Accepted Manuscript

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

Zeyad Q.H. Al-Zaydi, David L. Ndzi, Yanyan Yang, Munirah L. Kamarudin

PII: S1047-3203(16)30085-2
DOI: <http://dx.doi.org/10.1016/j.jvcir.2016.05.018>
Reference: YJVC I 1769

To appear in: *J. Vis. Commun. Image R.*

Received Date: 28 November 2015
Revised Date: 31 March 2016
Accepted Date: 30 May 2016

Please cite this article as: Z.Q.H. Al-Zaydi, D.L. Ndzi, Y. Yang, M.L. Kamarudin, An adaptive people counting system with dynamic features selection and occlusion handling, *J. Vis. Commun. Image R.* (2016), doi: <http://dx.doi.org/10.1016/j.jvcir.2016.05.018>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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

Zeyad Q. H. Al-Zaydi^{1*}, David L. Ndzi¹, Yanyan Yang¹, Munirah L. Kamarudin²

¹School of Engineering, University of Portsmouth, Portsmouth, PO1 3DJ, UK

²School of Computer and Communication Engineering, University Malaysia Perlis, Perlis, Malaysia

Tel: +44(0)23 9284 2551

E-mails: {zeyad.al-zaydi, david.ndzi, linda.yang}@port.ac.uk,
latifahmunirah@unimap.edu.my

Corresponding Author:

Zeyad Q. H. Al-Zaydi

School of Engineering, University of Portsmouth, Portsmouth, PO1 3DJ, UK

Tel: +44(0)23 9284 2551

Email: zeyad.al-zaydi@port.ac.uk

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.

Keywords: Crowd counting, Surveillance systems, Image processing, Computer vision.

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 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 Figure 1.

Figure 1: Flow diagram of the proposed framework.

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]:

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

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

$$y_* = K_*K^{-1}y \quad (2)$$

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

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

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]:

$$k(x, x') = \alpha (x^T x' + 1) \quad (4)$$

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]:

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

α_1, α_2 and α_3 are the kernels 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 and its sparseness. The level of occlusion that is measured will be compared with a predefined threshold to choose which regression model works. A simple equation has been derived to measure crowd density (number of people) [58]:

$$\text{Level of crowd} = \frac{\text{No of keypoints}}{\text{No of keypoints per person}} \quad (6)$$

A simple equation is also used to measure the level of occlusion:

$$\text{Level of occlusion} = \frac{\text{No of keypoints}}{\text{No of foreground pixels}} \quad (7)$$

Scale Invariant Feature Transform (SIFT) points are used as keypoints in the proposed system. SIFT points are defined as maxima/minima of the difference of Gaussians in scale-space [59]. SIFT keypoints is better than Features from Accelerated Segment Test (FAST) and Speeded-up Robust Features (SURF) because they are more invariant to scale, rotation, and affine transformations [59].

The output of the occlusion-level model of each frame is compared with a suitable threshold. The thresholding stage involves separating frames into two groups; high and low occluded frames. There is no technical definition of the low and high occluded frames because there are no clear boundaries between them. The range of crowd sizes, resolution and the area of the camera view are potential factors that can affect the choice of a suitable threshold because they specify the range of occlusion. The highest occluded frames in one environment can be the lowest occluded frames in another environment depending on the crowd size. In conclusion, choosing a suitable threshold depends on the type of environment in real-time applications or datasets in offline applications. In addition, the use of a fixed threshold for all environments would be problematic since the threshold would need to be adjusted depending on the crowd size and the area of the camera view. In the proposed system, the threshold is experimentally determined by using a multi-stage thresholding method. The range of the crowd is normalised to 0 - 1 range. In the first stage, the range is divided into ten equal intervals which are used as potential thresholds. Those thresholds are used to measure the accuracy of the system. In the second stage, the interval with the highest accuracy is divided into ten equal intervals which are used as potential thresholds. The threshold with the highest accuracy from both stages is selected as the best threshold.

In real time applications, the computational efficiency is important. In comparison to the state of the art methods, the training implemented in this paper is repeated 31 times using the classical GPR method to find the best combination of features. In addition, the best combination of features is used to train the proposed system with 19 potential thresholds. This takes approximately 143 and 190 seconds for each training in the UCSD and Mall datasets, respectively. The system training is performed only at the installation of the system so computational complexity can be neglected when the system starts working.

The occlusion-level model and thresholding method are a simple division and relational operators, respectively. They add small computation complexity to the working system in comparison to frame extraction, background segmentation using GMM algorithm, prospective normalization, features extraction and prediction using the GPR algorithm. One of the regression models (the low or high level regression models) has been used with each frame so the computation complexity of the regression stage is not changed.

In conclusion, the computational efficiency of the proposed system is a little less than classical GPR algorithms in the testing stage. However, the computation complexity is significantly increased in the training stage due to the threshold and features selections, it is not important because training is performed only at the installation stage of the counting system.

2.3. Feature representation and selection

Features is a general term used to describe low-level visual properties in an image or video such as colour, size, shape, intensity, edge and texture [5], [60]. Different features can be used as intermediate inputs to a regression model for people counting. A popular approach is to combine several features to achieve higher accuracy. The performance of various features and combinations of features for crowd counting depend on the type of environments in a real-time applications or datasets in offline applications [60]. As a consequence, the optimal combination of features for one environment may not be optimal for the others. In conclusion, an adaptive people counting method can be implemented which is capable of dynamically identifying the best features that can be used to find the number of people. Features can be categorised under the following headings:

- 1- Foreground features: they are common features in people counting that are obtained through a background subtraction algorithm. Foreground features are extracted to capture segment properties and can be categorised into two groups based on size and shape [60]. Size features include the number of foreground pixel (area), the total pixels count on the segment perimeter (perimeter), the complexity of the segment shape (perimeter-area ratio) and the number of blobs in a frame (blob count). Shape features refer to the orientation of the perimeter pixels, which include Perimeter orientation histogram [5].

- 2- Edge features: they refer to the relative change in pixel intensities across a frame [1]. They have a strong relationship with the number of people because there is a strong dependency between the number of people and the complexity of crowds. Low density crowds tend to present coarse edges while high density crowds tend to present complex edges [5]. Some common edge features are Total edge pixels, Edge orientation histogram and Minkowski dimension, which refer to how many pre-defined structure elements are required to fill the edge space [61].
- 3- Texture features: there is a strong relationship between the number of people and the texture of crowds, which refer to general description of a frame [5]. Gray-level co-occurrence matrix and local binary pattern are usually used to find texture features [5], [62], [63]. Texture features include homogeneity (texture smoothness), energy (total sum-squared energy), entropy (texture randomness) and contrast [5], [37], [60].
- 4- Keypoints: they refer to specific pixels of interest in an image or video [60]. The results of using moving keypoints to find the number of people show that they have a strong relationship [55]–[58]. Many people counting studies have been carried out using FAST, SIFT and SURF points [24], [60].

The optimal combination of features for any environment can be selected by training the regression model with different potential combination of features. There are only 31 combinations of features [60]. Multi stage training has been used in this paper to train a regression model to find the optimal combination of features.

Table 1: The potential features to be optimal (S = Size, P = Shape, E = Edges, K = Keypoints, T = Texture) [60].

All Combinations of Features		
S	PT	PET
P	EK	PKT
E	ET	EKT
K	KT	SPEK
T	SPE	SPET
SP	SPK	SPKT
SE	SPT	SEKT
SK	SEK	PEKT
ST	SET	SPEKT
PE	SKT	
PK	PEK	

2.4. Geometric correction

The size of a person changes depending on the distance of the person to the camera. As a consequence, features extracted from the person at different depths in frames would have

significantly different values. To solve this problem, different weights are used for pixels in the frames. Figure 2 shows the size of the same person at different distances to the camera. The weight of pixels at (ab) line is 1, whereas the weights of any other line can be found using the following equation [5]:

$$weight_{line} = \frac{h_{ab}w_{ab}}{h_{line}w_{line}} \quad (8)$$

Where h_{ab} and h_{line} are the heights of a person on (ab) line and the height of the same person on the line of interest, respectively. w_{ab} and w_{line} are the widths of the rectangle at (ab) line and at the line of interest, respectively.

Figure 2: The difference of size for the same person at different position.

3. RESULTS AND DISCUSSION

The UCSD and Mall datasets have been used for testing and evaluation [26], [34]. Three metrics are used to measure the performance of the proposed method; mean deviation error (MDE), mean absolute error (MAE) and mean squared error (MSE) [5]. For the UCSD and Mall datasets, the datasets are split into a training set, for learning the high and low occlusion models, and a test set, for validation. We followed the same training and testing partition as in [26], [34], [64], 800 frames are used for training and 1200 frames for testing. The proposed system is implemented using Matlab software and it is running on a PC with 2.5GHz core I5 processor and 4 GB memory. The mean deviation error is given as:

$$MDE = \frac{1}{N} \sum_{n=1}^N \frac{|y_n - \hat{y}_n|}{y_n} \quad (9)$$

The mean absolute error is given as:

$$MAE = \frac{1}{N} \sum_{n=1}^N |y_n - \hat{y}_n| \quad (10)$$

The mean squared error is given as:

$$MSE = \frac{1}{N} \sum_{n=1}^N (y_n - \hat{y}_n)^2 \quad (11)$$

Where N is the total number of the test frames, y_n is the actual count, and $\hat{y}_n$ is the estimated count.

3.1. Evaluation of the proposed system performance using the Mall dataset

The Mall dataset contains 2000 annotated frames inside a cluttered indoor shopping centre. As shown in Table 2, the MAE, MSE and MDE are 2.90, 13.70 and 0.095, respectively. From the results, it can be seen that the error of the proposed method is lower than other state of the art methods. Figure 3 shows some frames from the Mall dataset with their true number of people (TC), which are annotated with red dots, the estimated number of people (EC), MAE, MSE and MDE.

Table 2: Performance comparison between different algorithms (Mall dataset).

Algorithm	Mall dataset		
	MAE	MSE	MDE
The proposed method	2.90	13.70	0.095
Cumulative attribute based model (CA-RR) [64]	3.43	17.70	0.105
Squares Support Vector Machine Regression (LSSVR) [64]	3.51	18.20	0.108
Kernel Ridge Regression (KRR) [64]	3.51	18.10	0.108
Random Forest Regression (RFR) [64]	3.91	21.50	0.121
Gaussian Process Regression (GPR) [26], [64]	3.72	20.10	0.115
Ridge regression (RR) [26], [64]	3.59	19.00	0.110
Multi Output Ridge Regression (MORR) [26]	3.15	15.70	0.098
Multiple Localised Regression (MLR) [26]	3.90	23.90	0.119
Weighted Ridge Regression (WRR) [25]	3.44	18.00	0.105

TC=36
EC=38
MAE=2
MSE=4
MDE=0.055

TC=26
EC=29
MAE=3
MSE=9
MDE=0.115

(a)

(b)

TC=19
EC=20
MAE=1
MSE=1
MDE=0.051

TC=29
EC=31
MAE=2
MSE=4
MDE=0.069

(c)

(d)

Figure 3: Examples of the Mall dataset frames and their results.

3.2. Evaluation of the proposed system performance using the UCSD dataset

The UCSD dataset contains 2000 annotated frames of people moving in two directions along a walkway. As shown in Table 3, the MAE, MSE and MDE are 1.63, 4.32 and 0.066, respectively. From the results, it can be seen that the error of the proposed method is lower than other state of the art methods. Figure 4 shows some frames from the UCSD dataset with their true number of people (TC), which are annotated with red dots, the estimated number of people (EC), MAE, MSE and MDE.

Table 3: Performance comparison between different algorithms (UCSD dataset).

Algorithm	UCSD dataset		
	MAE	MSE	MDE
The proposed method	1.63	4.32	0.066
Improved Iterative Scaling -Label Distribution Learning (IIS-LDL) [65]	2.08	7.25	0.098
Kernel Ridge Regression (KRR) [65]	2.16	7.45	0.107
Random Forest Regression (RFR) [65]	2.42	8.47	0.116
Gaussian Process Regression (GPR) [25], [65]	>2.24	>7.97	>0.112
Ridge Regression (RR) [25], [65]	2.25	7.82	0.110
Multi Output Ridge Regression (MORR) [65]	2.29	8.08	0.109
Cumulative attribute based model (CA-RR) [64], [65]	2.07	6.86	0.102
Weighted Ridge Regression (WRR) [25]	2.05	6.75	0.102
Linear regression (LR), Partial Least Squares Regression (PLSR), KRR, LSSVR, GPR and RFR [5]	>2.02	>6.67	>0.100

TC=18
EC=19
MAE=1
MSE=1
MDE=0.055

(a)

TC=23
EC=23
MAE=0
MSE=0
MDE=0

(b)

TC=15
EC=15
MAE=0
MSE=0
MDE=0

(c)

TC=23
EC=21
MAE=2
MSE=4
MDE=0.086

(d)

Figure 4: Examples of the UCSD dataset frames and their results.

3.3. Computation efficiency

To study the computational efficiency of the proposed method, comparisons between the state of the art GPR algorithm and the proposed algorithm are made. GPR algorithm is chosen because it uses the same regression model which makes the comparison more reliable. The comparison with the GPR algorithm that uses the same regression model will also show the extra computational complexity of the proposed system that arises due to the extra steps. In addition, the comparison can be considered to be reliable as the classical GPR algorithm is implemented using the same hardware (PC) and software (Matlab 2014a). Comparison with other reported algorithms will not be effective because authors who reported any computational efficiency results could have been using different hardware or software. As shown in Table 4, the processing speeds of the proposed system are 10.48 fps and 11.43 fps for the MALL and UCSD datasets, respectively. The difference in the processing speeds between the classical GPR method and the proposed system is acceptable.

Table 4: The computation efficiency of the proposed system.

Algorithm	Processing speed (fps)	
	Mall	UCSD
The proposed method	10.48	14.88
Classical Gaussian Process Regression (GPR)	11.43	16.82

3.4. Threshold selection

To study the effect of different choices of threshold on the proposed system efficiency, the results of each stage of the multi-stage thresholding method are shown in Table 5 and 6. The multi-stage thresholding method is used to improve the accuracy by selecting the best threshold experimentally. 19 potential thresholds are used with each dataset to measure the accuracy of the system. The threshold with highest accuracy of the stages is selected to be the best threshold. Table 5 and 6 show the results of the Mall and UCSD datasets at each stage. Nine and ten thresholds are shown at the first and second stages, respectively. The best threshold for each dataset is marked in bold.

Threshold	Stage 1			Threshold	Stage 2		
	MAE	MSE	MDE		MAE	MSE	MDE
0.1	2.9417	13.8217	0.0965	0.55	2.9508	14.0292	0.0968
0.2	3.0333	14.6517	0.0995	0.56	2.9342	13.9742	0.0963
0.3	2.9725	14.1325	0.0975	0.57	2.9200	13.7533	0.0958
0.4	3.0292	14.6708	0.0994	0.58	2.9450	13.8867	0.0966
0.5	2.9367	14.2800	0.0963	0.59	2.9033	13.6233	0.0953
0.6	2.9083	13.6217	0.0954	0.61	2.9033	13.6433	0.0953
0.7	2.9333	13.9950	0.0962	0.62	2.9025	13.7042	0.0952
0.8	2.9317	13.9000	0.0962	0.63	2.9258	13.9092	0.0960
0.9	2.9408	13.8442	0.0965	0.64	2.9383	14.1167	0.0964
-	-	-	-	0.65	2.9392	14.1208	0.0964

Table 5: The performance at each stage (Mall dataset).

Table 6: The performance at each stage (UCSD dataset).

Threshold	Stage 1			Threshold	Stage 2		
	MAE	MSE	MDE		MAE	MSE	MDE
0.1	1.7100	4.8767	0.0695	0.15	1.6592	4.6042	0.0680
0.2	1.6417	4.5900	0.0672	0.16	1.6542	4.5575	0.0678
0.3	1.8758	5.9442	0.0768	0.17	1.6567	4.5333	0.0679
0.4	1.7267	4.9667	0.0707	0.18	1.6442	4.4808	0.0673
0.5	1.6733	4.8417	0.0685	0.19	1.6308	4.3275	0.0668
0.6	1.6917	4.8067	0.0693	0.21	1.6508	4.4708	0.0676
0.7	1.7892	5.3325	0.0733	0.22	1.6800	4.6583	0.0688
0.8	1.7825	5.3325	0.0733	0.23	1.7075	4.8342	0.0699
0.9	1.8150	5.4500	0.0743	0.24	1.7550	5.4417	0.0719
-	-	-	-	0.25	1.7775	5.2775	0.0728

3.5. Features selection

The relationship between the datasets and their optimal combination of features is fuzzy and there is no standard criteria to explain why a particular combination of features is appropriate for a given dataset. Environments can be described by different characteristics, e.g. frame rate, resolution, colour, location (indoor and outdoor), shadows, reflections, loitering, crowd size, occlusion level, background texture and background complexity. All these characteristics have an effect on the nature of the combination of features that is appropriate for a given dataset.

On the other hand, different features can be selected as the optimal set for a given dataset, e.g. foreground (shape and size), texture, edge and keypoints. This can lead to a high or low correlation with an environment depending on the characteristics of that environment. For instance, edge features can work better than size features in crowded environments because size features are reduced by occlusions while the edge features become stronger due to the overlapping body parts, differing skin tones and clothing [60]. Texture features can achieve high performance in environments with high textured backgrounds [60]. Some keypoints features are more appropriate for the high perspective distortion because some of them are scale invariant such as SIFT keypoints. Although there are some potential reasons that explain why such a combination of features is particularly appropriate for a given dataset, there is no standard criteria for the selection. In conclusion, training people counting systems with all potential combinations of features is the best solution to this problem.

The appropriate combination of features with the highest accuracy for the Mall and UCSD datasets are SPKT and SPEKT, respectively, where S = Size, P = Shape, E = Edges, K = Keypoints and T = Texture. The potential justification for this selection is that edge features are highly inaccurate in environments with complicated backgrounds and uneven textures of human clothes [24]. Mall dataset has a high complicated background, shadows and reflections than the UCSD dataset [60]. In addition, using different kinds of features can help to mitigate the non-linearities that arise from occlusion, segmentation errors and pedestrian configuration [34].

4. CONCLUSIONS

An adaptive and accurate people counting system is proposed and implemented which is capable of dynamically identifying the best set of features. In addition, two Gaussian regression models are used to improve the accuracy, which the most suitable regression model for each frame is selected depending on the level of occlusion. Experimental results based on two crowd datasets (UCSD & Mall datasets) have demonstrated that the proposed technique outperforms state-of-the-art methods reported in open literature. They achieve good results under situations of heavy occlusions and perspective distortions. By means of comparisons with other existing low-level features regression methods, our results demonstrate the ability of the proposed system to

outperform the others methods with respect to MAE, MSE and MDE metrics. The MAE, MSE and MDE of the proposed method are lower than those of comparable methods (2.90, 13.70 and 0.095, respectively, for the Mall dataset and 1.63, 4.32 and 0.066, respectively, for UCSD dataset). The computational efficiency results of the proposed system show that the processing effect of the extra steps is small with 10.48 fps and 11.43 fps for the MALL and UCSD datasets, respectively.

REFERENCES:

- [1] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Scene invariant multi camera crowd counting," *Pattern Recognit. Lett.*, vol. 44, pp. 98–112, 2014.
- [2] A. Technology, "Our customers," 2013. [Online]. Available: www.peoplecounting.co.uk/our-customers. [Accessed: 23-Mar-2015].
- [3] M. Wang, "Data Assimilation for Agent-Based Simulation of Smart Environment," 2014.
- [4] D. A. Ryan, "Crowd Monitoring Using Computer Vision," Queensland University of Technology, 2013.
- [5] C. Loy, K. Chen, S. Gong, and T. Xiang, *Crowd counting and profiling: Methodology and evaluation*. 2013.
- [6] ShopperTrak, "ShopperTrak Solutions," 2013. [Online]. Available: www.shoppertrak.com/products. [Accessed: 23-Mar-2015].
- [7] Biodata Ltd, "Use CCTV to Count People," 2013. [Online]. Available: www.videoturnstile.com. [Accessed: 23-Mar-2015].
- [8] DILAX Intelcom, "Public Transport," 2015. [Online]. Available: www.dilax.net/electronic-people-counter-passenger-counter-customer-counter/passenger-counting-systems-for-public-transportation. [Accessed: 23-Mar-2015].
- [9] F. Z. F. Zhu, X. Y. X. Yang, J. G. J. Gu, and R. Y. R. Yang, "A New Method for People-Counting Based on Support Vector Machine," *2009 Second Int. Conf. Intell. Networks Intell. Syst.*, vol. 1, no. 5, pp. 5–8, 2009.
- [10] N.-N. Li, J. Song, R.-Y. Zhou, and J.-H. Gu, "A People-Counting System Based on BP Neural Network," *Fourth Int. Conf. Fuzzy Syst. Knowl. Discov. (FSKD 2007)*, vol. 3, no. Fskd, 2007.
- [11] M. Nakatsuka, H. Iwatani, and J. Katto, "A Study on Passive Crowd Density Estimation using Wireless Sensors," *4th Intl.-Conf.-on Mob. Comput. Ubiquitous Netw. (ICMU 2008)*, no. 2, pp. 1–6, 2008.
- [12] Z. Ma and A. B. Chan, "Crossing the line: Crowd counting by integer programming with local features," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 2539–2546, 2013.
- [13] Y. Yuan, C. Qiu, W. Xi, and J. Zhao, "Crowd Density Estimation Using Wireless Sensor Networks," *2011 Seventh Int. Conf. Mob. Ad-hoc Sens. Networks*, pp. 138–145, 2011.
- [14] R. Shbib, S. Zhou, D. Ndzi, and K. Al-kadhimi, "Distributed Monitoring System Based On Weighted Data Fusing Model," *Am. J. Soc. Issues Humanit.*, vol. 3, no. March, pp. 53–62, 2013.
- [15] W. Xi, J. Zhao, X. Y. Li, K. Zhao, S. Tang, X. Liu, and Z. Jiang, "Electronic frog eye: Counting crowd using WiFi," *Proc. - IEEE INFOCOM*, pp. 361–369, 2014.
- [16] W.-C. Lin, W. Seah, and W. Li, "Exploiting radio irregularity in the Internet of Things for automated people counting," *IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC*, pp. 1015–1019, 2011.
- [17] V. Gandhi, J. Čech, and R. Horaud, "High-resolution depth maps based on TOF-stereo fusion," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 4742–4749, 2012.
- [18] Y. L. Hou and G. K. H. Pang, "People counting and human detection in a challenging situation," *IEEE Trans. Syst. Man, Cybern. Part A Systems Humans*, vol. 41, no. 1, pp. 24–33, 2011.

- [19] T. Tikkanen, "People detection and tracking using a network of low-cost depth cameras," 2014.
- [20] J. Li, L. Huang, and C. Liu, "Robust people counting in video surveillance: Dataset and system," *2011 8th IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2011*, pp. 54–59, 2011.
- [21] J. Tu, C. Zhang, and P. Hao, "Robust real-time attention-based head-shoulder detection for video surveillance," *Image Process.*, pp. 3340–3344, 2013.
- [22] A. Adegboye, G. Hancke, and G. H. Jr, "Single-pixel approach for fast people counting and direction estimation," *Satnac.Org.Za*.
- [23] C. Norris, M. Mccahill, and D. Wood, "Editorial . The Growth of CCTV : a global perspective on the international diffusion of video surveillance in publicly accessible space .," *Surveill. Soc.*, vol. 2, no. 2/3, pp. 110–135, 2004.
- [24] S. A. M. Saleh, S. A. Suandi, and H. Ibrahim, "Recent survey on crowd density estimation and counting for visual surveillance," *Eng. Appl. Artif. Intell.*, vol. 41, pp. 103–114, 2015.
- [25] K. Chen and J.-K. Kamarainen, "Learning to Count with Back-Propagated Information," *Pattern Recognit.*, pp. 4672–4677, 2014.
- [26] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature Mining for Localised Crowd Counting," *Proceedings Br. Mach. Vis. Conf. 2012*, pp. 21.1–21.11, 2012.
- [27] H. Foroughi, N. Ray, and H. Zhang, "People counting with image retrieval using compressed sensing," *Acoust. Speech Signal Process. (ICASSP), IEEE*, pp. 4354–4358, 2014.
- [28] O. Sidla, Y. Lypetsky, N. Brändle, and S. Seer, "Pedestrian detection and tracking for counting applications in crowded situations," *Proc. - IEEE Int. Conf. Video Signal Based Surveill. 2006, AVSS 2006*, 2006.
- [29] H. Ma, C. Zeng, and C. X. Ling, "A Reliable People Counting System via Multiple Cameras," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 2, pp. 1–22, 2012.
- [30] M. Rodriguez, E. N. Superieure, I. Laptev, J. Sivic, and J.-Y. Audibert, "Density-aware person detection and tracking in crowds," *Comput. Vis. (ICCV), IEEE*, pp. 2423–2430, 2011.
- [31] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
- [32] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *CVPR '05 Proc. 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Vol. 1*, pp. 886–893, 2005.
- [33] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, 2010.
- [34] A. B. Chan, Z. S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," *26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR*, 2008.
- [35] J. Zhang, B. Tan, F. Sha, and L. He, "Predicting pedestrian counts in crowded scenes with rich and high-dimensional features," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1037–1046, 2011.
- [36] T. W. S. Chow, J. Y. F. Yam, and S. Y. Cho, "Fast training algorithm for feedforward neural networks: Application to crowd estimation at underground stations," *Artif. Intell. Eng.*, vol. 13, no. 3, pp. 301–307, 1999.
- [37] A. B. Chan and N. Vasconcelos, "Counting people with low-level features and bayesian regression," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2160–2177, 2012.
- [38] A. B. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," *Proc. IEEE Int. Conf. Comput. Vis.*, no. Iccv, pp. 545–551, 2009.
- [39] A. Chan, M. Morrow, and N. Vasconcelos, "Analysis of Crowded Scenes using Holistic Properties," *9th Int. Symp. PETS*, no. Pets, 2009.
- [40] S.-Y. Cho, T. Chow, and C. Leung, "A neural-based crowd estimation by hybrid global learning

- algorithm,” *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 29, no. 4, pp. 535–541, 1999.
- [41] T. W. Cho, S. Y., & Chow, “A fast neural learning vision system for crowd estimation at underground stations platform,” *Neural Process. Lett.*, vol. 10(2), pp. 111–120, 1999.
 - [42] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, “Crowd counting using multiple local features,” *DICTA 2009 - Digit. Image Comput. Tech. Appl.*, pp. 81–88, 2009.
 - [43] I. S. Topkaya, H. Erdogan, and F. Porikli, “Counting People by Clustering Person Detector Outputs,” pp. 313–318, 2014.
 - [44] R. Ma, L. Li, W. Huang, and Q. Tian, “On pixel count based crowd density estimation for visual surveillance,” *IEEE Conf. Cybern. Intell. Syst. 2004.*, vol. 1, pp. 1–3, 2004.
 - [45] A. Davies, J. H. Yin, and S. Velastin, “Crowd monitoring using image processing,” no. February, 1995.
 - [46] G. J. Brostow and R. Cipolla, “Unsupervised Bayesian Detection of Independent Motion in Crowds,” *Comput. Vis. Pattern Recognition*, 2006.
 - [47] V. Rabaud and S. Belongie, “Counting crowded moving objects,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, pp. 705–711, 2006.
 - [48] A. M. Cheriadat, B. L. Bhaduri, and R. J. Radke, “Detecting multiple moving objects in crowded environments with coherent motion regions,” *2008 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. CVPR Work.*, 2008.
 - [49] D. Merad, K. E. Aziz, and N. Thome, “Fast people counting using head detection from skeleton graph,” *Proc. - IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2010*, pp. 233–240, 2010.
 - [50] C. K. I. Williams and C. E. Rasmussen, “Gaussian processes for regression: A Quick Introduction,” no. August, 2008.
 - [51] T. Y. Lin, Y. Y. Lin, M. F. Weng, Y. C. Wang, Y. F. Hsu, and H. Y. M. Liao, “Cross camera people counting with perspective estimation and occlusion handling,” *2011 IEEE Int. Work. Inf. Forensics Secur. WIFS 2011*, 2011.
 - [52] J. Mei and Y. Zhao, “An Improved Method of Crowd Counting Based on Regression,” pp. 143–150, 2013.
 - [53] R. Jin and H. Liu, “SWITCH: A Novel Approach to Ensemble Learning for Heterogeneous Data,” *Mach. Learn. ECML 2004*, pp. 560–562, 2004.
 - [54] R. Jin and H. Liu, “A Novel Approach to Model Generation for Heterogeneous Data Classification,” *IJCAI’05 Proc. 19th Int. Jt. Conf. Artif. Intell.*, pp. 746–751, 2005.
 - [55] A. Albiol, M. J. Silla, J. M. Mossi, and A. Albiol, “Video analysis using corner motion statistics,” *11th IEEE Int. Work. Perform. Eval. Track. Surveill. (PETS 2009)*, pp. 31–38, 2009.
 - [56] A. Albiol, A. Albiol, and J. Silla, “Statistical video analysis for crowds counting,” *Proc. - Int. Conf. Image Process. ICIP*, pp. 2569–2572, 2009.
 - [57] D. Conte, P. Foggia, G. Percannella, F. Tufano, and M. Vento, “Counting moving people in videos by salient points detection,” *Proc. - Int. Conf. Pattern Recognit.*, pp. 1743–1746, 2010.
 - [58] S. W. Jeong, C. Y., Choi, S., & Han, “A method for counting moving and stationary people by interest point classification,” *Image Process.*, pp. 4545–4548, 2013.
 - [59] H. Fradi and J. Dugelay, “People counting system in crowded scenes based on feature regression,” *Signal Process. Conf.*, no. Eusipco, pp. 136–140, 2012.
 - [60] D. Ryan, S. Denman, S. Sridharan, and C. Fookes, “An evaluation of crowd counting methods, features and regression models,” *Comput. Vis. Image Underst.*, vol. 130, pp. 1–17, 2015.
 - [61] A. N. Marana, L. D. F. Costa, R. A. Lotufo, and S. A. Velastin, “Estimating crowd density with Minkowski fractal dimension,” *1999 IEEE Int. Conf. Acoust. Speech, Signal Process. Proceedings. ICASSP99 (Cat. No.99CH36258)*, vol. 6, pp. 3521–3524, 1999.
 - [62] S. Ershad, “Texture Classification Approach Based on Combination of Edge & Co-occurrence and Local Binary Pattern,” *arXiv Prepr. arXiv1203.4855*, pp. 626–629, 2012.

- [63] B. Wang, H. Bao, S. Yang, and H. Lou, "Crowd Density Estimation Based on Texture Feature Extraction," *J. Multimed.*, vol. 8, no. 4, pp. 331–337, 2013.
- [64] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, no. 2, pp. 2467–2474, 2013.
- [65] Z. Zhang, M. Wang, and X. Geng, "Crowd counting in public video surveillance by label distribution learning," *Neurocomputing*, vol. 166, pp. 151–163, 2015.

Zeyad Q. H. Al-Zaydi received his B.S. in Computer Engineering, from University of Technology, Iraq in 2003. He received his M.S. in Computer Engineering from University of Baghdad, Iraq in 2012. He is currently a PhD student in the school of Engineering, University of Portsmouth, UK. His research interests are in the areas of image processing, computer vision, image representation, pattern recognition, artificial intelligence, people detection and crowd counting.

Dr. David L. Ndzi graduated with BSc (Joint Honours) in Electronics and Mathematics from Keele University in 1994, and a PhD in Telecommunications from the University of Portsmouth in 1998. He has been a lecturer since 1999 and International Coordinator of the Faculty of Technology since 2007. His research focuses on wireless sensor networks and mesh networks for applications in precision agriculture, environmental monitoring, behavioural economics, security, building control and energy management, etc.

Dr. Yanyan Yang is a Senior Lecturer at University of Portsmouth. She has been involved in teaching in the area of computer network, Web technologies and Internet applications. She is a member of the IET, WES and HEA. Her research interests include intelligent information retrieval, context-aware computing, mobile computing, information recommendation as well as their applications in Internet, social networks, health-informatics and business.

Dr. Munirah L. Kamarudin graduated with Ph.D in Computer Engineering, Universiti Malaysia Perlis, UniMAP (2012). M.Sc Communication Network Management and Planning, University of Portsmouth, UK (2008). B.Eng (Hons) Computer Science and Media Engineering, University of Yamanashi, Japan (2006). Her research interests include network architecture, wireless sensor network, mobile communications and information and communication technology.

Highlight of the Paper:

- A pair of collaborative Gaussian process models (GP) with different kernels is used to handle occlusion.
- A principled technique is proposed to measure the level of occlusion in a frame.
- It proposes a method of choosing the best combination of features depending on their environment.
- The system is comprehensively evaluated using two benchmark datasets, the Mall and University of California, San Diego (UCSD) datasets.