Performance analysis and formative assessment of visual trackers using PETS critical infrastructure surveillance datasets

Longzhen Li[1], **Tahir Nawaz**[1], **James Ferryman**[1,*]
[1]Computational Vision Group, Department of Computer Science, University of Reading, Reading, RG6 6AF, United Kingdom

Abstract. Surveillance of perimeter of critical infrastructures (CIs) has gained a particular attention worldwide due to rising threats of terrorist incidents. However, the generation and dissemination of real-world challenging datasets concerning CIs for performance assessment of surveillance tasks, particularly tracking, has been widely overlooked in the community. Recent PETS workshops have been aimed to address this issue by introducing ample CI surveillance datasets. This paper presents effectiveness of these publicly released real-world PETS CI visual datasets by providing a comprehensive statistically significant performance analysis as well as formative assessment of several state-of-the-art multi-target trackers using well-known and recent performance assessment criteria.

Keywords: surveillance, critical infrastructure, visual tracking, performance evaluation.

*James Ferryman, j.m.ferryman@reading.ac.uk

1 Introduction

Video surveillance is a widely-researched area in computer vision field.[1,2] In particular, the surveillance of perimeter of critical infrastructures (CIs) has particularly gained an enhanced importance worldwide due to increased threats of terrorist incidents in the recent past.[3] Therefore, the need remains to devise robust surveillance systems that enable effective monitoring of the region outside the perimeter of a CI and generate an early warning before a threat has approached.

Generally, one of the key tasks in surveillance applications is to reliably perform simultaneous visual tracking of multiple targets over time, multi-target tracking. Indeed, for several years, a substantial research has been aimed at devising robust multi-target tracking algorithms to deal with different scenarios and varying challenges.[4–11] Inevitably, in order to assess the effectiveness

(PRE-PRINT) L. Li, T. Nawaz, J. Ferryman. "Performance analysis and formative assessment of visual trackers using PETS critical infrastructure surveillance datasets", Journal of Electronic Imaging 28(4), 043004 (2019), doi: 10.1117/1.JEI.28.4.043004.

and suitability of trackers in CI surveillance scenarios, the availability of appropriate datasets is of paramount importance.

For nearly two decades, the goal of the Performance Evaluation of Tracking and Surveillance (PETS) workshops has been to foster the emergence of computer vision technologies particularly for tracking by providing a plethora of datasets and evaluation metrics that allow an accurate assessment and comparison of such methods.[12] In recent years, PETS workshops (PETS'14, PETS'15, PETS'16, PETS'17, PETS'18) have had a special focus in disseminating to community the real-world surveillance datasets concerning the protection of CIs, which has otherwise been lacking in other similar evaluation campaigns and datasets.

This paper presents a thorough performance analysis and formative assessment of several state-of-the-art multi-target visual trackers on real-world challenging datasets released generally as a part of PETS'14-18 workshops and particularly as a part of PETS'15 tracking challenge to enable development and testing of intelligent surveillance systems, which work robustly under a wide range of challenges and conditions for CI perimeter protection (e.g. power plants, prisons, communication sites).

The remainder of this paper is organized as follows. Section 2 reviews the related work, which is followed by description of datasets in Sec. 3 and trackers in Sec. 4. The experimental validation is presented in Sec. 5. The paper is finally concluded in Sec. 6.

2 Related work

The challenges in creating benchmark datasets for the performance evaluation of automated visual surveillance methods are broad.[12,13] The main aim of automated surveillance is frequently to locate and track objects of interest or to determine specific events and/or behaviors involving

the objects and/or the environment. In creating datasets whose content may be analyzed by algorithms/systems, these objects, events and behaviors must be presented within the recorded scenarios in a realistic and meaningful way. These range from, but are not limited to, varying scene conditions such as weather and illumination/lighting (including moving shadows and reflections) to the number (density), size and dynamics of objects present within the monitored scene. Such variation may be captured as a range of recorded scenarios, with increasing levels of complexity. Since the drive behind creation of such benchmark datasets is to evaluate the performance of developed surveillance methods/systems, attention must be paid as to how the evaluation will be carried out when recording the scenarios.[12] The creation of ground truth for the evaluation of detection, tracking and event/behavior analysis can be extremely time consuming and therefore the scenario content and length, as well as the level(s) of annotation to be made, must be carefully considered.[14] Furthermore, in addition to producing a dataset that may be used as an evaluation benchmark for a broad spectrum of developed automated surveillance methodology, adopting well known and established metrics may assist more readily the researchers tasked to establish comparisons in the performance of state of the art visual surveillance systems.[15,16]

Over the years, several evaluation campaigns have been introduced, providing a wealth of datasets to facilitate testing and evaluation of video surveillance algorithms.[17] These include the Context Aware Vision using Image-based Active Recognition (CAVIAR) project[1], Evaluation du Traitement et de l'Interpretation de Sequences vidEO (ETISEO)[2], imagery Library for Intelligent Detection Systems (i-LIDS)[3], Classification of Events, Activities and Relationships (CLEAR)[4],

[1] http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/. Accessed May 2019.
[2] http://www-sop.inria.fr/orion/ETISEO/index.htm. Accessed May 2019.
[3] http://www.ilids.co.uk. Accessed May 2019.
[4] http://www.clear-evaluation.org/. Accessed May 2011.

Visual Object Tracking (VOT) challenge[5], Multiple Object Tracking (MOT) challenge[6], and PETS workshops. CAVIAR focused on indoor city surveillance scenarios and made contributions by releasing several datasets for building entrance lobby and shopping mall scenes. ETISEO provided numerous datasets covering indoor and outdoor scenarios (building corridor and entrance, road, parking areas for cars and aircrafts, metro) for the evaluation of different surveillance tasks including detection, tracking, classification, and event recognition. i-LIDS was another important evaluation campaign that was introduced by the UK's Centre for Applied Science and Technology (CAST) in collaboration with the Centre for the Protection of National Infrastructure (CPNI). It focused on the evaluation of surveillance systems on numerous datasets including real-world CCTV video footage for scenarios such as underground station, traffic, and airport. CLEAR was aimed at evaluating detection, tracking, person identification, head pose estimation, and acoustic event classification algorithms. In collaboration with the Computers in the Human Interaction Loop (CHIL) project[18] and i-LIDS, CLEAR made the dataset available online that covered indoor (meeting room, lecture room) and outdoor (traffic) scenarios. VOT challenge is dedicated for the evaluation of single-target trackers and being conducted since 2013 using the existing commonly-used datasets covering varying scenarios including indoor scenes, sport, concert, and traffic. Unlike the VOT challenge, MOT is aimed at the evaluation of multi-target trackers using commonly-used existing datasets. MOT challenge is being organized since 2014. PETS is probably the longest running evaluation campaign for tracking as well other surveillance tasks such as people counting, hand posture classification, and event detection. Since 2000, PETS workshops are being organized while providing a large amount of datasets for a wide variety of scenarios and applications.

[5]http://www.votchallenge.net/index.html. Accessed May 2019.
[6]https://motchallenge.net/. Accessed May 2019.

There is however still an absence of ample publicly available challenging datasets for the evaluation of approaches on realistic CI perimeter scenarios. Those researching the creation of automated visual surveillance systems require benchmark datasets that provide realistic settings, environmental conditions and scenarios. It can be very time consuming to create enough scenarios to adequately test some approaches.

3 Datasets

Two challenging datasets concerning real-world CI perimeter protection scenarios released as part of recent PETS workshops are as follows. The *ARENA dataset* is a multi-sensor dataset and includes a selection of video sequences that were recorded as part of the EU project ARENA[7], which addresses the design of a flexible surveillance system to enable situational awareness and determination of potential threats on critical mobile assets in transit. The *P5 dataset* is a multi-modal multi-sensor dataset recorded as part of the EU project P5[8] and involves different staged activities around the perimeter of a nuclear plant. The selected scenarios from ARENA and P5 datasets are grouped into 'Normal', 'Warning' and 'Alarm' categories. 'Normal' alludes to activities that do not pose any threat. 'Warning' refers to abnormal activities that may potentially develop into a threat. 'Alarm' refers to activities that cause a threat in the scene and hence require immediate action.

Between ARENA and P5 datasets, varying illumination and weather conditions are covered. ARENA dataset has been recorded under clearer sunny weather conditions, whereas P5 dataset has been recorded under more adverse (cloudy/rainy) weather conditions with a comparatively poorer

[7] http://www.arena-fp7.eu/. Accessed May 2019.
[8] http://p5-fp7.eu/. Accessed May 2019.

Table 1 Sensor properties for ARENA dataset

ID	Model	Resolution (pxl)	Frame Rate
ENV_RGB_3	PTZ Axis 233D	768x576	7
TRK_RGB_1	Basler BIP2-1300c-dn	1280 x 960	30
TRK_RGB_2	Basler BIP2-1300c-dn	1280 x 960	30

visibility. For visualisation of datasets, please visit the webpage[9]. The datasets are described next.

3.1 ARENA dataset

The ARENA dataset comprises of a series of multi-camera video recordings where the main subject is detection and understanding of human behavior around a parked vehicle. The main objective is to detect and understand different behaviors from visual (RGB) cameras mounted on the vehicle itself.

One visual camera ENV_RGB_3 is used (Table 1) that is installed at the location as shown in Fig. 1 to cover an approximate area of 100m x 30m. This camera is not onboard the CI (i.e. a vehicle carrying assets) but mounted in the environment to provide a global view of the monitored area. Moreover, two onboard non-overlapping visual cameras (TRK_RGB_1, TRK_RGB_2) are mounted at corners of a truck in ARENA dataset at the locations shown in Fig. 1. Table 1 lists sensors while describing their respective characteristics. The dataset scenarios ('Normal', 'Warning', 'Alarm') are listed in Table 2.

3.2 P5 dataset

The dataset contains sequences with different activities staged around the perimeter of a nuclear power plant in Sweden. The dataset was recorded using multiple types of surveillance sensors including visual and thermal cameras.

[9] http://www.cvg.reading.ac.uk/PETS2015/a.html. Accessed May 2019.

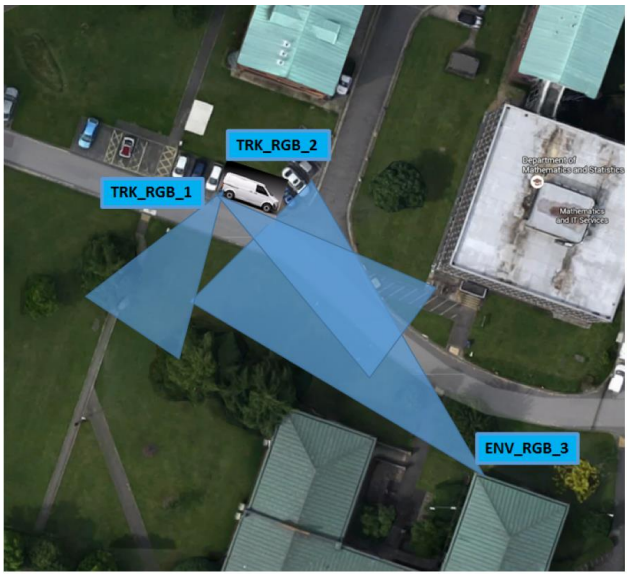


Fig 1 Sensor locations and their FOVs for ARENA dataset

There are five visual and thermal sensor positions covering a large area with 550m from one end to the other on the land side (see Fig. 2)[10]. It takes 10-15 minutes to walk from one end to the other. Three visual cameras (VS_1, VS_2, VS_3) at the locations shown in Fig. 2 are selected to mainly cover the road along the water area. Most of the scenarios take place in the monitored region. Two of the thermal sensors (TH_3, TH_4) as shown in Fig. 2 are installed side by side with visual cameras, with the aim to provide similar field of views (FOVs) to that of visual cameras. The main benefit of the joint use of thermal and visible sensors is that different modalities provide complementary information of the scene captured by thermal infrared spectrum and visible light spectrum, respectively. Two more thermal sensors TH_1 and TH_2 are installed at the locations shown in Fig. 2, which mainly cover the long road along the fence outside the nuclear plant. The sensor properties are summarized in Table 3. The dataset scenarios ('Normal', 'Warning', 'Alarm') are listed in Table 4.

[10]Note that the focus of this study is visual tracking only.

Table 2 List and description of scenarios for ARENA dataset. Key. SC: scale changes; Occ: occlusions; PC: pose changes; PR: person running; Cl: clutter; VS: varying speed.

Scenario type	ID	Description	Challenges
Normal	N1_ARENA	Persons walking in a group	SC, Occ, PC
Warning	W1_ARENA	Driver falls after being hit by someone	Occ, SC, PR
Alarm	A1_ARENA	Driver involved in a fight with someone	SC, PC, Occ, Cl
	A2_ARENA	Driver attacked by someone from a car	SC, VS, Occ

4 Trackers

For nearly two decades, a large number of multi-target visual trackers have been proposed to deal with varying challenges.[16] Among them, those trackers that use the so-called tracking-by-detection paradigm have shown very promising results,[4,5,7,10] partly thanks to the great improvements in detection accuracy using pre-trained object models.[19,20] In this study we specifically use the following state-of-the-art trackers that have been published in prestigious venues: LP2D,[4] DP-NMS,[5] SORT,[7] ELP,[10] CEM,[9] DCO,[6] DCO-X,[8] and SegTrack.[11] We used publicly available software implementations of trackers with default parameters to generate their tracks. Additionally, for a fair comparison, all trackers are fed with the same detection results which are obtained by running the well known DPM detector.[20] Two types of pre-trained models are used to generate detection bounding boxes: one for *person* targets and the other one for *vehicle* targets. The initial detections are pruned by applying non-max suppression (NMS) with the same threshold value as in the

Fig 2 Sensor locations and their FOVs for P5 dataset

Table 3 Sensor properties for P5 dataset (VS: Visual, TH: Thermal)

ID	Model	Resolution (pxl)	Frame Rate
VS_1	Basler BIP2-1300c-dn	1280 x 960	25
VS_2	Basler BIP2-1300c-dn	1280 x 960	15
VS_3	Basler BIP2-1300c-dn	1280 x 960	25
TH_1	FLIR SC655	640x480	25
TH_2	FLIR SC655	640x480	12.5
TH_3	FLIR SC655	640x480	25
TH_4	FLIR A65	640x512	30

original work.[5]

The selected trackers generally formulate the task of tracking multiple targets as a data association problem: it links a set of target hypotheses produced by object detector at a frame level into a set of consistent trajectories each with a unique ID. Due to the discrete nature of the distinct detections in frame and the disjunct frames in video sequence, a Directed Acyclic Graph (DAG) is normally applied in these tracking frameworks to model the entire spatio-temporal space in a consistent way. Various optimization strategies can then be applied thereafter to solve for globally optimal solution, ranging from using linear programming solved with the simplex algorithm,[4] the dynamic programming with non-maximal suppression (termed DP-NMS),[5] the classical Joint Probabilistic Data Association technique,[7] or the minimum-cost network flow algorithm,[10] to the energy minimization solution using conjugate gradient descent[9] or the conditional random field.[6,8,11] Additionally, different image features and/or dynamic models have also been utilized to handle more complicated tracking challenges[4–11] especially when the scene gets crowded or when human interactions are involved.

Table 4 List and description of scenarios for P5 dataset. Key. SC: scale changes; PC: pose changes; VS: varying speed; Cl: clutter; Occ: occlusions; PV: poor visibility.

Scenario type	ID	Description	Challenges
Normal	N1_P5	A vehicle driving across the scene	SC, PC, VS, Cl
Warning	W1_P5	A group of 6 people walking across the scene	Occ, PV, SC, Cl, VS
Alarm	A1_P5	An abandoned bag is picked up suspiciously	SC, PC, Cl, VS

5 Experimental validation

This section presents the experimental validation by first describing the performance assessment criteria (Sec. 5.1) and test sequences (Sec. 5.2), which are to be used in the experimental analysis of trackers (Sec. 5.3). The section ends with a discussion (Sec. 5.4).

5.1 Performance assessment criteria

To enable a precise quantitative performance assessment and ranking of various tracking algorithms, an accurate and detailed annotations of datasets have been generated. Indeed, the ground truth is obtained for every single frame of the sequences used. The annotation is obtained in the format of bounding box that effectively encloses each object in every frame. Note that, in the case of occlusion, only the visible part of the object is annotated. Next we describe the tracking assessment criteria that use the generated ground truth information to quantify performance in this paper.

Performance evaluation of tracking algorithms is a non-trivial task. In fact, for a thorough tracking assessment, several aspects needs to be assessed for which numerous metrics have been proposed over the years.[15,16,21–23] The choice of approrpiate metrics is quite challenging and could depend mainly on the application under consideration. Tracking evaluation accounts for the three key aspects including tracking accuracy (extent of match between an estimation and the corresponding ground truth), cardinality error (difference between the number of estimated targets and the number of ground-truth targets) and ID change (wrong associations between estimated and ground-truth targets).[16] We used the widely-used Multiple Object Tracking Accuracy (MOTA),[15] which takes into account the *cardinality error* (in the form of false positives and false negatives)

and *ID changes* without explicitly considering accuracy. MOTA is defined as follows:

$$\text{MOTA} = 1 - \frac{\sum_{k=1}^{K}(c_1|FN_k| + c_2|FP_k| + c_3|IDC_k|)}{\sum_{k=1}^{K} v_k}, \tag{1}$$

where the parameters c_1, c_2 and c_3 determine the contributions from the number of false negatives ($|FN_k|$), number of false positives ($|FP_k|$) and number of ID changes ($|IDC_k|$) at a frame k, respectively, and v_k is the number of ground-truth targets at frame k. $c_1 = 1, c_2 = 1, c_3 = \log_{10}$ as described in the paper.[24] False negatives are the missed targets at frame k and false positives are the estimated targets with overlap $O_{k,t} < \bar{\tau}$ for a pre-defined threshold, $\bar{\tau}$, such that

$$O_{k,t} = \frac{|\bar{A}_{k,t} \cap A_{k,t}|}{|\bar{A}_{k,t} \cup A_{k,t}|}, \tag{2}$$

for a tth pair of ground-truth and estimated bounding boxes at frame k. $\bar{A}_{k,t}$ and $A_{k,t}$ denote the occupied regions on image plane for ground-truth and estimated bounding boxes, respectively. $\bar{\tau}$ is often set to 0.5.[25] MOTA ≤ 1: the higher MOTA, the better the performance. To evaluate tracking *accuracy*, a more recent measure, Multiple Extended-target Lost-Track ratio (MELT),[16] is used. MELT provides accuracy evaluation using information about lost-track ratio. Let N_i be the total number of frames in ith ground-truth track and N_i^{τ} is the number of frames with overlap score below a threshold τ, then the lost-track ratio λ_i^{τ} is computed as follows:

$$\lambda_i^{\tau} = \frac{N_i^{\tau}}{N_i}. \tag{3}$$

MELT for a particular τ is computed as follows:

$$\text{MELT}_\tau = \frac{1}{V}\sum_{i=1}^{V}\lambda_i^\tau, \tag{4}$$

where V is the total number of ground-truth tracks, and

$$\text{MELT} = \frac{1}{S}\sum_{\tau\in[0,1]}\text{MELT}_\tau, \tag{5}$$

provides the overall tracking accuracy for a full variation of τ, where S is the number of sampled values of τ. MELT $\in [0,1]$: the lower the value the better the performance.

The set of measures described above primarily focus on evaluating end performance that is important particularly for ranking trackers. To obtain a deeper insight as why a certain end performance is achieved, it would also be desirable to analyze the factors (false positives, false negatives, ID changes) that contribute to the attainment of a certain end performance.[26] Therefore, to complement the evaluation using above metrics and further aid the performance analysis, we also adopt a recently proposed method[26] that enables revealing a dissected picture of the performance of trackers based on the analysis of probability density functions (PDFs) of different *fault types* (i.e. false positives, false negatives, ID changes) in a sequence. Inspired from analysis of PDFs, the method offers a more detailed picture of a tracker's performance by revealing two aspects: tracker's *robustness* and *per frame concentration* corresponding to each fault type, both are quantitatively accounted for in the form of following two performance scores. The first score tells the ability of a tracker to track without producing a fault across a sequence, and is called *robustness to a fault*

type (R):

$$R_{fp} = 1 - \frac{K_{fp}}{K};\ R_{fn} = 1 - \frac{K_{fn}}{K};\ R_{idc} = 1 - \frac{K_{idc}}{K}; \tag{6}$$

such that K_{fp} is the number of frames containing false positive(s), K_{fn} is the number of frames containing false negative(s), and K_{idc} is the number of frames containing ID change(s). $R_{fp} \in [0,1]$, $R_{fn} \in [0,1]$, $R_{idc} \in [0,1]$: the higher the value (R_{fp} / R_{fn} / R_{idc}), the better the ability. The second score tells the tendency of a tracker to produce a fault type per frame, and is called *per frame concentration of a fault type* (PFC):

$$PFC_{fp} = \frac{1}{K}\sum_{k=1}^{K} FP_k; PFC_{fn} = \frac{1}{K}\sum_{k=1}^{K} FN_k; \tag{7}$$
$$PFC_{idc} = \frac{1}{K}\sum_{k=1}^{K} IDC_k,$$

where FP_k, FN_k and IDC_k are the number of false positives, false negatives and ID changes, respectively, at frame k of a sequence that has a total of K frames.

5.2 Sequences

Table 5 provides a summary of the sequences (S1 to S11) that are selected from both P5 and ARENA datasets. The sequences belong to normal, warning and alarm scenarios, and cover key tracking challenges including occlusions, scale changes, pose changes, clutter, poor visibility, and varying target speed. Note that this paper focuses on visual tracking only, with vehicle and person target types. Figures 3 and 4 show visualization of camera views as used in this study for ARENA and P5 datasets, respectively.

Table 5 Summary of sequences.

Sequence	ID	No. of frames	No. of targets	Target type
S1	N1_P5-Tg-VS_1	400	1	Vehicle
S2	N1_P5-Tg-VS_3	387	1	Vehicle
S3	W1_P5-Tg-VS_3	180	6	Person
S4	W1_P5-Tg-VS_1	180	6	Person
S5	N1_ARENA-Tg_ENV_RGB_3	289	5	Person
S6	N1_ARENA-Tg_TRK_RGB_1	513	5	Person
S7	N1_ARENA-Tg_TRK_RGB_2	684	5	Person
S8	W1_ARENA-Tg_ENV_RGB_3	155	3	Person
S9	W1_ARENA-Tg_TRK_RGB_1	240	3	Person
S10	A1_ARENA-Tg_ENV_RGB_3	295	4	Person
S11	A1_ARENA-Tg_TRK_RGB_2	670	4	Person

5.3 Results and analysis

Figures 5, 6 and 7 show performance scores of all trackers on every sequence (S1-S11) using different assessment criteria (MOTA, MELT, R_{fp}, R_{fn}, R_{idc}, PFC_{fp}, PFC_{fn}, PFC_{idc}). On S1 CEM, DCO, DCO-X and DP-NMS consistently achieves the best performance based on most assessment criteria (Fig. 5, 6 and 7). The worst performance on S1 has however been mostly reported by either SegTrack or LP2D. This is apparently due to presence of clutter in S1 that has caused distractions for SegTrack and LP2D (see Fig. 8(a,b)). S2 has substantial background clutter, pose changes and scale changes of target (Fig. 8(c,d)). DP-NMS has coped with these challenges comparatively better than other trackers to achieve the best performance based on most assessment criteria, although

(a) (b) (c)

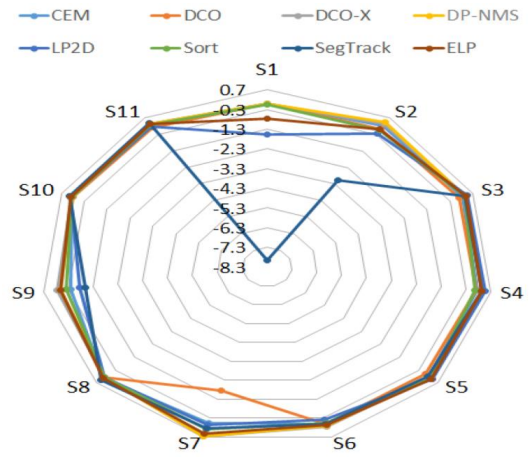
Fig 3 Visualization of ARENA camera views under consideration. (a) ENV_RGB_3; (b) TRK_RGB_1; (c) TRK_RGB_2.

(a) (b)

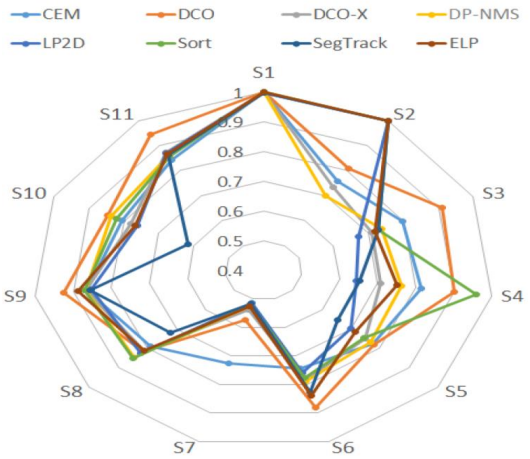
Fig 4 Visualization of P5 camera views under consideration. (a) VS_1; (b) VS_3.

the tracker still has difficulty maintaining the unique target ID across the sequence (see Fig. 8(c,d)).

S3 is particularly challenging due to severe occlusions and poor visibility due to cloudy/rainy con-

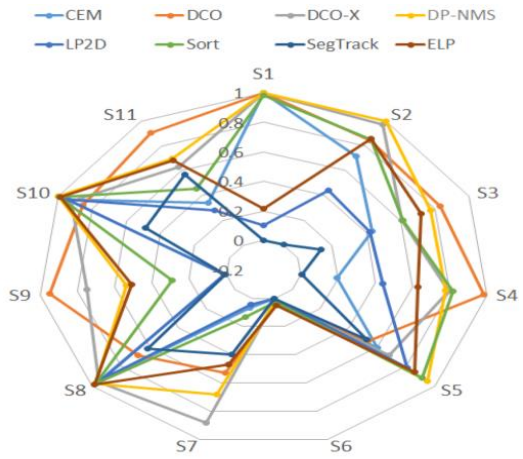


(a)

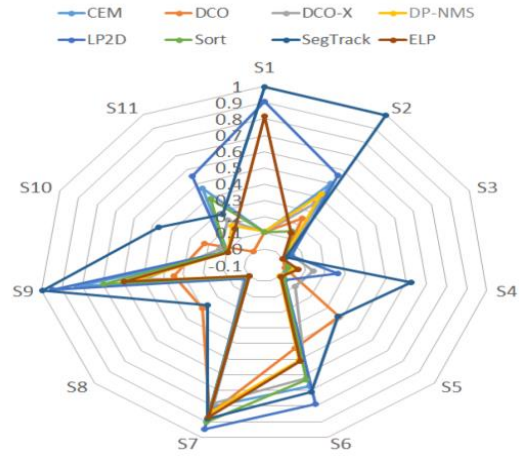


(b)

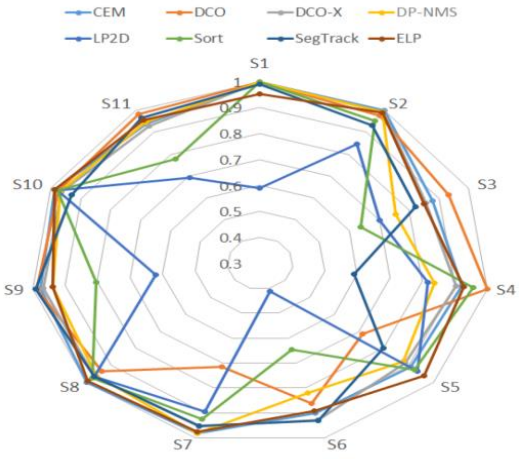
Fig 5 Radar charts showing the computed performance scores of trackers on every sequence (S1-S11) based on (a) MOTA and (b) MELT.



(a)



(b)



(c)

Fig 6 Radar charts showing the computed performance scores of trackers on every sequence (S1-S11) based on (a) R_{fp}, (b) R_{fn}, and (c) R_{idc}.

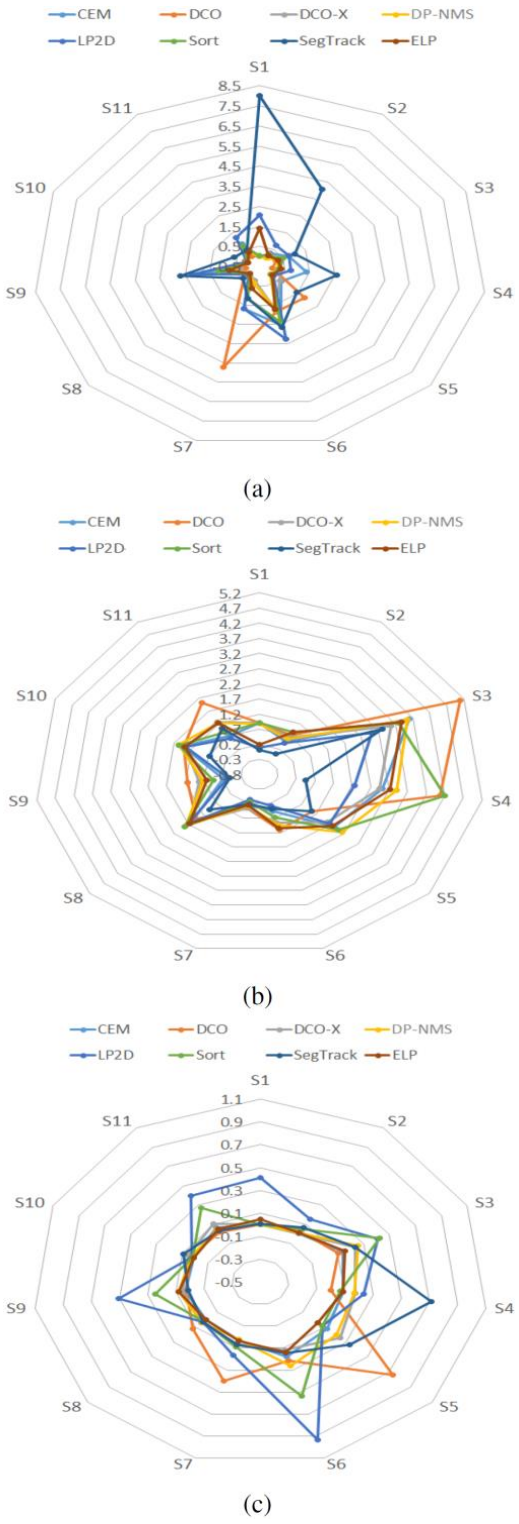


Fig 7 Radar charts showing the computed performance scores of trackers on every sequence (S1-S11) based on (a) PFC_{fp}, (b) PFC_{fn}, and (c) PFC_{idc}.

ditions making to it very hard to distinguish among targets. LP2D has performed comparatively better than others to track under these challenging conditions (see Fig. 8(e)). S4 also suffers from poor visibility. SegTrack has found particularly difficult to track and has obtained the worst performance based on several assessment criteria (Fig. 6(a,c), Fig. 7(a,c)). On the other hand, DCO has perfomed reasonably better than others; see a sample qualitative result of DCO that is able to correctly track two out of five targets under such challenging conditions (Fig. 8(f)). S5 possesses challenges of target scale and pose changes, and occlusions. DCO has consistenly shown the worst performance based on most assessment criteria due to its inability to distinguish among IDs of different targets and high concentration of false positives (see Fig. 9(a)); for a comparison, we show corresponding qualitative result of ELP (see Fig. 9(b)) that has performed much better than DCO. On S6, DCO-X has more often outperformed others across different assessment criteria due to its better ability to deal with challenges present in this sequence that are occlusions, scale changes and pose changes; see a sample qualitative result in which this tracker is able to track under a partial occlusion of a target (Fig. 9(c)). LP2D, on the other hand, has shown the worst performance based on majority of assessment criteria (Fig. 5(a), 6(a,c), 7(a,c)). On S7, the best performance has been reported by either DCO-X, LP2D or DP-NMS, whereas DCO has consistenly achieved the worst performance based on majority of assessment criteria (Fig. 5(a), 6(c), 7(a,b,c)). See a sample qualitative result for DCO-X on S7 in Fig. 9(d). The key challenge present in S8 is occlusion. Like S7, trackers have shown mixed performance trends on S8 with the best performance achieved either by SegTrack, CEM, or ELP, whereas SORT has shown the worst performance more often than others. Fig. 9(e) shows a sample frame where CEM has coped very well with a severe occlusion. S9 possesses challenge of occlusion, scale changes and a target running, and SegTrack has more often shown a better performance than other trackers; see the visualization of scene with

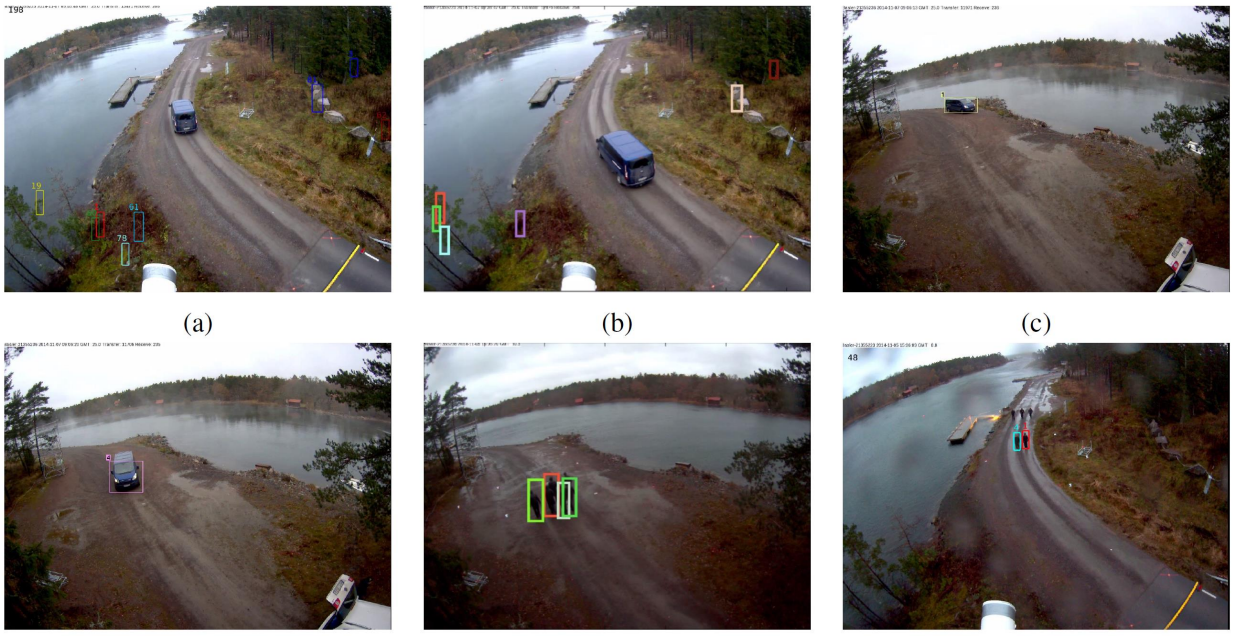


Fig 8 Qualitative results of trackers on key frames from P5 sequences (S1-S4). (a) S1 (SegTrack); (b) S1 (LP2D); (c) S2 (DP-NMS); (d) S2 (DP-NMS); (e) S3 (LP2D); (f) S4 (DCO).

a sample qualitative result of SegTrack in Fig. 9(f). S10 is challenging due to presence of scale changes, pose changes, severe occlusions, and clutter. On this sequence, SegTrack has more often performed better than others by achieving the best performance based on four assessment criteria (Fig. 5(a,b), Fig. 6(b), Fig. 7(b)). See a sample qualitative result for SegTrack in the scene with a target (driver) involved in fight with other targets (Fig. 9(g)). S11 captures the same scenario as S10, but from a different viewpoint. On S11, the best performance has been obtained by different trackers depending on assessment criteria, but the worst performance has been consistently shown by either LP2D or DCO. See the visualization of scene with a sample qualitative result for LP2D in Fig. 9(h).

To infer the overall performance trends of trackers, Table 6 provides the average performance scores computed across the test sequences (S1-S11) using each assessment criteria (the best scores are highlighted in bold). Moreover, the performance rankings (1-8) obtained based on each of the

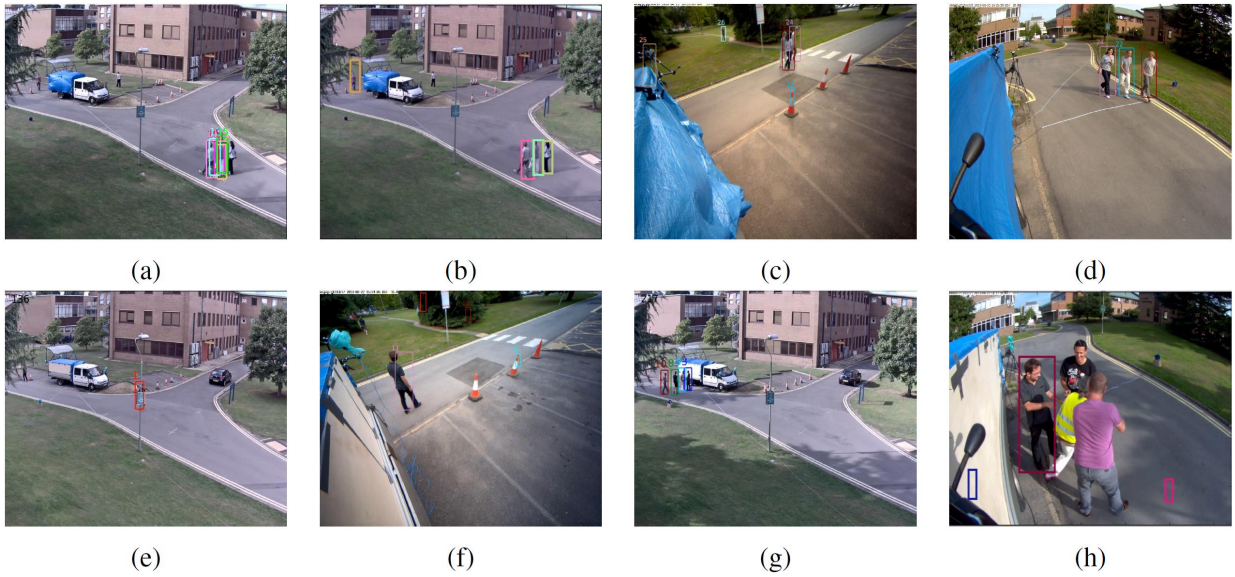


Fig 9 Qualitative results of trackers on key frames from ARENA sequences (S5-S11). (a) S5 (DCO); (b) S5 (ELP); (c) S6 (DCO-X); (d) S7 (DCO-X); (e) S8 (CEM); (f) S9 (SegTrack); (g) S10 (SegTrack); (h) S11 (LP2D).

criteria are presented in Fig. 10. A rank of '1' is the best and a rank of '8' is the worst. The best MOTA is reported by DCO-X followed by DP-NMS, ELP, SORT, CEM, DCO, LP2D and SegTrack (Table 6, Fig. 10). Interestingly, MELT ranks SegTrack the best followed by DCO-X, LP2D, DP-NMS, ELP, CEM, SORT and DCO. The reason behind disagreements between rankings obtained with MOTA and MELT is that the former provides an end performance by quantifying cardinality error and ID changes, and the latter instead provides an end performance by quantifying tracking accuracy. A deeper insight and understanding of performance can be provided by means of R and PFC scores. DP-NMS is the best among all trackers in terms of R_{fp}, indicating its enhanced robustness to track over extended period of time without producing any false positive. See Fig. 10 for a full ranking of trackers based on R_{fp}. Likewise, DP-NMS outperforms other trackers based on PFC_{fp} too, which shows its lesser tendency of producing a higher per frame concentration of false positives than other trackers. Additionally, based on R_{fn} and PFC_{fn}, SegTrack is the best, thus showing its better ability to deal with false negatives as compared to remaining trackers.

Table 6 Average performance scores of trackers across test sequences (S1-S11).

Tracker	MOTA	MELT	R_{fp}	R_{fn}	R_{idc}	PFC_{fp}	PFC_{fn}	PFC_{idc}
CEM	0.119	0.811	0.478	0.317	**0.954**	0.906	1.273	**0.060**
DCO	-0.066	0.853	0.742	0.262	0.917	0.856	1.759	0.153
DCO-X	**0.294**	0.777	0.747	0.278	0.941	0.387	1.361	0.101
DP-NMS	0.271	0.790	**0.765**	0.252	0.916	**0.326**	1.504	0.103
LP2D	-0.067	0.781	0.402	0.469	0.769	1.188	0.962	0.304
SORT	0.130	0.830	0.641	0.276	0.870	0.588	1.522	0.164
SegTrack	-0.888	**0.766**	0.249	**0.601**	0.898	2.282	**0.688**	0.159
ELP	0.168	0.807	0.632	0.298	0.947	0.529	1.378	**0.060**

Moreover, in terms of dealing with ID changes, CEM is the best both based on R_{idc} and PFC_{idc}. We checked the statistical significance of performance rankings of trackers across all sequences using Welch ANOVA test. Statistical significance is achieved at the standard 5% significance level.

We also compared the rankings obtained by trackers in this study against their rankings obtained on MOT Challenge. We noticed that out of eight trackers evaluated in this study, most (DCO-X, DP-NMS, ELP, CEM, LP2D and SegTrack) can be found under 2D MOT 2015 Challenge[11]; hence we used 2D MOT 2015 for comparison here. On 2D MOT 2015, ELP is ranked the best followed by SegTrack, LP2D, DCO-X, CEM, DP-NMS based on MOTA (the common metric used in this study as well as in 2D MOT 2015), which is different from the ranking obtained in this study that declares DOC-X the best followed by DP-NMS, ELP, CEM, LP2D and SegTrack (see Fig. 10). The disagreement in rankings is apparently due to the differences in test scenarios of datasets used in this study and MOT datasets. Unlike MOT, the datasets in this study are primarily designed for surveillance scenarios involving perimeter protection of critical infrastructures. Consequently, in the present datasets the cameras are all mounted at high (top-downish) positions, whereas MOT datasets mostly account for video footages filmed from chest or eye level, thus offering different challenges, applications, and complexity levels.

[11] https://motchallenge.net/results/2D_MOT_2015/. Accessed May 2019.

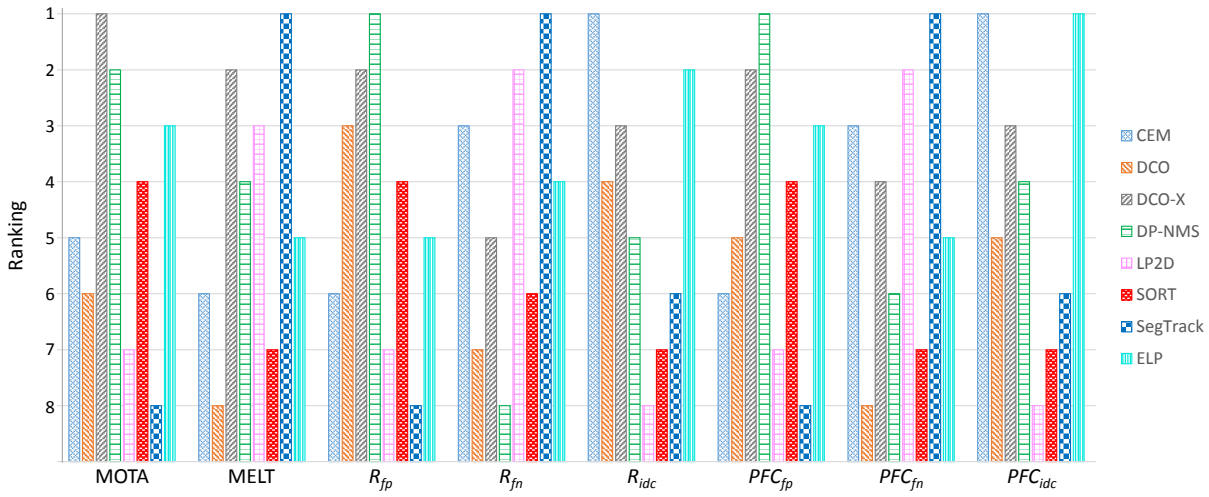


Fig 10 Performance rankings (1-8) of trackers based on each performance assessment criterion. $Rank = 1$ is the best.

Fig. 11 shows the average performance ranking of trackers across all eight performance assessment criteria (MOTA, MELT, R_{fp}, R_{fn}, R_{idc}, PFC_{fp}, PFC_{fn} and PFC_{idc}). Overall, DCO-X is ranked the best tracker based on all criteria followed by ELP, CEM, DP-NMS, SegTrack, LP2D, DCO and SORT.

It is relevant to mention that the presented performance metrics are expected to have varying (higher/lower) impact in the evaluation depending on application under consideration. For example, in an autonomous driving application in which the goal might be to avoid collision with objects present on/off the road (e.g. oncoming vehicles, pedestrians, traffic lights/signs, etc.), tracking accuracy and cardinality error would be more critical and important to compute than maintaining unique target IDs. Therefore, for tracking evaluation in this application, the metrics such as MELT, R_{fp}, R_{fn}, PFC_{fp}, and PFC_{fn} are expected to have a higher impact. On the other hand, in a sport (e.g. Football) application where the goal might be to track and analyze match performance of individual player(s), maintaining unique target IDs would be very important to assess. Therefore, for tracking evaluation in this application, the metrics such as MOTA, R_{idc} and PFC_{idc} are expected to

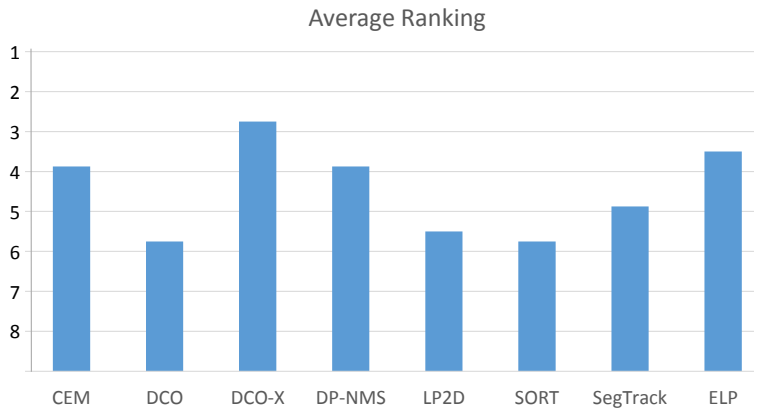


Fig 11 Average performance ranking of trackers across all eight performance assessment criteria. $Rank = 1$ is the best.

have a higher impact.

5.4 Discussion

The R and PFC scores also enable to provide formative feedback to aid in addressing limitations of trackers.[26] Indeed, the analysis based on false positives could aid in analyzing the effect on tracking performance originating from the detection stage. For instance, SegTrack shows the worst PFC_{fp} and R_{fp} as compared to remaining trackers (Table 6, Fig. 10), indicating a particular need of improvement at its detection stage. Similarly, inferior performance in terms of false negatives (i.e., PFC_{fn}, R_{fn}) can point toward improvement at the detection stage, and/or inability to effectively temporally connect small tracks ('tracklets'). For example, DCO, DP-NMS and SORT have shown inferior PFC_{fn} and R_{fn}, which is likely because of lack of an effective dedicated strategy to connect tracklets in these trackers as compared to the remaining ones. Likewise, the analysis based on ID changes (PFC_{idc}, R_{idc}) provide a formative feedback regarding the tracking stage. For example, CEM and ELP are the best in terms of PFC_{idc} and R_{idc}, showing they have a more effective ID management mechanism than other trackers; whereas LP2D has shown the worst PFC_{idc} and R_{idc}, indicating the need on improving its ability to distinguish among IDs of different trackers.

We believe the recent advancements in MOT Challenge could potentially help in addressing identified trackers shortcomings at detection and tracking stages. Generic neural network-based pre-trained target models have shown promising results in terms of an enhanced detection accuracy and robustness. For example, many recent trackers[27–29] have adopted Convolutional Neural Networks (CNN) based frameworks to show significant performance improvements on MOT Challenges. Moreover, other trackers have used Recurrent Neural Networks (RNN) using multiple cues over a temporal window for performing long-term tracking by effectively resolving data association problem even under occlusions[30] or using an instance-aware tracking approach integrating single-target tracking techniques for multi-target tracking to better manage and distinguish targets IDs.[31]

6 Conclusions

This paper presents statistically significant performance ranking and comparison as well as formative assessment of several state-of-the-art multi-target visual trackers on real-world surveillance datasets concerning perimeter protection of critical infrastructures (CI), released publicly as a part of recent PETS workshops and cover a wide range of challenges and conditions. We used well-known and recent performance assessment criteria for a thorough experimental analysis and comparison.

Overall, the results show DCO-X to be the best-ranked tracker, making it more suitable to be used in CI perimeter protection applications. SegTrack is identified to be the tracker requiring a particular improvement at its detection stage. DCO, DP-NMS and SORT lack an effective tracklet linking strategy as compared to other trackers. LP2D is found to be needing improvement in its ID management strategy.

Acknowledgments

This work was supported by funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 312784. The first and second authors did this work at the University of Reading, UK.

References

1 S. Fleck and W. Strasser, "Smart camera based monitoring system and its application to assisted living," *Proceedings of the IEEE* **96**(10), 1698–1714 (2008).

2 T. Nawaz, A. Berg, J. Ferryman, *et al.*, "Effective evaluation of privacy protection techniques in visible and thermal imagery," *Journal of Electronic Imaging* **26**(5), 051408 (2017).

3 A. T. Rath and J.-N. Colin, "Protecting personal data: Access control for privacy preserving perimeter protection system," in *IFIP Annual Conference on Data and Applications Security and Privacy*, (Fairfax, VA) (2015).

4 L. Leal-Taixe, M. Fenzi, A. Kuznetsova, *et al.*, "Learning an image-based motion context for multiple people tracking," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, (Columbus, OH) (2014).

5 H. Pirsiavash, D. Ramanan, and C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (Colorado Springs, CO) (2011).

6 A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (Providence, RI) (2012).

7 A. Bewley, Z. Ge, L. Ott, *et al.*, "Simple online and realtime tracking," in *Proceedings of IEEE International Conference on Image Processing*, (Phoenix, AZ) (2016).

8 A. Milan, K. Schindler, and S. Roth, "Multi-target tracking by discrete-continuous energy minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(10), 2054–2068 (2016).

9 A. Milan, K. Schindler, and S. Roth, "Continuous energy minimization for multitarget tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(1), 58–72 (2014).

10 N. McLaughlin, J. M. D. Rincon, and P. Miller, "Enhancing linear programming with motion modeling for multi-target tracking," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, (Waikoloa, HI) (2015).

11 A. Milan, L. Leal-Taixe, K. Schindler, *et al.*, "Joint tracking and segmentation of multiple targets," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (Boston, MA) (2015).

12 J. Ferryman and A.-L. Ellis, "Performance evaluation of crowd image analysis using the pets2009 dataset," *Pattern Recognition Letters* **44**, 3–15 (2014).

13 L. Leal-Taixe, A. Milan, I. Reid, *et al.*, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *CoRR* , http://arxiv.org/abs/1504.01942 (2015).

14 T. Biresaw, T. Nawaz, J. Ferryman, *et al.*, "ViTBAT: Video Tracking and Behavior Annotation Tool," in *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance*, (Colorado Springs) (2016).

15 K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP Journal on Image and Video Processing* **2008**, 1–10 (2008).

16 T. Nawaz, F. Poiesi, and A. Cavallaro, "Measures of effective video tracking," *IEEE Transactions on Image Processing* **23**(1), 376–388 (2014).

17 T. H. Nawaz, *Ground-truth-based trajectory evaluation in videos*. PhD thesis, Queen Mary University of London, UK (2014).

18 A. Waibel, R. Stiefelhagen, R. Carlson, *et al.*, *Handbook of Ambient Intelligence and Smart Environments*, ch. Computers in the Human Interaction Loop, 1071–1116. Springer (2010).

19 N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (San Diego, CA) (2005).

20 P. F. Felzenszwalb, R. B. Girshick, D. McAllester, *et al.*, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(9), 1627–1645 (2010).

21 J. Black, T. Ellis, and P. Rosin, "A novel method for video tracking performance evaluation," in *Proceedings of Joint IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, (2003).

22 B. Ristic, B.-N. Vo, D. Clark, *et al.*, "A metric for performance evaluation of multi-target tracking algorithms," *IEEE Transactions on Signal Processing* **59**(7), 3452–3457 (2011).

23 T. Nawaz and A. Cavallaro, "A protocol for evaluating video trackers under real-world conditions," *IEEE Transactions on Image Processing* **22**(4), 1354–1361 (2013).

24 R. Kasturi, D. Goldgof, P. Soundararajan, *et al.*, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(2), 319–336 (2009).

25 B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (Colorado Springs) (2011).

26 T. Nawaz, A. Ellis, and J. Ferryman, "A method for performance diagnosis and evaluation of video trackers," *Signal, Image and Video Processing* **11**(7), 1287–1295 (2017).

27 L. Chen, H. Ai, C. Shang, *et al.*, "Online multi-object tracking with convolutional neural networks," in *Proceedings of IEEE International Conference on Image Processing*, (Beijing) (2017).

28 Q. Chu, W. Ouyang, H. Li, *et al.*, "Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism," in *Proceedings of IEEE International Conference on Computer Vision*, (Venice) (2017).

29 J. Son, M. Baek, M. Cho, *et al.*, "Multi-object tracking with quadruplet convolutional neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (Honolulu, HI) (2017).

30 A. Sadeghian, A. Alahi, and S. Savarese, "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," in *Proceedings of IEEE International Conference on Computer Vision*, (Venice) (2017).

31 P. Chu, H. Fan, C. Tan, *et al.*, "Online multi-object tracking with instance-aware tracker and dynamic model refreshment," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, (Waikoloa Village, HI) (2019).

Longzhen Li received M.S. degree in Signal and Information System from The University of Electronic Science and Technology of China (UESTC) in 2002 and MSc. degree in Computer

Science from the University of Reading in 2005. He worked as a Research Assistant since 2005 in Computational Vision Group at the same university. His research interests include Computer Vision and Machine Learning. He is currently working as senior research engineer in the private industry.

Tahir Nawaz received a PhD in 2014 jointly from Queen Mary University of London, UK, and Alpen-Adria University of Klagenfurt, Austria, and an MSc in 2009 jointly from Heriot-Watt University, UK, University of Girona, Spain, and University of Burgundy, France. He has a substantial research experience both in academic and industrial sectors. He has published several journal and conference papers, is a reviewer of well-known journals, and was Co-organizer of PETS 2015.

James Ferryman is Professor of Computational Vision and leads CVG housed within the Department of Computer Science, University of Reading, UK. He has over 20 years of experience in image and video analysis and has co-authored over 100 papers related to computer vision. He has been a PI on several EPSRC and EC projects. He also leads the IEEE International series of PETS workshops and has been a Director of both BMVA and SITC.

List of Figures

List of Tables