PETS 2015: Datasets and Challenge

Longzhen Li, Tahir Nawaz and James Ferryman
Computational Vision Group, School of Systems Engineering, University of Reading, UK
{ longzhen.li | t.h.nawaz | j.m.ferryman }@reading.ac.uk

Abstract

This paper presents the two datasets (ARENA and P5) and the challenge that form a part of the PETS 2015 workshop. The datasets consist of scenarios recorded by using multiple visual and thermal sensors. The scenarios in ARENA dataset involve different staged activities around a parked vehicle in a parking lot in UK and those in P5 dataset involve different staged activities around the perimeter of a nuclear power plant in Sweden. The scenarios of each dataset are grouped into 'Normal', 'Warning' and 'Alarm' categories. The Challenge specifically includes tasks that account for different steps in a video understanding system: Low-Level Video Analysis (object detection and tracking), Mid-Level Video Analysis ('atomic' event detection) and High-Level Video Analysis ('complex' event detection). The evaluation methodology used for the Challenge includes well-established measures.

1. Introduction

Video surveillance is a widely-researched field presently. Several techniques have been designed and tested for the tasks of object detection and tracking as well as for detection of events of interest. However it is still difficult to compare or evaluate such algorithms because of the lack of standard metrics and benchmarks that indicate how detection, tracking and threat analysis system perform against a common database. The goal of the PETS workshop has been to foster the emergence of computer vision technologies for detection and tracking by providing evaluation datasets and metrics that allow an accurate assessment and comparison of such methodologies. PETS 2015 is sponsored by the EU project P5, the Privacy Preserving Perimeter Protection Project, that aims to develop an intelligent perimeter proactive surveillance system that works robustly under a wide range of weather and lighting conditions for the protection of critical infrastructures[1].

PETS 2015 workshop includes a Challenge and provides two datasets (ARENA and P5) to enable the community to test and rank the algorithms on[2].The ARENA dataset includes a selection of video sequences from PETS 2014 dataset [4] that are made available by another EU project ARENA[3], which addresses the design of a flexible surveillance system to enable situational awareness and determination of potential threats on mobile assets in transit. The P5 dataset contains multi-modal, multi-sensor recordings involving different staged activities around the perimeter of the OKG nuclear plant outside Oskarshamn, Sweden and was recorded jointly by the P5 project partners. Overall, the two datasets cover a variety of tasks (to constitute the PETS 2015 Challenge) involving low-level video analysis (object detection and tracking), mid-level analysis ('atomic' or simple event detection) and high-level analysis (complex 'threat' event detection).

The remainder of this paper is organised as follows. Section 2 presents the PETS 2015 datasets in detail. The PETS 2015 Challenge is described in Section 3. The evaluation framework used in the Challenge is described in detail in Section 4. Section 5 concludes this paper.

2. Datasets

PETS 2015 consists of two datasets:

- **ARENA dataset:** a multi-sensor dataset as used for the PETS2014 challenge which addresses protection of critical mobile assets.
- **P5 dataset:** a multi-modal multi-sensor dataset addressing the application of multi sensor surveillance to protect a nuclear power plant.

The ARENA dataset was used in its full form in PETS 2014 Challenge [4]. For PETS 2015, a selection of the ARENA dataset is used by including fewer scenarios that are more relevant to PETS 2015 Challenge. The selected

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

[1] http://www.p5-fp7.eu

[2] http://pets2015.net

[3] http://www.arena-fp7.eu

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

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. A detailed description of the two datasets are given below.

2.1. ARENA dataset

2.1.1 Overview

The ARENA dataset comprises of a series of multi-camera video recordings where the main subject is the detection and understanding of human behaviour around a parked vehicle. The main objective is to detect and understand the different behaviours from visual (RGB) cameras mounted on the vehicle itself. With this dataset already available for download from PETS 2014 workshop, PETS 2015 workshop provides the opportunity for researchers and industry to submit methodological advances and results obtained using this dataset since the 2014 workshop.

2.1.2 Camera setup and characteristics

Environmental camera: One visual camera ENV_RGB_3 is used (Table 1) that is installed at the location as shown in Figure 1 to cover an approximate area of 100m x 30m. This camera provides a global view of the monitored area.

On-board cameras. Originally four non-overlapping visual cameras were mounted at each corner of a truck in ARENA dataset. The selective part of ARENA dataset used for PETS 2015 Challenge includes two on-board cameras (TRK_RGB_1, TRK_RGB_2) at the locations shown in Figure 1. Table 1 lists the sensors while describing their respective characteristics.

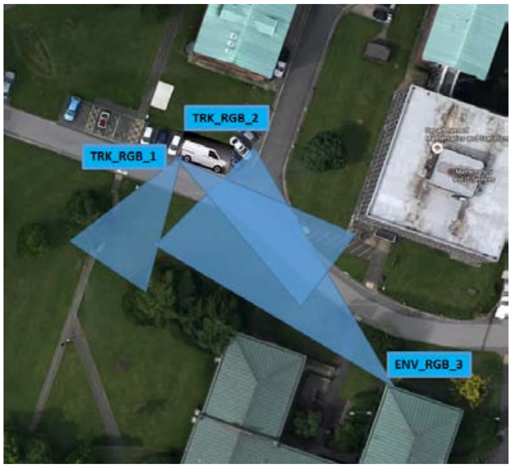

Figure 1. Sensor locations and their FOVs for ARENA dataset

2.1.3 Scenarios

The dataset scenarios ('Normal', 'Warning', 'Alarm') are listed in Table 2.

2.2. P5 dataset

2.2.1 Overview

The dataset contains sequences with different activities staged around the perimeter of the OKG nuclear plant outside Oskarshamn, Sweden. The dataset was recorded by P5 partners collectively by using multiple types of surveillance sensors including digital IP cameras and thermal sensors.

2.2.2 Camera setup and characteristics

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 Figure 2). It takes 10-15 minutes to walk from one end to the other.

Visible sensors. Three visual cameras (VS_1, VS_2, VS_3) at the locations shown in Figure 2 are selected to mainly cover the road along the water area. Most of the scenarios take place in the monitored region.

Thermal sensors. Two of the thermal sensors (TH_3, TH_4) as shown in Figure 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 Figure 2, which mainly cover the long road along the fence outside the nuclear plant.

The sensor properties are summarised in Table 3.

2.2.3 Scenarios

The dataset scenarios ('Normal', 'Warning', 'Alarm') are listed in Table 4.

Sample images from the cameras for both ARENA and P5 datasets are shown in Figure 3.

3. Challenge

The PETS 2015 Challenge addresses the application of automated sensor surveillance for the protection of critical

Table 2. List and description of scenarios for ARENA dataset

Scenario type	**ID**	**Description**	**Challenges**
Normal	N1_ARENA	Persons walking in a group	Scale change; occlusion; pose change
Warning	W1_ARENA	Driver falls after being hit by someone	Occlusion; scale change; person running
Alarm	A1_ARENA	Driver involved in a fight with someone	Scale change; pose change; occlusion; clutter
	A2_ARENA	Driver attacked by someone from a car	Scale change; speed change; occlusion

infrastructure. It aims to bring together researchers, practitioners and students from computer vision and surveillance-related fields to share knowledge on methodologies, features and results related to the evaluation, modelling and understanding of object motion and behaviour from video analytics.

Submissions are solicited that either:

- Describe an approach to low-level video analysis / mid-level video analysis / high-level video analysis (see below) and report results based on the datasets provided for this workshop. The actual results are also to be submitted in XML format which will be described in Section 4.2.
- Contribute to general performance evaluation methodology for detection, tracking and behaviour (threat) analysis. It is not necessary to explicitly consider the datasets provided for the workshop, however one is encouraged to use the PETS datasets that are made available.

The PETS 2015 datasets (i.e. ARENA dataset and P5 dataset) are designed to accommodate different categories in a typical video surveillance system: Low-level video analysis, Mid-level video analysis and High-level video analysis. Within each category, one or more specific vision tasks are further defined to address the diversity of the challenges as well as different level of complexity. These tasks are described next.

Object tracking

The task involves detecting and/or tracking objects in all frames of the video sequences specified for this task.

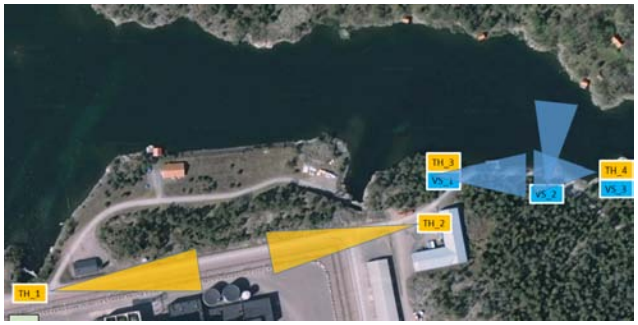

Figure 2. Sensor locations and their FOVs for P5 dataset

'Group walking' event detection

The task involves detection of the occurrence of the event of a group of people walking in a segment of the specified video sequences. A group is defined to consist of more than two people walking together.

'Person running' event detection

The task involves detection of the occurrence of the event of a person running in a segment of the specified video sequences.

'Threat' event detection

The task involves detection of the occurrence of a threat event in a segment of the specified video sequences. A threat event generally consists of a sequence of simpler atomic events.

The paper submission for PETS 2015 Challenge can take place for any (one or more) of the above tasks using ARENA and/or P5 datasets. The PETS 2015 Challenge indeed focuses on *single-camera processing only*. This means for a task, a participant must run its algorithm independently on each of the corresponding single-camera video sequences specified for this task. All sequences under each task for each category are listed in Table 5.

4. Evaluation Methodology

4.1. Ground truth

To enable a precise quantitative comparison and ranking of various algorithms, efforts are made to provide accurate

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

Table 4. List and description of scenarios for P5 dataset

Scenario type	**ID**	**Description**	**Challenges**
Normal	N1_P5	A vehicle driving across the scene	Scale change; pose change; speed change; clutter
Warning	W1_P5	A group of 6 people walking across the scene	Occlusion; scale change; clutter; speed change
Alarm	A1_P5	An abandoned bag is picked up suspiciously	Scale change; pose change; clutter; speed change

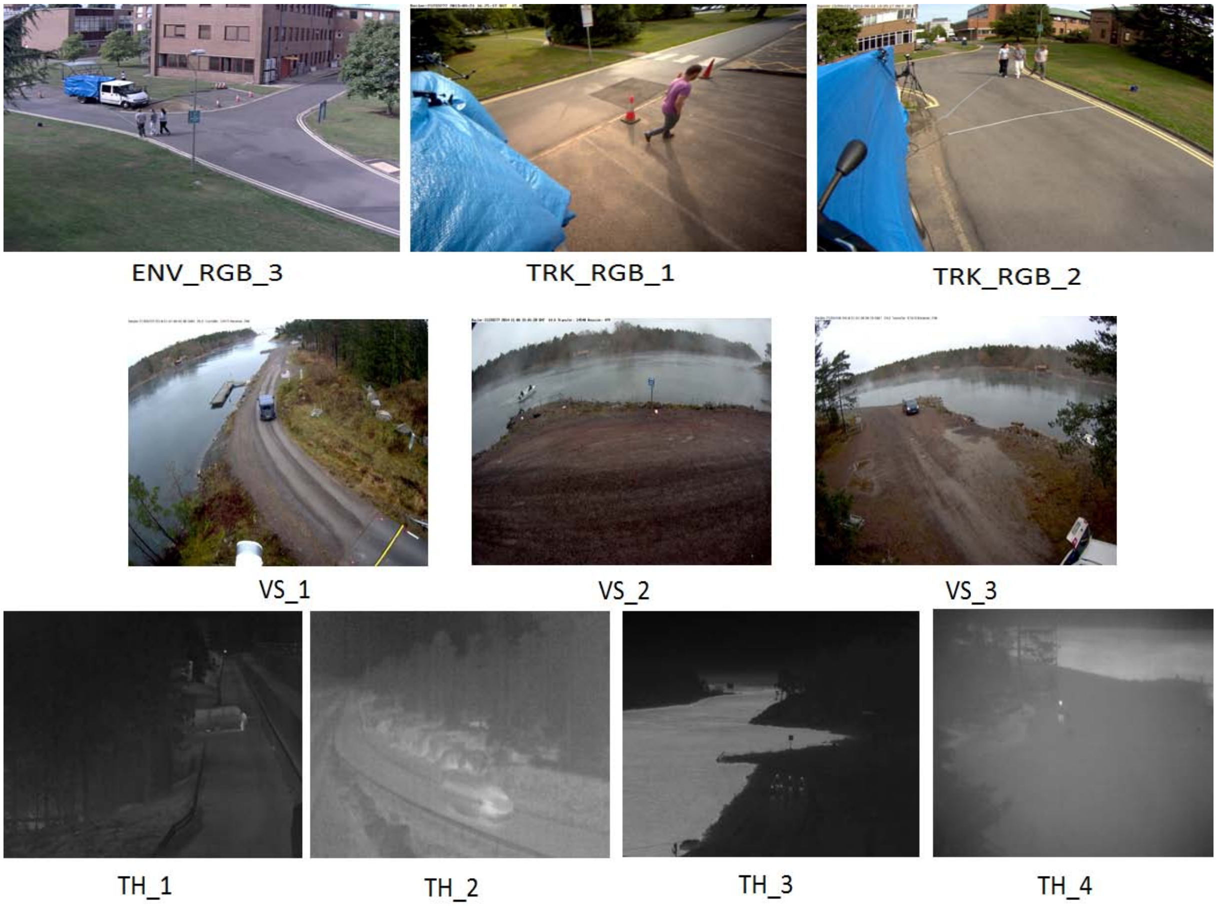

Figure 3. Top row: sample images from ARENA dataset; middle row: sample images for visual sensors from P5 dataset; bottom row: sample images for thermal sensors from P5 dataset.

Table 5. List of sequences under each task for PETS 2015 Challenge

Task	**Category**	**Sequences (ARENA dataset)**	**Sequences (P5 dataset)**
Object tracking	Low-level analysis	N1_ARENA-Tg_ENV_RGB_3; N1_ARENA-Tg_TRK_RGB_1; N1_ARENA-Tg_TRK_RGB_2; W1_ARENA-Tg_ENV_RGB_3; W1_ARENA-Tg_TRK_RGB_1; A1_ARENA-Tg_ENV_RGB_3; A1_ARENA-Tg_TRK_RGB_2	N1_P5-Tg_VS_1; N1_P5-Tg_VS_3; N1_P5-Tg_TH_1; N1_P5-Tg_TH_2; W1_P5-Tg_VS_1; W1_P5-Tg_VS_3; W1_P5-Tg_TH_3; A1_P5-Tg_VS_2; A1_P5-Tg_TH_3
’Group walking’ event detection	Mid-level analysis	N1_ARENA-Gp_ENV_RGB_3; N1_ARENA-Gp_TRK_RGB_1; N1_ARENA-Gp_TRK_RGB_2	W1_P5-Gp_VS_1; W1_P5-Gp_VS_3; W1_P5-Gp_TH_3
’Person running’ event detection	Mid-level analysis	W1_ARENA-Rg_ENV_RGB_3; W1_ARENA-Rg_TRK_RGB_1; A2_ARENA-Rg_ENV_RGB_3	W1_P5-Rp_VS_1; W1_P5-Rp_VS_3; W1_P5-Rp_TH_3
’Threat’ event detection	High-level analysis	A1_ARENA-Tt_ENV_RGB_3; A1_ARENA-Tt_TRK_RGB_2 A2_ARENA-Tt_ENV_RGB_3; A2_ARENA-Tt_TRK_RGB_2	A1_P5-Tt_VS_2; A1_P5-Tt_TH_3 A1_P5-Tt_TH_4

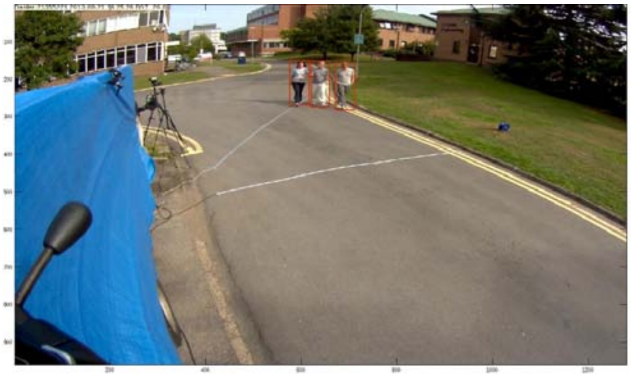

Figure 4. Object tracking annotation: A sample annotated image for a camera view with a red bounding box overlaid around each object.

and detailed annotations for PETS 2015 datasets. Effectively two types of ground truth annotation are obtained, corresponding to object tracking and event detection tasks as discussed in Section 3. The annotation process and format are describe in detail in the following sub sections.

4.1.1 Object tracking

For the Object Tracking task, the aim is to detect and track objects in all frames of the video sequences specified for this task. Consequently, the ground truth is abtained for every single frame of all the sequences within this task. The annotation is obtained in the format of bounding box which effectively encloses each object in each frame. Figure 4 shows a sample annotated image for a camera view with red bounding boxes overlaid around each object. Note in the case of occlusion, only the visible part of the object is annotated. Table 6 presents a summary for the annotation task performed for object tracking.

Table 6. Annotation summary for object tracking task

Sequence	# Frames	# Object	Object Type
N1_P5-VS_1	400	1	Vehicle
N1_P5-VS_3	387	1	Vehicle
N1_P5-TH_1	600	1	Vehicle
N1_P5-TH_2	220	1	Vehicle
W1_P5-VS_1	180	6	Person
W1_P5-VS_3	180	6	Person
W1_P5-TH_3	740	6	Person
A1_P5-VS_2	720	1	Boat
A1_P5-TH_3	1000	1	Boat
N1_ARENA-Tg_ENV_RGB_3	289	5	Person
N1_ARENA-Tg_TRK_RGB_1	513	5	Person
N1_ARENA-Tg_TRK_RGB_2	684	5	Person
W1_ARENA-Tg_ENV_RGB_3	155	3	Person
W1_ARENA-Tg_TRK_RGB_1	240	3	Person
A1_ARENA-Tg_ENV_RGB_3	295	4	Person
A1_ARENA-Tg_TRK_RGB_2	670	4	Person

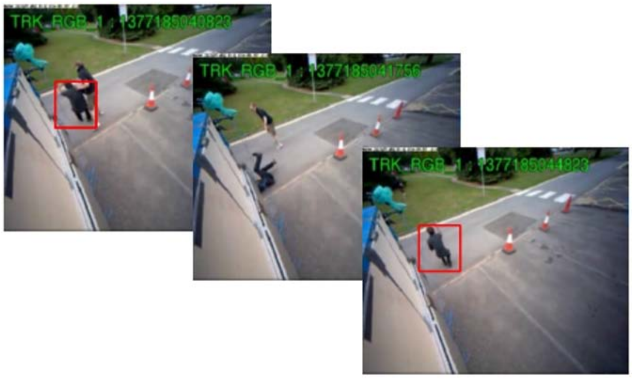
TRK_RGB_1 : 1377185040823
TRK_RGB_1 : 1377185041756
TRK_RGB_1 : 1377185044823

Figure 5. Event detection annotation: An example with start and final frame numbers of an event annotated and the bounding box overlaid around the object under consideration in the two frames.

4.1.2 Event detection

For the event detection task, the event is annotated by defining the start and end frame of the event, as well as the bounding box(es) for the main target(s) involved in the event. An example of the annotation of an event detection annotation is illustrated in Figure 5, with a red box overlaid around the person in the start and final frames of the defined event.

4.2. Submission format

Collectively researchers develop their algorithms on different computing platforms, use a variety of programming languages and typically store their algorithm results in their own data structures. A standard file format for the submission of results is therefore required in order to evaluate different researchers' methods. For PETS 2015 workshop, a requirement was set that submitted papers were accompanied with algorithmic results in XML format. This way, the object detection and tracking results can be reconstructed from the XML files, which provided for a qualitative comparison of a number of algorithms operating on the same PETS video sequences.

The submitted XML file must conform to the PETS 2015 XML Schema which can be found on the PETS 2015 website [5]. An XML Schema is a definition on how to construct a valid XML file.

4.3. Evaluation metrics

Object tracking: 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) [3]. For the Challenge, the widely-used Multiple Object Tracking Accuracy (MOTA) [2] is used, 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 [2]. 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 [1]. MOTA $\leq$ 1: the higher MOTA, the better the performance. To evaluate tracking accuracy, a recently-introduced measure, Multiple Extended-target Lost-Track ratio (MELT) [3], is used. MELT provides 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

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

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.

Event detection: Let K_{ini}^{est} and K_{end}^{est} be start and final frames, respectively, for a detected event by a candidate algorithm in a sequence, and K_{ini}^{gt} and K_{end}^{gt} be the corresponding ground-truth information for the start and final frames, respectively, of the same event. The event detection error, E, is then computed as a distance between the estimated and ground-truth information:

$$E = \sqrt{(K_{ini}^{est} - K_{ini}^{gt})^2 + (K_{end}^{est} - K_{end}^{gt})^2}. \tag{3}$$

$E \geq 0$: the lower the value, the better the performance.

5. Conclusions

In this paper, we have described the datasets and the Challenge that are part of PETS 2015 workshop. Two datasets are presented that are the ARENA and P5 datasets. Each dataset is divided into three categories: Normal, Warning and Alarm. The ARENA dataset contains visible imagery recorded from multiple sensors installed onboard a critical asset (a truck) or in the environment. The P5 dataset contains both visible and thermal imagery recorded from multiple sensors installed outside the perimeter of a critical infrastructure (a nuclear power plant). The datasets account for the key challenges that are of interest within the community. The Challenge includes different surveillance tasks including object tracking, group walking event detection, person running event detection and threat event detection. The Challenge allows participants to run their algorithms for the specified tasks using the sequences provided under two datasets, and to submit the results that are then evaluated using the well established measures.

References

[1] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *Proc. of IEEE CVPR*, pages 3457–3464, 2011.

[2] 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.

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

[4] L. Patino and J. Ferryman. PETS 2014: Dataset and Challenge. In *Proc. of IEEE Int. Conf. on AVSS*, pages 355–360, Seoul, August 2014.

[5] IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS). http://pets2015.net/.