2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance

PETS2010: Dataset and Challenge

J. Ferryman and A. Ellis
Computational Vision Group
School of Systems Engineering
University of Reading
Whiteknights, Reading, RG6 6AY, UK
{j.m.ferryman|a.shahrokni}reading.ac.uk

Abstract

This paper describes the crowd image analysis challenge that forms part of the PETS 2010 workshop. The aim of this challenge is to use new or existing systems for i) crowd count and density estimation, ii) tracking of individual(s) within a crowd, and iii) detection of separate flows and specific crowd events, in a real-world environment. The dataset scenarios were filmed from multiple cameras and involve multiple actors.

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 automatic 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 has not been met with complementary systematic performance evaluation of developed techniques using a common benchmark. It is especially difficult to make comparisons between algorithms if they have been tested on different datasets under widely varying conditions.

PETS 2010 continues the theme of the highly successful PETS workshop series [2]. For PETS 2010, the theme is multi-sensor event recognition in crowded public areas. As part of this workshop a challenge was set to evaluate an approach to one or more of people counting and density estimation, tracking, and flow estimation and event recognition, and to report results based on annotated datasets made available on the workshop website [1]. This paper details the datasets and the challenge that the contributing authors had to present solutions for.

2. The PETS2010 Challenge — Crowd Image Analysis

The aim of the PETS 2010 challenge is to detect one or more of three types of crowd surveillance characteristics/events within a public space outdoor scene. Automatic detection of, for example, a high density of people in a restricted area will allow public space personnel to respond quickly to potentially hazardous situations, improving the security and safety of public areas.

3. The Datasets

Three datasets were recorded for the workshop at Whiteknights Campus, University of Reading, UK. The datasets comprise multi-sensor sequences containing crowd scenarios with increasing scene complexity. Dataset S1 concerns person count and density estimation. Dataset S2 addresses people tracking. Dataset S3 involves flow analysis and event recognition.

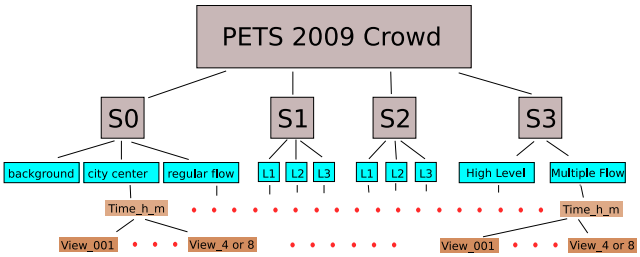

Figure 1. Hierarchy of recorded sequences and related views.

3.1. S0: Training Data

This dataset contains three sets of training sequences from different views and is provided to help researchers obtain the following models:

1. BACKGROUND. Background model for all cameras.

978-0-7695-4264-5/10 $26.00 © 2010 IEEE
DOI 10.1109/AVSS.2010.90

Figure 2. Left-to-right, top-to-bottom: the 8 camera views used in the data collection.

The frames may contain people or other moving objects. The frames in the set are not necessarily synchronised. For Views 1-4 different sequences corresponding to the following timestamps: 13-05, 13-06, 13-07, 13-19, 13-32, 13-38, 13-49 are provided. For Views 5-8 144 nonsynchronised frames are provided.

2. CITY CENTRE. Includes random, walking crowd flow. Sequence 9 with timestamp 12-34 using Views 1-8 and Sequence 10 with timestamp 14-55 using Views 1-4.

3. REGULAR FLOW. Includes regular walking pace crowd flow. Sequences 11-15 with timestamps 13-57, 13-59, 14-03, 14-06, 14-29 for Views 1-4.

3.2. S1: People Count and Density Estimation

Crowd density is based on a maximum occupancy (100%) of 40 individuals in $10m^2$ on the ground. One individual is assumed to occupy $0.25m^2$ on the ground.

Figure 3. Regions R0, R1 and R2.

Dataset S1 (L1 - Walking - Sequences 1 & 2) Medium density crowd, overcast, subjective difficulty level 2.

The requirement of this task is to report the count of the number of individuals within Region R0 for each frame of the sequence, for View 1 only. Additionally, the crowd density (%) for Regions R1 and R2 may also be reported (based on ground occupancy). Figure 3 depicts the regions R0, R1 and R2 overlaid on View 1 and Figure 4 shows representative frames. This sequences exhibit regular crowd movement in a relatively dense queue in two directions.

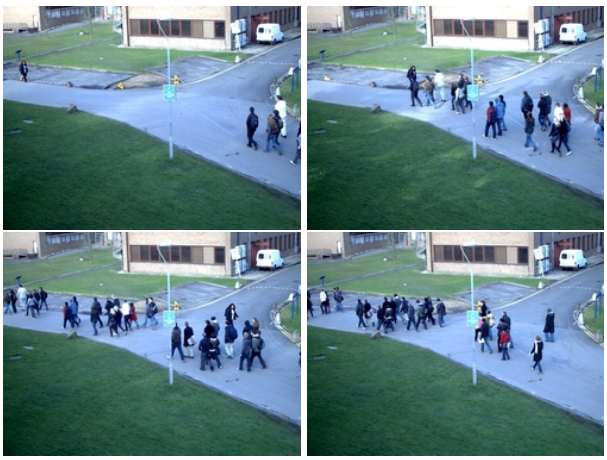

Figure 4. Dataset S1 L1, frames 0, 50, 100 and 150 (left-to-right, top-to-bottom).

Dataset S1 (L2 - Walking - Sequences 1 & 2) high density crowd, overcast, subjective difficulty level 2.

The task for Sequence 1 is to report the crowd density in both Regions R1 and R2 for each frame of the sequence. The task for Sequence 2 is to report, for each frame of the sequence, the total number of individuals who have passed the entry point in the scene (brown line) AND the total number of individuals who have passed the exit points (purple and red lines respectively) for View 1. Table 1 defines the image line coordinates for the entry and exit points and Figure 6 depicts the lines. Figure 5 shows representative frames from the sequence. The first sequence exhibits regular crowd movement in a dense queue and sequence 2 contains movement in diverging directions.

Table 1. Coordinates of Image Lines for Person Count

Region	Top-Left	Bottom-Right
R0	(10,10)	(750,550)
R1	(290,160)	(710,430)
R2	(30,130)	(230,290)

Dataset S1 (L3 - Running - Sequences 1 & 2) Medium density crowd, nright sunshine and shadows, subjective difficulty level 3.

Figure 5. Dataset S1 L2, frames 25, 75, 125 and 175 (left-to-right, top-to-bottom).

Figure 6. Entry (brown Line) and Exit (rurple and red) lines for Dataset S1 L2 Sequence 2.

The requirement of this task is to report the crowd density (%) in Region R1 for each frame of the sequence. Figure 7 shows representative frames from the sequence. The first sequence shows a walking crowd whose individuals start to run in the middle of the sequence. The second sequence show converging walking people whom form a dense stationary group of people at the end of the sequence.

3.3. S2: People Tracking

Dataset S2 (L1 - Walking - Sequence 1) Sparse crowd, subjective difficulty level 1.

The task is 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 in View 2 only. This is based on the assumption that two or more views are used, and that the result is re-projected into View 2 based on the calibration details provided. If the tracking algorithm used is monocular based, one can alternatively submit results based on the particular view used. The origin (0,0) of the image is the top-left corner. Figure 8 shows representative frames from the sequence. This sequence exhibits a randomly walking sparse crowd.

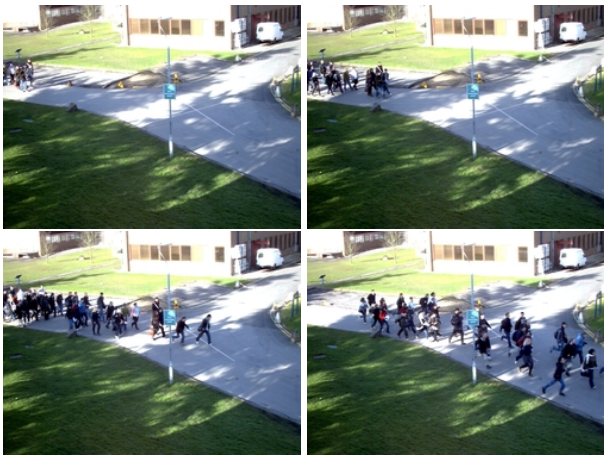

Figure 7. Dataset S1 L3, frames 0, 26, 65 and 90 (left-to-right, top-to-bottom).

Figure 8. Dataset S2 L1, frames 0, 50, 100 and 150 (left-to-right, top-to-bottom).

Dataset S2 (L2 - Walking - Sequence 1) Medium density crowd, subjective difficulty level 2.

The task is to track two labelled individuals and to report the bounding box (2D, and optionally, 3D) of both individuals for every frame of the sequence in view 2 only. This is based on the assumption that two or more views are used, and that that the result is re-projected into View 2 based on the calibration details provided. If the tracking algorithm

used is monocular based, one can alternatively submit results based on the particular view used. The origin (0,0) of the image is the top-left corner. Figure 9 shows representative frames from the sequence. This sequence exhibits a randomly walking dense crowd.

Figure 9. Dataset S2 L2, frames 50, 70, 300 and 400 (left-to-right, top-to-bottom).

Dataset S2 (L3 - Walking - Sequence 1) Dense crowd, subjective difficulty level 3.

The task is to track two labelled individuals and to report the bounding box (2D, and optionally, 3D) of both individuals for every frame of the sequence in view 2 only. This is based on the assumption that two or more views are used, and that that the result is re-projected into View 2 based on the calibration details provided. If the tracking algorithm used is monocular based, one can alternatively submit results based on the particular view used. The origin (0,0) of the image is the top-left corner. Figure 10 shows representative frames from the sequence. This sequence shows two individuals which are bystanders in an empty scene which later join a moving crowd walking in a the same direction.

Figure 10. Dataset S2 L3, frames 80, 185, 210 and 230 (left-to-right, top-to-bottom).

3.4. S3: Multiple Flow

Dataset S3 (Sequences 1-5) Dense crowd, running, subjective difficulty level 2.

This dataset contains five sequences respectively with timestamps 12-43, 14-13, 14-37, 14-46 and 14-52. The task is to detect and report the separate, individual flows in one or more of the sequences. For each frame of the sequence, each flow should be reported. Figure 11 shows representative frames from the sequence. The first sequence shows a sparse queue of people walking regularly along a linear path and slightly curving to avoid a virtual obstacle. The second sequence shows a more dense crowd walking in a queue while traversing around a human wall. Sequence 3 shows merging groups of people walking together. Sequence 4 shows two individuals walking and advancing their way against a dense queue of people walking in the opposite direction. Sequence 5 is similar to sequence 4 except that the 3 individuals walking against the crowd are wearing bright jackets.

Figure 11. Dataset S3 Flow, frames 1, 26, 52 and 78 (left-tight, top-bottom).

3.5. S3: Event Recognition

Dataset S3 (Sequences 1-4) Dense crowd, subjective difficulty level 3.

This dataset contains four sequences respectively with timestamps 14-16, 14-27, 14-31 and 14-33. Each of the sequences may contain one or more of the following set of events. The onset of an event should be reported at the earliest time that the conditions below are met.

Walking corresponds to the onset of a crowd (significant number of individuals) moving at "typical" walking pace.

Running corresponds to the onset of a crowd (significant number of individuals) moving at "typical" running pace.

Evacuation corresponds to rapid dispersion of the crowd, moving in different directions. This is defined as the onset of rapidly moving, multiple diverging flows.

Local Dispersion corresppponds to localised movement of people within a crowd away from a given threat (e.g. unattended object). The crowd, as a whole, remains located in approxiamtely the same area. This is defined as the onset of multiple, localised, diverging flows.

Crowd Formation - Gathering/Merging corresponds to formation of a crowd of individuals, whereby the individuals approach from different directions. This is defined as the convergence of multiple flows. The crowd is formed at the time point where the flows converge at the same physical location.

Crowd Dispersal - Splitting corresponds to a cohesive crowd of individuals which split into two or more entities. The is defined as multiple diverging flows. The crowd splits at the time point when the crowd starts to separate into distinct flows.

For each sequence, and for each frame, a probabilistic measure of each event (multi-class, all events) should be reported. In other words, all events above should be reported for each frame, even in the absence of recognition of a given event. When one or more events are detected, the probability of each event should be reported. Figure 11 shows representative frames from the sequence. The first sequence shows a walking crowd whose individuals start to run in the middle of the sequence and shows both directions. Sequence 2 depicts a stationary crowd with occasional local dispersions. Sequence 3 shows a dense walking crowd which splits into three directions. Finally, Sequence 4 contains a merging crowd of people which forms a stationary group of people. This crowd eventually evacuates the scene, dispersing in a chaotic and random manner.

3.6. Calibration and XML Results Schema

Geometric patterns on the ground were used to calibrate the cameras using the Tsai model [3]. All spatial measurements are in metres. All datasets were filmed using the cameras detailed in Table 2.

The camera installation points are shown in Figure 13. The GPS coordinates of the centre of the recording are: $51^0 26' 18.5N\ 000^0 56' 40.00W$. The resolution of all sequences are PAL standard and compressed as JPEG image sequences (approx. 90% quality). All sequences (except one) contain Views 1-4. A few sequences also contain

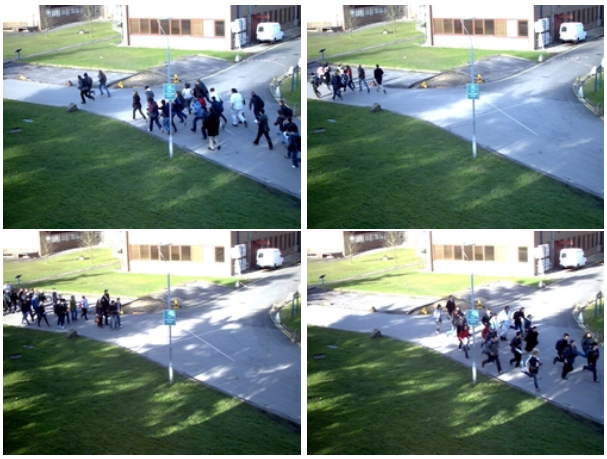

Figure 12. Dataset S3 Events, frames 50, 100, 150 and 200 (left-to-right, top-to-bottom).

Table 2. Cameras' Specification. PS = Progressive Scan. DI = Deinterlaced.

View	Model	Resolution	Frame Rate	Scan
1	Axis 223M	768x576	∼7	PS
2	Axis 223M	768x576	∼7	PS
3	Axis 233D	768x576	∼7	PS
4	Axis 233D	768x576	∼7	PS
5	Axis 223M	720x576	∼7	DI
6	Axis 223M	720x576	∼7	DI
7	Canon MV1	720x576	∼7	PS
8	Canon MV1	720x576	∼7	PS

Views 5-8. While every effort has been made to ensure the frames from different views are synchronised, there may exist slight delays and frame drops in some cases. In particular, View 4 suffers from frame rate instability and should only be used as a supplementary source of information.

All detection results are reported based on a specified XML Schema. Further details can be found on the PETS 2010 website [1].

4. Discussion

In this paper we have described the PETS 2010 datasets and challenge. This workshop is addressing the problem of crowd image analysis within a public space. The PETS 2010 challenge provide researchers with the opportunity to evaluate new or existing detection algorithms on datasets captured in a real-world environment.

Acknowledgements

This work is supported by the EU (SUBITO Grant Agreement No. 218004 and Co-Friend (214975)) [1]Thanks

[1]This paper does not necessarily represent the opinion of the EU; the EU is not responsible for any use which may be made of its contents.

Figure 13. Plan view showing the location and direction of the 8 cameras on the Whiteknights Campus.

to members of the Computational Vision Group, School of Systems Engineering, University of Reading, and students from the University of Reading, for their help in compiling the challenge datasets. Thanks to Google for the image shown in Figure 13.

Legal Note

The PETS 2009 datasets described herein have been made publicly available for the purposes of academic research. The video sequences are copyright University of Reading and permission for the publication of these sequences is hereby granted provided that the source is acknowledged.

References

[1] Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. *http://pets2009.net*. 1, 5

[2] PETS: Performance Evaluation of Tracking and Surveillance. *http://www.cvg.rdg.ac.uk/slides/pets.html*. 1

[3] R. Tsai. An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 323–344, 1986. 5