J. Vis. Commun. Image R. 39 (2016) 218–225

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

An adaptive people counting system with dynamic features selection and occlusion handling ☆

Zeyad Q.H. Al-Zaydi [a,*], David L. Ndzi [a], Yanyan Yang [a], Munirah L. Kamarudin [b]

[a] School of Engineering, University of Portsmouth, Portsmouth PO1 3DJ, UK
[b] School of Computer and Communication Engineering, University Malaysia Perlis, Perlis, Malaysia

ARTICLE INFO

Article history:
Received 28 November 2015
Revised 31 March 2016
Accepted 30 May 2016
Available online 3 June 2016

Keywords:
Crowd counting
Surveillance systems
Image processing
Computer vision

ABSTRACT

This paper presents an adaptive crowd counting system for video surveillance applications. The proposed method is composed of a pair of collaborative Gaussian process models (GP) with different kernels, which are designed to count people by taking the level of occlusion into account. The level of occlusion is measured and compared with a predefined threshold for regression model selection for each frame. In addition, the proposed method dynamically identifies the best combination of features for people counting. The Mall and UCSD datasets are used to evaluate the proposed method. The results show that the proposed method offers a higher accuracy when compared against state of the art methods reported in open literature. The mean absolute error (MAE), mean squared error (MSE) and the mean deviation error (MDE) for the proposed algorithm are 2.90, 13.70 and 0.095, respectively, for the Mall dataset and 1.63, 4.32 and 0.066, respectively, for UCSD dataset.

1. Introduction

People counting is an important task for operational, safety and security purposes. Systems with these functions can be highly effective tools for establishing awareness [1–5]. Information about the number and distribution of people in a given space can be used to develop business intelligence, such as the interest in any product based on the number of customers visiting the area, counting the number of a store's visitors and other applications in behavioural economics [2,6,7]. In addition, there are other applications such as crowd management [2], transport [8], staff planning which are related to the density of visitor traffic or to indicate congestion. This kind of information can also be utilised to improve energy efficiency by optimising air conditioning, lighting and heating, or to develop emergency evacuation procedures [3].

Different technologies are often used to count people, such as tally counter, infrared beams, thermal imaging, computer vision, Service Set Identifier (SSID) from mobile phones, wireless sensor networks and Wi-Fi based counters [9–22]. The choice of system depends on different priorities which may include accuracy, flexibility, cost and acquiring people distribution information. Even though different techniques can be used for people counting, a method based on computer vision is one of the best choices because cameras have already become ubiquitous and their uses are increasing. For example, there were an estimated 4.2 million CCTV installed in the United Kingdom in 2004 [23]. People counting system is one of the most challenging systems in computer vision to implement [4,5,18,24]. In comparison with computer vision based technology, the problem with other technologies are that they need to be carefully planned and deployed for specific purposes. In addition, their cost is prohibitive for many organisations and the accuracy is often less than a computer vision based technology. Most of these systems are also ineffective for acquiring people distribution without high cost.

Different vision-based people counting methods have been developed to increase accuracy for both outdoor and indoor environments [12,14,18,20–22]. People counting based on computer vision can be classified into line of interest (LOI) and region of interest (ROI) [20]. LOI algorithms involve counting people who cross a virtual line in a certain period of time [12] whereas, ROI algorithms count people in a given space [21]. Video counters can also be classified into three categories; counting by detection, regression and clustering [25–27].

People counters based on detection involve detecting all people in a frame-to-frame analysis individually. The number of people and their location are then obtained [28]. The detection process can depend on an entire person's body, face, eyes, head, head

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: zeyad.al-zaydi@port.ac.uk (Z.Q.H. Al-Zaydi), david.ndzi@port.ac.uk (D.L. Ndzi), linda.yang@port.ac.uk (Y. Yang), latifahmunirah@unimap.edu.my (M.L. Kamarudin).

http://dx.doi.org/10.1016/j.jvcir.2016.05.018
1047-3203/

and shoulder or shape matching using ellipses or Bernoulli shapes. They can also use multiple cameras or density aware information to improve accuracy [29,30]. Different features can be used to represent the appearance of people, such as Haar like features [31] and histogram of oriented gradient features [32]. Different classifiers are also used for learning how to detect people such as support vector machine, neural network and AdaBoost [32,33]. People counters based on detection are significantly affected by occlusion, varying lighting and have long processing time [26]. In low crowd density scenarios they produce more accurate results, whereas the accuracy decreases significantly in high crowd density scenarios [18]. In addition, they require high-resolution videos to achieve good accuracies.

Low level features regression algorithms usually involve a background subtraction that is applied to a frame-to-frame analysis and then extracting useful features from the foreground such as foreground segment features [18,34–41], edge features [26,36,40–42], texture features [26,34,37–39] and keypoints [34,37–39,42]. A regression model is then trained using the extracted features to find the relationship between those features and the number of people without detecting each person individually [43]. Different regression models have been proposed that include linear regression [44,45], Neural networks [36,40–42] and Gaussian process regressions [1,37,38]. Low level features regression algorithms preserve privacy and their accuracy is better than that of detection based and feature trajectories clustering algorithms in crowded environments [22].

In feature trajectories clustering based algorithms, useful features are tracked in a frame-to-frame analysis and then cluster the trajectories using spatial and temporal consistency heuristics or use other factors to find the unique track for each person [43,46–48]. The number of clusters represents the number of people [49]. The accuracy significantly decreases in crowded scenarios with frequent inter-object occlusion. A complex trajectory management is required due to occlusions and requires a robust method to assess similarities between trajectories of different lengths [14]. In addition, accuracy can be affected by errors of coherently moving features that do not fit to the same person [14].

This work distinguishes itself with the following four main contributions. First, a pair of collaborative Gaussian process models (GP) with different kernels is used to handle occlusion. Second, a principled technique is proposed to measure the level of occlusion in a frame. Third, it proposes a method of choosing the best combination of features depending on their environment. Fourth, the system is comprehensively evaluated using two benchmark datasets, the Mall and University of California, San Diego (UCSD) datasets.

2. System design

This section provides the detailed description of the proposed system starting with the description of the low-level and high-level occlusion regression models. Secondly, the method to measure the level of occlusion in occlusion-level model is described. Thirdly, the feature representation and selection is presented which is followed by a description of the mechanisms for handling variations of scales and appearances in cameras. An overview of the proposed system is given in Fig. 1.

2.1. The low-level and high-level occlusion regression models

Low-level features regression algorithms usually consists of three steps: (a) background subtraction that is applied in a frame-to-frame analysis; (b) extraction of useful features from foreground such as foreground segment features, edge features, texture features and keypoints, and (c) a regression model trained to find the relationship between the number of people and the extracted features which is used to estimate the number of people.

Two independent Gaussian process regression (GPR) models with different kernels are used in the proposed system. The first regression model (low-level occlusion regression model) is trained with low occlusion frames and the second (high-level occlusion regression model) is trained with high occlusion frames. Mathematically, estimation of the number of people in GPR follows the Gaussian distribution [50]:

$$y_*|y \sim N(K_*K^{-1}y, K_{**} - K_*K^{-1}K_*^T) \tag{1}$$

and the best estimate for y_* is the mean of this distribution [50]:

$$y_* = K_*K^{-1}y \tag{2}$$

and the uncertainty in the estimate is captured in its variance [50]:

$$var(y_*) = K_{**} - K_*K^{-1}K_*^T \tag{3}$$

where y and y_* are the function values of the training and testing sets, respectively. K, K_* and K_{**} are the covariance functions (kernels) of the training, training–testing and testing inputs, respectively. There are different kernels that can be used with a Gaussian process regression. In low level occlusion scenarios, feature values are expected to grow linearly with respect to the number of people so a linear kernel is used in the regression model [51]. The linear kernel on two inputs x and x', represented as feature vectors is given by [37]:

$$k(x, x') = \alpha(x^T x' + 1) \tag{4}$$

α is the kernel parameter. In high level occlusion scenarios, the relationship between the features and the number of people follows a linear trend roughly while the data fluctuates non-linearly due to occlusion [52]. A combination of linear and radial basis function (RBF) kernels are used in a high-occlusion regression model. The linear kernel can capture the linear main trend well and the RBF kernel can be used to model the fluctuation of the data points [52]. Mathematically, a combination of linear and RBF kernels is given by [37,50]:

$$k(x, x') = \alpha_1(x^T x' + 1) + \alpha_2^2 \exp\left[\frac{-1}{2\alpha_3^2}\|x - x'\|^2\right] \tag{5}$$

α_1, α_2 and α_3 are the kernals parameters. In addition, we can use an ensemble learning method that first partitions the heterogeneous training data into linear and non-linear homogeneous sections (low-level occlusion frames and high-level occlusion frames) and then build a regression model for each homogeneous section. Unlike most existing ensemble learning methods where different models are combined linearly, the proposed method uses a switch approach between the regression models that automatically determines which regression model should be applied to input frame. In conclusion, dividing heterogeneous training data into a number of homogeneous partitions will likely generate reliable and accurate regression models over the homogeneous partitions that may increase the accuracy of the proposed method [53,54]. In the next section, the method of measuring the level of occlusion is explained.

2.2. The occlusion-level model

Many studies have used keypoints to find the level of the crowd (number of people) due to their strong inter-dependance [55–58]. It is worth noting that although there is a degree of correlation relationship between the level of the crowd and the level of occlusion in a frame, this relationship is not always valid in all scenarios. As a consequence, there is a need to develop a method to measure the level of occlusion that takes into account the level of the crowd