Pattern Recognition Letters 44 (2014) 98–112

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Scene invariant multi camera crowd counting

David Ryan *, Simon Denman, Clinton Fookes, Sridha Sridharan

Image and Video Laboratory, S1101, Queensland University of Technology, 2 George St, Brisbane 4000, Australia

ARTICLE INFO

Article history:
Available online 12 October 2013

Communicated by Simone Calderara

Keywords:
Crowd counting
Multi camera
Scene invariant

ABSTRACT

Automated crowd counting has become an active field of computer vision research in recent years. Existing approaches are scene-specific, as they are designed to operate in the single camera viewpoint that was used to train the system. Real world camera networks often span multiple viewpoints within a facility, including many regions of overlap.

This paper proposes a novel scene invariant crowd counting algorithm that is designed to operate across multiple cameras. The approach uses camera calibration to normalise features between viewpoints and to compensate for regions of overlap. This compensation is performed by constructing an 'overlap map' which provides a measure of how much an object at one location is visible within other viewpoints. An investigation into the suitability of various feature types and regression models for scene invariant crowd counting is also conducted. The features investigated include object size, shape, edges and keypoints. The regression models evaluated include neural networks, *K*-nearest neighbours, linear and Gaussian process regression.

Our experiments demonstrate that accurate crowd counting was achieved across seven benchmark datasets, with optimal performance observed when all features were used and when Gaussian process regression was used. The combination of scene invariance and multi camera crowd counting is evaluated by training the system on footage obtained from the QUT camera network and testing it on three cameras from the PETS 2009 database. Highly accurate crowd counting was observed with a mean relative error of less than 10%.

Our approach enables a pre-trained system to be deployed on a new environment without any additional training, bringing the field one step closer toward a 'plug and play' system.

1. Introduction

Automated crowd counting has become an active field of computer vision research in recent years. Crowd size is the most common indicator of security threats such as rioting, violent protest and mass panic, and it can also indicate congestion and other abnormal events within peaceful crowds. Crowd information can also be used to provide operational analytics for business intelligence.

Existing approaches to crowd counting are scene specific, as they are designed to operate in the same environment that was used to train the system. In a facility containing numerous cameras, this requires each viewpoint to be trained independently, which can be an arduous and time consuming task. It is not practical to supply hundreds of frames of ground truth for every viewpoint. In this paper, a novel algorithm is proposed which utilises camera calibration to achieve *scene invariance* by scaling features appropriately between viewpoints. This enables the system to be deployed on different training and testing sets, including those captured at different locations.

In practice, this means that a system can be trained on a bank of 'reference viewpoints', and then deployed on any number of *new* viewpoints without any additional ground truth annotations being required, greatly reducing the time and cost of configuring a crowd counting system. To facilitate the development of this technique, an investigation into various features and regression models for scene invariant crowd counting is conducted to determine the best combination in practice.

Another limitation of existing methods is that they are designed to count crowds within a *single camera viewpoint*, whereas real-world camera networks span multiple viewpoints within a facility, including some regions of overlap. Since some individuals will be detected across multiple cameras, it is necessary to compensate for this overlap to avoid over-estimation of the total number of people. This paper extends crowd counting across multiple cameras by utilising camera calibration parameters. An 'overlap map' is calculated which provides a measure of how much an object at one location is visible within other viewpoints, and this is used to modify crowd density on a pixelwise basis.

* Corresponding author. Tel.: +61 7 3138 9329.
E-mail addresses: david.ryan@qut.edu.au, d23.ryan@student.qut.edu.au (D. Ryan), s.denman@qut.edu.au (S. Denman), c.fookes@qut.edu.au (C. Fookes), s.sridharan@qut.edu.au (S. Sridharan).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.10.002

The proposed algorithm is tested on seven datasets which utilise camera calibration: PETS 2009, Views 1 and 2 (PETS, 2009); PETS 2006, Views 3 and 4 (PETS, 2006); and QUT, Cameras 3, 5 and 8 (Section 5.2). These datasets feature crowds of size 1 to 43 people in various lighting conditions and differing camera angles. The system is demonstrated to be scene invariant and capable of supporting multiple cameras, with accurate crowd counting results.

The details of the scene invariant crowd counting algorithm have been previously published in Ryan et al. (2012). This paper makes a number of additional contributions:

1. A comprehensive investigation into optimal feature sets and regression models for scene invariant crowd counting.
2. An extension to multi camera environments, allowing the total number of people across a scene to be estimated.
3. A combination of scene invariance and multi camera crowd counting algorithms, which is evaluated on a three-camera setup.

The remainder of this paper is structured as follows: Section 2 reviews the existing crowd counting literature; Section 3 describes the proposed scene invariant crowd counting algorithm; Section 4 extends this algorithm to operate across multiple cameras; Section 5 presents the evaluation protocol and benchmark datasets; Section 6 presents the experimental results of our algorithm; and Section 7 presents conclusions and directions for future work.

2. Background

Current approaches to crowd counting generally employ supervised machine learning techniques to map between the image feature space and the crowd size estimate. Regression is performed at either the holistic level of an image (Davies et al., 1995; Kong et al., 2006; Chan et al., 2009) or at a local scale (Kilambi et al., 2008; Ryan et al., 2009; Lempitsky and Zisserman, 2010).

Holistic image features include textural statistics (Marana et al., 1997), Minkowski fractal dimension (Marana et al., 1999) and translation invariant orthonormal Chebyshev moments (Rahmalan et al., 2006). Holistic textural features such as these are sensitive to external changes, and for outdoor environments the natural fluctuations in lighting between morning and afternoon have been shown to reduce system performance (Rahmalan et al., 2006). A number of algorithms use background modelling techniques (Stauffer and Grimson, 1999; Denman, 2009) in order to identify pedestrians in the foreground. Davies et al. (1995) modelled the relationship between foreground pixels and crowd size using linear regression, and subsequent approaches have attempted to deal with perspective and occlusion. Paragios and Ramesh (2001) introduced the use of a density estimator to account for perspective and Ma et al. (2004) computed a density map which weighted each pixel by the area it represented on the ground plane. The sum of *weighted* foreground pixels is used as a measure of crowding.

Kong et al. (2006) proposed the use of histogram based features to capture the various levels of occlusion present in a scene. Foreground 'blob' segments were aggregated into size-based histograms, and an edge orientation histogram was constructed based on the gradient directions. The edge orientation histogram is used to help distinguish between humans and other structures in the scene (Kong et al., 2006). Similar features have been used in other visual surveillance research, such as the histogram of oriented gradients employed by Dalal and Triggs (2005) for the explicit purpose of human detection.

A unique segmentation technique was used by Chan et al. (2008) to identify foreground motion in two directions, based on the mixture of dynamic textures. A large number of holistic image features were extracted including foreground area, perimeter pixel count, edge orientation histogram and textural features. In total, 29 features were extracted and Gaussian process regression was used to predict the number of pedestrians walking in each direction.

Local approaches to crowd counting utilise detectors or features which are specific to individuals or groups of people within an image. Lin et al. (2001) has proposed the use of head detection for crowd counting. The Haar wavelet transform was used in conjunction with the support vector machine to classify head-like contours as human.

Celik et al. (2006) proposed a person-counting algorithm which did not require training: it assumes proportionality between the number of pixels within a blob segment and the number of people represented by that segment, in order to obtain an estimate for each group. Kilambi et al. (2008) models a group of pedestrians as an elliptical cylinder, assuming a constant spacing between people within the group. Lempitsky and Zisserman (2010) proposed an object counting algorithm which sought to estimate a density function of the pixels in an image, so that integrating the density over any region would yield the number of objects in that region. This is a localised approach in which every pixel is represented by a feature vector containing foreground and gradient information, and a linear model is used to obtain the density at each pixel.

These approaches rely on scene-specific training data which requires a system to be trained and tested on the same viewpoint, using potentially hundreds (Kong et al., 2006) or thousands (Chan et al., 2009) of annotated training frames. Even though large-scale CCTV networks are becoming increasingly common, automated crowd counting is not widely deployed. One of the largest barriers to full deployment of this technology is the requirement to train each camera independently, which is both time-consuming and expensive.

3. Scene invariant crowd counting

This section describes a scene invariant crowd counting algorithm which can be trained and tested on different cameras. The system is trained on a bank of 'reference viewpoints' before being deployed on any number of unseen viewpoints, without any additional training requirements.

The approach is based on camera calibration, which is used to normalise features across viewpoints. This algorithm was originally published in Ryan et al. (2012, 2011) and the details of the approach are discussed in this section. Furthermore, this paper extends the approach by utilising additional image features, and Section 6.1 investigates various combinations of these features to determine the best approach in practice. The algorithm is extended to multi camera environments in Section 4.

This section is structured as follows: Section 3.1 describes how scene invariance is achieved through the use of a 'density map' to normalise features; Section 3.2 describes the feature extraction process; and Section 3.3 outlines the procedure used to train the system.

3.1. Scene invariance

Scene invariance is achieved by scaling the features extracted from each pixel to normalise for camera position and orientation. A density map, S, is constructed based on camera calibration parameters. The density map supplies a weight, $S(i,j)$, to each pixel, which is used to scale the features extracted from that pixel. This approach has been used previously to compensate for the effects of perspective within a single image (Section 2), however in this