Computer Vision and Image Understanding 130 (2015) 1–17

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

An evaluation of crowd counting methods, features and regression models ☆

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

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

ARTICLE INFO

Article history:
Received 8 October 2013
Accepted 31 July 2014
Available online 9 August 2014

Keywords:
Crowd counting
Holistic features
Local features
Histogram features
Regression

ABSTRACT

Existing crowd counting algorithms rely on holistic, local or histogram based features to capture crowd properties. Regression is then employed to estimate the crowd size. Insufficient testing across multiple datasets has made it difficult to compare and contrast different methodologies. This paper presents an evaluation across multiple datasets to compare holistic, local and histogram based methods, and to compare various image features and regression models. A *K*-fold cross validation protocol is followed to evaluate the performance across five public datasets: UCSD, PETS 2009, Fudan, Mall and Grand Central datasets. Image features are categorised into five types: size, shape, edges, keypoints and textures. The regression models evaluated are: Gaussian process regression (GPR), linear regression, *K* nearest neighbours (KNN) and neural networks (NN). The results demonstrate that local features outperform equivalent holistic and histogram based features; optimal performance is observed using all image features except for textures; and that GPR outperforms linear, KNN and NN regression.

1. Introduction

Crowd size estimation is an important task for both operational and security purposes. The distribution of people throughout a public space can be used to gather business intelligence, such as consumer shopping patterns, or to ensure that normal operating conditions are maintained. Overcrowding may be an indicator of congestion, delay or security-related abnormalities such as fighting and rioting.

As closed-circuit television (CCTV) becomes ubiquitous, it grows increasingly difficult for human operators to monitor all of the available data due to the sheer number of cameras installed. For example, there are estimated to be between 1.85 million [40] and 4.2 million [64] CCTV cameras installed in the United Kingdom alone. In most cases, security footage is used to investigate events after they occur, rather than to generate real-time alerts during an evolving situation.

In recent years, researchers have turned to computer vision based surveillance technologies to monitor crowds automatically from CCTV. Existing crowd counting algorithms are predominantly holistic in nature, employing machine learning techniques to perform regression between image features and crowd size [71,24,59,65,45,53,48,83,43,8]. In recent years a number of local systems have also been proposed, although many of these algorithms are detection based and rely on assumptions about camera placement or visibility of human features such as head, face or body parts [51,85,15,90]. Other local approaches divide an image into a number of subregions and perform counting locally [47,5,50,13,22,75]. Histogram based approaches have also been proposed in which local information is accumulated into histogram bins and represented on a holistic level [48,49].

Insufficient testing across multiple datasets has made it difficult to compare and contrast different methodologies. A comprehensive analysis across multiple datasets is required to compare local and holistic methods, and to compare various image features and regression models.

This paper uses a cross validation protocol to evaluate the performance of various methods, features and regression models across five public datasets. Image features are categorised into five types: size, shape, edges, keypoints and textures. The regression models evaluated are: Gaussian process regression (GPR), linear regression, *K* nearest neighbours (KNN) and neural networks (NN). The following methods are evaluated: holistic (in which features are extracted across an image and regression is performed globally); local (in which foreground segmentation is used to localise groups and to perform feature extraction and regression locally); and a histogram based approach [48].

☆ This paper has been recommended for acceptance by R. Bergevin.

* Corresponding author.
E-mail addresses: david.ryan@qut.edu.au (D. Ryan), s.denman@qut.edu.au (S. Denman), s.sridharan@qut.edu.au (S. Sridharan), c.fookes@qut.edu.au (C. Fookes).

http://dx.doi.org/10.1016/j.cviu.2014.07.008
1077-3142/

Our experiments demonstrate that local features outperform equivalent holistic features and histogram based features; best performance is observed using all image features except for textures; and that Gaussian process regression outperforms linear, *K*-nearest neighbours and neural network regression.

The remainder of this paper is structured as follows: Section 2 presents the literature review; Section 3 introduces the benchmark datasets used in this evaluation; Section 4 describes the system design; Section 5 presents the experimental results of the evaluation; and Section 6 discusses the conclusions of this research.

2. Literature review

Crowd counting algorithms are generally categorised into two groups: holistic and local. Holistic approaches use global image features to describe each frame in a video sequence, and a classifier or regression model is used to map between the feature space and the crowd size estimate. Local approaches, by contrast, utilise local image features to detect, track or count pedestrians within local regions of an image. In this case the crowd size is the sum of its parts. An intermediate approach has also been proposed [49,48] which utilises blob size histograms based on local segments and expresses this information on a holistic level.

Table 1
A taxonomy of crowd counting methods, image features and regression models used in this evaluation.

System component	Parameters evaluated
Counting method	Holistic Histograms (intermediate) Local
Image features	Size Shape Edges Keypoints Texture
Regression model	Gaussian process regression (GPR) Linear *K*-nearest neighbours (KNN) Neural network (NN)

Section 2.1 describes the holistic approaches; Section 2.2 discusses the intermediate approach; and Section 2.3 describes local approaches. Table 1 presents a taxonomy of system components used in this evaluation and Table 2 summarises the regression based algorithms discussed in the following literature review.

2.1. Holistic approaches

Holistic crowd counting algorithms use global image features to estimate the size of a crowd. They may also be described as "mapping-based" approaches because they map directly between the feature space and the crowd size estimate. Features used by these systems include textures [59], foreground pixels [24] and edge features [48], amongst others, while the classification and regression strategies have included linear regression [24], neural networks [59,48] and Gaussian process regression [8].

Textural approaches are based on the notion that low density crowds exhibit course textures and high density crowds exhibit fine textures. Rather than estimate the number of people directly, these approaches classify the crowd density using a four or five point scale.

Marana [59,57] proposed the use of grey level cooccurrence matrix (GLCM) based statistics [41] for crowd density estimation. Marana also proposed the Minkowski fractal dimension [60]. Xiaohua [83] proposed the use of the 2D discrete wavelet transform (DWT) as a basis for extracting textural features, while Rahmalan [69] proposed Translation Invariant Orthonormal Chebyshev Moments (TIOCM). Rahmalan's evaluation observed superior performance of textural features on an afternoon dataset, "because the afternoon data has smaller variation of illumination when compared with morning data". When morning and afternoon datasets were combined to form a larger mixed set, performance decreased compared to the afternoon dataset alone due to these illumination changes over time. This highlights the principle limitation of textural features: they are sensitive to the scene background, and are thus impractical for real world use as they would need to be re-trained after any significant background change.

Table 2
High level summary of regression based crowd counting systems. See the main text (Section 2) for a full description.

Method	Reference	Image features					Model
		Size	Shape	Edges	Keypoints	Texture	
Holistic	Regazzoni [71]			✓			EKF/BBN
	Davies [24]	✓		✓			Linear
	Marana [59,58,61,57]					✓	NN
	Marana [60]					✓	NN
	Cho [17,16,18]	✓		✓			NN
	Paragios [65]	✓					Linear
	Huang [45]	✓					NN
	Ma [53]	✓					Linear
	Rahmalan [69]					✓	NN
	Xiaohua [83]					✓	SVM tree
	Hou [43,44]	✓					NN
	Chan [8,6,10,7]	✓	✓	✓	✓	✓	GPR
	Zhang [84]	✓	✓	✓	✓	✓	GPR/ Ensemble (KNN + NN)
	Tan [80]	✓	✓	✓	✓	✓	Linear
Intermediate	Kong [49,48]	✓	✓				NN/Linear
Local							
Motion regions	Conte [19,20,22,21]				✓		ϵ-SVR
	Ryan [75,76]	✓	✓	✓	✓		GPR/Linear
	Celik [5]	✓					Linear
	Kilambi [46,47]	✓	✓				Linear (Cylinder model)
	Fehr [32]	✓	✓				Linear (Cylinder model)
Grid	Chen [13]	✓	✓	✓		✓	Linear
Pixelwise	Lempitsky [50]	✓	✓				Linear