Tracking performance evaluation on PETS 2015 Challenge datasets

Tahir Nawaz, Jonathan Boyle, Longzhen Li and James Ferryman
Computational Vision Group, School of Systems Engineering
University of Reading, Whiteknights, Reading RG6 6AY, UK
{t.h.nawaz,j.n.boyle,longzhen.li,j.m.ferryman}@reading.ac.uk

Abstract

This paper presents a quantitative evaluation of a tracking system on PETS 2015 Challenge datasets using well-established performance measures. Using the existing tools, the tracking system implements an end-to-end pipeline that include object detection, tracking and post-processing stages. The evaluation results are presented on the provided sequences of both ARENA and P5 datasets of PETS 2015 Challenge. The results show an encouraging performance of the tracker in terms of accuracy but a greater tendency of being prone to cardinality error and ID changes on both datasets. Moreover, the analysis show a better performance of the tracker on visible imagery than on thermal imagery.

1. Introduction

The International Workshop on Performance Evaluation of Tracking and Surveillance (PETS) workshops have been aimed to foster the emergence of computer vision technologies for object detection and tracking by providing evaluation datasets and metrics that allow an accurate assessment and comparison of such methodologies. PETS 2015 workshop includes a Challenge and provides datasets for participants to test and rank their algorithms [1]. The datasets cover a variety of tasks involving low-level video analysis (detection, tracking), mid-level analysis (simple event detection) and high-level analysis (complex 'threat' event detection).

This paper focuses on the performance evaluation of an existing tracking system [6] on the PETS 2015 Challenge datasets. The tracking system uses various existing techniques to implement an end-to-end pipeline including detection, tracking and post-processing stages. Images are fed into the motion/change detectors, and the result fused and filtered. The resulting detections are directed to the tracking stage. Before making use of these detections, the tracker exploits information in the image, including motion information from the optical flow field computed by the motion detector, to track and optimise the location of known targets to suit the current image. The tracker now breaks down all of the detections and existing tracking objects into what are termed "atomic regions" or simply atoms. These atoms are then used to determine the association between detections and existing targets. Finally, new targets are created as appropriate and appearance models of existing targets updated ready for subsequent frames.

The remainder of this paper is organised as follows. Sec. 2 describes the tracking system. This is followed by experimental results in Sec. 3. Sec. 4 concludes the paper.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 312784.

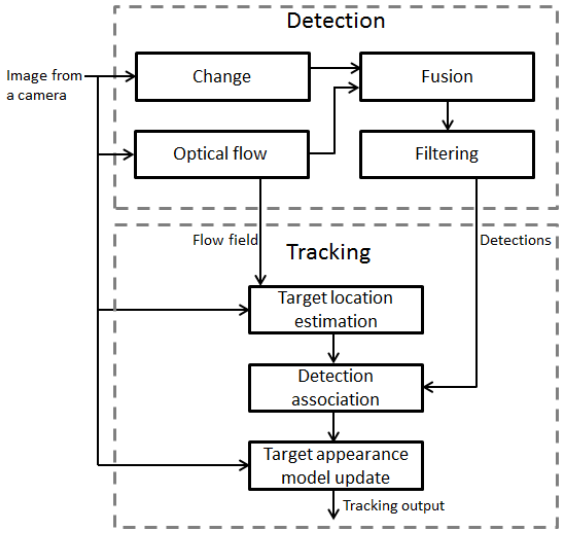

Figure 1. Block diagram of the tracking system.

2. Tracking system

Fig. 1 shows the block diagram of the tracking system. The tracking system is designed to run in real-time, or near

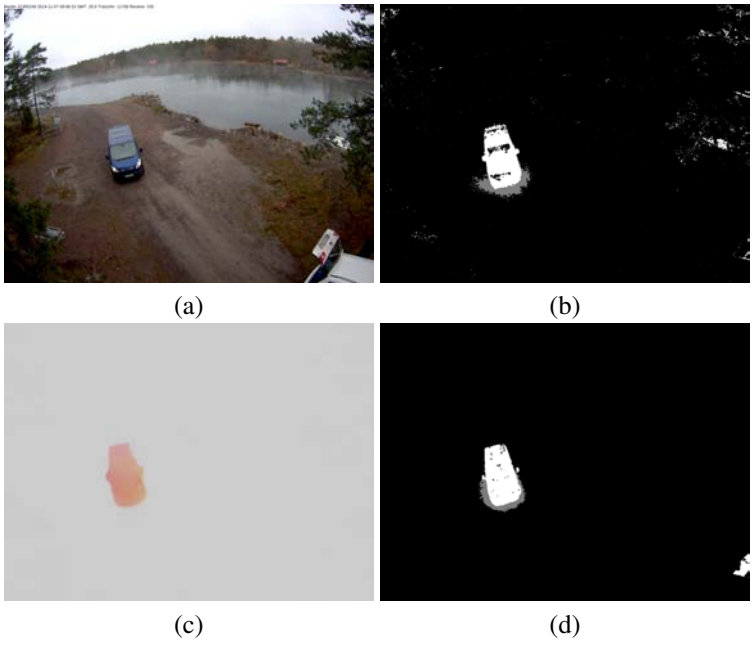

Figure 2. (a) A sample image from PETS 2015 dataset; (b) the foreground mask estimated using adaptive Gaussian mixture model; (c) the color-coded optical flow field indicating image motion; (d) final foreground mask obtained by combining (b) and (c).

real-time, at upwards of 5 frames per second. It uses input from a combination of change and motion detectors, and performs reasoning regarding the current state of the scene, how existing tracked targets should be associated to current detections, and how to update the location of current tracked targets.

2.1. Target detection

The detector provides a foreground mask indicating pixels of the image where objects are believed to be, as well as an optical flow field indicating object motion in the images. It also reports a set of detections as bounding boxes interpreted from the foreground mask, along with an associated "label map" linking foreground pixels to specific detections. Indeed, the foreground mask is obtained by combining the individual mask generated using the adaptive Gaussian mixture model [7] with the one generated using the optical flow estimation method [3] (see Fig. 2).

Detections that are not associated to existing tracked targets are upgraded to new tracked targets. A tracked target consists of a bounding box (directly taken from the detection bounding box), a colour appearance model (an image the same size as the bounding box initialised to the pixels inside of the detection bounding box), and an extents mask (a greyscale image the same size as the bounding box that is initialised to white pixels for the foreground mask of the detection). Once a target exists, it must be tracked into a new frame, and for this the optical flow field calculated by the detector is used. Given the previous location of the tracked target, the pixels inside the targets bounding box can have their motions accumulated giving a good indication of how the target has moved between image frames. Once this initial motion estimate has been computed, the appearance model can be used to further optimise the location of the target in the image and verify its continued presence. This is achieved by computing the difference between the appearance model of the target and the current image for a given location in the image, weighting the significance of the pixels using the extents mask of the target. Using this difference error, a search can be undertaken to determine the location in the image that minimises the difference between the template and the image.

The tracker depends heavily upon having a reliable detection input, however motion/change detection are extremely prone to producing undesirable noise detections, or simply inaccurate detections. Typical failures in detection consist of situations where the detector is unable to correctly handle the presence of a shadow, merges multiple objects into a single detection region, fails to handle sudden lighting changes, reports detections on non-salient objects, or suffers from unintended camera motion. To improve the detection results filtering processes are considered to verify good detections or remove improbable detections. These filtering approaches include simply considering the size of the detection or masking out areas of the image (with a manual intervention) likely to produce noise detections and preventing detection in that region.

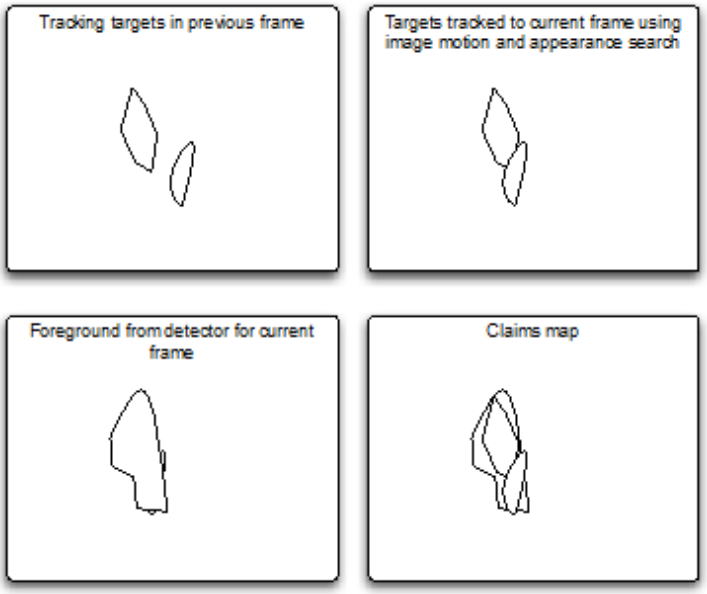

Figure 3. An example of the atomization process permitting known tracked targets to break a merged foreground region into sensible sub-regions.

2.2. Detection association and atomisation process

Once targets have been tracked to an estimated position in the current image, the new set of detections must be associated to the existing targets to determine if targets are still being detected, or if new objects of interest have entered the scene. Often the detector produces some errors which the tracker needs to identify and compensate for. These include the merging of multiple objects into a single detection region, as well as the partial detection of objects, or fragmented detection of objects. To this end, a process of "atomisation" is undertaken.

The atomisation process consists of breaking existing targets and foreground regions into segments, with existing targets "claiming" sections of foreground regions they overlap with. This results in a set of *atomic regions* that can be described as either an "undetected claim" (a target region that does not overlap with a foreground region), a "detected claim" (a target region that does overlap with a foreground region), or an "unclaimed detection" (a foreground region which is unclaimed by a target). A diagrammatic example of this is given in Fig. 3. Should the detector produce a large foreground region corresponding to multiple objects, then existing tracked targets will be able to claim their portion of that foreground region, and recover detections of the individual objects. Similarly if a single object produces a fragmented detection, then the existing tracked target will be able to claim multiple detections, and merge those fragments to a unified object.

The association process consists of building a table denoting the overlap between each atomic region and each tracked target. Atoms are then associated uniquely to a single tracked target, iterating through the table from the most certain association to the weakest. This results, potentially, in a many-to-one detection to target association result, and does not permit a one-to-many situation. In the event of a many-to-one association, consideration must be given to whether a tracked target actually consists of multiple objects producing those multiple detections. This is discussed in the Sec. 2.4 below.

2.3. Target updating

Once detections have been associated to a tracked target, the final stage of tracking is to update the information of the tracked target, that is to say, update the appearance and extent models and the size of the bounding box. The bounding box is resized to ensure that all associated atoms fit inside. The appearance and extents models are then updated as a running average, updated using the values of the pixels in the current image beneath the updated location of the tracked target. The running average must be performed with some care to compensate for any resizing of the bounding box.

2.4. Target splitting and merging

Typical problems that must be considered for any tracker are instances of splitting or merging, when a single object splits to become multiple objects (for instance a group of people breaking apart into the individuals), or the merging

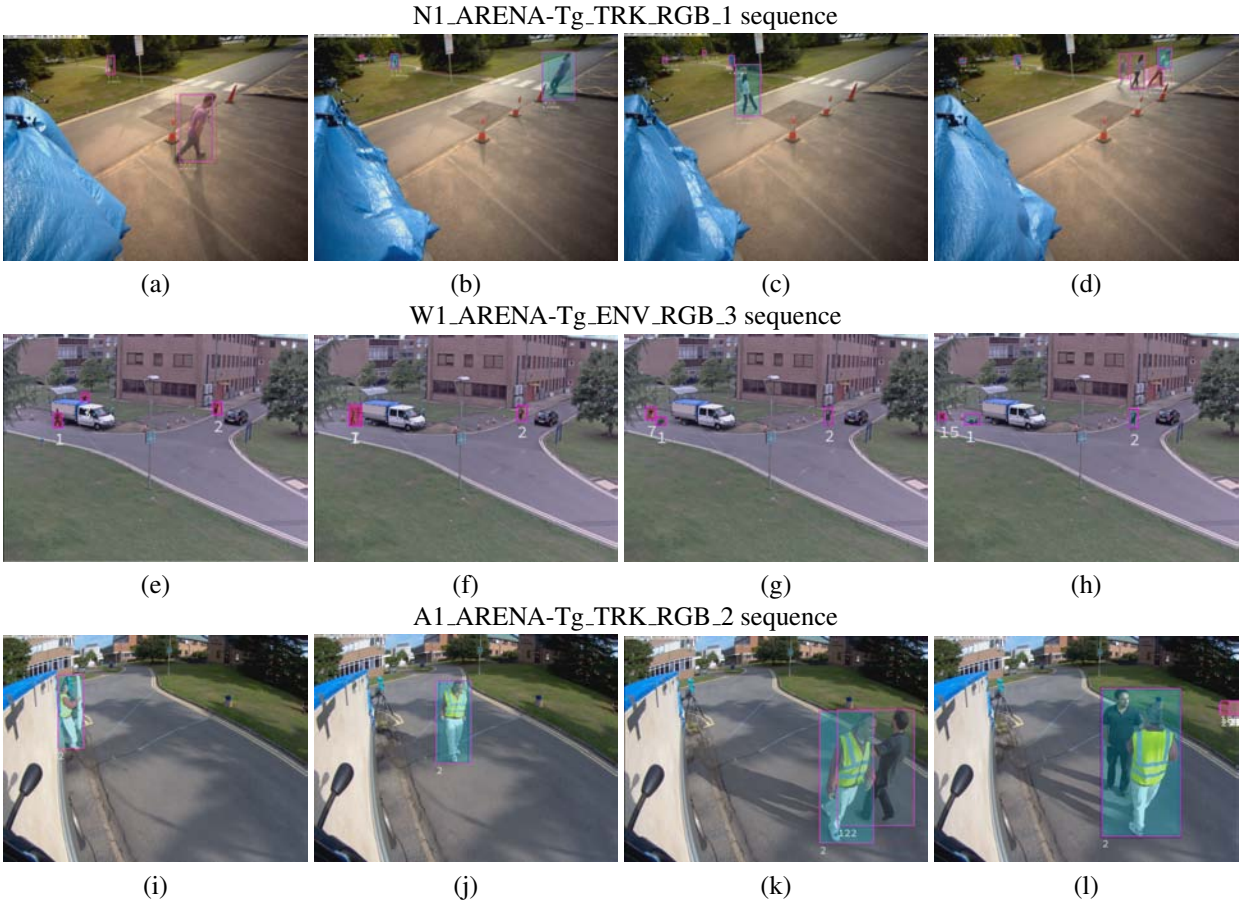

Figure 4. Qualitative tracking results shown in the form of magenta bounding boxes (with ID displayed underneath) on key frames of representative sequences from ARENA dataset: N1_ARENA-Tg_TRK_RGB_1 (a-d), W1_ARENA-Tg_ENV_RGB_3 (e-h) and A1_ARENA-Tg_TRK_RGB_2 (i-l).

of two or more objects into one. Arguably, when two objects merge it is desirable to maintain track of the original individuals until such time as that is no longer possible. The atomisation process already allows for this - the merging of two or more objects into one object will result in a merged detection, and the atoms will allow each target to continue to claim that part of the detection that is appropriate to them.

Splitting is a more interesting proposition. In a single frame, the presence of multiple detections associated to a single target could be the result of a fragmentation of the detection, or it could be a genuine splitting of a group. To resolve this, a many-to-one atoms-to-target association situation results in the target initialising sub-regions for each associated atom. These sub-regions are treated as tracked targets in their own right. If they persist over time and move away from the main tracked target, then they are considered to be a new object and split. Otherwise, they are merged back into the main object (which is to say, the sub region is simply removed from tracking).

Fig. 4 and 5 show sample qualitative tracking results on different sequences.

3. Experimental results

This section describes the results and analysis of the tracking system described above on the provided PETS 2015 Tracking Challenge datasets [1]. PETS 2015 Challenge uses two datasets that are ARENA and P5 datasets. ARENA dataset for the Tracking Challenge contains seven sequences all with visible imagery, whereas P5 dataset contains nine sequences - five with visible imagery and four with thermal imagery. Table 1 presents a summary of the video sequences. The parameters of the tracker are fixed for the experiments. Next we first describe the measures used for the evaluation followed by the results.

We quantitatively evaluated the proposed tracking system on all the datasets using the measures prescribed in PETS 2015 Challenge. The evaluation accounts for the three key aspects including tracking accuracy (extent of match between an estimation and the corresponding ground

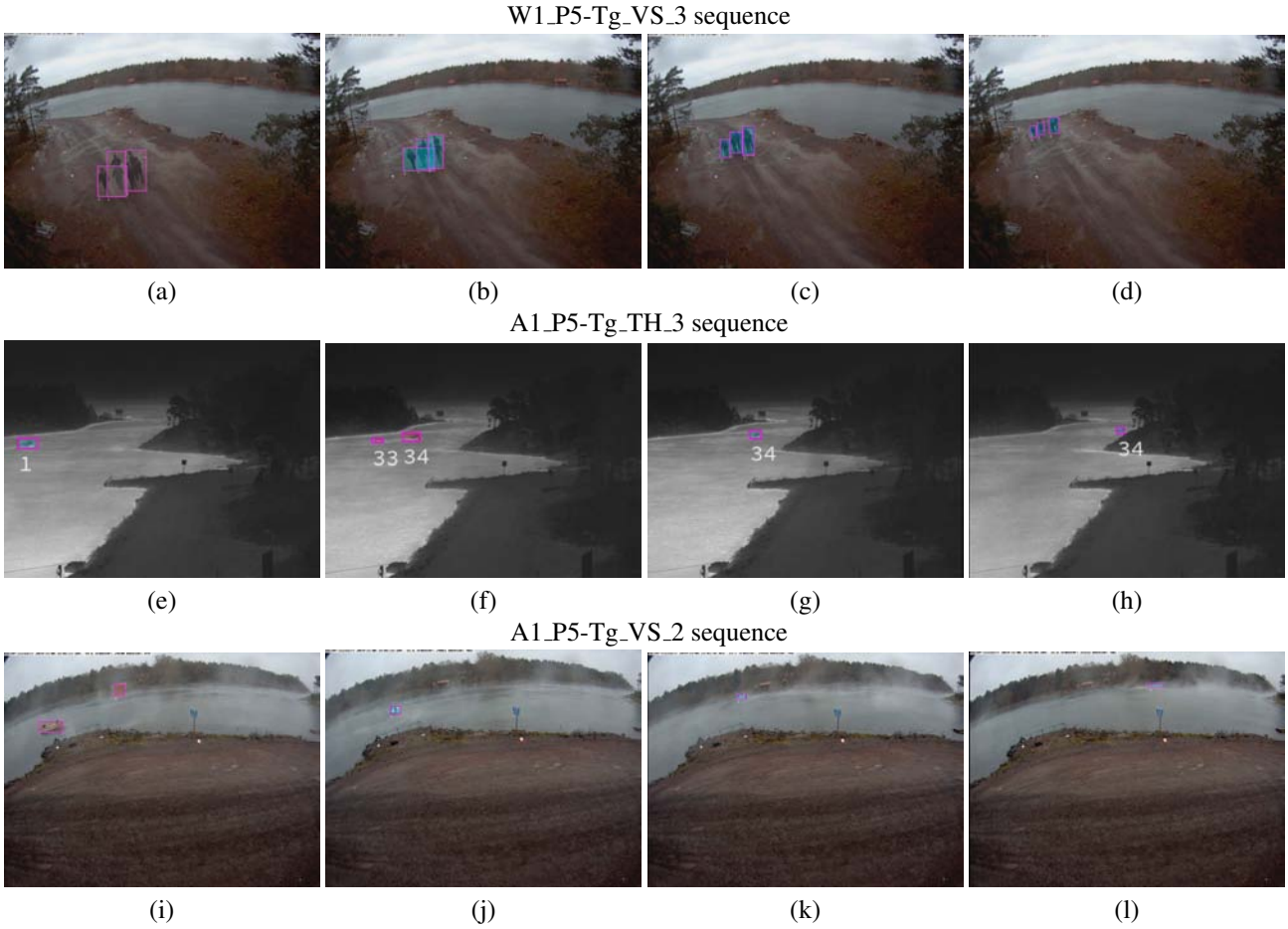

Figure 5. Qualitative tracking results shown in the form of magenta bounding boxes (with ID displayed underneath) on key frames of representative sequences from P5 dataset: W1_P5-Tg_VS_3 (a-d), A1_P5-Tg_TH_3 (e-h) and A1_P5-Tg_VS_2 (i-l).

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) [5]. PETS 2015 Challenge uses two evaluation measures that cover these three aspects together including the widely-used Multiple Object Tracking Accuracy (MOTA) [4] and the recently-introduced Multiple Extended-target Lost-Track ratio (MELT) [5]. MOTA takes into account the cardinality error (in the form of false positives and false negatives) and ID changes without explicitly considering accuracy, and 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 [4]. False negatives are the missed targets at frame k and false positives are the estimated targets with overlap $O_{k,t} < \bar{\tau}$ such that $\bar{\tau}$ is a pre-defined threshold and $O_{k,t} = \frac{|\bar{A}_{k,t} \cap A_{k,t}|}{|\bar{A}_{k,t} \cup A_{k,t}|}$ 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 the image plane for the ground-truth and estimated bounding boxes, respectively. $\bar{\tau}$ is often set to 0.5 [2]. MOTA $\leq$ 1: the higher MOTA, the better the performance. MELT provides tracking accuracy evaluation using the information about lost-track ratio. Let N_i be the total number of frames in the ith ground-truth track and N_i^τ is the number of frames with the overlap score below a threshold τ, then the lost-track ratio λ_i^τ is computed as follows: $\lambda_i^\tau = \frac{N_i^\tau}{N_i}$. MELT for a particular τ is computed as follows: $\text{MELT}_\tau = \frac{1}{V}\sum_{i=1}^{V} \lambda_i^\tau$, where V is the total number of ground-truth tracks, and

$$\text{MELT} = \frac{1}{S} \sum_{\tau \in [0,1]} \text{MELT}_\tau, \tag{2}$$

Table 1. Summary of the PETS 2015 Tracking Challenge sequences.

Sequence	Sensor	Frame Size	No. of frames
N1_ARENA-Tg_ENV_RGB_3	Visible	600×800	289
N1_ARENA-Tg_TRK_RGB_1	Visible	960×1280	513
N1_ARENA-Tg_TRK_RGB_2	Visible	960×1280	684
W1_ARENA-Tg_ENV_RGB_3	Visible	600×800	155
W1_ARENA-Tg_TRK_RGB_1	Visible	960×1280	240
A1_ARENA-Tg_ENV_RGB_3	Visible	600×800	295
A1_ARENA-Tg_TRK_RGB_2	Visible	960×1280	670
N1_P5-Tg_VS_1	Visible	960×1280	400
N1_P5-Tg_VS_3	Visible	960×1280	387
N1_P5-Tg_TH_1	Thermal	480×640	600
N1_P5-Tg_TH_2	Thermal	480×640	220
W1_P5-Tg_VS_1	Visible	960×1280	180
W1_P5-Tg_VS_3	Visible	960×1280	180
W1_P5-Tg_TH_3	Thermal	480×640	740
A1_P5-Tg_VS_2	Visible	960×1280	720
A1_P5-Tg_TH_3	Thermal	480×640	1000

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.

We evaluated the tracker on all the sequences of ARENA and P5 datasets. Tables 2 and 3 provide quantitative performance of the tracker on ARENA and P5 datasets, respectively, in the form of MELT and MOTA.

On ARENA dataset, the tracker has generally shown encouraging accuracy that is reflected in the obtained mean MELT=0.582 (Table 2) on all the sequences, whereas MOTA is generally lower with a mean MOTA=-0.800 showing tendency of producing a greater cardinality error and ID changes. On P5 dataset, similar trends are shown with the tracker obtaining a mean MELT=0.558 and mean MOTA=-1.798 on all sequences (Table 3). Additionally, these results show a better performance of the tracker on ARENA dataset than on P5 dataset in terms of MOTA and a better performance in terms of MELT on P5 dataset than on ARENA dataset. Moreover, we also analysed the combined performance of the tracker on ARENA and P5 datasets separately for the sequences with visible imagery and separately for those with thermal imagery. For thermal sequences, mean MELT=0.616 and mean MOTA $= -4.324$; for visual sequences, mean MELT=0.553 and mean MOTA $= -0.374$. These scores show that the tracker performed better on visible imagery than thermal imagery based on both MELT and MOTA.

Table 2. MELT and MOTA scores of the tracker on all sequences of ARENA dataset.

Sequence	MELT	MOTA
N1_ARENA-Tg_ENV_RGB_3	0.729	-0.740
N1_ARENA-Tg_TRK_RGB_1	0.477	-0.232
N1_ARENA-Tg_TRK_RGB_2	0.658	-0.179
W1_ARENA-Tg_ENV_RGB_3	0.535	-0.055
W1_ARENA-Tg_TRK_RGB_1	0.341	-0.584
A1_ARENA-Tg_ENV_RGB_3	0.497	-1.270
A1_ARENA-Tg_TRK_RGB_2	0.834	-2.539
Mean	0.582	-0.800

Table 3. MELT and MOTA scores of the tracker on all sequences of P5 dataset.

Sequence	MELT	MOTA
N1_P5-Tg_VS_1	0.262	0.994
N1_P5-Tg_VS_3	0.460	-0.827
N1_P5-Tg_TH_1	0.588	0.220
N1_P5-Tg_TH_2	0.560	0.183
W1_P5-Tg_VS_1	0.596	0.454
W1_P5-Tg_VS_3	0.812	0.112
W1_P5-Tg_TH_3	0.798	-0.388
A1_P5-Tg_VS_2	0.431	0.376
A1_P5-Tg_TH_3	0.520	-17.312
Mean	0.558	-1.798

4. Conclusions

In this paper we presented the evaluation results of a tracker on PETS 2015 Challenge (ARENA and P5) datasets using the well-known measures that are MELT and MOTA. On both datasets, the tracker showed an encouraging performance in terms of a mean MELT=0.582 on ARENA dataset and a mean MELT=0.558 on P5 dataset; hence showing an increased ability to track with an encouraging accuracy. On the other hand, the tracker obtained a mean MOTA=-0.800 on ARENA dataset and mean MOTA=-1.798 on P5 dataset, which shows a relatively greater tendency of the tracker to produce cardinality error and ID changes. Moreover, the performance analysis showed better MELT and MOTA scores for the tracker on visual sequences than on thermal sequences.

References

[1] PETS 2015. pets2015.net. Accessed June 2015.

[2] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *Proc. of IEEE CVPR*, 2011.

[3] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. of ECCV*, 2004.

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

[5] T. Nawaz, F. Poiesi, and A. Cavallaro. Measures of effective video tracking. *IEEE TIP*, 23(1):376–388, 2014.

[6] L. Patino and J. Ferryman. Pets 2014: Dataset and challenge. In *Proc. of IEEE Int. Conf. on AVSS*, 2014.

[7] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proc. of ICPR*, 2004.