Manuscript Number: PRLETTERS-D-13-00285R1

Title: Performance Evaluation of Crowd Image Analysis using the PETS2009 Dataset

Article Type: Special Issue: SIPRCA

Keywords: Surveillance; Detection; Tracking; Performance Evaluation; Dataset

Corresponding Author: Dr. James Ferryman, Ph.D.

Corresponding Author's Institution: University of Reading

First Author: James Ferryman, Ph.D.

Order of Authors: James Ferryman, Ph.D.; Anna Ellis, Ph.D.
**Pattern Recognition Letters**  
**Authorship Confirmation**

Please save a copy of this file, complete and upload as the “Confirmation of Authorship” file.

As corresponding author I, Dr, James Ferryman, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal. If presented at a conference, the conference is identified. If published in conference proceedings, the publication is identified below and substantial justification for re-publication must be presented.

2. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.

3. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Date 28th February 2013

Some of the material describing the PETS2009 dataset were presented in the paper:

**Previous conference presentation**

As above.

**Previous conference proceedings publication**

As above.

**Justification for re-publication**

This paper contains details of significant new performance evaluation results based on the PETS2009 dataset.
Re: Submission of manuscript to SI: PRCA

Dear Sirs,

Please find enclosed a revised version of the manuscript entitled “Performance Analysis of Crowd Image Analysis using the PETS2009 Dataset” by Ferryman et al., for submission to Pattern Recognition Letters Special Issue on Pattern Recognition and Crowd Analysis (PRCA).

In this work, the PETS2009 outdoor crowd image analysis surveillance dataset is presented and the performance evaluation of people counting, detection and tracking results, using the dataset, submitted to five IEEE Performance Evaluation of Tracking and Surveillance (PETS) workshops.

The evaluation highlights the detection and tracking performance of the authors’ methods in areas such as precision, accuracy and robustness, allowing a comparative analysis to be performed.

I look forward to hearing from you in due course,

Yours sincerely,

Dr. James Ferryman
Re: Submission of manuscript to SI: PRCA

Dear Sirs,

Please find enclosed a manuscript entitled “Performance Analysis of Crowd Image Analysis using the PETS2009 Dataset” by Ferryman et al., for submission to Pattern Recognition Letters Special Issue on Pattern Recognition and Crowd Analysis (PRCA).

In this work, the PETS2009 outdoor crowd image analysis surveillance dataset is presented and the performance evaluation of people counting, detection and tracking results, using the dataset, submitted to five IEEE Performance Evaluation of Tracking and Surveillance (PETS) workshops.

The evaluation highlights the detection and tracking performance of the authors’ methods in areas such as precision, accuracy and robustness, allowing a comparative analysis to be performed

I look forward to hearing from you in due course,

Yours sincerely,

Dr. James Ferryman
Title: Performance Evaluation of Crowd Image Analysis using the PETS2009 Dataset

The authors thank the reviewers and editor for their positive and constructive comments which have been carefully taken into consideration in the revision. The minor revisions made to the article, based on the reviewers’ and editors’ comments, are summarized in the table below.

<table>
<thead>
<tr>
<th>Editor/Reviewer Comment</th>
<th>Authors’ Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emphasize the problems in designing benchmarks.</td>
<td></td>
</tr>
<tr>
<td>I would encourage the authors to revise [the paper] in a way that will emphasize</td>
<td></td>
</tr>
<tr>
<td>on the abstract problem of benchmark design.</td>
<td></td>
</tr>
<tr>
<td>The discussed protocol that generates the benchmark dataset seems to be very specific</td>
<td></td>
</tr>
<tr>
<td>to the problem of crowd image analysis.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A new section 2.1 on Benchmark Design has been added to include details on the</td>
</tr>
<tr>
<td></td>
<td>general challenges in designing benchmarked datasets for the surveillance community.</td>
</tr>
<tr>
<td>Table 3 on page 3 should be Table 2.</td>
<td></td>
</tr>
<tr>
<td>Under Acknowledgements there is a strange reference named XXX.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>These minor errors have been corrected.</td>
</tr>
</tbody>
</table>

In addition further details have been added to Sections 4.1 and 4.2.

Further, a new Section 6.3 has been added to evaluate the metrics themselves and to provide further insight into the authors’ algorithms. The authors would value the reviewers’ opinion on whether these details represent an informative addition to the paper.
Pattern Recognition Letters

Title: Performance Analysis of Crowd Image Analysis using the PETS2009 Dataset

Article Type: SI: PRCA

Keywords: surveillance, detection, tracking, performance evaluation, dataset

Corresponding author: J. Ferryman, email: james@computer.org

Highlights:

- Presents PETS2009 crowd image analysis dataset
- Comparative performance evaluation of analysis submitted to five PETS workshops
- Highlights detection and tracking performance on PETS2009 dataset
Performance Evaluation of Crowd Image Analysis using the PETS2009 Dataset

James Ferryman\textsuperscript{a,}\textsuperscript{*}, Anna-Louise Ellis\textsuperscript{a},

\textsuperscript{a}\textit{Computational Vision Group, School of Systems Engineering, Whiteknights, University of Reading, Reading RG6 6AY, UK}

Abstract

This paper presents the PETS2009 outdoor crowd image analysis surveillance dataset and the performance evaluation of people counting, detection and tracking results using the dataset submitted to five IEEE Performance Evaluation of Tracking and Surveillance (PETS) workshops. The evaluation was carried out using well established metrics developed in the Video Analysis and Content Extraction (VACE) programme and the Classification of Events, Activities, and Relationships (CLEAR) consortium. The comparative evaluation highlights the detection and tracking performance of the authors’ systems in areas such as precision, accuracy and robustness and provides a brief analysis of the metrics themselves to provide further insights into the performance of the authors’ systems.

Keywords:
surveillance, detection, tracking, performance evaluation, dataset

1. Introduction

Visual surveillance is a major research area in computer vision. The large number of surveillance cameras in use has led to a strong demand for auto-
matic methods of processing their outputs. The scientific challenge in crowd
image analysis is to devise and implement methods for obtaining detailed
information about the number, density, movements, and actions involving
people observed by a single camera or by a network of cameras. The growth
in the development of the field, however, has not been met with comple-
mentary systematic performance evaluation of developed techniques using a
common benchmark. It has been especially difficult to make comparisons be-
tween algorithms if they have been tested on different datasets under widely
varying conditions.

To address this need a new crowd image analysis dataset called PETS2009
was devised, collected and disseminated to the wider community. In addi-
tion, a series of five dedicated consecutive workshops (PETS2009, Winter-
PETS2009, PETS2010, PETS2012, and PETS2013) were held with the pre-
requisite that paper submissions on crowd image analysis be accompanied
by XML results based on processing the dataset.

Section 2 reviews the issues in designing benchmarks as well as benchmark
datasets that are readily available to the computational vision community
for development and testing of crowd image analysis methodology. Section 3
describes the PETS2009 dataset and the ground truth annotation. Section 4
provides an overview of the techniques the participating authors used to
address the challenges presented within the dataset. A brief description of
the evaluation methodology follows in Section 5, and analytic discussion of
the overall performances is provided in Section 6. Concluding remarks and
future work are given in Section 7.
2. Related Work

There is a rising demand for quantitative performance evaluation of automated video surveillance. To advance research in this area, it is essential that comparisons in detection and tracking approaches may be drawn and improvements in existing methods can be measured. There are a number of challenges related to the proper evaluation of detection, tracking, event recognition, and other components of a crowd image analysis system that are unique to the video surveillance community. These include the volume of data that must be evaluated, the difficulty in obtaining ground truth data, the definition of appropriate metrics, and achieving meaningful comparison of diverse systems.

This section reviews the issues in designing benchmarks as well as existing representative datasets that are publicly available to assist in the evaluation of performance crowd image analysis, focussing on robust detection and tracking solutions. Most provide a set of data for training a system, if necessary, and a separate set for testing. There are some that require the evaluation set to remain unseen.

2.1. Benchmark Design

The challenges in creating benchmark datasets for the performance evaluation of automated visual surveillance methods are broad. The main aim of automated surveillance is frequently to locate and track objects of interest or to determine specific events and/or behaviours involving the objects and/or the environment. In creating datasets whose content may be analysed by algorithms/systems, these objects, events and behaviours must be presented
within the recorded scenarios in a realistic and meaningful way. These in-
clude, but are not limited to, varying scene conditions such as weather and
illumination/lighting (including moving shadows and reflections) - which are
not of interest and generally hamper detection/tracking methods - 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. The creation of ground truth for
the evaluation of detection, tracking and event/behaviour 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.
Furthermore, in addition to producing a dataset that may be used as an
evaluation benchmark for a broad spectrum of developed automated surveil-
lance methodology, adopting well known and established metrics may assist
more readily researchers tasked to establish comparisons in the performance
of state of the art visual surveillance systems.

2.2. Dataset Based Challenges

As depicted in Table 1, both real world and simulated surveillance footage
datasets are available to the computer vision community. They bring differ-
ent strengths to an automated crowd image analysis research project. Real
world footage highlights the numerous challenges faced in developing ro-
 bust solutions, such as rapidly shifting light levels and shadows due to ever
changing cloud cover and reflective surfaces. Simulated footage offers exact
control over content, can provide numerous camera angles and a means of automating time consuming ground-truthing. Automated visual surveillance from the footage shot with a single camera can be affected by shortcomings such as occlusions (where one person moves in front of some other person or object), illumination differences and complex movements. Systems commonly require a multicamera configuration approach that may be used to overcome the limitations of the single camera configurations. Occasionally, however, the footage from a single camera is the only appropriate solution and is discussed later in this section. Whilst quality cameras may produce higher resolution images at a fast frame rate, this may not be representative of actual CCTV surveillance footage. Additionally the availability of ground truth is one of the biggest barriers in assessing the performance of new and innovative techniques for automated visual surveillance.

The collective datasets of Project ETISEO (Nghiem et al., 2007) consist of indoor and outdoor scenes, corridors, streets, building entries, a subway station and an airport apron. For some scenarios, the researchers providing the available datasets recognised that in addition to multiple cameras it is entirely possible that the use of multiple image modalities may bring further benefits towards developing robust solutions. The ETISEO project presents many of its scenes as multicamera datasets and some include additional imaging modality such as infrared footage.

The VIRAT video dataset (Oh et al., 2011) was designed to contain a wide range of human activity/event categories than previously released datasets. The dataset includes a wide range of resolutions and frame rates, realistic and natural scenes, diverse types of human actions as well as vehicles and both
ground camera views and aerial views. The TREC Video Retrieval Evaluation (TRECvid) series (TRECvid) is sponsored by the National Institute of Standards and Technology (NIST) and other U.S. government agencies. It promotes progress in content-based analysis of and retrieval from digital video; including automatic segmentation, indexing, summarisation and content-based retrieval of digital video broadcast news, documentary, and education programming.

Addressing the need to investigate the ability to automate visual surveillance at night, the Multicamera Human Action Video Data (MuHAVi) (Singh et al., 2010) was created as a part of the EPSRC funded REASON project. It is set in large laboratory and uses real night time street light illumination, and uneven paved surfaces. The video footage has sequences of actions, performed by actors, such as walk and turn back, run and stop, punch, and collapse.

In contrast to the aforementioned datasets, Virtual Human Action Silhouette DataSoftware (ViHASi) (Hossein et al. (2008)) places emphasis on the actions performed. Software developed for animation and film industry was employed to generate videos of 20 different actions such as run, walk, punch, and collapse and was created using 9 different virtual actors. The motion data correspond to the actions performed previously by human actors using optical or magnetic motion capture in order to produce realistic results. One of the biggest benefits of the ViHASi datasets, which uses only virtual cameras, is the copious amount of viewpoints (up to 40) available for the captured actions.

The image library for intelligent detection systems, i-LIDS (i-LIDS, 2007)
is the UK government's benchmark for Video Analytic (VA) systems developed in partnership with the Centre for the Protection of National Infrastructure. These datasets make up five scenarios which include sterile zone monitoring, amongst others. The footage includes an outside fenced area where objects of interest may appear in the detection zone. Further footage of Gatwick airport, UK, was made available to the TREC Vid project. While the dataset is useful for evaluating detection algorithms it remains limited because some parts of the dataset are monocular and it also does not contain examples of specific behavioural interactions.

The Context Aware Vision using Image-based Active Recognition (CAVIAR) (CAVIAR) project datasets include two sets of video clips filmed at separate locations; the first being a building's entrance lobby and the second an indoor city shopping centre. The footage was shot to address analysis of city centre surveillance, both from the view of antisocial behaviour and that of potential customers in a commercial setting. The CAVIAR benchmark datasets collectively show people walking alone, meeting with others, window shopping, and entering and exiting shops, fighting, passing out and leaving a package in a public place. The shopping centre dataset includes two separate viewpoints that are time synchronised.

Finally, the series of international workshops on Performance Evaluation of Tracking and Surveillance (PETS) endeavours to provide a platform for the objective evaluation of all participants' systems on the same dataset. There are various datasets available to download including, but not limited to: PETS2001 concentrates on an outdoor car park and street setting displaying pedestrians and moving vehicles. It is unique in that it includes the
view from a catadioptric camera alongside four independent standard views. In addition to these, an alternative scenario provides front and rear view footage from a moving vehicle. PETS2006 (PETS2006) provides seven indoor scenarios recorded at a busy city train station and include unattended luggage of various shapes and sizes left by actors and provide 4 separate views; 2 at ground level and 2 elevated shots. The scenarios involve an increasing number of people and passers-by. The simplicity of the scenarios allows very limited situation awareness and was designed mainly to test if the low level processing stages are sufficient to cope with real-world scenarios. The PETS2007 dataset (PETS2007) focuses on two additional indoor scenarios: theft and loitering, with increasing scene complexity. The videos are much more challenging from the tracking point of view as the scenes are more crowded. There are 8 scenarios, each viewed from 4 cameras (2 at ground level and 2 elevated shots). The situation awareness in the PETS2007 challenge is relatively simple - reduced to comparing the distance of a bag to its owner (abandoned bag, theft) or measuring the time for which a person stays in the scene (loitering).

2.3. Limitations of Existing Approaches

The computer vision community lacks a standard database of videos that can be used for the evaluation of crowd image analysis approaches. 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.
Table 1: Summary of Surveyed Surveillance Datasets

<table>
<thead>
<tr>
<th>Project &amp; Reference</th>
<th>Real World Activities</th>
<th>Simulated</th>
<th>#Cameras</th>
<th>Footage Quality</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIRAT</td>
<td>Natural Scenes/Human Actions</td>
<td>No</td>
<td>Multiple</td>
<td>Various</td>
<td>No</td>
</tr>
<tr>
<td>ETISEO</td>
<td>Inside Metro/Airport Apron</td>
<td>No</td>
<td>4,2,1</td>
<td>Mixed Sensors</td>
<td>Permission Required</td>
</tr>
<tr>
<td>i-LIDS</td>
<td>Various</td>
<td>No</td>
<td>4,1</td>
<td>PAL or less</td>
<td>No</td>
</tr>
<tr>
<td>PETS’01</td>
<td>Street Scene</td>
<td>No</td>
<td>4,2</td>
<td>768x576 25fps</td>
<td>No</td>
</tr>
<tr>
<td>PETS’06</td>
<td>Train Station</td>
<td>No</td>
<td>4</td>
<td>768x576 25fps</td>
<td>No</td>
</tr>
<tr>
<td>PETS’07</td>
<td>Airport Terminal</td>
<td>No</td>
<td>4</td>
<td>768x576 25fps</td>
<td>No</td>
</tr>
<tr>
<td>PETS’09</td>
<td>Crowds, Outdoors</td>
<td>No</td>
<td>8</td>
<td>768x576 25fps</td>
<td>Yes</td>
</tr>
<tr>
<td>MUHAVI</td>
<td>Human Actions</td>
<td>Street Lights</td>
<td>4</td>
<td>720x576; 704x576; 25fps</td>
<td>Yes</td>
</tr>
<tr>
<td>ViHASi</td>
<td>No</td>
<td>Avatar Actions</td>
<td>40</td>
<td>640x480 25fps</td>
<td>Yes</td>
</tr>
<tr>
<td>CAVIAR</td>
<td>Inside Shops/Offices</td>
<td>No</td>
<td>2</td>
<td>374x288 25fps</td>
<td>Yes</td>
</tr>
<tr>
<td>TRECVid</td>
<td>Various</td>
<td>No But Some Edited Footage</td>
<td>Varies</td>
<td>Multiple (Broadcast Quality)</td>
<td>Some</td>
</tr>
</tbody>
</table>


3. PETS2009 Dataset

To address the need for outdoor surveillance style footage of the behaviour of individuals within crowds, using multiple views, three datasets were recorded at Whiteknights Campus, University of Reading, UK. Unlike the datasets reviewed in Section 2, these datasets offer footage of an outdoors scene with real world unpredictable weather patterns present whilst providing various crowd behaviours using surveillance quality cameras. The scenes themselves contain between one to forty individuals performing a multitude of behaviours and comprise multi-sensor sequences with increasing scene complexity. These included scenarios specifically created for person count and density estimation (Dataset S1), people detection and tracking (Dataset S2) and flow analysis and event recognition (Dataset S3). The dataset has 8 camera views in total with 4 of these positioned at ground level and 4 placed in an elevated position, as shown in Figure 1.

Training sequences were also provided to allow researchers to create the following models: Background including frames containing people or other
moving objects; City Centre including random, walking crowd flow; Regular Flow including regular walking pace crowd flow.

3.1. Challenges

3.1.1. Dataset S1 - People Count and Density Estimation

Challenge S1.L1 required reporting the count of the number of individuals within Regions R0, R1 and R2 for each frame of the sequence, for View 1 only. Figure 2(a) depicts the regions R0, R1 and R2 overlaid on View 1. This particular sequence exhibited regular crowd movement in a relatively dense queue in two directions. Challenge S1.L2 was to report the total number of walking individuals who had passed an entry point in the scene and the total number of individuals who have passed exit points, specified in for View 1 for each frame of the sequence. Challenge S1.L3 was to report the crowd density in Region R1 for each frame of the sequence where individuals start to run, from walking, in the middle of the first sequence and converge from walking into a group of dense stationary people at the end of a second sequence. For dataset S1.L3 crowd density was based on a maximum occupancy of 40 individuals in 10m² on the ground. One individual was assumed to occupy 0.25m² on the ground.

The detection and tracking challenges within Dataset S2 included tracking all of the individuals within a sequence and reporting the bounding box of each individual for every frame of that sequence. Results could be submitted for each individual view. Figure 2(b) provides representative snapshots depicting a randomly walking sparse crowd.
3.1.2. Dataset S2 - People Detection and Tracking

Challenge S2.L1 was to track all of the individuals within Sequence 1, reporting the bounding box (2D, and optionally, 3D) of each individual for every frame of the sequence exhibiting a randomly walking sparse crowd. Challenge S2.L2 was to detect and track two labelled individuals and to report the bounding box of both individuals for every frame from a sequence exhibiting a randomly walking dense crowd. Challenge S2.L3 again was to detect and track two labelled individuals and to report the bounding box (2D, and optionally, 3D) of both individuals for every frame, however there are two separate individuals that are bystanders in an empty scene which later join the very dense moving crowd walking in the same direction.

The dataset also included (but was not limited to) additional challenges such as: multiple occlusions of individuals by others that were to be detected and tracked; partial camouflage of individuals by others occluding them and
wearing similar coloured clothing or of stationary objects in the scene of a
similar colour; the movement of objects that were not part of the challenge
yet present in the scene; changes in lighting due to rapid movements of
clouds; static objects partially obscuring those that were to be detected and
tracked in certain views. These additional challenges are representative of
those found in many outdoor surveillance scenes. The specific additional
challenges per view for both object detection and tracking are listed in Table
2.

3.1.3. Dataset S3 - Flow Analysis and Event Recognition

Finally multiple flow and event recognition challenges were presented in
Dataset S3. For multiple flows the challenge was to detect and report the
separate, individual flows in one or more of the sequences. The sequences pro-
vided were: a sparse queue of people walking regularly along a linear path and
curving to avoid a virtual obstacle; a denser crowd walking in a queue while
traversing around a human wall; merging groups of people walking together;
two individuals walking and advancing their way against a dense queue of
people walking in the opposite direction; 3 individuals walking against the
crowd were wearing bright jackets. For the event recognition challenge the
following dense crowd scenarios were provided: walking, running, evacuation
(rapid dispersion of the crowd moving in different directions), local disper-
sion (localised movement of people within a crowd), crowd formation (gathering/merging) and crowd dispersal-splitting (cohesive crowd which splits into
two or more entities). For each sequence, and for each frame, a probabilistic
measure of each event (multi-class, all events) could be reported.
<table>
<thead>
<tr>
<th>View</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>Fluttering tape in the background crossing 4/5 of the scene width. Car boot being opened by individual.</td>
</tr>
<tr>
<td>Three</td>
<td>Leafless tree partially occluding 1/3 of the scene, Vehicles moving in the background.</td>
</tr>
<tr>
<td>Four</td>
<td>Fluttering tape in the foreground crossing 9/10 of the scene width (motion closest to camera). Car boot being opened by individual. Movement behind translucent window. Reflective windows. Occluding leafless tree.</td>
</tr>
<tr>
<td>Five</td>
<td>Shadows. Individuals appear both very large and close to the camera and small and in the far distance.</td>
</tr>
<tr>
<td>Six</td>
<td>Blue tint and lack of natural lighting to entire footage. Fluttering tape in the background crossing 4/5 of the scene width. Shadows. Individuals appear both very large and close to the camera and small and in the far distance.</td>
</tr>
<tr>
<td>Seven</td>
<td>Vehicles moving in the background, Trees moving in the wind, Small section of tape moving in the wind.</td>
</tr>
<tr>
<td>Eight</td>
<td>Very prominent shadows.</td>
</tr>
</tbody>
</table>
3.2. Ground Truth

The ground truth for a sub sampled set of frames was obtained for each sequence with the average sampling frequency being 1 frame in every 7 frames. Each of the seven independent 2D camera views (views 1, 3, 4, 5, 6, 7, 8) were ground truthed using the Video Performance Evaluation Resource (ViPER) ground truth tool. This provided the necessary bounding boxes and their identifying key and location for accurate evaluations. This set of ground truth was chosen as errors in calibration due to the approximation of the ground surface as a plane, in addition to radial distortion are found to occur in ground truth created from a 3D perspective. The ground truth is available in ViPER XML format (object based as opposed to frame based) for views 1, 3, 4, 5, 6, 7 and 8 and may be downloaded from pets2009.net

4. People Counting, Detection and Tracking

This section introduces the techniques that the participating authors used to address the challenges set in the detection, tracking and people counting tasks. In this paper the focus is on evaluating the people counting (Dataset S1) challenge, and the detection and tracking challenge (Dataset S2.L1.12-34) due to the fact that the majority of the submitted evaluations and papers were dedicated to these tasks. An overview of each authors system is given and a summary of the main methods used for people counting is given in Table 3 and detection and tracking in Table 4. Details of how these various systems performed against the described PETS 2009 dataset and an analysis of this performance, in relation to the techniques used, are given in Section 6. Greater detail of each authors methodology may be found in their corre-
Table 3: Participating Authors and Summary of Their Methods for PETS09 People Counting Challenge

<table>
<thead>
<tr>
<th>Author</th>
<th>People Counting Method</th>
<th>Single/Multi Camera</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alahi et al. (2009)</td>
<td>Foreground segmentation based on motion; Generative person model; Sparsity modelling; General Scene Surface (can handle dense crowd)</td>
<td>Single, Multiple</td>
</tr>
<tr>
<td>Albiol et al. (2009)</td>
<td>Salient Points motion estimation</td>
<td>Single</td>
</tr>
<tr>
<td>Chan et al. (2009)</td>
<td>Mixture models to capture global feature attributes; Regression function</td>
<td>Single</td>
</tr>
<tr>
<td>Choudri et al. (2009)</td>
<td>Learning-based using foreground pixel-maps; Human region of interest detector</td>
<td>Single</td>
</tr>
<tr>
<td>Sharma et al. (2009)</td>
<td>Pedestrian detection based on Cluster-Boosted-Trees and non-parametric tracking model</td>
<td>Single</td>
</tr>
<tr>
<td>Conte et al. (2010)</td>
<td>SURF-based feature extraction; Regression function</td>
<td>Single</td>
</tr>
<tr>
<td>Paetzold et al. (2010)</td>
<td>Shape (HOG-based), motion model, optical flow, probability map</td>
<td>Single</td>
</tr>
<tr>
<td>Morerio et al. (2012)</td>
<td>Blob estimation; Projection; Mapping to density based on calibration</td>
<td>Single</td>
</tr>
<tr>
<td>Subburaman et al. (2012)</td>
<td>Head detector based on Boosted Integral Features</td>
<td>Single</td>
</tr>
<tr>
<td>Eiselein and Senst (2013)</td>
<td>PHD Filter; Multi-object tracking; RLOF</td>
<td>Single</td>
</tr>
</tbody>
</table>

285 sponding paper publication and a reference to this publication is provided in Tables 3 and 4.
4.1. People Counting

For people counting, Alahi et al. (Alahi et al., 2009) creates degraded foreground silhouettes from some binary silhouette image and its approximation, using rectangular and ellipse shapes. These then help form the input to a Multi-Silhouette Dictionary which is made up of atoms modelling the presence of individuals at given locations on an occupancy grid. The atoms are generated using homography mapping points in a three-dimensional scene to their two-dimensional coordinates in the planar view. An occupancy vector is constructed from the observed data using a re-weighted Lasso method. A non-regular spaced sampling process is used to discretize the ground in order to resolve any differences in camera resolutions between viewpoints, and the sparsity of the people present.

Albiol et al.s (Albiol et al., 2009) system is based on the statistical analysis of the motion vectors of corners. The corner points are detected using a technique inspired by the Harris method. A motion vector with respect to the previous frame (backward estimation) is estimated using a multi-resolution block matching technique for each detected corner. In order to estimate the number of people it is assumed that a person contains a constant average number of corners on it.

Chan et al. (Chan et al., 2009) segment the video into crowd regions moving in different directions, using a mixture of dynamic textures. For each crowd segment a perspective map is applied to weight each location according to its approximate size in the real scene. Various features are extracted from the crowd segments and the number of people per segment is estimated from the feature vector with Gaussian process regression.
Choudri et al.’s (Choudri et al., 2009) method is based on scale-weighted pixel counting to approximate the number of people in a region of interest. The Colour Mean and Variance algorithm, applied to HSV colour space provides the initial segmentation. The foreground pixels are explicitly classified and pixels that are not of interest are updated as background. A human region of interest detector is implemented with heads being the chosen feature for human detection. Head detections which are smaller than a defined region of interest depth and encompass a foreground detection area are used as acceptance regions for the foreground mask.

Sharma et al. (Sharma et al., 2009) use a cluster boosted tree classifier for pedestrian detection which consists of edgelet features. Tracking is achieved by Data Association based Tracking which involves an affinity model that measures the likelihood of two tracklets/detections belonging to the same target and an optimization framework. Both aspects are enhanced by a hierarchical association framework and learned non-parametric affinity models. People are counted by analysing the final tracks within any specific region.

In Conte et al. (Conte et al., 2010) SURF based feature extraction is performed followed by application of a regression function to estimate the number of people present in the scene.

Paetzold et al. (Paetzold et al., 2010) trained a detector to find the upper body region of a human and combined the shape model with a uniform motion model, generated using optical flow information. This produces a probability map with uniform motion and characteristic shape. Discrete detections, which are associated to trajectories using the motion information, are obtained from the map. An algorithm is applied that which validates the
trajectories associated with a human and a count is made of the resulting
detections to identify the number of individuals in a scene.

Morerio et al. (Morerio et al., 2012) perform blob detection and projec-
tion, mapping to a density estimate. The method utilises camera calibration
and a training phase to learn the model parameters.

Subburaman et al. (Subburaman et al., 2012) employ a head detector
based on boosted integral features. Background subtraction and an inter-
est point detector are also applied to reduce the search region for the head
detector.

Eiselein and Senst (Eiselein and Senst, 2013) employ a novel motion-
enhanced (to account for missed detections) probability hypothesis density
for real-time tracking and inference of the number of people present in the
scene.

4.2. Detection and Tracking

Bolme et al. (Bolme et al., 2009) approaches the detection challenge with
the filtering method Average of Synthetic Exact Filters which considers the
entire output of the filter under a full convolution operation. The approach
is compared to two other detectors: Viola and Jones et al."s cascade classifier
with both visual and motion features used for detection, and the deformable
parts model system created by Felzenswalb et al.

Zweng et al. (Zweng et al., 2013) perform person detection using a Rela-
tional Feature Model (RFM) combined with histogram similarity functions.
Detection is performed using a multi-scale overlapping sliding window ap-
proach. Relational features using the HOG descriptor compute the simi-
larity between histograms of the HOG descriptor. The obtained histogram
similarities reduces the number of false positives while maintaining positive detections.

Arsic et al. (Arsic et al., 2009) employs a multi-layer homography framework, relying on the fusion of previously segmented foreground regions visible from multiple views, to detect the presence of people. The foreground segmentations are produced by finding the median of pixel values and composing a reference image for simple background subtraction. A graph cut optimisation algorithm is optionally carried out to fill in small holes in foreground silhouettes. A simple Kalman filter with a linear motion model is used to associate detections across frames.

Berclaz et al. (Berclaz et al., 2009) cast tracking as a discrete linear programming problem. First, an object detector is employed based on a probabilistic occupancy grid, using a set of probabilities of the presence of objects, at a discrete set of locations, at each time step. The generative model used represents people as cylinders that project down to rectangles in individual camera views. Tracking is achieved by creating flow variables that are linked to provide complete trajectories and the complexity is solved by using flow linear programming.

Breitenstein et al. (Breitenstein et al., 2009) presents an algorithm for multi-person tracking-by-detection using a particle filtering framework. The approach requires the two dimensional input from a single non-calibrated camera. A HOG object detector produces the input for the observation model of a particle filter, which includes not only the objects detected, but their confidence density of that detection (represented as a colour heat map). Each object has its own particle filter initialised which includes its position
and velocity. The particles are propagated with a constant velocity motion model and their weight is updated using the observation model. This observation model uses the following: the possible detection of an object, the confidence density of this detection and finally by evaluating bounding boxes which are created by a boosted ensemble of weak classifiers employing colour histograms. A final decision on detection is only used to guide a particle filter if it is very likely to belong to its respective track and only one detection object may be assigned to one track.

Ge et al. (Ge et al., 2009) regard people in a crowd scene as a realisation of a Marked Point Process that consists of a random set of people in a bounded region. Each person is associated with a random mark that specifies their location and size within the frame. A binary foreground mask is obtained by an adaptive background subtraction method and is subjected to further morphological processing. This then becomes the input to the detector. A reverse jump Markov Chain Monte Carlo sampling method is used to detect individuals and their location. This is extended to multiple views by taking the foreground masks from views 1 and 2 and projecting these to a centroid plane which is half the average person's height above the ground plane. A fused proposal map is created by warping the individual foreground masks to the centroid and averaging them. The hypothesised individuals are back projected to the individual views.

Yang et al. (Yang et al., 2009) use a probabilistic appearance model to track multiple people through complex situations. Dynamic appearance models are employed with single Gaussians for foreground descriptions, and a Gaussian background model. Tracking is achieved by maximising the ap-
pearance model likelihood near the neighbourhood of a Kalman filter predicted centre. A tracking centre shift technique makes the appearance models robust to changing scales within the view and a drifting tracker retrieval system allows the system to recover from mistakes. The identification of the reappearance of individuals is attempted with a weighted partial histogram matching method.

Conte et al. (Conte et al., 2010) utilise an adaptive background image difference algorithm to detect moving objects. In order to make the system robust in realistic environments this has been extended to included processes that handle illumination, camouflage detection, noise filtering, shadow filtering and reflection removal. The tracking algorithm, based on an assignment framework, employs an appearance model to cope with the failure of short-term tracking and deals with occlusions and where one object appears to split into two.

Badie et al. (Badie et al., 2012) address single view tracking and the problem of temporary loss of tracks of people. The main contribution is the application of Mean Riemannian Covariance Grid (MRCG) descriptor for short-term error recovery and for long-term re-acquisition by linking several tracklets belonging to the same person in real time.

Nie et al. (Nie et al., 2012) apply a part-based human detector to locate people, followed by spatiotemporal tracking. Tracklets are generated on a frame-to-frame basis using part-based visual similarity. Then, tracklets are associated based on both spatial and temporal knowledge using constrained lineage programming, to produce a complete trajectory for a single person.

Xu et al. (Xu et al., 2012) employ different approaches to single and
multiple-view tracking. Single-view tracking is formulated by HOG based
detector combined with an optimized observation model to address ID switching
and/or tracking drift. For multiple views, a multi-view Bayesian Network
(MBN) is employed to reduce phantom detections in a tracking-by-detection
framework.

Hofmann et al. (Hofmann et al., 2013) formulate a unified hierarchical
multi-object tracking architecture. Tracking of multiple objects simultaneously is cast as a MAP problem which is solved via a three-stage framework.
Explicit occlusion reasoning is also considered.

Heili et al. (Heili et al., 2013) also examine multiple object tracking. Statistical labelling is solved using a Conditional Random Field (CRF) model which involves detection pairs and their corresponding hidden labels. Emphasis is paid to parameter estimation via tracklets and local context.

Foggia et al. (Foggia et al., 2013) simultaneously track individuals and
groups in single views through contextual graph-based data association. The history of moving objects is modelled by a Finite State Automaton. The spatiotemporal aspects of the objects are encoded into an efficient and robust graph-based approach.

5. Evaluation Methodology

The authors of the representative algorithms submitted their results (the identity and location of bounding boxes) in XML format using the PETS2009 published XML Schema available at http://www.pets2009.net. The detection and tracking evaluation was based on the framework by Kasturi et al. (Kasturi et al., 2009), a well established protocol for performance evalu-
Table 4: Participating Authors and Summary of Their Methods for PETS09 Detection and Tracking Challenge

<table>
<thead>
<tr>
<th>Author</th>
<th>Detection/Tracking Method</th>
<th>Single, Multiple Camera</th>
<th>Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bolme et al. (2009) (ASEF)</td>
<td>Adaptive background image difference, Appearance model</td>
<td>Single</td>
<td>Detection</td>
</tr>
<tr>
<td>Bolme et al. (2009) (Cascade)</td>
<td>Shape model, motion model, optical flow, probability map</td>
<td>Single</td>
<td>Detection</td>
</tr>
<tr>
<td>Bolme et al. (2009) (Parts)</td>
<td>Edgelet features, Cluster Boosted Tree classifier, data association based tracking</td>
<td>Single</td>
<td>Detection</td>
</tr>
<tr>
<td>Zweng et al. (2013)</td>
<td>Person detection; Relational feature model based on HOG + histogram similarity</td>
<td>Single</td>
<td>Detection</td>
</tr>
<tr>
<td>Alahi et al. (2009) (Olasso), Alahi (Ogreedy)</td>
<td>Corner points estimation, motion vectors</td>
<td>Single</td>
<td>Detection, Tracking</td>
</tr>
<tr>
<td>Arsic et al. (2009)</td>
<td>Multi-layer foreground model-based Tracking using Graph Cuts and Kalman filtering</td>
<td>Multiple</td>
<td>Detection, Tracking</td>
</tr>
<tr>
<td>Herclaz et al. (2009)</td>
<td>Generative model-based global optimization; Linear Programming for trajectory inference</td>
<td>Single, Multiple</td>
<td>Detection, Tracking</td>
</tr>
<tr>
<td>Breitenstein et al. (2009)</td>
<td>Particle filter; Continuous Detection Output to Robustify Tracking</td>
<td>Single</td>
<td>Detection, Tracking</td>
</tr>
<tr>
<td>Ge et al. (2009)</td>
<td>Gaussian mixture model (forwards and backwards), Reverse Jump Markov Chain Monte Carlo</td>
<td>Single, Multiple</td>
<td>Detection, Tracking</td>
</tr>
<tr>
<td>Yang et al. (2009)</td>
<td>Appearance-based model-based tracking; Kalman filtering</td>
<td>Single</td>
<td>Detection, Tracking</td>
</tr>
<tr>
<td>Conte et al. (2010)</td>
<td>Basic segmentation; Tracking based on Assignment Problem; Appearance modelling to recover lost IDs</td>
<td>Single, Multiple</td>
<td>Detection, Tracking</td>
</tr>
<tr>
<td>Badie et al. (2012)</td>
<td>Person detection; Short-term tracking (multi feature); Long-term tracking (tracklet linkage)</td>
<td>Single</td>
<td>Detection, Tracking</td>
</tr>
<tr>
<td>Nie et al. (2012)</td>
<td>Part-based person detection; Tracklet generation and association</td>
<td>Single</td>
<td>Detection, Tracking</td>
</tr>
<tr>
<td>Xu et al. (2012)</td>
<td>Tracking by detection framework</td>
<td>Single, Multiple</td>
<td>Detection, Tracking</td>
</tr>
<tr>
<td>Hofmann et al. (2013)</td>
<td>3-stage multiple object architecture; MAP formulation; Explicit occlusion handling</td>
<td>Single</td>
<td>Detection, Tracking</td>
</tr>
<tr>
<td>Heili et al. (2013)</td>
<td>Multiple object tracking; Statistical labelling solved using CRF; Contextual adaptation step</td>
<td>Single</td>
<td>Detection, Tracking</td>
</tr>
<tr>
<td>Foggia et al. (2013)</td>
<td>Simultaneous tracking of individuals and groups; Contextual graph-based data association</td>
<td>Single</td>
<td>Detection, Tracking</td>
</tr>
</tbody>
</table>
ation of object detection and tracking in video sequences. These metrics were formally used by the Video Analysis and Content Extraction (VACE) programme and the CLassification of Events, Activities, and Relationships (CLEAR) consortium. For both detection and tracking metrics, for the following descriptions the accuracy metrics provide a measure of the correctness of the detections or tracks. The precision metrics provide the measure of, in the instance where there has been a correct detection or track, how close to the ground truth that detection or track may be. The metrics used were Multiple Object Detection Accuracy (MODA), Multiple Object Detection Precision (MODP), Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), Sequence Frame Detection Accuracy (SFDA), and Average Tracking Accuracy (ATA) are used by the CLEAR and VACE consortiums and are defined and described further in [4]. In addition to the evaluation of detection and tracking, a simple comparison of the people count per region, against a ground truth count per region for the sampled frames, produced the average percentage error in counting per region, for each sequence.

Notation.

- $G^t_i$ denotes $i^{th}$ ground-truth object in frame $t$; $G_i$ denotes the $i^{th}$ ground-truth object at the sequence level; $N_{frames}$ is the number of frames in the sequence
- $D^t_i$ denotes the $i^{th}$ detected object in frame $t$; $D_i$ denotes the $i^{th}$ detected object at the sequence level
- $N^t_G$ and $N^t_D$ denote the number of ground-truth objects and the num-
ber of detected objects in frame $t$, respectively; $N_G$ and $N_D$ denote the number of unique ground-truth objects and the number of unique detected objects in the given sequence, respectively.

- $N_{frames}^i$ refers to the number of frames where either ground-truth object $(G_i)$ or the detected object $(D_i)$ existed in the sequence.

- $N_{mapped}$ refers to sequence level detected object and ground truth pairs, $N_{mapped}^t$ refers to frame $t$ mapped ground truth and detected object pairs.

- $m_t$ represents the missed detection count, $(fp_t)$ is the false positive count, $c_m$ and $c_f$ represent respectively the cost functions for missed detects and false positives, and $c_s = \log_{10}ID - SWITCHES_t$.

### 5.1. Sequence Frame Detection Accuracy (SFDA)

SFDA uses the number of objects detected, the number of missed detections, the number of falsely identified objects, and the calculation of the spatial alignment between the algorithm’s output for detected objects and that of the ground truthed objects. It is derived from a Frame Detection Accuracy (FDA) measure. The FDA is calculated using a ratio of the spatial intersection and union of an output object and mapped ground truthed objects.

\[
OverlapRatio = \frac{\sum_{i=1}^{N_{mapped}^t} \frac{|G_i^t \cap D_i^t|}{|G_i^t \cup D_i^t|}}{N_G^t + N_D^t}
\]

\[
FDA(t) = \frac{OverlapRatio}{\left[\frac{N_G^t + N_D^t}{2}\right]}
\]
SFDA = \frac{\sum_{t=1}^{N_{frames}} FDA(t)}{\sum_{t=1}^{N_{frames}} \exists (N^t_G \lor N^t_D)} \quad (3)

For this study although the annotation of the ground truth was challenging, as described in Section 1, an overlap threshold of 100 percent for the intersection over union scores, was used.

5.2. Average Tracking Accuracy (ATA)

ATA is obtained from the Sequence Track Detection Accuracy (STDA). The STDA is a measure of the tracking performance over all of the objects in the sequence and from this ATA is defined as the sequence track detection accuracy per object. The mapping between ground truth objects and detected objects is performed so as to maximise the measure score. This metric is implemented with a hash function due to the fact that the track correspondence matrix to be mapped is reasonably sparse.

\[
STDA = \sum_{i=1}^{N_{mapped}} \frac{\sum_{t=1}^{N_{frames}} \left[ \frac{|G^i_t \cap D^i_t|}{|G^i_t \cup D^i_t|} \right]}{N_{(G^i \cup D^i) \neq \emptyset}} \tag{4}
\]

\[
ATA = \frac{STDA}{\left[ \frac{N_G + N_D}{2} \right]} \tag{5}
\]

For both detection and tracking metrics in the following descriptions the accuracy metrics provide a measure of the correctness of the detections or tracks. The precision metrics provide the measure of, in the instance where there has been a correct detection or track, how close to the ground truth that detection or track may be.
5.3. Multiple Object Detection Accuracy (MODA)

MODA is an accuracy measure that uses the number of missed detections and the number of falsely identified objects. Cost functions to allow weighting to either of these errors are included, however for the sake of both PETS 2009 evaluations they were equally set to 1.

\[
MODA = 1 - \frac{c_m(m_t) + c_f(f_p_t)}{N^t_G}
\]  

(6)

5.4. Multiple Object Detection Precision (MODP)

MODP gives the precision of the detection in a given frame. Again, with this metric, an overlap ratio is calculated as previously defined in (1), and, in addition to a count of the number of mapped objects, the MODP is defined as:

\[
MODP(t) = \frac{\text{OverLapRatio}}{N^t_{\text{mapped}}}
\]  

(7)

5.5. Multiple Object Tracking Accuracy (MOTA)

MOTA uses the number of missed detections, the falsely identified objects, and the switches in an algorithm’s output track for a given ground truth track. These switches are calculated from the number of identity mismatches in a frame, from the mapped objects in its preceding frame.

\[
MOTA = 1 - \frac{\sum_{t=1}^{N_{\text{frames}}} (c_m(m_t) + c_f(f_p_t) + c_s)}{\sum_{t=1}^{N_{\text{frames}}} N^t_G}
\]  

(8)
5.6. Multiple Object Tracking Precision (MOTP)

MOTP is calculated from the spatio-temporal overlap between the ground truthed tracks and the algorithm’s output tracks.

\[
MOTP = \frac{\sum_{i=1}^{N_{\text{mapped}}} \sum_{t=1}^{N_{\text{frames}}} \frac{|G_t^i \cap D_t^i|}{|G_t^i \cup D_t^i|}}{\sum_{t=1}^{N_{\text{frames}}} N_{\text{mapped}}} \tag{9}
\]

6. Evaluation Results

This section presents the objective evaluation of the submitted results by contributing authors of the PETS2013, PETS2012, PETS2010, Winter-PETS2009, and PETS2009 workshops on the challenges defined on the PETS2009 crowd dataset. As stated the central theme of these PETS workshops was multi-sensor event recognition in crowded public areas and part of this challenge was to evaluate the authors’ approaches to people counting, object detection and tracking and to report their performance, based on annotated datasets made available on the workshop website. An analysis of the overall performance of the submitted results, using the defined metrics in Section 5, is described. The submitted results were diverse in terms of the sequences and views used and people counting, people detection and tracking challenges were considered. Nevertheless, the evaluations presented in this Section lead to helpful insights about the effectiveness of different methodologies.

6.1. People Counting

Figure 3 provides the evaluation of the counting people per region task. Note that the y axis on this graph represents the average error in number of people per frame, where the lower the value, the better the performance per
frame. Table 3 gives the corresponding publication reference, for each label, for Figure 3. A wide variety of methods have been proposed and tested in this category and from Figure 3 it can been seen that the majority of the methods and their variants have consistent and comparable performance. The algorithms proposed by Albiol et al. (Albiol et al., 2009) and Conte et al. (Conte et al., 2010) performed robustly throughout each time sequence. Several methods such as Alahi et al. (Alahi et al., 2009), Chan et al. (Chan et al., 2009), Subburaman et al. (Subburaman et al., 2012) and Choudri et al. (Choudri et al., 2009) also performed well on the more challenging sequence 14-17. Further details of the variant of each method can be found in their companion publication.
6.2. Detection and Tracking

The vast majority of participants in the five PETS workshops which used the PETS2009 dataset reported results from the detection and tracking Dataset S2.L1 at time sequence 12-34, for the first camera view.

Figure 4 shows how the individual algorithms performed for the detection task and Figure 5 for the tracking task, according to various VACE and CLEAR metrics on this single representative camera view. Note that in the case of these metrics, higher values indicate better performance. It is clear that for this sequence for the detection task (Figure 4), using SODA, SFDA, MODA and MOPD as a measure, the systems described by Breitenstein et al. (Breitenstein et al., 2009), Yang et al. (Yang et al., 2009), Xu et al.
(Xu et al., 2012) MultiView and Hoffman et al.s (Hofmann et al., 2013) MC performed strongly at multiple object detection and tracking. Berclaz et al. (Berclaz et al., 2009) and Ge and Collins’s (Ge et al., 2009) detection accuracy(MODA) results suggested a robust performance for this particular area. The system created by Heili et al.s (Heili et al., 2013) highlights robust detection precision (MODP). Equally systems created by Alahi et al. (Alahi et al., 2009), Conte et al. (Conte et al., 2010) and Berclaz et al. (Berclaz et al., 2009) are noteable for this measure. For the tracking task (Figure 5) for this sequence, using MOTP and MOTA as a measure the systems described by Hofman et al. (Hofmann et al., 2013), Yang et al. (Yang et al., 2009) and Xu et al.s (Xu et al., 2012) MultiView performed strongly. In addition the system described by Yang et al. (Yang et al., 2009) showed a promising performance in ATA which has been shown throughout the PETS 2009-2013 series of workshops to be the most challenging metric. Again Heili et al.s (Heili et al., 2013) systems precision, this time for the tracking task, using MOTP as a measure is highlighted. Finally the system described by Foggia et al. (Foggia et al., 2013) also suggests a robustness to tracking accuracy using the MOTA metric.

6.3. Metric Evaluation

Figure 6 shows the median of each metric value, for all the computed views, excluding View 2 which was not provided in the dataset, from each author. The performance measures highlight the algorithms provided by Xu et al.s MultiView system (Xu et al., 2012) for all defined metrics and Ge and Sikora and Berclaz et al. (Berclaz et al., 2009) for multiple object detection accuracy and tracking. In addition the detection and tracking
Figure 5: Performance of authors’ systems per tracking metric, camera view 1, dataset: S2.L1, time sequence: 12.34.

Precision metric measurements (MODP and MODA) for the system of Foggia et al. (Foggia et al., 2013) are notable. From this figure it can be seen that although there are variations per metric per author, the results for MODP and MOTP indicated a general consensus of accuracy.

To estimate the consistency of the metrics themselves another evaluation is illustrated here. Figure 7 (top) showed, for each view, the median value of each metric for all authors. It highlights the relationships between two pairs of metrics. SODA and SFDA present extremely similar median values across all camera views, as do MODP and MOTP. These relationships are emphasised further in the evaluation of the results in Figure 4, where the relationship between each metric measurement pairing (SODA+SFDA) can
be clearly seen. In addition, the tracking accuracy metric values appear to be strongly related to the detection accuracy metric values. The figure suggests that camera view one was the simplest task for the participating authors and camera view six presented a challenge.

As the final evaluation, a view for each metric which corresponds to the median value of the metric for all authors, was used. The results are shown in Figure 7 (bottom). From this figure a fair overall performance comparison of each algorithm and their variant forms can be inferred. Due to its robustness to outliers, this visualisation gives a clear indication of how different algorithms perform relative to each other.
Figure 7: Metric evaluation. Top: Median metric result of all authors. Bottom: Median metric per view per author.
7. Conclusions and Future Work

This paper has presented the performance evaluation of state of the art crowd image analysis visual surveillance methods to enable comparisons to be drawn. It is essential that authors are able to objectively evaluate their detection and tracking algorithms with standardised metrics. The ability to compare results, with others, whether anonymous or not, provides a realistic and encouraging research technique towards advanced, robust, real time visual systems. In addition the latest results highlight the need for careful consideration regarding the ground truthing of data sets and the subsequent evaluation. The use of these metrics and this study provides a mechanism to highlight the strengths of the individual systems, such as accuracy, precision and robustness. It may be used for future decisions for systems placement. For example, those that require a high degree of precision may benefit from techniques described by authors whose systems performed well using precision metrics. Conversely where the accuracy of the tracking is more important than the precision, those requiring such may benefit from techniques shown to perform well in this paper. During this study the challenge of performance evaluation was ever present. An important part of the future of research within the computational vision community and that of those applying their research to real world problems is that of being able to adequately evaluate their approaches with realistic and achievable methods. Datasets that are popular and/or readily available and appropriate to detection and tracking tasks, for automated visual surveillance, have been presented. The choices of which datasets to use when objectively analysing the performance of any solution is restricted to those which have ground truth readily avail-
able and published evaluation results of systems participating in their use.

Open source/web-based performance evaluation tools, and a standardisation
of benchmark datasets and metrics for specific challenges is a sensible way
forward and some indicators of this trend have been presented. The published
evaluation of any participating systems’ performances against a set of pub-
licly available benchmark data, using widely accepted metrics for provided
challenges proved a valuable tool in assessing the robust nature of motion
segmentation and tracking algorithms. Future workshops such as those pro-
vided by the PETS series are an invaluable resource to researchers requiring
an object evaluation of their approaches to automated visual surveillance.

Acknowledgements

This work was supported in part by respectively the EU SUBITO and
ARENA project grant agreements 218004 and 261658. Any opinions ex-
pressed in this paper do not necessarily reflect the views of the European
Community. The Community is not liable for any use that may be made of
the information contained herein.

References

A. Alahi, L. Jacques, Y. Boursier and P. Vandergeynst., 2009. Sparsity-
Driven People Localization Algorithm: Evaluation in Crowded Scenes En-
of Tracking and Surveillance (PETS-Winter), pp. 1-8, DOI:10.1109/PETS-
WINTER.2009.5399487


M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. van Gool., 2009. Markovian Tracking-by-Detection from a Single, Uncalibrated Cam-


shop on Performance Evaluation of Tracking and Surveillance (PETS), pp. 22-28., DOI: 10.1109/PETS.2013.6523791


S. Singh, S. Velastin and R. Hossein., 2010. Muhavi: A Multicamera Human Action Video Dataset for the Evaluation of Action Recognition Meth-


http://trecvid.nist.gov/ [last accessed: 01/03/2013]


