(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.

Table 2: Properties (name, sensor, resolution, no. of frames, no. of bits per pixel, tracked object and status) of all sequences included in the LTIR dataset. PUP = Previously Used in Publication, PPA = Previously Publicly Available.

ID	Name	Sensor	Resolution	#Frames	Bpp	Object	PUP/PPA
1	rhino behind tree	FLIR A35	320 × 256	619	8/16	Rhinoceros	N/N
2	running rhino	FLIR A35	320 × 256	763	8/16	Rhinoceros	N/N
3	garden	FLIR Tau 320	324 × 256	676	8/16	Human	Y/Y
4	horse	FLIR Tau 320	324 × 256	348	8/16	Horse	Y/Y
5	hiding	FLIR Photon 320	320 × 240	358	8	Human	Y/N
6	mixed distractors	FLIR Photon 320	320 × 240	270	8	Human	Y/N
7	saturated	AIM QWIP	640 × 480	218	8	Human	Y/N
8	street	AIM QWIP	640 × 480	172	8	Human	Y/N
9	car	FLIR A655SC	640 × 480	1420	8/16	Car	N/N
10	crouching	FLIR A655SC	640 × 480	618	8/16	Human	N/N
11	crowd	FLIR A65	640 × 512	71	8/16	Human	N/N
12	soccer	3×AXIS Q-1922	1920 × 480	775	8	Human	Y/N
13	birds	FLIR T640	640 × 480	270	8	Human	N/N
14	crossing	FLIR A655SC	640 × 480	301	8/16	Human	N/N
15	depthwise crossing	FLIR A655SC	640 × 480	851	8/16	Human	N/N
16	jacket	FLIR A655SC	640 × 480	1451	8/16	Human	N/N
17	quadrocopter	FLIR T640	640 × 480	178	8	Quadrocopter	N/N
18	quadrocopter2	FLIR A655SC	640 × 480	1010	8/16	Quadrocopter	N/N
19	selma	FLIR A655SC	640 × 480	235	8/16	Dog	N/N
20	trees	FLIR A655SC	640 × 480	665	8/16	Human	N/N

The background is urban and the objects to track are humans. In the *saturated* sequence the human pixels are saturated, implying that there are no spatial structure that can be utilized for tracking.

9–11 Sequences *car*, *crouching* and *crowd* were recorded by the EU FP7 project P5[1] at an undisclosed location in the UK. *Car* contains large variations in scale and viewpoint, *crowd* has cluttered background and *crouching* includes both occlusions and changes in aspect ratio.

12 *soccer* is a panorama of three static cameras provided by Aalborg University. The sequence was originally recorded for the purpose of evaluating tracking of sports players [8].

13–20 The final eight sequences have been provided by the company Termisk Systemteknik AB. The sequences mainly address surveillance applications where different objects (human, quadrocopter, dog) are to be tracked. Two sequences, *birds* and *quadrocopter* were recorded using a moving, hand-held camera.

3.4. Benchmark annotations

All benchmark annotations have been performed in accordance with the VOT2013 annotation process [12]. One object within each sequence has been 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 anno-

[1] http://foi.se/P5/

tations, global attributes have been per-sequence annotated and local attributes per-frame annotated.

Global attributes The per-sequence global attributes from VOT had to be adapted to the properties of thermal infrared 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 is 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.

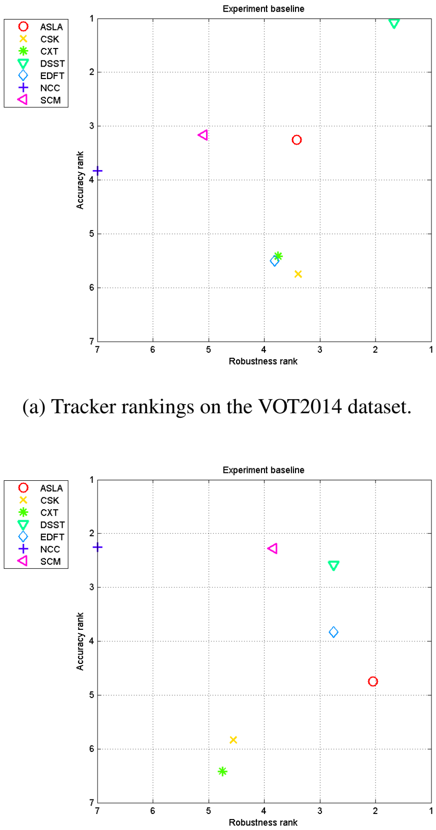
3.5. Evaluation methodology

We have chosen to use the same evaluation methodology as in the VOT Challenge. In short, each tracker is evaluated in terms of accuracy and robustness, and then ranked (*i.e.*, the best tracker getting rank 1, and so on). In order to compensate for individual sequences having an impact on tracker rankings, the results are weighted so that the attributes above should have equal significance. The final result is an accuracy and a robustness ranking for each tracker.

4. Evaluation

In order to evaluate whether benchmarking of trackers on visual and TIR imagery will give different results, seven different trackers have been evaluated on the VOT 2014 and LTIR datasets respectively. The VOT toolkit[2] was used to perform the evaluation.

The correlation coefficients for the local attribute rankings can be found in Table 3 and the accuracy/robustness-plots (AR-plots) in Fig. 2. The correlation coefficients and distribution of tracker ranking results in the AR-plots indicates that the ranking between trackers are not equal for the two benchmarks. Furthermore, the mean ranking change is 1.37, *i.e.*, each tracker has on average changed its ranking with 1.37 between visual and thermal.



(a) Tracker rankings on the VOT2014 dataset.

(b) Tracker rankings on the LTIR dataset.

Figure 2: AR-plots of ranking results for seven trackers evaluated on the (a) VOT2014 and (b) LTIR datasets.

	Acc.	Rob.	Acc + Rob
ASLA [10]	0.55	-0.37	0.03
CSK [9]	0.47	0.27	0.66
CXT [5]	0.60	0.65	0.81
DSST [2]	0.19	-0.56	-0.29
EDFT [6]	0.55	-0.56	0.48
NCC [1]	0.54	1.00	0.85
SCM [21]	-0.61	0.33	0.21

Table 3: Correlation coefficients for ranking of different trackers.

When inspecting the two AR-plots in Fig. 2, the DSST [2] tracker is not as superior to the others in thermal as it is in visual. This is probably due to the fact that DSST partly relies on HOG-features and there is less (high-resolution) spatial structure in thermal compared to visual. ASLA [10]

[2] https://github.com/vicoslab/vot-toolkit

has a lower accuracy ranking while SCM [21] has a higher one. EDFT [6] receives an increased ranking in both accuracy and robustness. In Table 3, it is confirmed that DSST, ASLA and SCM have the least correlated ranking results.

5. Conclusion

We have described a new benchmark for evaluation of short-term single-object tracking methods on thermal infrared imagery. The benchmark consists of: a) 20 sequences originating from multiple sources. b) Object annotations. c) Local and global attribute annotations. d) An evaluation metric (same as used in VOT). The LTIR dataset and annotations can be downloaded from `http://www.cvl.isy.liu.se/en/research/datasets/ltir/`.

We have also performed an evaluation of seven tracking methods designed for visual sequences on both visual and thermal data. The evaluation shows that different methods are optimal for visual and thermal tracking respectively. Trackers based on spatial structure and/or sparse representations (ASLA, DSST, SCM) are ranked better on visual imagery than they are on thermal imagery. For the tracker based on distribution of pixel values (EDFT), it is the other way around.

Future work include a more detailed analysis of which features/principles are beneficial for thermal tracking. Considering that the rescaling ability of ASLA, SCM and DSST is lacking in EDFT, our next step would be to investigate how distribution-based tracker with rescaling ability would perform on thermal imagery.

6. Acknowledgements

The research was funded by the Swedish Research Council through the project Learning Systems for Remote Thermography, grant no. D0570301, as well as by the European Community Framework Programme 7, Privacy Preserving Perimeter Protection Project (P5), grant agreement no. 312784. We also gratefully acknowledge all who has contributed with sequences to the LTIR dataset.

References

[1] K. Briechle and U. D. Hanebeck. Template matching using fast normalized cross correlation. In *Proc. SPIE*, 4387:95–102, 2001.

[2] M. Danelljan et al. Accurate scale estimation for robust visual tracking. In *BMVC*, 2014.

[3] J. W. Davis and M. A. Keck. A two-stage template approach to person detection in thermal imagery. In *WACV/WMVC*, 1:364–369, 2005.

[4] J. W. Davis and V. Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *CVIU*, 106(2-3):162–182, 2007.

[5] T. B. Dinh et al. Context tracker: Exploring supporters and distracters in unconstrained environments. In *CVPR*, pp. 1177–1184, 2011.

[6] M. Felsberg. Enhanced Distribution Field Tracking using Channel Representations. In *ICCV Workshops*, pp. 121–128, 2013.

[7] R. Gade and T. Moeslund. Thermal cameras and applications: A survey. *Mach. Vis. & App.*, 25(1), 2014.

[8] R. Gade and T. B. Moeslund. Thermal tracking of sports players. *Sensors*, 14(8):13679–13691, 2014.

[9] J. a. F. Henriques et al. Exploiting the circulant structure of tracking-by-detection with kernels. In *ECCV*, IV:702–715, 2012.

[10] X. Jia, H. Lu, and M.-H. Yang. Visual tracking via adaptive structural local sparse appearance model. In *CVPR*, pp. 1822–1829, 2012.

[11] K. Jüngling and M. Arens. Local feature based person detection and tracking beyond the visible spectrum. In R. Hammoud et al. (eds), *Machine Vision Beyond Visible Spectrum*, pp. 3–32, Springer 2011.

[12] M. Kristan et al. The Visual Object Tracking VOT2013 challenge results. In *ICCV Workshops*, 2013.

[13] M. Kristan et al. The Visual Object Tracking VOT2014 challenge results. In *VOT 2014 (ECCV Workshop)*, pp. 1–27, 2014.

[14] R. Miezianko. *IEEE OTCBVS WS series bench; Terravic research infrared database.*

[15] J. Portmann et al. People Detection and Tracking from Aerial Thermal Views. In *ICRA*, 2014.

[16] M. Talha and R. Stolkin. Particle filter tracking of camouflaged targets by adaptive fusion of thermal and visible spectra camera data. *Sensors*, 14(1):159–166, 2014.

[17] A. Torabi et al. An iterative integrated framework for thermal-visible image registration, sensor fusion, and people tracking for video surveillance applications. *Comp. Vis. and Im. Underst.*, 116(2):210 – 221, 2012.

[18] Y. Wu et al. Online object tracking: A benchmark. In *CVPR*, pp. 2411–2418, 2013.

[19] Z. Wu et al. A thermal infrared video benchmark for visual analysis. In *CVPR Workshops*, 2014.

[20] D. P. Young and J. M. Ferryman. PETS metrics: Online performance evaluation service. In *ICCCN*, pp. 317–324, 2005.

[21] W. Zhong. Robust object tracking via sparsity-based collaborative model. In *CVPR 2012*, pp. 1838–1845.

[22] E. Gundogdu. Comparison of Infrared and Visible Imagery for Object Tracking: Toward Trackers with Superior IR Performance. In *CVPR Workshops*, 2015.