Pattern Recognition Letters 44 (2014) 3–15

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Performance evaluation of crowd image analysis using the PETS2009 dataset ☆

James Ferryman *, Anna-Louise Ellis

Computational Vision Group, School of Systems Engineering, University of Reading, Whiteknights, Reading RG6 6AY, UK

ARTICLE INFO

Article history:
Available online 27 January 2014

Keywords:
Surveillance
Detection
Tracking
Performance evaluation
Dataset

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.

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, however, has not been met with complementary systematic performance evaluation of developed techniques using a common benchmark. It has been especially difficult to make comparisons between 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 addition, a series of five dedicated consecutive workshops (PETS2009, Winter-PETS2009, PETS2010, PETS2012, and PETS2013) were held with the prerequisite 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.

☆ This paper has been recommended for acceptance by Rita Cucchiara.

* Corresponding author. Tel.: +44 118 3786697; fax: +44 118 9751994.
E-mail address: james@computer.org (J. Ferryman).

http://dx.doi.org/10.1016/j.patrec.2014.01.005
0167-8655/

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 include, 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 surveillance 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 different strengths to an automated crowd image analysis research project. Real world footage highlights the numerous challenges faced in developing robust 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 [23] 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.

Daimler's Pedestrian Detection Benchmark Data Set [14] presents an alternative automated visual surveillance task. It provides a video sequence captured from a vehicle during a 27 min drive through city traffic. The dataset consists of a specified training set containing pedestrian and non-pedestrian samples and a test set containing a sequence with more than 21,790 images with 56,492 pedestrians labelled. The VIRAT video dataset [25] 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 [32] is sponsored by the National Institute of Standards and Technology (NIST) and other US 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) [30] 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) [19] 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 (upto 40) available for the captured actions.

The image library for intelligent detection systems, i-LIDS [20] is the UK governments 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

Table 1
Summary of surveyed surveillance datasets.

Project & reference	Real world	Simulated	#Cameras	Footage quality	Ground truth
Daimler [14]	Street scene	No	1	640 × 480	No
VIRAT [25]	Natural scenes/human actions	No	Multiple	Various	No
ETISEO [23]	Inside metro/airport apron	No	4,2,1	Mixed sensors	Permission required
i-LIDS [20]	Various	No	4,1	PAL or less	No
PETS'01	Street scene	No	4,2	768 × 576; 25 fps	No
PETS'06 [27]	Train station	No	4	768 × 576; 25 fps	No
PETS'07 [28]	Airport terminal	No	4	768 × 576; 25 fps	No
MUHAVI [30]	Human actions	Street lights	4	720 × 576, 704 × 576; 25 fps	Yes
ViHASi [19]	No	Avatar actions	40	640 × 480; 25 fps	Yes
CAVIAR [8]	Inside shops/offices	No	2	374 × 288; 25 fps	Yes
TRECVid [32]	Various	No but some edited footage	Varies	Multiple (broadcast quality)	Some