VIDEO AND IMAGE PROCESSING BASED TECHNIQUES FOR PEOPLE DETECTION AND COUNTING IN CROWDED ENVIRONMENTS

by

Zeyad Qasim Habeeb Al-zaydi

This thesis is submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy of the University of Portsmouth.

September 2017

Dedicated to my lovely autistic son Sajjad Al-zaydi

Abstract

Different technologies are used to count people but people counting systems based on computer vision are good choices due to different priorities. These priorities may include accuracy, flexibility, cost and acquiring people distribution information. People counting systems based on computer vision can use closed circuit television (CCTV) cameras that have already become ubiquitous and their uses are increasing. This thesis aims to develop people counting systems that can be incorporated with existing CCTV cameras. People counting is a useful task for safety, security and operational purposes and can be important for improving awareness.

This thesis presents two intelligent people counting systems; pixel-wise optimisation based and features regression based people counting systems. Each system works independently to count people and may be more appropriate for particular scenarios.

The pixel-wise optimisation based people counting system based on two algorithms that estimate the density of each pixel in each frame and use it as a basis for counting people. One algorithm uses scale-invariant feature transform (SIFT) features and clustering to represent pixels of frames (SIFT algorithm) and the other uses features from accelerated segment test (FAST) corner points with SIFT features (SIFT-FAST algorithm). Both algorithms are designed using a novel combination of pixel-wise, motion edges, grid map, background subtraction using Gaussian mixture model (GMM).

The features regression based people counting system is composed of a pair of collaborative Gaussian process regression (GPR) model 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, this system dynamically identifies the best combination of features for people counting.

The University of California (UCSD), Mall and New York Grand Central Station datasets have been used to test and evaluate the proposed systems. These datasets have been chosen because they cover a wide range of variation of characteristics. They cover a variation of frame rate (fps), resolution, colour, location, shadows, loitering, reflections, crowd size and frame type.

By means of comparisons with state of the art methods, the results of the proposed systems outperform the others methods with respect to mean absolute error (MAE), mean squared error (MSE) and the mean deviation error (MDE) metrics. The MAE, MSE and MDE of the proposed systems are 2.83, 13.92 and 0.092, respectively, for the Mall dataset; 1.63, 4.32, and 0.066, respectively, for UCSD dataset; and 4.41, 25.62 and 0.029, respectively, for New York Grand Central dataset. The computational efficiency results of the proposed systems are 20.76 fps, 38.47fps and 19.23 fps for the Mall, UCSD and New York Grand Central datasets, respectively.

Acknowledgement

I would like to express my profound gratitude to the supervisors Dr David Ndzi and Dr Branislav Vuksanovic for their continuous commitment and support to this work. Their guidance, support and encouragement throughout my study were invaluable. Their belief in my abilities and incredible depth of knowledge provided invaluable support for me through the tough times. I am also very grateful for the help, advice and encouragement of my colleagues.

I would like to thank my siblings, particularly to my elder brother Nasir Al-zaydi, who provides a foundation in my life on which I know I can always depend on. I would like to thank my wife for supporting me throughout this study. I would also like to thank our children: Sajjad and Zainab, who they have always been the source of our joy.

Lastly, but most importantly, I would like to express my profound gratitude to my parent, who made me understand that "Have faith and hope. Because night is darkest just before dawn". It is impossible to put into words everything I appreciate about them but I can say that they are my shining stars in my night sky.

Dissemination

Journal Papers:

- Al-Zaydi, Z. Q., Ndzi, D. L., Yang, Y., & Kamarudin, M. L. (2016). An adaptive people counting system with dynamic features selection and occlusion handling. *Journal of Visual Communication and Image Representation*, *39*, 218-225 (impact factor 2.16).

- Al-Zaydi, Z. Q., Ndzi, D. L., Kamarudin, M. L., Zakaria, A., & Shakaff, A. Y. (2016). A robust multimedia surveillance system for people counting. *Multimedia Tools and Applications*, 1-28 (impact factor 1.53).

Conference Proceedings:

- Al-Zaydi, Z. Q., Ndzi, D., & Sanders, D. (2016). Cascade method for image processing based people detection and counting. *Proceedings of 2016 International Conference on Image Processing, Production and Computer Science (ICIPCS'2016),* 30-36.

Table of Contents

Declaration

Whilst registered as a candidate for the above Degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

Word count: 32104

List of Figures

List of Tables

List of Acronyms and Abbreviations

BRISK	Binary Robust Invariant Scalable Keypoints
FAST	Features from accelerated segment test
FPDW	Fastest pedestrian detection in the west
FPPI	False positive per image
GLCM	Gray-level co-occurrence matrix
GPR	Gaussian process regression
HOG	Histogram of oriented gradients
LBP	Local binary pattern
LIDAR	Light detection and ranging
LOI	Line of interest
LR	Linear regression
LSVM	Linear support vector machine
MAE	Mean absolute error
MDE	Mean deviation error
MSE	Mean squared error
NN	Neural network
RBF	Radial basis function
RFID	Radio-frequency identification
ROI	Region of interest
SIFT	Scale-invariant feature transform
SURF	Speeded-up robust features
SVR	Support vector regression
TSE	Taylor series expansion

Chapter 1: Introduction

An introduction to people counting systems is provided in this chapter. It presents the main problems and limitations of the current people counting systems based on conventional cameras. Occlusion handling, features selection, improving performance and the system practicality are some issues that need to be addressed. As a consequence, new accurate and practical computer vision based systems are required to count people which is the focus of this work. The objectives and the author's major contributions of this thesis are presented in this chapter.

1.1 Background and Research Motivation

Information about the number and distribution of people is important for operational, safety and security purposes. Therefore, systems with this kind of functionality can be highly effective tools for establishing ambient awareness (Ryan et al. 2014; Technology 2013; Wang 2014; Ryan 2013; Loy et al. 2013). This information can also 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 (ShopperTrak 2013; Technology 2013; Biodata Ltd 2013). In addition, there are other applications such as crowd management (Longo & Cheng 2015; Technology 2013), transport (DILAX Intelcom 2015; Lumentut et al. 2015; Garcia-Bunster et al. 2012), 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 developing emergency evacuation procedures (Wang 2014).

Different methods are used for people counting, such as tally counters, cameras, differential weight, sensitive carpet, infrared beams, Bluetooth, audio

tones, radio-frequency identification (RFID), wireless fidelity network (Wi-Fi) and wireless sensor network (WSN) based counters (Zhu et al. 2009; Li et al. 2007; Nakatsuka et al. 2008; Ma & Chan 2013; Yuan et al. 2011; Shbib et al. 2013; W. C. Lin et al. 2011; Gandhi et al. 2012; Hou & Pang 2011; Tikkanen 2014; Li et al. 2011; Tu et al. 2013; Adegboye et al. 2012; Xi et al. 2014). Each method has some advantages and disadvantages but people counting systems based on conventional camera are one of the best choices because CCTV cameras are widely used and their uses are increasing. For example, there were an estimated 4.2 million CCTV installed in the United Kingdom in 2004 (Norris et al. 2004) and an estimated up to around 5.9 million in 2015 (Xing et al. 2015).

Different types of cameras can be used with camera based people counting counters e.g. 3D, smart, thermal and conventional camera. 3D cameras extract depth information from frames, which has great potential for improving the performance of people detection and counting systems. Depth information provide size and shape information that can be used to distinguish people from other objects. It also allows occlusions of people by each other or by background objects to be handled more explicitly. However, 3D camera based people counting systems are challenged by their substantial amounts of noise and unreliable data.

Smart cameras support integrated visual processing. The video processing is implemented using advanced software, which are developed by the provided companies. GeoVision smart cameras are one of the best choices that can be used for people counting. These can intelligently process on-board video analytics detecting intruders, loitering, people counting, unattended object, missing object, and tampering of alarms to identify motion, find and trace objects and even produce alarms for unusual activities. Axis smart cameras are also a good option for people counting. They are flexible mini domes that give you a built-in IR illumination and HDTV 1080p video quality. It is suitable for outdoor use and is perfect for indoor environments. Their software is

completely autonomous with all the counting done on the processing units of the cameras to maintain privacy.

Thermal cameras detect and count people by their body heat profile and can therefore count bidirectionally even when a number of people are passing simultaneously. Thermal counters are unaffected by light and can achieve good accuracy. Irisys Gazelle Thermal Counters are one of the best choices because they are ideal for general people traffic counting, measuring live occupancy and a range of security applications, they are widely used to monitor footfall in the transport, banking, retail, security and leisure industries. They also can be linked from a number of entrances to bring data to a central location. They use power over ethernet, keeping wiring simple.

People counting system is one of the most challenging systems in computer vision to implement (Saleh et al. 2015; Wang et al. 2014; Ryan 2013; Hu et al. 2015; Hou & Pang 2011). In comparison with the conventional camera based method, the problem with other technologies is 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 the conventional camera based method. Most of these technologies are also ineffective for acquiring people distribution such as tally, differential weight, sensitive carpet, infrared beams and WSN counters which make them an inappropriate option for various types of applications.

Surveillance systems based on CCTV cameras are usually human operated. Therefore, there is a need for a numerous number of human operators to monitor all installed surveillance cameras (Rao et al. 2015). In addition, the number and distribution of people are easy to count by the human operators in sparse environments while it is very difficult or impossible to count crowd of people. Crowds are a large numbers of people in closed or open areas such as stadiums, railway stations, universities, airports, parks, walkways and public spaces (Saleh et al. 2015). It is impossible for an operator to periodically count hundreds of moving people in a short period. In addition, it is an error-prone

task which makes this information unreliable. The short attention period of human operators and a lack of the sufficient training and knowledge are other limitations of manual counting. Automatic and accurate people counting systems are required to work in those types of environments to provide an accurate and reliable information. Although, a lot of research has been carried on to find an accurate visual based people counting systems, there are still many problems and limitations that need to be addressed. These may include improving accuracy, handling occlusion and working in complex environments efficiently (Loy et al. 2013; Hu et al. 2015; Saleh et al. 2015; Adegboye et al. 2012). The practicality of the people counting systems is another challenge that needs to be improved (Zeyad Q.H. Al-Zaydi et al. 2016). Practicality is measured by the percentage of the training frames minimisation. This minimisation leads to a reduction of the installation time of the people counting systems, which makes them easy to deploy.

This thesis is inspired by the need for further improvements in the conventional cameras based people counting systems. This is particularly important in crowded environments where traditional counting systems struggle or fail to work effectively.

1.2 Research Questions

Research Question 1: How would the accuracy of people counting systems improve if a pair of regression models is used to consider the level of occlusion? What is the best method for combining them? Is homogeneous data better than heterogeneous data for training those regression models?

Research Question 2: What is the effect of using dynamic features selection methods, instead of static methods, to select the best combination of features? If dynamic methods can improve the accuracy, what are the essential types of features should be considered?

Research Question 3: What are the best people counting techniques can be used to preserve the privacy of people? What is the justification for this selection?

1.3 Author's Main Contribution of This Thesis

This thesis presents a number of the main contributions to the conventional cameras based people counting systems. The main contributions of the authors' work are summarised as follows:

1. A new method to measure the level of occlusion has been proposed. A simple but effective equation has been derived which takes into account the density of the crowd size and its sparseness. The level of occlusion information can be used to improve the performance of the counting systems.

2. Instead of using a single GPR model, a pair of GPR models with different kernels has been used to improve the performance. They are designed to count people by considering different levels of occlusion individually.

3. A multi-stage thresholding method has been proposed to determine the best threshold. The threshold with the highest accuracy from all stages of this method has been selected as the best threshold. An analysis of the effects of different choices of thresholds on the performance has been performed and presented.

4. A new method to train the GPR models has been proposed. An ensemble training method has been used that first partitions the heterogeneous data into linear and non-linear homogeneous groups (low-level occluded frames and high-level occluded frames) and then train a GPR model for each homogeneous group.

5. A method for selecting the best combination of low-level features for the existing regression based people counting systems has been

developed. It is based on the characteristics of each individual environment because there is no standard criteria that can be used to choose the appropriate combination of features for each environment. Therefore, a dynamic method is used instead of static ones.

6. A new method has been proposed to analyse the performance of proposed systems in crowded and sparse situations. No partition of the training dataset has been used to ensure that the people counting system is practical and robust because there is no technical definition of the boundary that separates the crowded and sparse frames. In addition, dividing the training dataset into two groups would require two training stages.

7. A new combination of SIFT and FAST features has been proposed as input to the proposed pixel-wise optimisation based people counting system. This combination is better than using one of them for improving the performance of counting.

8. Moving edge pixels have been proposed instead of foreground pixels with the proposed pixel-wise optimisation based people counting system. It reduces the number of SIFT descriptors required and reduces the time required to cluster them which in turn reduces the processing time.

9. A new combination of pixel-wise technique and grid map has been proposed. It is used to improve the Cluster classification in the proposed pixel-wise optimisation based people counting system. As a consequence, different densities are assigned to the same clusters identification depending on their location in the frame.

1.4 Thesis Organisation

In this chapter, the background of people counting systems has been discussed. The author described the importance of people counting for

different applications. The main objectives and contributions of the work have also been presented. The rest of the thesis is organised as follows:

Chapter 2 reviews the previous research and development of people counting systems. The related work done by other researchers is reviewed. In addition, background subtraction algorithms and perspective normalisation method are also presented. The main gaps in the current people counting systems are also presented.

Chapter 3 presents the methodology and experimental design of the proposed counting systems. Two independent systems are proposed; low-level features regression based and pixel-wise optimisation based systems.

Chapter 4 describes the experimental setup and evaluation metrics used in this thesis. It presents the outcomes from experiments using two datasets. Comparison of the proposed systems and some existing methods are presented. An extensive discussion of the outcome of the experiments is presented.

In chapter 5, the efficiency of the proposed systems is tested and validated using a very crowded and challenging dataset collected from the New York Grand Central Station. This chapter describes the experimental dataset and presents the results of the experiments.

Chapter 6 concludes the main findings of this thesis. Summay of the contribution of this study is also reviewed and presented. Suggestions for the future work are also given in this chapter.

Chapter 2: Literature Review and Related Works

The development of people counting systems is described in this chapter. Related works are reviewed and some current methods are discussed. The review begins with a general overview of visual and non-visual people counting methods. The main categories of people counting based on conventional camera are introduced which include; people detection algorithms, features trajectories clustering algorithms, features regression algorithms and pixel-wise algorithms. The background subtraction algorithm and perspective normalisation are also explained. Finally, description of gaps in research literature are presented

2.1 People Counting Systems

The people counting task has been studied extensively using non-visual methods. In recent years, many researchers have turned to visual technologies to count people automatically using cameras. Automated visual people counting is an active area of research due to a large number of unsolved limitations and problems. For further development of people counting systems and to provide more accurate performance estimation, there is an increasing need for a good understanding of their key characteristics, problems and limitations. Although current research into people counting in sparse environments is well established, there are still many challenges and limitations in crowded environments.

This thesis focuses on visual people counting systems based on conventional cameras. People counting systems based on conventional camera often involve features extraction stage which is followed by classification, regression, optimisation or trajectories clustering stage. It is important to choose appropriate features that correspond accurately to the number of people. A combination of features is typically used instead of a single feature

type to achieve a higher performance because it can help to minimise the non-linearity that arise from segmentation errors, occlusion and pedestrian configuration (Chan et al. 2008).

This thesis uses existing methods from the field of image processing and computer vision, e.g. background subtraction, regression, optimisation and features extraction algorithms. However, the focus of this thesis is not to improve these algorithms. Instead, this thesis focuses on the bigger picture of people counting systems by a new employment of these algorithms, combining between them and improving the design and performance of people counting systems.

2.1.1 Non-visual people counting systems

Non-visual based people counting systems use different methods which may include tally counters, differential weight, sensitive carpet, infrared beams, Bluetooth, audio tones, RFID, Wi-Fi and WSN counters (Zhu et al. 2009; Li et al. 2007; Nakatsuka et al. 2008; Ma & Chan 2013; Yuan et al. 2011; Shbib et al. 2013; W. C. Lin et al. 2011; Gandhi et al. 2012; Hou & Pang 2011; Tikkanen 2014; Li et al. 2011; Tu et al. 2013; Adegboye et al. 2012; Xi et al. 2014).

Tally counters provide easy and accurate counts (Lev et al. 2008). They are mechanical, or electronic devices that incrementally count people. They can be used to manually count the number of people walking in and out of a venue. The main disadvantages of tally counters are inflexibility and in some instances inability to detect people distribution. Furthermore, they are not suitable for detailed analysis and can be a bottleneck in crowded situations.

Differential weight counters estimate the number of people by evaluating the weight variations using load cells (Vasco Dantas dos Reis 2014). They may be useful for carriage environments such as trains. buses or lifts. These counters are only suitable for a few types of environments. They also assume a fixed weight for each person. That is not always reliable due to the significant

difference in weight between children and adults or between fat and thin people.

Sensitive carpet counters are an accurate option but involve severe modifications of the environment and they are prone to wear (Vasco Dantas dos Reis 2014). They use sensitive electronic sensors to count the steps of people. They are particularly useful for indoor environments. Furthermore, people stand with two feet but walking with one or both feet which lead to error in counting.

Infrared beams counters can be also used to count the number of people (Li et al. 2007; Zhu et al. 2009). One or more horizontal infrared beams are usually used across an entrance. If the beam is broken, the counter counts a 'tick'. Multiple beams are used by many researchers to find the direction of people or to improve the accuracy. Infrared counters are still widely used due to their low cost and simplicity of installation. On the other hand, simple infrared beams counters are non-directional and their main disadvantages are that they cannot discern people walking side-by-side and they can be blocked by people standing in front of the beam. In addition, they are not suitable to work in open areas where no particular entrances and exits exist.

Bluetooth, audio tones, RFID and Wi-Fi are used as device-based methods to count or localise people (Kannan et al. 2012; Lionel et al. 2003; Weppner & Lukowicz 2011; Xi et al. 2014). Device-based methods require people to carry mobile devices. They also require people to enable the Bluetooth units, use speakers or to use extra hardware such as RFID tags. The main disadvantage of this technique is that some people carry more than one mobile device and not everyone carries a device which affects the accuracy significantly.

WSN and Wi-Fi are also used as device-free techniques to count the number of people (Domenico et al. 2016; Yuan et al. 2013; Xi et al. 2014; Yuan et al. 2011; Nakatsuka et al. 2008). Device-free methods do not require people to carry certain devices to be counted. They usually depend on the variation of the wireless signal to find the relationship between it and the number of people.

These techniques are easily affected by environmental dynamics, noise, fading and other factors that may influence signal. In addition, their application is mainly limited to indoor environments.

2.1.2 Visual people counting systems

Different kinds of cameras can be used for people counting. Visual based people counting systems can be classified into four categories; 3D, smart, thermal and conventional cameras.

The 3D camera counter is a technology used in people counting which can help to identify the depth information of the people (Tikkanen 2014), (Del Pizzo et al. 2015).The release of Microsoft Kinect (Microsoft 2011) in 2010 increased the interest in the field of 3D camera counter because Microsoft Kinect provides good quality depth images at a lower price compared to previous technologies. However, the information from Microsoft Kinect can still contain a lot of noise (Zhang et al. 2012). In addition, the practical sensors range of Microsoft Kinect is 3.5 metres which makes it useless to count people in the large areas (Han et al. 2013). It also cannot sense objects that are illuminated by direct sunlight so it does not work in outdoor environments (Tikkanen 2014). Image depth can also be obtained using time-of-flight (TOF), light detection and ranging (LIDARs) and stereo cameras methods (Gandhi et al. 2012). However, stereo cameras are affected by changing illumination and cannot operate in the dark. Another problem emerges when monitoring a large area with a similar colour and little edges, because it may be difficult to find features (Sensors et al. 2008). In addition, developing a stereo based depth sensing system is more complex and would, therefore, require a significant amount of knowledge and computational power (Tikkanen 2014). On the other hand, the luminance sensitivity of TOF cameras is poor and their depth range is limited (Gandhi et al. 2012) and the size of LIDARs cameras are large (Tikkanen 2014). In addition, they are expensive and their accuracies are lower than Microsoft Kinect (Tikkanen 2014; Han et al. 2013).

Smart cameras (intelligent cameras) can also be used for people counting. They refer to cameras that have in-built processing capabilities so there is no need for an external processing unit such as computers (Valera & Velastion 2005). The main disadvantage of these cameras is the cost because they are expensive. In addition, this is not a very convenient option because most of the current surveillance systems use conventional cameras. To use this option, the current CCTV cameras would have to be replaced which make this option not very practical due to the scalability.

Thermal cameras are also used to count people (W. C. Lin et al. 2011; Tikkanen 2014). They are usually positioned at an entrance or a gate and they detect people's body heat. Accuracy can be affected if the ambient temperature within the counting area is above a certain threshold. Heat sources and external weather conditions may affect the accuracy of detecting the emitted heat from people. In addition, they have narrow fields and may not cover wide spaces. Thermal counters have the advantage that they are not affected by changing illumination and do not need background subtraction algorithm, therefore have a shorter processing time (Tikkanen 2014).

For counters based on conventional camera, different algorithms have been introduced to increase the accuracy of counting (Ma & Chan 2013; Shbib et al. 2013; Hou & Pang 2011; Li et al. 2011; Tu et al. 2013; Adegboye et al. 2012). Most of them are proposed to work in both indoor and outdoor environments whereas some algorithms are proposed to only work in indoor environments (Luo et al. 2016; Cetinkaya & Akcay 2015). Conventional camera based people counters can be classified depending on the area of view into the line of interest (LOI) and region of interest (ROI) (Li et al. 2011). LOI algorithms involve counting the number of people who cross a real or virtual line during a certain period of time (Ma & Chan 2013) whereas ROI algorithms involve counting the number of people in a specific region during a certain period of time (Tu et al. 2013). These counters can also be classified into four categories: people detection based, features trajectories based, features regression based and pixel-wise based algorithms (Zeyad Q.H. Al-Zaydi et al.

2016). People detection based algorithms involve detecting all people in the frame-to-frame analysis individually and then counting them (Hou & Pang 2011). These algorithms lack scalability when working in crowded environments. Features trajectories based algorithms count people by tracking and identifying their features over time (Yoshinaga et al. 2010). The feature trajectories of each person are then clustered so the number of clusters represents the number of people. Features regression based algorithms involve extracting useful features from the frame-to-frame analysis. These features are then used to count people without detecting each person individually (Hou & Pang 2011). These algorithms preserve privacy and are more robust (Adegboye et al. 2012). In pixel-wise optimisation based algorithms, the density of each pixel in a frame is determined and then integrated (Lempitsky & Zisserman 2010). Optimisation algorithms are used with these algorithms. These categories will be presented and discussed in more details in the following sections.

2.2 People Detection Based Algorithms

People detection algorithms involve detecting all people in a frame-to-frame analysis individually and then counting them (Adegboye et al. 2012). Entire or parts of person's body are used in the detection process such as the head, face or head-shoulder (Adegboye et al. 2012). The main advantages of these algorithms are that they can find the number of people and their location simultaneously. Therefore they are important in people tracking applications (Hou & Pang 2011). In addition, the tracking information may be used to improve the accuracy of detection. The accuracy of these algorithms are significantly affected by occlusion, varying lighting and a long processing time (Tu et al. 2013). They produce more accurate results when the crowd density is low whereas the accuracy decreases significantly in high crowd density scenarios (Hou & Pang 2011). In addition, they require high-resolution videos to achieve good accuracies (Hou & Pang 2011). In the last 10 years,

researchers have introduced different methods to improve the performance of people detection algorithms (Tikkanen 2013). Dalal and Triggs introduced histogram of oriented gradients (HOG) technique thereby creating a basis for the development of fast appearance-based detection (Dalal & Triggs 2005). Many improvements of the HOG technique have been proposed. One of the most promising variants is the fastest pedestrian detection in the west (FPDW) which has significantly increased the speed of detection (Dollar et al. 2010).

Pedestrian detection is constrained to horizontal, vertical or tilted downwards camera angles. In people counting, horizontal camera angles can be used, but a vertical is often preferred to avoid occlusions (Sidla et al. 2006). The majority of commercial people counting products use cameras that are placed on the ceiling pointing vertically down to get the best view. However, this is not the optimal set-up if the detection area needs to be maximised. On the other hand, it is important to develop systems that can work with the CCTV cameras with most of them tilted downwards.

People detection algorithms can be classified into five categories: **full body detection** (Tuzel et al. 2008; Leibe et al. 2005; Dalal & Triggs 2005); **part body detection** (Felzenszwalb et al. 2010; Lin et al. 2001; Wu & Nevatia 2007); **shape matching detection**, where ellipse or Bernoulli shapes are used to identify the number of people in each blob (Li et al. 2011), (Ge & Collins 2009); **multi-camera detection**, which is used to avoid occlusion (Ma et al. 2012); and **density-aware detection**, which is used to reduce the false positive per image (FPPI) in low crowd density locations and decreases the miss rate in high crowd density locations in the frame (Rodriguez et al. 2011).

2.2.1 Full body detection algorithms

Those are direct approaches to counting the number of people in a scene through detection (Al-zaydi et al. 2016). The algorithms are trained using the full body appearance of a set of people (Tuzel et al. 2008; Leibe et al. 2005; Dalal & Triggs 2005). They suffer from large pose variations and partial

occlusion as the number of people increases (Yuk et al. 2006). Different features are used to represent the full body appearance such as Haar-like features (Viola & Jones 2004), HOG features (Dalal & Triggs 2005), Local binary patterns (LBP) (Mu et al. 2008), LBP-HOG combination (Zeng & Ma 2010) shape context (Mori et al. 2005), edgelets (Wu & Nevatia 2007) and Shapelets (Sabzmeydani & Mori 2007). Different linear and nonlinear classifiers are also used to find the relationship between the features and the number of people such as support vector machine (SVM) and adaboost (Dalal & Triggs 2005; Viola et al. 2005). The accuracy of full body detectors is acceptable in sparse environments, whereas the accuracy significantly decreases in crowded environments due to full and partial occlusion.

2.2.2 Part body detection algorithms

Many studies have been carried out to mitigate the partial occlusion by detecting only part of the body such as heads, faces, eyes and head-shoulders (C. Gao et al. 2016; Felzenszwalb et al. 2010; Lin et al. 2001; Wu & Nevatia 2007). The shape of people's heads is different due to hair styles and head coverings, hence head based human detection is not robust enough for counting people (Yuk et al. 2006). On the other hand, a head-shoulder region occupies a larger region in a human body than a head alone and they are more likely to be detected even in highly occluded cases (Yuk et al. 2006). Faces and eyes are rarely used to count the number of people because many people do not look at the cameras when passing. Faces and eyes are also easy occluded. Finally, tracking can also be used to increase detection and accuracy.

2.2.3 Shape matching detection algorithms

Ellipses are used by some researchers to count people (Li et al. 2011). In this method, the background subtraction method is applied to segment the foreground blobs (Kim et al. 2005; Ilyas et al. 2009) and ellipse detection is

applied to identify the number of people in each blob. Other shapes such as Bernoulli shapes have been used by other researchers to count the number of people (Ge & Collins 2009). The accuracy of shape matching detectors is acceptable in sparsely occupied environments but it decreases significantly in crowded environments.

2.2.4 Multi-camera detection algorithms

Much research has focused on counting the number of people using a single camera, which can fail in crowded environments where heavy occlusions occur (Ryan et al. 2014; Ma et al. 2012; Mehmood 2016). Some researchers have used multiple cameras to count people to avoid occlusion (Ma et al. 2012). The cost of hardware and the incorporation of the multi-camera set-up is the main disadvantages of this approach. In addition, it is required to fuse information from all cameras which require a consistency across all the views of people (Mehmood 2016). The fusion is used to determine the absence or the presence of each person.

2.2.5 Density-aware detection algorithms

This approach combines people detection algorithms and crowd density estimation (Rodriguez et al. 2011). Full body, head and head-shoulder detection algorithms can be improved and the accuracy can be increased by using a density-aware information (Rodriguez et al. 2011). The aim of this approach is to reduce the FPPI in low crowd density locations in the frame, which happens when it falsely detects the presence of people when there is actually nobody. In addition, this approach decreases the miss rate in high crowd density locations in the frame.

2.3 Features Trajectories Clustering Based Algorithms

These algorithms track useful features in a frame-to-frame analysis over time and the trajectories of these features are clustered into unique tracks per person using spatial and temporal consistency heuristics or other factors (Brostow & Cipolla 2006; Rabaud & Belongie 2006; Topkaya et al. 2014; Cheriyadat et al. 2008). The number of people is found by counting the number of clusters which each cluster represent one person (Merad et al. 2010). Different methods can be used to measure the similarities between trajectories. The Kanade–Lucas–Tomasi (KLT) feature tracker can be used to track the trajectories (Rabaud & Belongie 2006). KLT employs spatial intensity information to find the new location of the same person based on the best match. The Bayesian clustering algorithm is proposed by Brostow et al. (Brostow & Cipolla 2006) to track simple frame features such as corners and Tomasi-Kanade features and then cluster them.

Clustering based algorithms often avoid supervised learning or model appearance features as in the people detection based algorithms. These methods are efficient when the size of people are large in a frame due to the presence of enough frame pixels depicting the people to track them effectively whereas the performance decreases when the camera is far and the size of people is small (Mukherjee 2014). However, their accuracies significantly decrease in crowded scenarios with complicated background and frequent inter-object occlusion. A complex trajectory management due to occlusions or assessing similarities between the trajectories of different lengths is another limitation of these algorithms (Shbib et al. 2013). In addition, errors in the number of people due to the cohesiveness of features that belong to different people also affect their accuracy (Shbib et al. 2013). High video frame rate is required for these algorithms to work efficiently because motion information must be extracted reliably (Chen et al. 2012). Features trajectory clustering based algorithms can be used to estimate the number of people who passed

within a specific time, but real-time processing is difficult to achieve due to the long processing time (Yoshinaga et al. 2010; Vasco Dantas dos Reis 2014).

Features trajectories clustering based algorithms may be useful for some particular environments such as an entrance of a station or a corridor. They may suffer in outdoor environments where people move at variable speeds and in different directions.

2.4 Features Regression Based Algorithms

Regression based algorithms usually consist of three steps, starting with a background subtraction that is used in the frame-to-frame basis to detect the foreground information (Zeyad Q.H. Al-Zaydi et al. 2016). Low-level features are then extracted from the foreground such as edge features (Chen et al. 2012; Chow et al. 1999; Ryan et al. 2009; Cho et al. 1999; Cho, S. Y., & Chow 1999), segment features (Chan et al. 2008; Zhang et al. 2011; Hou & Pang 2011; Chow et al. 1999; Chan & Vasconcelos 2012; Chan & Vasconcelos 2009; Chan et al. 2009; Cho et al. 1999; Cho, S. Y., & Chow 1999), texture features [45], [51], [80], [81] and keypoints (Hashemzadeh & Farajzadeh 2016; Ma et al. 2004). A regression model is then trained using these features to find the relationship between the number of people and the extracted features which it is then used to estimate the number of people (Topkaya et al. 2014). Various types of regression models have been used e.g. support vector machine tree (Conte et al. 2010a; Xiaohua et al. 2006), linear (Shimosaka et al. 2011; Davies et al. 1995; Ma et al. 2004), neural networks (Chow et al. 1999; Ryan et al. 2009; Cho et al. 1999; Cho, S. Y., & Chow 1999; Zhang et al. 2016; Fu et al. 2015) and Gaussian process algorithms (Chan et al. 2008; Merad et al. 2010; Chan & Vasconcelos 2012; Chan & Vasconcelos 2009; Chan et al. 2009). A significant amount of research has been carried out to improve these algorithms by varying the number of features. Some other researchers have tried to improve them by using more than one regression model and then choosing the best fitting features (Fradi & J. L. Dugelay 2012).

Computer vision systems that involve humans raise important privacy concerns. Privacy of the people should be preserved by any people counting system based on CCTV cameras. Individuals' privacy rights should not be infringed by CCTV cameras based people counting systems because they may be used in public environments (Chan et al. 2008). Features regression based algorithms are fully privacy preserving method because they do not based on people detection or tracking (Vasco Dantas dos Reis 2014).

Some new contributions have also been presented to improve the accuracy, handle occlusions and adapt to new environments. Recent technique in people counting has been tested using mid-level and high-level crowd static pictures (Hu et al. 2016). A deep learning approach that uses convolutional neural networks to predict the number of people has been proposed in that technique. Three datasets have been used to test and validate this system achieving better results than some current methods.

Hafeezallah et al. (Hafeezallah & Abu-Bakar 2016) have proposed a new method to extract the features from frames. In this method, frames are converted using a curvelet transform in the first stage. The differences between every two sequential frames are then calculated at every subband. Statistical features of all subbands have been extracted and used to train the people counting system using a neural network.

A random projection forest, as a regression model, has also been proposed by other researchers to increase the maximum number of features that is used for training (Xu & Qiu 2016). The authors have noticed from the current research that a richer set of frame features can improve the performance of many computer vision applications including people counting. A small number of features can be handled by traditional regression models which can negatively affect the performances of people counting systems.

Multi-cameras knowledge transfer technique has been used by Nick et al. (Tang et al. 2015) to provide different views of the crowd which are used to minimise occlusion and improve performance. Calibration-based methods

have been used to achieve the correspondence between multiple cameras thereby enabling multiple cameras to share visual knowledge. In addition, a pair of collaborative regression models has been used. The first regression model has been used to count people based on extracted features from the first camera, while the second one has been used to compensate the residual from the conflicts between multi-cameras.

Finally, a support vector regression (SVR) model is used to train the people counting system (L. Gao et al. 2016). This method takes into account the temporal domain of a series of frames to build the network flow constraints of people. This network is then used as an input to a linear or quadratic programming model to improve the accuracy of people counting. The authors conclude that quadratic and linear programming method is not restricted to use with the SVR model only but it can also be used with any regression model. Quadratic programming model performs better than linear programming model in the most experiments of this system.

The accuracy of these algorithms are higher than feature trajectory clustering and detection based algorithm in crowded scenarios, and the computational time is shorter (Chen et al. 2012; Fradi & J. L. Dugelay 2012; Tu et al. 2013). Therefore, a comparison between the results of some of these algorithms and the proposed systems will be used in the chapter 4 to prove the efficiency of the proposed systems. The algorithms that used the same datasets, training set and testing set will only be selected in this comparison which makes it more reasonable.

There are three main categories of regression based algorithms; local, histograms (intermediate) and holistic algorithms (Ryan et al. 2015). A brief description of each category with their advantages and disadvantages will be presented in the following subsections.

2.4.1 Holistic algorithms

Global frame features and a single regression model are used by these algorithms over the whole frame space (Chan et al. 2008; Chow et al. 1999; Chan & Vasconcelos 2012; Chan & Vasconcelos 2009; Chan et al. 2009; Cho et al. 1999; Cho, S. Y., & Chow 1999). The features used by these algorithms may include textures, foreground, keypoints and edge features. The regression models that are used to find the relationship between these features and the crowd density may include linear regression, GPR and artificial neural network. These algorithms do not suffer scalability limitation in large environments because there is no need for multiple regression models. Using a single regression model over the whole frame is the main limitation of these algorithms because of the high variation in crowd behaviour, density and distribution in different regions of the frame (Ryan et al. 2015).

2.4.2 Histograms algorithms

Histograms algorithms use histogram features on a holistic level such as blob size histogram, edge orientation histogram and HOG (Xu et al. 2016; Kong et al. 2006a; Kong et al. 2006b). These features are usually used to count people using one global regression model. These algorithms use histogram bin of edge direction and blobs size to distinguish people and to avoid noise, respectively. The smallest blob size histogram bins are usually affected by noise whereas people contribute the larger angle bins. In regard to edge orientation histogram, eight angle bins between 0˚ and 180˚ are often used to distinguish people from other objects because the edges of people are usually in the vertical direction. Histograms algorithms ignore the high variation in crowd behaviour, density and distribution in different regions of the frame (Ryan et al. 2015). In addition, it is difficult to choose the appropriate bin width of the blob size histogram because it depends on the frame resolution and how

far is the camera position (Ryan 2013). The background subtraction errors also negatively affect the magnitudes of histogram bins.

2.4.3 Local algorithms

Local algorithms count the number of people by dividing the image into regions and using separate regression models for each region to find the total number of people. The regions can be cells having regular or irregular sizes (Chen et al. 2012) or can be foreground blobs. The total number of people is counted by summing the blob-level counts (Çelik et al. 2006; Kilambi et al. 2008; Ryan et al. 2009; Conte et al. 2010b; Jeong, C. Y., Choi, S., & Han 2013). The main limitation of the cell approach is the difficulty in annotating people in each region for the training stage. Bodies of many people will exist in two or more regions if the frame is divided into cells. In the crowded environments, the negative effect of this problem is higher than the sparse one. Although, using foreground blobs can solve this problem, the background subtraction errors can produce the following problems;

1. One person exists in multiple foreground blobs which lead to the difficulty of annotating people in each blob.
2. Noise can lead to a very large number of regression models required due to each foreground blob requiring a regression model.

Local algorithms also suffer a scalability limitation in the large environments because they required training an individual regression model for each separated region (Chen et al. 2012). In addition, the lack of the shared information between regions can decrease the performance of people counting (Chen et al. 2012).

2.5 Pixel-wise Optimisation Based Algorithms

Some researchers use the pixel-wise technique to find the number of people (Lempitsky & Zisserman 2010). In this technique, the density of each pixel is found and then integrated over an image region to find the number of people within that region (Lempitsky & Zisserman 2010). Instead of using regression algorithms, optimisation technique is used to train pixel-wise optimisation based algorithms (Zeyad Q. H. Al-Zaydi et al. 2016). It is very difficult to annotate people based on their actual shape because the annotation must include all people pixels. People are annotated using a dot annotation with a Gaussian kernel and the pixels' summation of each Gaussian kernel must equal to one. A Maximum excess over subarrays distance ($\mathrm{D}_{\mathrm{MESA}}$) is used as a cost function to measure the difference between the actual and the estimated number of people (Zeyad Q. H. Al-Zaydi et al. 2016).

This approach can be used to improve people detection algorithms by combining full body, head, head-shoulder detection based algorithms with the density-aware techniques (Pixel-wise techniques) (Rodriguez et al. 2011). The aim of this combination is to reduce the FPPI in low crowd density locations in the frames which happen when people detectors inaccurately detect the presence of people when there is actually nobody. In addition, this approach decreases the miss rate in high crowd density locations in the frames.

Pixel-wise optimisation based algorithms can be trained using a lower number of frames in comparison to the features regression based algorithms (Lempitsky & Zisserman 2010). Therefore, the installation time of the system can be reduced by more than 25% in comparison to the features regression based algorithms which also lead to low set-up cost. This is because the number of people in the training frames needs to be annotated manually for the training stage which is a very slow operation (may take days for the highly crowded and low-resolution frames). This can negatively affect the accuracy of the training because counting people manually is an error-prone task. The

accuracy of people counting systems is significantly affected by the errors of the training stage.

2.6 Background Subtraction Algorithm

Background subtraction is a process of extracting foreground information in the frame-to-frame basis. Background subtraction algorithms usually consist of three steps; background initialization, foreground detection and background maintenance (Sobral & Vacavant 2014). In the background initialization, various techniques such as statistical, fuzzy and neuro-inspired techniques are used to build a background model. In foreground detection, a comparison is implemented between the current frame and the background model. Updating a background model according to changes in the environment is processed in the background maintenance step. Background subtraction methods can be classified into recursive and non-recursive algorithms (Adegboye et al. 2012). In non-recursive algorithms, the background model is considered to be static and does not update, whereas in the recursive algorithm, it is a dynamic and changes depending on the change of environment (Adegboye et al. 2012). Figure 2.1 shows the general block diagram of background subtraction algorithms.

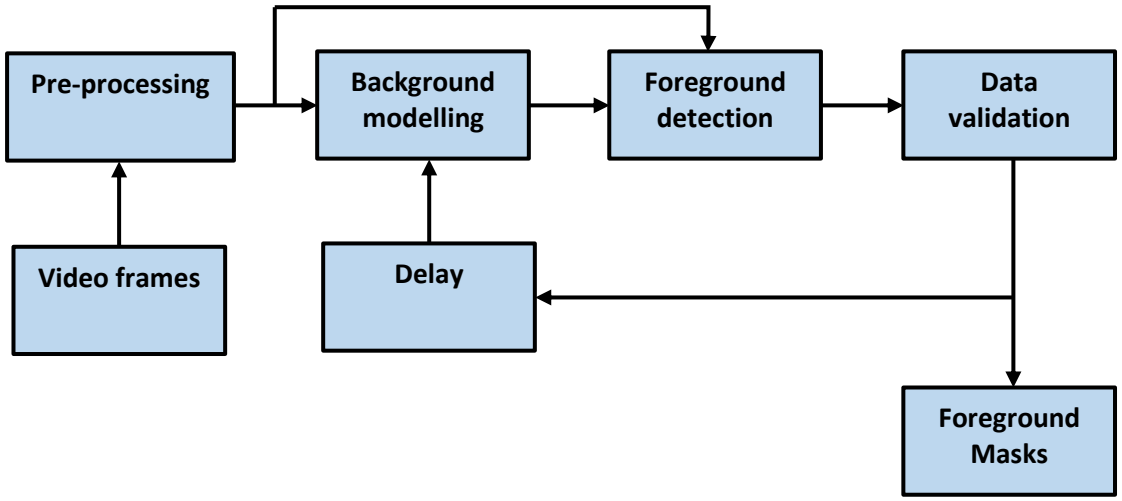


Figure 2. 1: General block diagram of background subtraction algorithms (Alawi et al. 2013).

GMM has been used in the proposed people counting systems. Each pixel in a background model is formed using a mixture of Gaussian distributions (normally from three to five distributions) rather than one Gaussian distribution (Stauffer & Grimson 1999; Shbib et al. 2014; Adegboye 2013).

$$p(x_t) = \sum_{i=1}^{K} w_{i,t} * f\left(x_t|\mu_{i,t}, \Sigma_{i,t}\right) \quad (2.1)$$

Where K is the number of Gaussian distributions and $w_{i,t}$ is the weight of the i^{th} distribution at time t. Each Gaussian distribution can be found using the probability density function;

$$f(x_t|\,\mu_t, \Sigma_t) = \frac{1}{\sqrt{(2\pi)^n|\Sigma_t|}} \exp\left(-\frac{1}{2}(x_t - \mu_t)^T {\Sigma_t}^{-1}(x_t - \mu_t)\right) \quad (2.2)$$

Where μ_t is the mean and Σ_t is the covariance matrix. If a pixel matches with one of the Gaussian distribution, then the background model is updated using (Stauffer & Grimson 1999; Nurhadiyatna et al. 2013; Benezeth et al. 2010);

$$w_{i,t} = \alpha + (1 - \alpha)w_{i,t-1} \quad (2.3)$$

$$\mu_{i,t} = \rho x_t + (1 - \rho)\mu_{i,t-1} \quad (2.4)$$

$$\Sigma_{i,t} = \rho(x_t - \mu_{i,t})^T(x_t - \mu_{i,t}) + (1 - \rho)\Sigma_{i,t-1} \quad (2.5)$$

$$\rho = \alpha\, f\left(x_t|\mu_{i,t}, \Sigma_{i,t}\right) \quad (2.6)$$

Where;

- α is the learning rate that controls the speed of the learning,
- x_t is the current pixels values.

In the case that all the Gaussian distribution do not match a pixel, then only the weight is updated using (Nurhadiyatna et al. 2013);

$$w_{i,t} = (1 - \alpha)w_{i,t-1} \quad (2.7)$$

2.7 Perspective Normalisation

The size of a person changes depending on the distance of the person from the camera. As a consequence, features extracted from the person at different depths in a frame will have significantly different values. To solve this problem, a density map is usually created to assign different weights to the pixels in a frame. These weights are applied to the features extracted from the frames. Two main techniques are used to generate the density map, as introduced by Ma (Ma et al. 2004) and Chan (Chan et al. 2008).

Figure 2.2 shows the method of the density map that presented by Ma. The author presented a method to weight each pixel based on the area it located on the ground plane. Two parallel lines with their four coordinates in a frame are used to find the weight of each pixel. C1, C2, C3 and C4 represente the coordinates of the two parallel lines in a frame and Cv represents the coordinate of a vanishing point (horizon). The weights of the pixels at each horizontal row of the frame are equal because the camera is assumed to be placed horizontally (Ryan 2013). The weight of the pixels at the reference line is assigned a value equal to one. The width between the two parallel lines at the reference line is Δx_r and the distance between the reference line and the vanishing point is Δy_r. The width between the two parallel lines at the line of interest is Δx_l and the distance between it and the vanishing point is Δy_l.

The width weight of the pixels at the line of interest is found by dividing the widths between the parallel lines at the reference line and the line of interest;

$$weight_{width} = \frac{\Delta x_r}{\Delta x_l} \quad (2.8)$$

Therefore, the weight of the pixels at the lines of interest is greater than one because Δx_r is always wider than Δx_l. By similar triangle formula,

$$weight_{width} = \frac{\Delta y_r}{\Delta y_l} = \frac{y_r - y_v}{y_l - y_v} \tag{2.9}$$

Equation (3.9) represents the perspective compensation of horizontal dimension. The full compensation (horizontal and vertical dimensions) is found by;

$$weight_{full} = \left(\frac{y_r - y_v}{y_l - y_v}\right)^2 \tag{2.10}$$

Equation (3.10) is applied for each row in a frame. The vanishing point is determined by equalling the line equation of the parallel lines;

$$y_{line1} = y_{line2} \tag{2.11}$$

Where

$$y_{line1,2} = m_{line1,2}x + c_{line1,2} \tag{2.12}$$

$$m_{line1,2} = \frac{y_2 - y_1}{x_2 - x_1} \tag{2.13}$$

$$c_{line1,2} = y_2 - m_{line1,2}x_2 \tag{2.14}$$

Therefore, at the vanishing point;

$$m_{line1}x_v + c_{line1} = m_{line2}x_v + c_{line2} \tag{2.15}$$

$$x_v = \frac{c_{line2} - c_{line1}}{m_{line1} - m_{line2}} \tag{2.16}$$

$$y_v = m_{line1}x_v - c_{line1} \tag{2.17}$$

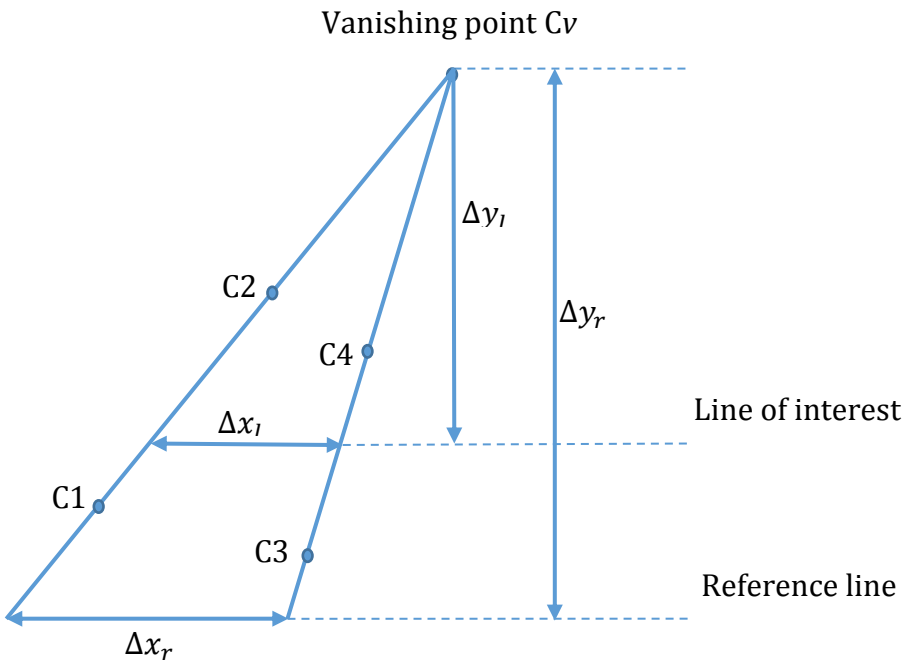


Figure 2. 2: The density map method that derived by Ma.

A similar linearly interpolating is used by Chan (Chan et al. 2008) to create the density map. Figure 2.3 illustrates the density map method that is introduced by Chan. The reference line at the line (ab) and the line of interest at the line (cd), the widths of them are w_r and w_l, respectively. The weight of pixels at the reference line is assigned value equal to one and the width weight of objects at the line of interest is determined by;

$$weight_{width} = \frac{w_r}{w_l} \tag{2.18}$$

The objects at the reference line appear wider than them at the lines of interest by factor $weight_{width}$. Similarly, the height weight of objects at the lines of interest can be found by;

$$weight_{height} = \frac{h_r}{h_l} \tag{2.19}$$

To rescale pixels in the horizontal and vertical dimensions, Equation (3.20) is used;

$$weight_{full} = \frac{h_r w_r}{h_l w_l} \tag{2.20}$$

Chan' method has been used in the proposed systems because it is accurate and its equation depends on the width of the walkway and the height of people that are easy to calculate using the user annotated reference values (Ryan 2013). The height of people is determined using a linear equation;

$$h_l = m_h y_l + c_h \tag{2.21}$$

Where

$$m_h = \frac{h_2 - h_1}{y_2 - y_1} \tag{2.22}$$

$$c_h = h_1 - m_h y_1 \tag{2.23}$$

The width of walkway is determined using a linear equation;

$$w_l = m_w y_l + c_w \tag{2.24}$$

Where m_w and c_w are found by;

$$m_w = \frac{w_2 - w_1}{y_2 - y_1} \tag{2.25}$$

$$c_w = w_1 - m_w y_1 \tag{2.26}$$

Figure 2. 3: The density map method that derived by Chan.

2.8 Description of Gaps in Research Literature

Different visual and non-visual technologies can be used to count people. This thesis focuses on the conventional camera based people counting technology because it is one of the best choices depending on different priorities e.g. accuracy, flexibility, cost and acquiring people distribution information. The literature review shows that there is a lack of knowledge of how to handle occlusion. Occlusion may slightly affect people counting in sparse environments but its effect increases considerably in crowded environments. Although a vertical angle camera can be used to minimise the occlusion problems (Sidla et al. 2006), it is important to develop people counting systems that can be incorporated with the CCTV cameras. Many researchers have shown that there is a correlation between the density of the crowd and the level of occlusion but this is not always correct in all scenarios due to the effect of sparseness. As a consequence, there is a need to develop a method to measure the level of occlusion thereby improving the performance of counting.

The second gap is the selection problem of the best combination of features to be used as an input to the features regression based algorithms. Although, different static combinations of features have been proposed by the researchers as described in the literature review, there is no one combination of features that works with all environments efficiently because the best accuracy of people counting systems for each environment has been achieved using a different combination of features. Adaptive method to select the features is required which may depend on the characteristics of each environment.

Another gap is the need to improve the performance of counting in complicated environments with severe light changing, shadows and reflections. In addition, developing systems that be able to work with a low-resolution camera (coloured or grey) efficiently because most of the CCTV cameras work in the low-resolution setting.

Practicality is another gap that needs to be resolved. A practical people counting system, with at least a comparable accuracy to the state-of-the-art methods, is required in particular scenarios such as large distribution monitoring systems. Practicality and accuracy are used to evaluate the performance of people counting systems. Low practical people counting systems are slow to deploy. In addition, those systems use a large number of training frames which can negatively affect the accuracy of the training because manual annotation of people is an error-prone task. In conclusion, the practicality of the people counting systems needs to be improved.

2.9 Chapter Summary

This chapter has reviewed recent related studies and developments in the people counting systems. Some current works have been discussed. From the literature review, the author has found that more detailed studies and solutions of occlusion handling problem, features selection, practicality and improving

the system performance are needed. This chapter has also presented the visual and non-visual people counting systems and the four main categories of visual systems have been introduced and discussed. This forms the main objectives of this study. Two people counting systems are proposed and their experiential results are presented in the following chapters.

Chapter 3: People Counting Systems

The experimental benchmark datasets used with the proposed people counting systems are described in this chapter. Two datasets are used to evaluate the proposed systems. Two people counting systems are proposed and a detailed description of each system is presented. The structure and the components of each system are introduced. In the first system, the pixel-wise optimisation based people counting system, a new method to improve the practicality of people counting and a new combination of features are proposed. In the second system, a low-level features regression based people counting system, a developed occlusion handling method are proposed and adaptive features selection methods are presented.

3.1 Methodology and Experimental Design

Different techniques can be used to detect and count people. Each technique has certain advantages and disadvantages. Conventional camera based people counting systems are selected as the best technique based on six main criteria;

1. Cost-efficiency

They may use the CCTV cameras that have already been installed for monitoring. The people counting software can be integrated into a standard CCTV system to form highly intelligent systems and there is no need for new or additional hardware. CCTV cameras are widely used and their uses are increasing. For example, there were an estimated 4.2 million CCTV cameras installed in the United Kingdom in 2004 (Norris et al. 2004) and an estimated up to around 5.9 million in 2015 (Xing et al. 2015).

2. High flexibility

They work efficiently with different types of CCTV cameras, camera settings and a wide range of variation of environments characteristics e.g.

- Low and high-resolution cameras.
- Grey or colour cameras.
- Vertical or tilted camera angles.
- Indoor and outdoor environments.
- Environments with a complicated background due to shadows, reflections and loitering.
- Small and large crowd size.
- They work without impeding traffic. People counting systems should not cause or increase a people bottleneck.
- They work without modifying the environment.

3. Acquiring distribution information

They can discover both the number of people and their distribution. Many of the other people counting techniques are unable to acquire information on the distribution of people. Distribution information is very important for medium and large environments such as large retailers, councils, universities, theme parks and stadiums.

4. Wide coverage

In monitoring, they can cover either a small or large area. Moreover, they can work with a distant camera, placed very high, thus making people look small and difficult to recognise. One example of this type of environment is the New York Grand Central Station dataset, which is used to test and evaluate the proposed systems.

5. Privacy preservation

Computer vision systems that involve humans raise important privacy concerns. The privacy of people should be preserved by any people counting

system. Individuals' privacy rights should not be infringed by people counting systems because they may be used in public environments. The regression and optimisation techniques that are used by the proposed systems are a comprehensive privacy preserving method because they do not rely on people recognition or tracking.

6. High accuracy

They produce an accurate estimation of the number of people. Most of them achieve more than 90% accuracy for crowded environments.

Two people counting systems: pixel-wise optimisation based and features regression based people counting systems are proposed in this thesis. The algorithms and methods used as components in these systems are carefully selected to improve the performance of people detection and counting.

GMM method has been used in the proposed people counting systems because it is one of the most widely used algorithms for background subtraction. In addition, this algorithm is a robust in light varying conditions and in environments with animated textures such as waves on the surface of water or trees being blown by the wind (Adegboye 2013). GMM method can also work under noise conditions, low contrast, camera automatic adjustments, dynamic background (Cuevas et al. 2016).

In regard to perspective normalisation, Chan's method has been used to create the density map. This map assigns different weights for the features that extract at different locations of frames. At long distances, people appear smaller than those closer to the camera. Therefore, the extracted features of the same person at different locations in the scene are significantly different. Chan's method has been used in the proposed systems because it is accurate and its equation depends on the width of the walkway and the height of people which are easy to calculate using the user-annotated reference values (Ryan 2013).

Maximum excess over subarrays distance (D_{MESA}) is used with the pixel-wise optimisation based people counting system to compare the predicted count and true count as a loss function. D_{MESA} is chosen for the proposed system because it is not significantly affected by jitter and noise but it has a strong relationship with the number and positions of people (Lempitsky & Zisserman 2010). Kadane's algorithm has been used to calculate the D_{MESA}. This algorithm searches for all positive contiguous segments of the array and keeps track of the maximum sum contiguous segment among all positive segments, representing maximum excess over the subarray.

In order to train the low-level features regression based people counting system, a regression function has to be learned using a set of training samples to find the relationship between the features and the number of people. GPR has been selected with this system. GPR does not use any prior assumptions about the relationship between the features and the crowd size and can achieve high accuracy so it has been chosen in the proposed system (Ryan 2013; Zeyad Q.H. Al-Zaydi et al. 2016; Chan & Vasconcelos 2012; Chan et al. 2008). A combination of kernels can be used with the GPR when the relationship between the extracted features and the number is linear, with some local nonlinearities due to occlusions and segmentation errors (Chan & Vasconcelos 2012).

k-means clustering is selected with the pixel-wise optimisation based people counting system. It is invariant to data order, guaranteed to converge, its time and memory complexity are basically linear to the input point, and it is easy to implement (Celebi et al. 2013). Clustering is used with the proposed system to reduce the number of descriptors (hundreds of thousands for 640x480 frame size) into a reasonable number of clusters (256 clusters in the SIFT features algorithm and 257 clusters in the SIFT-FAST features algorithm) that can be used with quadratic programming.

Edges can be extracted using different algorithms such as Sobel, Canny, Prewitt, Roberts and Fuzzy logic algorithms (Kaur & Virk 2014; Joshi &

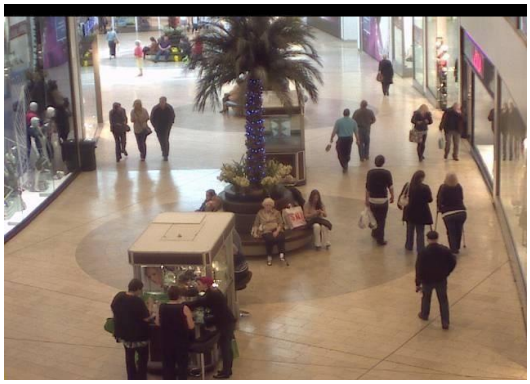
Choubey 2014). Canny edge detection is used in the pixel-wise optimisation based people counting system to detect the moving edge pixels while it is used in the low-level features regression based people counting system as an adaptive feature. It is a high-performance algorithm and can work efficiently under noise conditions (Chen et al. 2014; Shrivakshan & Chandrasekar 2012).

The pedestrian dataset from the University of California, San Diego (UCSD) and the Mall datasets have been used to evaluate the proposed systems (Chan et al. 2008; Chen et al. 2012). Figure 3.1 shows sample frames from the benchmark datasets. The UCSD dataset has been widely used for the testing and validation of people counting methods (C. Zhang et al. 2015). It has been collected using a fixed camera to monitor individuals' pathways. The Mall dataset was introduced by Chen (Chen et al. 2012). It is a newer and more comprehensive dataset because it covers a different range of crowd densities, different activity patterns (static and moving crowds), collected under a large range of illumination conditions at different times of day with a more severe perspective distortion. Thus, individual objects may exhibit larger variations in size and appearance at different depths of the scene (Loy et al. 2013). It has been collected inside a cluttered indoor environment and includes 2000 annotated frames.

The two datasets have the same length (2000 frames) but they have different features in terms of the frame rate (fps), resolution, colour, location, shadows, reflections, crowd size and frame type (Saleh et al. 2015; Ryan et al. 2015). Table 3.1 shows the features of each dataset.

Table 3. 1 The features of the benchmark datasets.

	Mall dataset	UCSD dataset
Year	2012	2008
Length (frames)	2000	2000
Frame rate (fps)	<2	10
Resolution	640x480	238x158
Colour	RGB	Grey
Location	Indoor	Outdoor
Shadows	Yes	No
Reflections	Yes	No
Loitering	Yes	No
Crowd size	11-45	13-53
Frame type	.jpeg	.png



(a) Mall dataset

(b) UCSD dataset

Figure 3. 1: Sample frames from the benchmark datasets that used with the proposed crowd counting systems.

3.1.1 Training setup and repeatability

The benchmark datasets are partitioned into a training set, for learning the proposed systems, and a test set, for validation. In the pixel-wise optimisation based people counting system, 100 frames from different locations of each dataset (Mall and UCSD datasets) are allocated individually for training and 1900 frames for testing. In the features regression based people counting system, the same training and testing partition as in (Chen et al. 2013; Chan et al. 2008; Chen et al. 2012) has been followed in the Mall and UCSD datasets, 800 frames are used for training and 1200 frames for testing. The proposed systems are implemented using MATLAB software 2016a and it is running on a PC with 3.2 GHz core I5 processor and 8 GB memory.

In the pixel-wise optimisation based people counting system, the VLFeat open source toolbox has been used to find the SIFT features. This library includes popular computer vision algorithms specialising in image understanding and local features extraction and matching. It is written in C language for efficiency and compatibility, with interfaces in MATLAB for ease of use. The FAST features are found using the computer vision system toolbox on the MATLAB software.

IPM CPLEX optimisation studio has been connected to the MATLAB software so the functions of both software can be used. This studio combines a fully featured integrated development environment that supports Optimization Programming Language (OPL) and the high-performance CPLEX optimiser solvers. This combination makes it easier to understand and see constraints, goals and costs. Choose from a large set of interfaces, programming languages or deployment scenarios. Deploy in MATLAB, Java, Python, C and C++ or with a client/server architecture. IPM cplexqp function has been used to implement a quadratic programming to find the density of each cluster in each cell.

In the low-level features regression based people counting system, the GPML toolbox has been used to implement the regression models. The GPML toolbox is a combination of an octave and MATLAB implementation of inference and prediction in Gaussian process regression (GPR) models. The strength of the function lies in its flexibility, simplicity and extensibility. The function is flexible because it allows specification of the properties of the GPR through selection of mean function and covariance functions. Extensibility is ensured by modular design allowing for easy addition of extension for the already fairly extensive libraries for inference methods, mean functions, covariance functions and likelihood functions.

The foreground, edge and texture features that are used with the low-level features regression based people counting system were extracted using the toolbox from Matlab software. Table 3.2 shows the features that extracted using the computer vision system toolbox. Table 3.3 shows the features that were extracted using the image processing toolbox.

Table 3. 2 The extracted features using computer vision toolbox.

Features	Description
Foreground segment	segment area segment perimeter perimeter orientation histogram (90 degrees) perimeter orientation histogram (120 degrees) perimeter orientation histogram (150 degrees) perimeter orientation histogram (0 degrees) perimeter orientation histogram (30 degrees) perimeter orientation histogram (60 degrees) perimeter-area ratio Blob count
Edge	internal edge length internal edge orientation histogram (90 degrees) internal edge orientation histogram (120 degrees) internal edge orientation histogram (150 degrees) internal edge orientation histogram (0 degrees) internal edge orientation histogram (30 degrees) internal edge orientation histogram (60 degrees)

Table 3. 3 The extracted features using image processing toolbox.

Features	Description
Texture	GLCM energy (0 degrees) GLCM homogeneity (0 degrees) GLCM entropy (0 degrees) GLCM energy (45 degrees) GLCM homogeneity (45 degrees) GLCM entropy (45 degrees) GLCM energy (90 degrees) GLCM homogeneity (90 degrees) GLCM entropy (90 degrees) GLCM energy (135 degrees) GLCM homogeneity (135 degrees) GLCM entropy (135 degrees)

The code implementation, hardware specifications, datasets and training setup of the proposed systems have been described and presented in this section to achieve the repeatability of this research. Repeatability in computer systems research is important because it measures whether an entire study or experiment can be reproduced in its entirety (Collberg et al. 2016). Some components of the proposed systems have been implemented using MATLAB toolboxes while external toolboxes have been used to implement the others. The IPM CPLEX optimization studio, GPML and VLFeat have been combined with the MATLAB software to implement the proposed systems. In conclusion, this thesis provides a high level of repeatability due to the detailed description of the components of the proposed systems and their implementations. Two people counting systems are proposed and a detailed description of each system is presented in the following sections.

3.2 System One: Pixel-wise Optimisation Based People Counting System

This system includes two algorithms that are based on the estimated density of pixels in a frame to count people. SIFT features and clustering are used in the first algorithm (SIFT algorithm) to represent pixels of frames. The second algorithm uses a combination of FAST corner points and SIFT features with clustering (SIFT-FAST algorithm). A new combination of pixel-wise, motion region, grid map, background subtraction using GMM, and edge detection are used with each algorithm. A fusion technique is proposed and used to validate the accuracy by combining the results of the algorithms at a frame level. The proposed system is more practical than the state of the art regression-based methods because it is trained with a small number of frames, so it is relatively easy to deploy. In addition, it reduces the training error, set-up time, and cost, and opens the door to developing more accurate people detection methods.

3.2.1 Maximum excess over subarrays distance ($\mathrm{D_{MESA}}$)

The proposed system depends on supervised learning to estimate the number of people. The training frames are annotated and Gaussian representation is used to represent people. Quadratic programming is used for learning the proposed system and $\mathrm{D_{MESA}}$ is used to measure the difference between the true and predicted count which represents the loss function as given by equation (3.28).

The proposed system assumes that each pixel (p) in a frame is represented by a SIFT or SIFT-FAST feature vector. The density function of each pixel is represented as a linear transformation of the pixel representation $(\mathrm{x_p})$ as given by;

$$F(p) = \mathrm{w^T x_p} \tag{3.1}$$

Where w^{T} is the weight of each pixel in the frame. At the learning stage, a training frames set with their ground truth (true count) are used to find the correct weight (w^{T}) of each pixel. Then the densities of all pixels in the frame are summed to find the predicted count. $\mathrm{D}_{\mathrm{MESA}}$ is used to compare between the predicted count and true count as a loss function. $\mathrm{D}_{\mathrm{MESA}}$ is defined as (Rodriguez et al. 2011);

$$D_{MESA}\,(F1, F2) = max\left|\sum_{p\in B} F1(p) - \sum_{p\in B} F2(p)\right| \tag{3.2}$$

Where $F1(p)$ and $F2(p)$ are the predicted count and true count of people in a frame. $\mathrm{D}_{\mathrm{MESA}}$ is chosen for the proposed system because it is not significantly affected by jitter and noise but it has a strong relationship with the number and positions of people (Lempitsky & Zisserman 2010). The ultimate goal of the learning stage is to find the best weight for each pixel that minimises the sum of the errors between the true counts and the predicted counts (the loss function) (Lempitsky & Zisserman 2010);

$$w = argmin_w\left(w^T w + \gamma \sum_{i=1}^{N} D_{MESA}\right) \tag{3.3}$$

Where γ is a scalar parameter to control the regularization strength, $argmin_w$ represents the best weight that minimises the D_{MESA}. Quadratic programming can be used to solve equation (3.29) by using;

$$\min_{w,\xi_1,\ldots\ldots,\xi_N}\left(w^T w + \gamma \sum_{i=1}^{N} \xi_i\right) \tag{3.4}$$

Subject to;

$$\xi_i \geq \sum_{p\in B}\big(F1(p) - F2(p)\big), \qquad \xi_i \geq \sum_{p\in B}\big(F2(p) - F1(p)\big) \tag{3.5}$$

Where ξ_i are the auxiliary variables of training frames. Quadratic programming uses iterations to optimise the results and find the best weight (w^{T}) of each

pixel. The iterations terminate when the right side of equation (3.31) is within (ξ_i + β) factor. β is a small constant (β << 1). It uses to decrease the number of iterations and faster convergence. Choosing β equal to 0 solves the equations (3.30) and (3.31) exactly. However, the convergence will finish faster if β is chosen to a very small value and that will not affect the performance of training (Lempitsky & Zisserman 2010). In the experiments of the proposed system, β has been chosen to be equal to 0.001. The flow diagram of the proposed system is illustrated in the Figure 3.5. It consists of two counting algorithms, one video source and one fusion model.

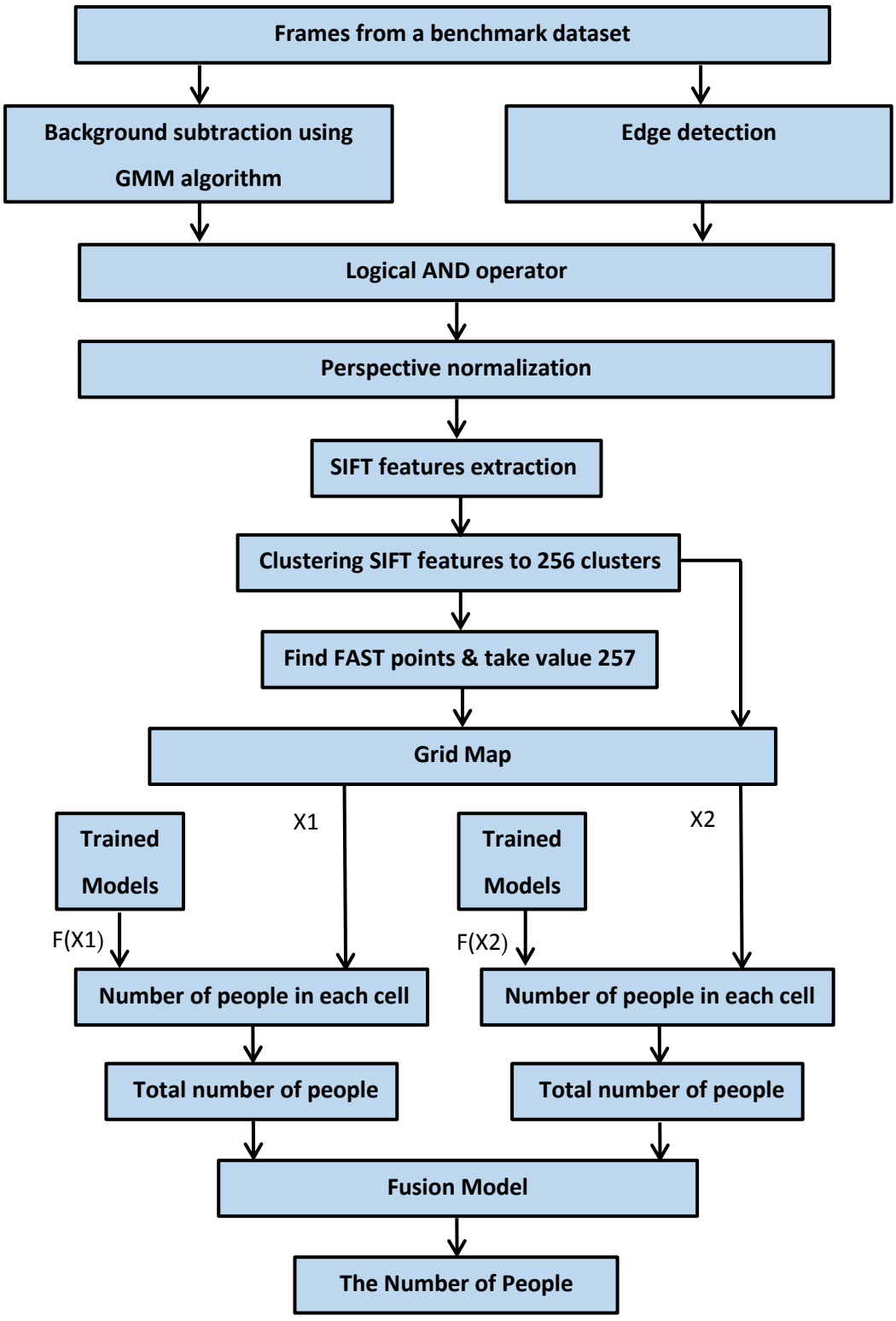


Figure 3. 2: Flow diagram of the proposed pixel-wise optimisation based crowd counting system.

3.2.2 Algorithm 1: SIFT features algorithm

This algorithm combines the following techniques to count the number of people; motion edges, SIFT descriptors, gird map and pixel-wise techniques. This combination that is used to find the density of each pixel, is novel. Edge pixels are used because their number is less than foreground pixels. As a consequence, the required time to find the SIFT descriptors and cluster them in a frame will be significantly reduced which makes the proposed system faster than other people counting techniques based on $\mathrm{D}_{\mathrm{MESA}}$ optimisation. There is a high correlation between SIFT descriptors and the number of people. This is difficult for quadratic programming to be used to find the density of a large number of SIFT descriptors (equal to the number of edge motion pixels). To solve this problem, clustering is used to reduce the number of SIFT descriptors to 256 clusters. The main disadvantage of using clustering is that many SIFT descriptors can be grouped into one cluster to reduce the problem space but they represent different densities. Grid map is used to improve the cluster classification in the frames which enables similar clusters in different cells to be assigned different densities depending on their location in the frame. The proposed algorithm can better adapt to high variations in crowd behaviours, distributions and densities. As a result, the accuracy is improved. Figure 3.6 shows the flow diagram of this algorithm. The procedure of the algorithm is illustrated in the following steps;

1. Implement Gaussian mixture model (GMM) to find the foreground information of the frame.

$$F_{GMM} = GMM(i,j) \tag{3.6}$$

Where F_{GMM} is the foreground pixels of the frame and $GMM(i,j)$ is the GMM of each pixel of the frame.

2. Implement edge detection to find the edges of the frame.

$$F_{Edge} = E(i,j) \tag{3.7}$$

Where F_{Edge} is the edge of the frame and $E(i,j)$ is the detected edge of each pixel of the frame.

3. Perform logical (AND) operation between the foreground pixels of the frame and the detected edge to find the motion edge of the frame.

$$F_{motion\ edge} = F_{GMM}(i,j)\ \&\&\ F_{Edge}(i,j) \tag{3.8}$$

Where $F_{motion\ edge}$ is the motion edge for the frame.

4. The pixels in each line of the frame are assigned different weight as a perspective normalisation.

5. Find the SIFT descriptor for each motion edge pixel. Then, cluster the SIFT descriptors to 256 clusters. The centres of SIFT features are used as criteria for clustering them.

$$F_{SIFT} = SIFT(i,j) \qquad (i,j) \in \text{motion edge} \tag{3.9}$$

$$F_{Cluster} = Cluster(F_{SIFT}) \qquad (i,j) \in \text{motion edge} \tag{3.10}$$

Where F_{SIFT} is the SIFT descriptors of the frame and $F_{Cluster}$ is the SIFT descriptors clustering.

6. Divided the frames into cells (as a grid map) and count the number of people in each cell.

$$F_{Grid} = \sum_{n} C_n \tag{3.11}$$

Where F_{Grid} is the grid map of each frame, C is a cell in the grid map and n is the number of cells in the grid map. Four cells configuration has been used in the proposed system which gives the best accuracies experimentally.

7. Use a quadratic programming to find the density of each cluster in each cell.

8. Integrate the densities of pixels over each cell to find the number of people in each cell.

$$N_{cell} = \sum_{(i,j)\in B_n} P_{density}\,(i,j) \tag{3.12}$$

Where N_{cell} is the number of people in each cell and $P_{density}(i,j)$ is the density of each pixel that belongs to this cell.

9. The summation of the number of people in all cells represents the total number of people in the frame.

$$N_{total} = \sum_{n} N_{cell} \tag{3.13}$$

Where N_{total} is the total number of people in a frame and n is the number of cells.

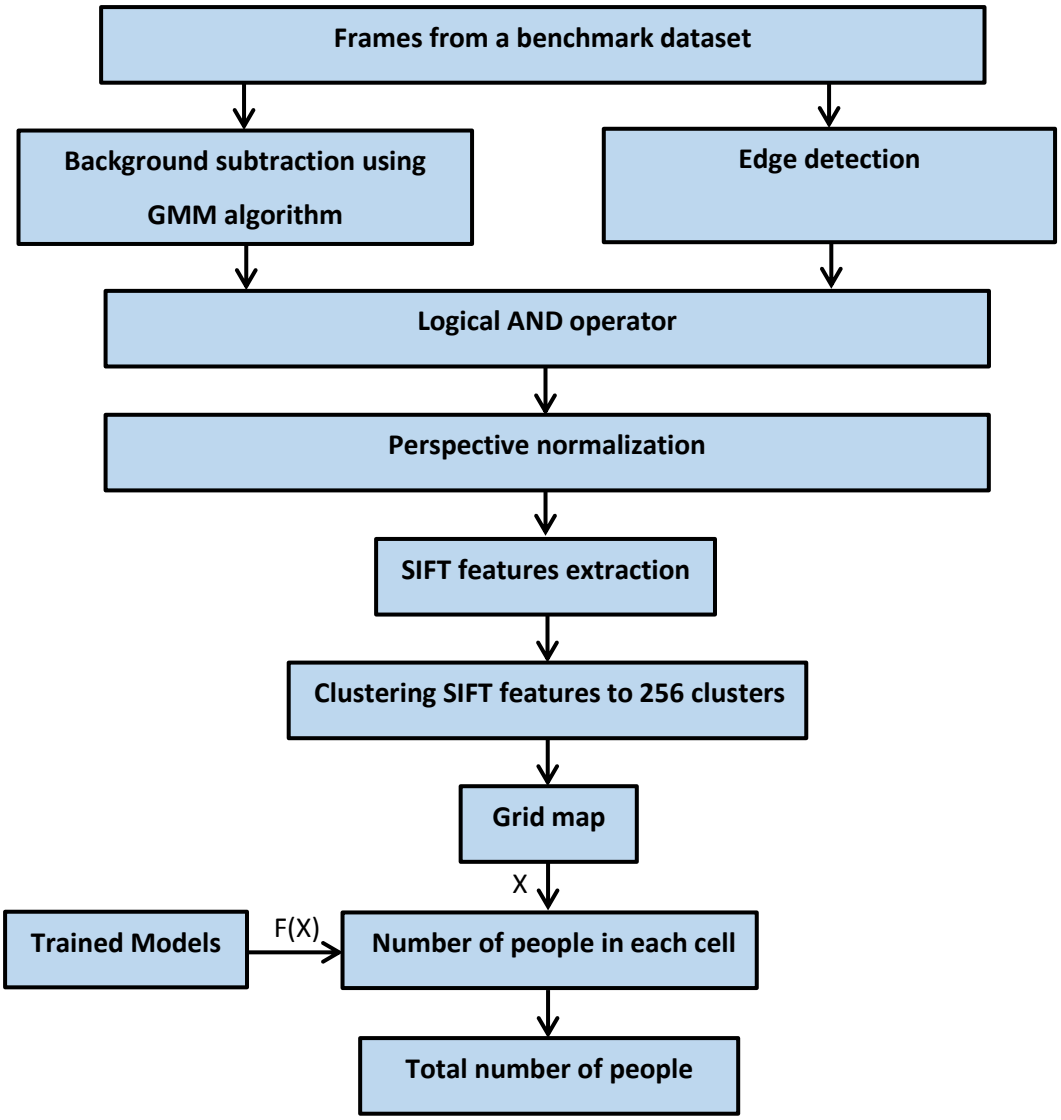


Figure 3. 3: Flow diagram of the SIFT Features Algorithm.

3.2.3 Algorithm 2: SIFT-FAST features algorithm

This algorithm uses two features; FAST and SIFT. This algorithm combines the following techniques to count the number of people; motion edges, grid map, SIFT & FAST features and pixel-wise techniques. Edge pixels are used because their number is less than those of foreground pixels. The same approach as for SIFT feature algorithm described in Section 3.4.2 is used. However, FAST corner points are used to improve the accuracy due to the high correlation between the number of people and FAST corner points. The algorithm can also better adapt to high variations due to crowd behaviours, distribution and density. Figure 3.7 shows the flow diagram of the algorithm. Steps 1 to 5 are the same as for SIFT feature algorithm and descriptions from step 6 are as follows:

6. Find FAST points in each frame within the motion region.

$$F_{FAST} = F(i,j) \qquad (i,j) \in \text{motion regions} \qquad (3.14)$$

Where F_{FAST} is the FAST corner points of a frame.

7. All pixels that are FAST corner points are assigned the value 257 so that quadratic programming can be used to find 257 density values instead of 256.

8. Divide the frame into cells (as a grid map) and the number of people in each cell is counted individually.

$$F_{Grid} = \sum_{n} C_n \qquad (3.15)$$

Where F_{Grid} is the grid map of the frames, C is a cell in the grid map and n is the number of cells in the grid map.

9. Use a quadratic programming to find the density value of each cluster in each cell.

10. Integrate the densities of pixels over each cell to find the number of people in each cell.

$$N_{cell} = \sum_{(i,j)\in B_n} P_{density}\,(i,j) \tag{3.16}$$

Where N_{cell} is the number of people in each cell and $P_{density}(i,j)$ is the density of each pixel that belongs to the cell.

11. The summation of the number of people in all cells represents the total number of people in each frame.

$$N_{total} = \sum_{n} N_{cell} \tag{3.17}$$

Where N_{total} is the total number of people in a frame, n is the number of cells.

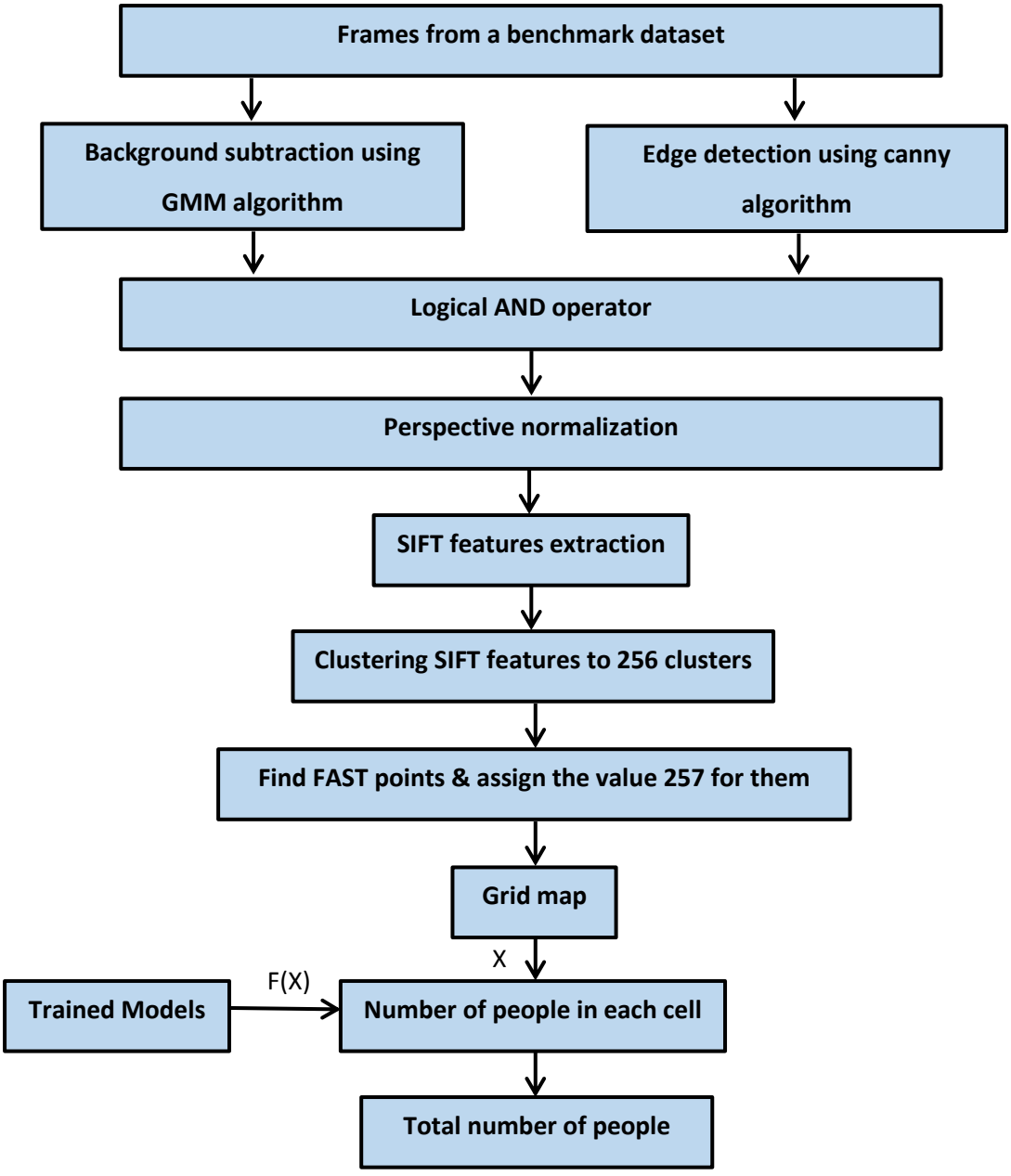


Figure 3. 4: Flow diagram of the SIFT-FAST features algorithm.

3.2.4 Edge detection algorithm

They refer to the process of localising pixel intensity transitions (Ershad 2012). There is a strong relationship between the complexity of edges and the number of people because crowded environments tend to produce complex edges, while sparse environments tend to produce coarse edges (Loy et al. 2013). Edges can be extracted using different algorithms such as Sobel, Canny, Prewitt, Roberts and Fuzzy logic algorithms (Kaur & Virk 2014; Joshi & Choubey 2014). Canny edge detection is used in the proposed systems. It is a high-performance algorithm and can work efficiently under noise conditions (Chen et al. 2014; Shrivakshan & Chandrasekar 2012). The following steps explain the procedure of Canny edge algorithm (Shrivakshan & Chandrasekar 2012):

1. Smooth the image using a Gaussian filter to minimise noise.

$$S(i,j) = G(i,j,\boldsymbol{\Sigma}) * I(i,j) \tag{3.18}$$

Where $G(i,j,\boldsymbol{\Sigma})$ is a Gaussian filter and $I(i,j)$ is a pixel.

2. Use derivative approximation by finite differences to find gradient magnitude and orientation. Firstly, partial derivatives $X(i,j)$ and $Y(i,j)$ are found by using the smoothed array $S(i,j)$:

$$X(i,j) \approx (S(i,j+1) - S(i,j) + S(i+1,j+1) - S(i+1,j))/2 \tag{3.19}$$

$$Y(i,j) \approx (S(i,j) - S(i+1,j) + S(i,j+1) - S(i+1,j+1))/2 \tag{3.20}$$

The partial derivatives $X(i,j)$ and $Y(i,j)$ are then used to find the magnitude and orientation of the gradient:

$$M(i,j) = \sqrt{X(i,j)^2 + Y(i,j)^2} \tag{3.21}$$

$$\theta(i,j) = \arctan\big(X(i,j), Y(i,j)\big) \tag{3.22}$$

3. Non-Maximal Suppression algorithm (NMS) is performed to thin out the edges. The edges are then detected using the double thresholding algorithm.

3.2.5 Scale-invariant feature transform (SIFT)

The SIFT algorithm is used to detect and describe local features within images (Zhong et al. 2015). SIFT descriptors are invariant to image scale, translations and rotations (Giveki et al. 2017). In addition, they are robust with moderate illumination variations and perspective transformations (Zhong et al. 2015). The SIFT algorithm consists of four stages: scale-space extrema extraction, keypoint localisation, orientation assignment and keypoint descriptor (Giveki et al. 2017; Prathap et al. 2016).

1. Scale-space extrema extraction:

This stage seeks to identify potential interest points identifiable under different views (Montazer & Giveki 2015). Different scales are used to find the locations of these points which are invariant to scale change (Zhong et al. 2015). A convolution operator between the original input image and a variable scale Gaussian function is used to produce the scale space (Giveki et al. 2017). The result of this operation is an octave that consisting of different scales;

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{3.23}$$

Where $*$ is the convolution operator, s is the scale value, $G(x, y, \sigma)$ is a variable scale Gaussian function and $I(x, y)$ is the input image. The variable scale Gaussian function is given by;

$$G(x, y, \sigma) = 1/2\pi\sigma^2 \ exp(-(x^2 + y^2)/2\sigma^2) \tag{3.24}$$

An octave is produced by this convolution. Frames are resized several times to produce other octaves and the convolution operator is used with each octave.

Different techniques that based on the scale space can be used to detect stable points. The difference of Gaussians is often used to find the local maxima and local minima points which are invariant to scale and rotation (Montazer & Giveki 2015). The difference of Gaussians of each consecutive scale $DoG(x, y, \sigma)$ is calculated using;

$$DoG(x, y, s) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{3.25}$$

Figure 3.8 shows the difference of Gaussian at the scale space. Figure 3.9 shows that the local maxima and minima points are found by comparing each point with its 26 neighbours, eight neighbours at the same scale and nine neighbours up and down the scale. The minimum or maximum value of the comparison are the extrema points (Giveki et al. 2017).

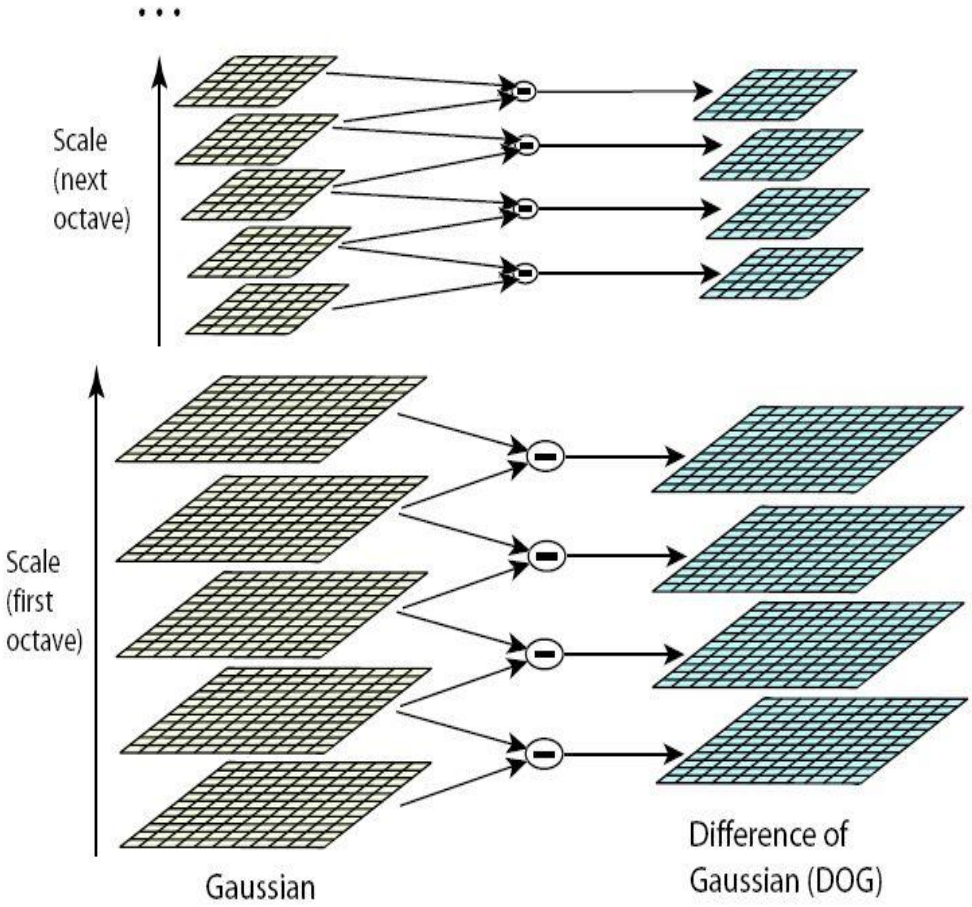


Figure 3. 5: Gaussian and difference of Gaussian (Prathap et al. 2016).

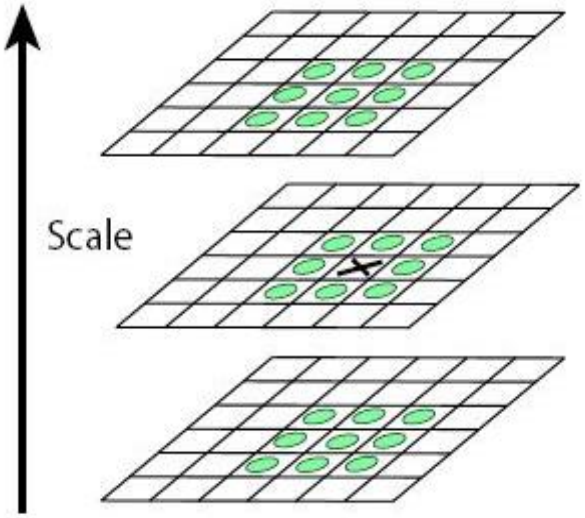


Figure 3. 6: The minimum or maximum extrema points calculation by comparison with its 26 neighbours. (Prathap et al. 2016).

2. Keypoint localisation

All the potential interest points calculated in the first stage are now filtered to remove the low contrast points or those that are localised on the edges of the image (Prathap et al. 2016). In addition, the locations of the calculated points from the first stage may not be accurate so there may be a need to correct their locations. To correct the location of these points, the Taylor series expansion (TSE) of the scale space function is used (Lowe 2004);

$$DoG(x) = DoG + \frac{\partial DoG^T}{\partial x}x + \frac{1}{2}x^T \frac{\partial^2 DoG}{\partial x^2}x \tag{3.26}$$

The location of extremum ($\hat{x}$) is found by taking its derivative and set to zero.

$$\hat{x} = -\frac{\partial^2 DoG^{-1}}{\partial x^2}\frac{\partial DoG}{\partial x} \tag{3.27}$$

The low contrast keypoints are removed using the scale space function value at the extremum (Lowe 2004);

$$DoG(\hat{x}) = DoG + \frac{1}{2}\frac{\partial DoG^T}{\partial x}\hat{x} \quad (3.28)$$

The value of each point is compared with a predefined threshold. The points with values less than the threshold are removed. The edge points are removed by measuring the levels of curvature of the keypoint and in the perpendicular direction of it using the difference of Gaussian function (Giveki et al. 2017). Keypoints are considered as edge point when the difference between these levels are large (Giveki et al. 2017).

3. Orientation assignment

In this stage, one or more dominant consistent orientations of each keypoint are assigned based on the local properties of the frame. Invariance to rotation task is achieved in this stage (Zhong et al. 2015). The gradient magnitude can be found by (Giveki et al. 2017);

$$m(x,y) = \sqrt{\big(L\,(x+1,y) - L(x-1,y)\big)^2 + (L(x,y+1) - L(x,y-1))^2} \quad (3.29)$$

The gradient orientation can be found by (Giveki et al. 2017);

$$\theta(x,y) = \tan^{-1}(L(x,y+1) - L(x,y-1), L(x+1,y) - L(x-1,y)) \quad (3.30)$$

The orientation histogram of each keypoint is constructed using the magnitude and the orientation of every pixel around the keypoints. The histogram consists of 36 bins with 10 degrees per bin (Prathap et al. 2016). The highest peak and any other peaks within 80% of the highest peak are assigned as the main orientation angles to the keypoint (Giveki et al. 2017). The three closest histogram values of each peak are used to interpolate a better accurate peak (Giveki et al. 2017).

4. Keypoint descriptor

The keypoints from the third stage are represented as descriptors in this stage. The descriptors are a vector of orientation histograms. Sixteen histograms are used to calculate each SIFT descriptor. These histograms are aligned in a 4 × 4 grid with eight orientation bins for each one. As a result, the keypoint descriptor is represented by a 128-dimensional vector (Prathap et al. 2016).

In this system, a dense-SIFT version is used instead of a standard SIFT version. Standard SIFT algorithm include detecting interest points and then representing them using descriptors. However dense-SIFT version does not detect interest points, it only calculates a descriptor for each pixel in a frame.

3.2.6 Features from accelerated segment test (FAST)

This algorithm was developed to detect corner points in an image for use in real-time applications due to its high computational speed. It is faster than most of the other corner detection methods (Ghosh & Kaabouch 2016). Corner points represent points containing two edges with different directions (Majumder et al. 2013). To detect FAST points, a circle of sixteen pixels around each pixel in a frame, is considered, as shown in figure 3.10. A pixel is considered as FAST points if twelve contiguous pixels in its circle are either brighter or darker than the intensity of the pixel by pre-threshold (Ghosh & Kaabouch 2016). All pixels in the circle are assigned one of three states; brighter, darker and same.

$$S_{brighter} = I_x \geq I_p + T$$

$$S_{darker} = I_x \leq I_p - T \tag{3.31}$$

$$S_{same} = I_p - T < I_x < I_p + T$$

Where I_p is a pixel at a frame. I_x is one pixel in the circle around I_p . These equations repeated 16 times to check the states of all pixels in the circle and depending on the results, the pixel is considered as FAST corner or not.

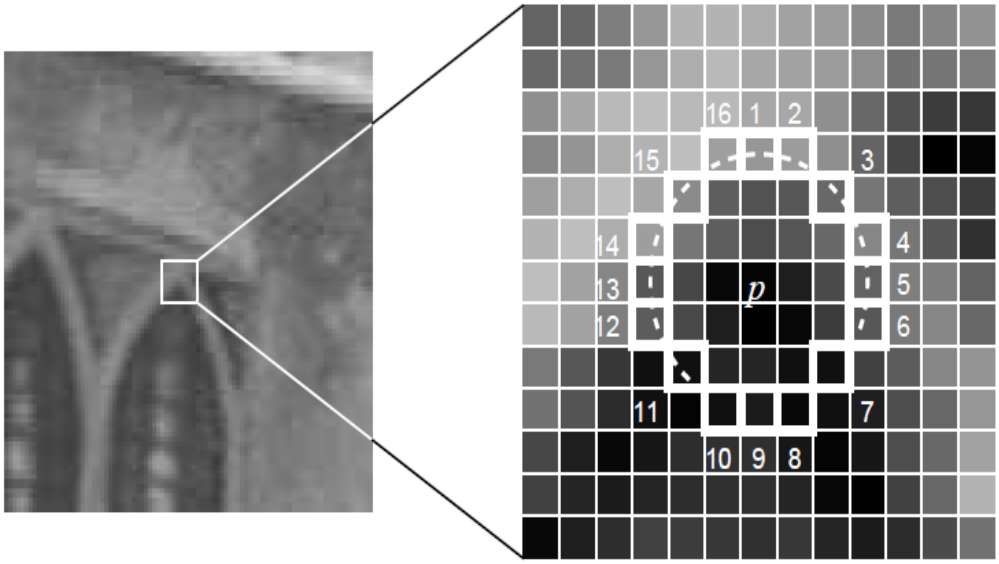


Figure 3. 7: A corner point based on FAST corner detector (Kitamura et al. 2015).

Choosing a large threshold reduces the number of detected FAST corner points whereas a small threshold yields a larger number of corner points (Biadgie & Sohn 2014). The speed of the FAST algorithm can be increased only by comparing the intensity of each pixel at a frame with pixels 1, 5, 9 and 13 of its circle. If at least three of them satisfy the criterion, then check all the 16 pixels in the circle to detect if there are 12 contiguous pixels are brighter or darker the pixels.

3.2.7 K-means clustering

Cluster analysis is very useful for different applications such as pattern recognition, statistics, machine learning, information retrieval, data mining, data compression, business and biology (Tan et al. 2006). The aim of the

clustering algorithms is that the observations within a cluster be similar to one another and different from the observations in other clusters. The good clustering algorithms produce highest homogeneity or similarity within a cluster and the highest difference between clusters.

The K-means clustering algorithm is based on the mean of a group of observations. It starts by choosing the number of clusters (K). The K parameter is chosen by a user, based on the number of clusters desired. If the number of observations is equal to or less than the number of clusters, then each observation is assigned to one cluster. If the number of observations is larger than the number of clusters, the observations are then assigned to the closest cluster based on the Euclidean distance from each observation to each cluster. The means of the clusters are then recomputed based on the new assigned observations of each cluster. Mean is measured for each dimension of the observations and then the means of all dimensions are combined to find the multi-dimensional mean. The K-means algorithm terminates when the mean of each cluster does not change. The convergence of the K-means clustering algorithm usually happens in the first few iterations. The procedure of the K-means clustering algorithm is illustrated in the following steps (Tan et al. 2006);

1. Select the number of clusters K.
2. **Repeat**
3. Form K clusters based on assigning the observations to their closest mean.
4. Recompute the mean of each cluster.
5. **Until** all means of the clusters do not change.

There are several reasons why k-means clustering is one of the most widely used clustering algorithms. It is invariant to data order, guaranteed to converge, its time and memory complexity are basically linear to the input point, and it is easy to implement (Celebi et al. 2013).

Clustering is used with the proposed system to reduce the number of descriptors (hundreds of thousands for 640x480 frame size) into a reasonable number of clusters (256 clusters in the SIFT features algorithm and 257 clusters in the SIFT-FAST features algorithm) that can be used with quadratic programming. K-means clustering is a method of vector quantisation and aims to partition *n* observations into $k \leq n$ clusters. In other words, it aims to find (Jain 2010):

$$argmin_c \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2 \quad (3.32)$$

Where x_i is the observation, c_k is the k^{th} cluster and μ_k is the mean of cluster c_k. In the proposed system, the k-means algorithm is used to cluster the SIFT descriptors of the datasets frames and produce a codebook of 256 entries. The codebook is constructed using only the descriptors of the training frames and then the descriptors of the testing frames are clustered by the K-means algorithm and the codebook. The SIFT descriptor of each pixel is represented using a vector. The length of this vector is equal to 256 items which one of them is equal to 1 while the other are equal to 0.

3.2.8 Fusion technique

The fusion model is updated periodically using the results of the both algorithms. Each algorithm works independently to count the number of people and then they update the fusion model. Fusion is used to improve accuracy by determining the average error for each frame and to increasing the confidence of the proposed system because the result of one algorithm is confirmed by that of another. This produces a cooperative paradigm and improves the confidence level of the results.

3.3 System Two: Features Regression Based People Counting System

The proposed system uses a pair of collaborative GPR models with different kernels instead of a single model. The calculated level of occlusion is used with these GPR models to improve the accuracy of counting. The level of occlusion is measured and compared with a predefined threshold to select the regression model that should work with each frame. In addition, the best combination of features is dynamically identified to improve the accuracy of people counting.

3.3.1 Adaptive and non-adaptive people counting systems

Adaptive people counting systems are adaptable to the level of occlusion of each frame, as well as the characteristics of each environment. These factors may affect the selection of features, regression models and kernels. This is why adaptive systems use an adaptive combination of features, regression models and kernels. Although a lot of research has been carried out to find the best, or most efficient, static combination of features for all types of environment, researchers were unable to do so, because the most accurate people counting system for each environment was achieved using a different combination of features. Moreover, heterogeneous training data has been widely used to train only one regression model and one covariance function (kernel). This can negatively affect the accuracy of counting. In conclusion, homogeneous training data, specific purpose regression models, and dynamic features selection, can be used to improve the accuracy of people counting systems.

This system is classified as an adaptive people counting system due to the following three main reasons. First, a pair of collaborative GPR models 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.

3.3.2 System structure

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 the 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 3.11.

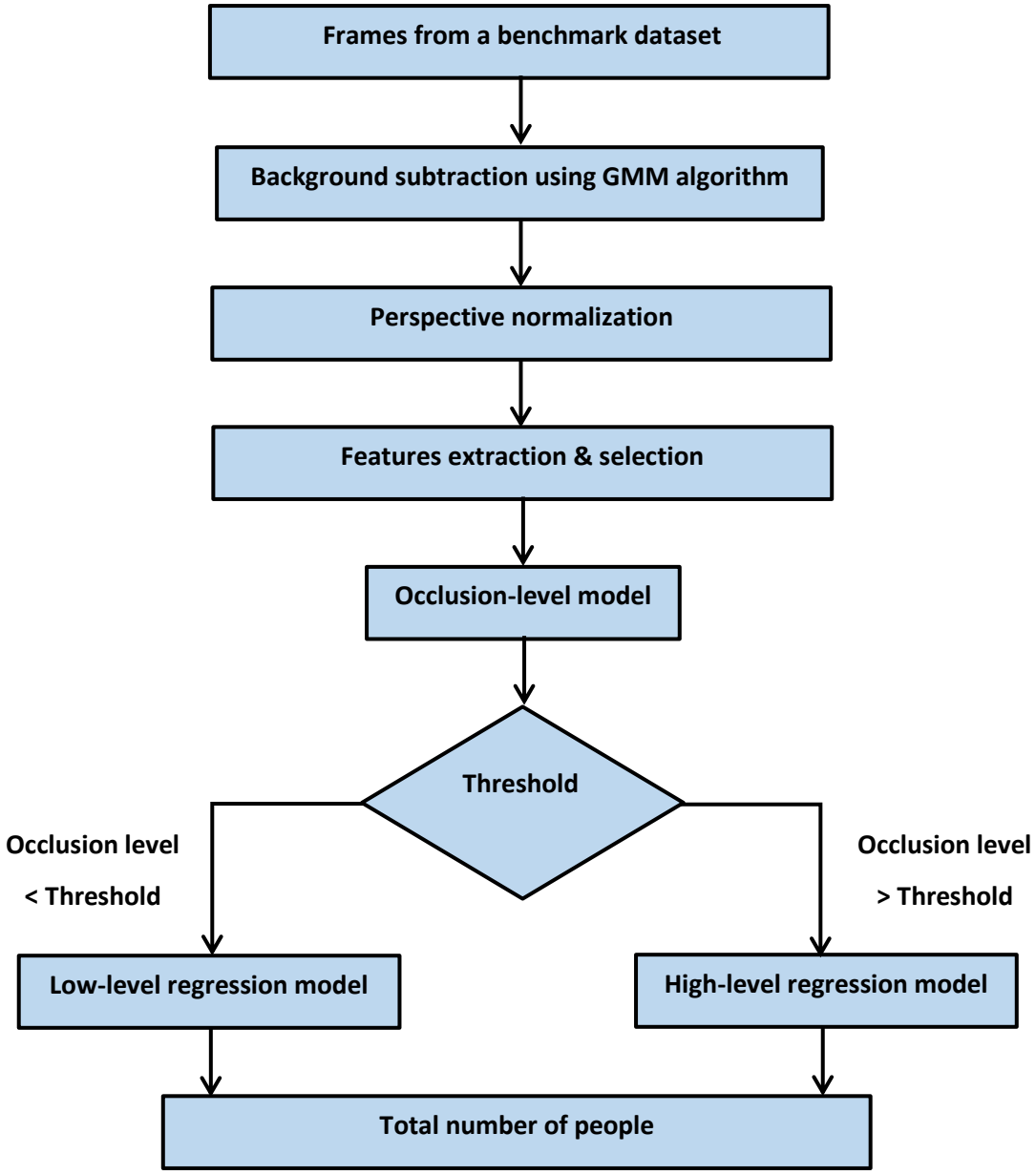


Figure 3. 8: Flow diagram of the proposed low-level features regression based crowd counting system.

3.3.3 Regression model selection

In order to train the people counting system, a regression function has to be learned using a set of training samples to find the relationship between the features and the number of people. GPR has been selected in this system. GPR does not use any prior assumptions about the relationship between the features and the crowd size and can achieve high accuracy so it has been chosen in the proposed system (Ryan 2013; Zeyad Q.H. Al-Zaydi et al. 2016; Chan & Vasconcelos 2012; Chan et al. 2008).

Two independent 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 (Williams & Rasmussen 2008):

$$y_*|y \sim N\left(K_* K^{-1} y, K_{**} - K_* K^{-1} {K_*}^T\right) \tag{3.33}$$

The best estimate for y_* is the mean of this distribution (Williams & Rasmussen 2008):

$$y_* = K_* K^{-1} y \tag{3.34}$$

The uncertainty in the estimate is captured in its variance (Williams & Rasmussen 2008):

$$var(y_*) = K_{**} - K_* K^{-1} {K_*}^T \tag{3.35}$$

Where y and y_* are the function values of the training and testing sets, respectively. $K, K_* \text{ and } K_{**}$ are the covariance functions (kernels) of the training, training-testing and testing inputs, respectively.

$$K = \begin{bmatrix} k(x1,x1) & \cdots & k(x1,xn) \\ \vdots & \ddots & \vdots \\ k(xn,x1) & \cdots & k(xn,xn) \end{bmatrix} \quad (3.36)$$

$$K_* = [k(x_*,x1) \quad \ldots \quad k(x_*,xn)] \quad (3.37)$$

$$K_{**} = (x_*,x_*) \quad (3.38)$$

Where $x1, x2, x3, \ldots\ldots\ldots xn$ are the training set. x_* is the test set. $k(x,x')$ is the covariance function (kernel). There are different kernels that can be used with a GPR model. 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 (T. Y. Lin et al. 2011). The linear kernel on two inputs x and x', represented as feature vectors is given by (Chan & Vasconcelos 2012):

$$k(x,x') = \alpha\,(x^T x' + 1) \quad (3.39)$$

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 (Mei & Zhao 2013). 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 (Mei & Zhao 2013). Mathematically, a combination of linear and RBF kernels is given by (Williams & Rasmussen 2008; Chan & Vasconcelos 2012):

$$k(x,x') = \alpha_1\,(x^T x' + 1) + {\alpha_2}^2 \exp\left[\frac{-1}{2{\alpha_3}^2}\,\|x - x'\|^2\right] \quad (3.40)$$

α_1, α_2 and α_3 are the kernals parametes. In addition, we can use an ensemble learning method that first partitions the heterogeneous training data into linear and non-linear homogeneous groups (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 (Jin & Liu 2004), 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 (Jin & Liu 2004; Jin & Liu 2005). In the next section, the method of measuring the level of occlusion is explained.

3.3.4 Occlusion handling

Keypoints has been used by many studies to find the number of people (the level of the crowd) due to their strong inter-dependence (Alberto Albiol et al. 2009; Antonio Albiol et al. 2009; Conte et al. 2010b; Jeong, C. Y., Choi, S., & Han 2013). Although there is a degree of correlation between the level of occlusion and the level of crowd in a frame, the validity of this relationship is not always correct in all scenarios due to the effect of sparseness. As a consequence, there is a need to develop a method to measure the level of occlusion that takes into account the sparseness and level of the crowd. Two independent GPR models with different kernels are used in the proposed system. The first regression model (high-level occlusion regression model) is trained using high occluded frames and the second (low-level occlusion regression model) is trained using low occluded frames. 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 the crowd density (number of people) (Jeong, C. Y., Choi, S., & Han 2013):

$$Level\ of\ crowd = \frac{No\ of\ keypoints}{No\ of\ keypoints\ per\ person} \tag{3.41}$$

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

$$Level\ of\ occlusion = \frac{No\ of\ keypoints}{No\ of\ foreground\ pixels} \quad (3.42)$$

The number of SIFT points are used to measure the level of occlusion in the proposed system. The level of occlusion is measured by dividing the number of SIFT points by the number of foreground pixels. SIFT points are defined as maxima/minima of the difference of Gaussians in scale-space (Fradi & J. Dugelay 2012). SIFT keypoints is better than the FAST and speeded-up robust features (SURF) because they are more invariant to scale, rotation, and affine transformations (Fradi & J. Dugelay 2012).

The output of the occlusion level model of each frame is compared with a predefined threshold. The thresholding stage involves classifying frames into two categories; high and low occluded frames. There is no technical definition of the separated level between the low and high occluded frames because there are no clear boundaries between them. The highest occluded frames in one environment can be the lowest occluded frames in another environment depending on the different potential factors. Those factors include the range of crowd sizes, resolution and the area of the camera view. In conclusion, choosing a suitable threshold depends on the nature of environments 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, resolution and the area of the camera view. In the proposed system, the threshold is experimentally selected by using a multi-stage thresholding method. The range of the crowd size is normalised to 0 - 1 range. In the first stage, the range is divided into nine equal intervals which are used as potential thresholds. Those thresholds are used to measure the accuracy of the system in terms of mean deviation error (MAE), mean squared error (MSE) and mean absolute error (MDE). 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 predefined threshold.

Fast people counting systems are particularly suitable for real-time applications where computational efficiency is important. In the proposed system, the training stage is repeated 31 times using classical GPR method to select the best combination of features. In addition, 19 potential thresholds are used with the selected combination of features to train the proposed system and find the best threshold that can be used to achieve the highest accuracy. The training stage is performed only at the installation of the system so the extra computational complexity can be neglected when the system starts working.

On the other hand, the occlusion level model and thresholding method are a simple division and relational operators, respectively. They add a very low computation complexity to the working system in comparison with the frame extraction, background subtraction using GMM algorithm, prospective normalisation, features extraction and estimating using GPR algorithm. The computation complexity of the regression stage will not be changed because one of the regression models (the low or high-level regression model) will be used with each frame.

In conclusion, the computational complexity of the proposed system is a little higher than classical GPR methods in the testing stage. However, the computation efficiency is significantly decreased in the training stage due to the threshold and features selections, it can be neglected when the system starts working because training is performed only at the installation stage of the counting system.

3.3.5 Features representation

The low-level feature 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 (Loy et al. 2013; Ryan et al. 2015). Different intermediate features can also be used as inputs to a regression model for people counting such as blob size histogram and edge orientation histogram.

In features regression based algorithms, a popular technique is to combine several features to achieve higher accuracy. The performance of various features and combinations of features for people counting depend on the type of environments in a real-time applications or datasets in offline applications (Ryan et al. 2015). Therefore, 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.

In pixel-wise optimisation based system, a combination of features is rarely used because the complexity of optimisation could be increased significantly. For single kind of feature, features descriptors are extracted based on a pixel basis. A very large number of feature descriptors that proportional to the number of pixels at a frame is extracted. In the case of a combination of features, the feature descriptors extraction is repeated for each kind of features so the delay of processing time will increase significantly. Features can be categorised under the following headings:

3.3.5.1. Foreground segment 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 (Ryan et al. 2015). 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 (Loy et al. 2013).

3.3.5.2. Texture features

There is a strong relationship between the number of people and the texture of crowds, which refer to the general description of a frame (Loy et al. 2013). Gray-level co-occurrence matrix (GLCM) and local binary pattern (LBP) are usually used to find texture features (Ershad 2012; Wang et al. 2013; Loy et al. 2013). Texture features include homogeneity (texture smoothness), energy (total sum-squared energy), entropy (texture randomness) and contrast (Loy et al. 2013; Chan & Vasconcelos 2012; Ryan et al. 2015).

Many people counting studies have used the GLCM whereas the LBP has been widely used in expression analysis and face recognition applications (Loy et al. 2013). Therefore, the GLCM has been used in the proposed systems. The first step to calculate the GLCM is that each frame is quantised into eight grey levels. The co-occurrence matrix ($P(i,j|\theta)$) is then created based on the grey levels which represents a joint probability distribution of pixels. θ represents the orientations of the co-occurrence which is assigned one of the four angles ($0°, 45°, 90°, 135°$). The symmetric co-occurrence matrix ($P_s(i,j|\theta)$) is found by (Ryan et al. 2015);

$$P_s(i,j|\theta) = P(i,j|\theta) + P(i,j|\theta)^T \tag{3.43}$$

The symmetric co-occurrence matrix is then normalised by;

$$P_n(i,j|\theta) = \frac{P_s(i,j|\theta)}{\sum_{i,j} P_s(i,j|\theta)} \tag{3.44}$$

Where $P_n(i,j|\theta)$ is the normalised co-occurrence matrix. Figure 3.9 shows how to produce the co-occurrence matrix, symmetric co-occurrence matrix and the normalised co-occurrence matrix using GLCM method.

4	5	6	3	5	4	1	1	1	1
4	5	4	8	5	3	1	1	1	1
4	7	7	7	2	7	1	1	1	1
3	7	5	5	3	2	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1

(a) Quantised level of image

	1	2	3	4	5	6	7	8
1	48	0	0	0	0	0	0	0
2	1	0	0	0	0	0	1	0
3	1	1	0	0	1	0	1	0
4	1	0	0	0	2	0	1	1
5	0	0	2	2	1	1	0	0
6	0	0	1	0	0	0	0	0
7	1	1	0	0	1	0	2	0
8	0	0	0	0	1	0	0	0

(b) $P(i,j|\theta)$

	1	2	3	4	5	6	7	8
1	96	1	1	1	0	0	1	0
2	1	0	1	0	0	0	2	0
3	1	1	0	0	3	1	1	0
4	1	0	0	0	4	0	1	1
5	0	0	3	4	2	1	1	1
6	0	0	1	0	1	0	0	0
7	1	2	1	1	1	0	4	0
8	0	0	0	1	1	0	0	0

(c) $P_s(i,j|\theta)$

	1	2	3	4	5	6	7	8
1	x_{96}	x_1	x_1	x_1	x_0	x_0	x_1	x_0
2	x_1	x_0	x_1	x_0	x_0	x_0	x_2	x_0
3	x_1	x_1	x_0	x_0	x_3	x_1	x_1	x_0
4	x_1	x_0	x_0	x_0	x_4	x_0	x_1	x_1
5	x_0	x_0	x_3	x_4	x_2	x_1	x_1	x_1
6	x_0	x_0	x_1	x_0	x_1	x_0	x_0	x_0
7	x_1	x_2	x_1	x_1	x_1	x_0	x_4	x_0
8	x_0	x_0	x_0	x_1	x_1	x_0	x_0	x_0

(d) $P_n(i,j|\theta)$

Figure 3. 9: GLCM with $\theta = 0°$ (a) a 10 by 10 image with quantised level; (b) the co-occurrence matrix of the image; (c) the symmetric co-occurrence matrix of the image; (d) the normalised co-occurrence matrix of the image where $x_0 = 0, x_1 = 0.00694, x_2 = 0.0139, x_3 = 0.0208, x_4 = 0.0278, x_{96} = 0.667$.

Different features can be derived for each θ such as contrast, energy, homogeneity and entropy;

$$Contrast_\theta = \sum_{i,j} (i-j)^2 \; P_n(i,j|\theta) \tag{3.45}$$

$$Energy_\theta = \sum_{i,j} P_n(i,j|\theta)^2 \tag{3.46}$$

$$Homogeneity_\theta = \frac{P_n(i,j|\theta)}{1+|i-j|} \tag{3.47}$$

$$Entropy_\theta = \sum_{i,j} -P_n(i,j|\theta) \log P_n(i,j|\theta) \tag{3.48}$$

3.3.5.3. Edge features

They refer to the relative change in pixel intensities across a frame (Ryan et al. 2014). 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 (Loy et al. 2013). 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 (Marana et al. 1999). The edge orientation histogram is used to help distinguish edges of the people with other structures in the scene such as noise (Ryan 2013).

3.3.5.4. Keypoints

They refer to specific pixels of interest in an image or video (Ryan et al. 2015). The results of using moving keypoints to find the number of people show that they have a strong relationship (Alberto Albiol et al. 2009; Antonio Albiol et al.

2009; Conte et al. 2010b; Jeong, C. Y., Choi, S., & Han 2013). Many people counting studies have been carried out using FAST, SIFT and SURF points (Saleh et al. 2015; Ryan et al. 2015). Harris corner and binary robust invariant scalable keypoints (BRISK) are also used to count people (Jeong & Choi 2016).

3.3.6 Features selection

The optimal combination of features for any environment can be selected by training the regression model with different potential combinations of features. There are only 31 potential combinations of features that can be used with features regression based people counting systems (Ryan et al. 2015). Table 3.10 shows all potential combinations of features. Multi-stage training has been used in this paper to train a regression model to find the optimal combination of features.

Table 3. 10: The potential features to be optimal (S = Size, P = Shape, E = Edges, K = Keypoints, T = Texture) (Ryan et al. 2015).

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	

3.4 Chapter Summary

CCTV cameras are already widely used for monitoring. Two people counting systems have been proposed and described in this chapter, both capable of using existing CCTV cameras to provide estimates on the number of people in a given space. In the pixel-wise optimisation-based people counting system, two algorithms were proposed using a novel combination of four techniques: motion edges, grid map, SIFT & FAST features, and pixel-wise techniques. Edge pixels is used in this system because their number in frames is smaller than the foreground pixels. SIFT and FAST features were chosen due to their high correlation with the number of people.

With the features regression-based people counting system, an adaptive and accurate people counting system was proposed and implemented, one that is capable of dynamically identifying the best set of features. Moreover, two GPR models were used to improve the accuracy. The most suitable regression model for each frame was selected depending on the level of occlusion.

Chapter 4: Experimental Results and Discussion

This chapter presents comprehensive empirical results of the proposed people counting systems. The UCSD and Mall datasets have been used to evaluate the proposed systems. The results have shown that the proposed systems achieve good results in heavily occluded environments with perspective distortions. Comparisons with the state of the art systems show that the proposed systems improve the accuracies based on MAE, MSE, and MDE metrics. In addition, sparse and crowded scenarios are used to test end evaluate the proposed systems. An attempt is made to explain the implications of these results.

4.1 Evaluation Metrics

The benchmark datasets are partitioned into a training set, for learning the proposed systems, and a test set, for validation. In the pixel-wise optimisation based people counting system, 100 frames from different locations of each dataset (Mall and UCSD datasets) are allocated individually for training and 1900 frames for testing. In the features regression based people counting system, the same training and testing partition as in (Chen et al. 2013; Chan et al. 2008; Chen et al. 2012) has been followed in the Mall and UCSD datasets, 800 frames are used for training and 1200 frames for testing.

Three metrics have been used as performance indicators for people counting; mean absolute error (MAE), mean squared error (MSE) and mean deviation Error (MDE) (Loy et al. 2013). The MAE is defined as;

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n| \tag{4.1}$$

The MSE is given as;

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2 \tag{4.2}$$

The MDE is given as;

$$MDE = \frac{1}{N}\sum_{n=1}^{N}\frac{|y_n - \hat{y}_n|}{y_n} \tag{4.3}$$

Where N is the total number of the test frames, y_n is the actual count, and $\hat{y}_n$ is the estimated count of n_{th} frames. MAE and MSE are indicative quantities of the error of the estimated people count but they contain no information about how crowded the environment is (Loy et al. 2013). MDE takes into account the crowdedness and gives an indication of how good a measurement is relative to the actual count (Hafeezallah & Abu-Bakar 2016).

4.2 Experimental Results

4.2.1 Evaluation of the proposed systems performance using the Mall dataset

As shown in Table 4.1, the MDE of the SIFT features algorithm is 0.096 and 0.092 for the SIFT-FAST features algorithm. The MDE of the features regression based people counting system is 0.095. The results are compared with results presented by other researchers for the same datasets as a measure of accuracies of the proposed systems. From the results, we can see that the accuracy of the SIFT-FAST features algorithm is better than that of

SIFT features algorithm and the features regression based people counting system. It shows that there is a reasonable improvement in the accuracy of the implemented systems when compared to those published by other researchers. Figure 4.1 shows the percentage of frames within the MDE distribution of the proposed systems. Figure 4.2 shows the true count (TC) of people from sample frames of the Mall dataset, which is annotated by red dots. EC1, EC2 and EC3 represent the estimated number of people using SIFT features algorithm, SIFT-FAST features algorithm and features regression based people counting system, respectively.

The performance of people counting systems is measured using the accuracy (MAE, MSE and MDE) and practicality. Practicality is measured by the percentage of the training frames minimisation (Ryan et al. 2009). People counting systems are practical if they are easy to deploy. In the real world, people counting systems are deployed in different environments which means they are individually trained for the location. Therefore, it is very important to reduce the number of the training frames required. The ground truth (the actual number of people) for each training frame is required when training people counting systems. Each environment needs several hundreds of frames (usually 400-800 training frames) for the training [50], [81]–[83], so the training process becomes time-consuming.

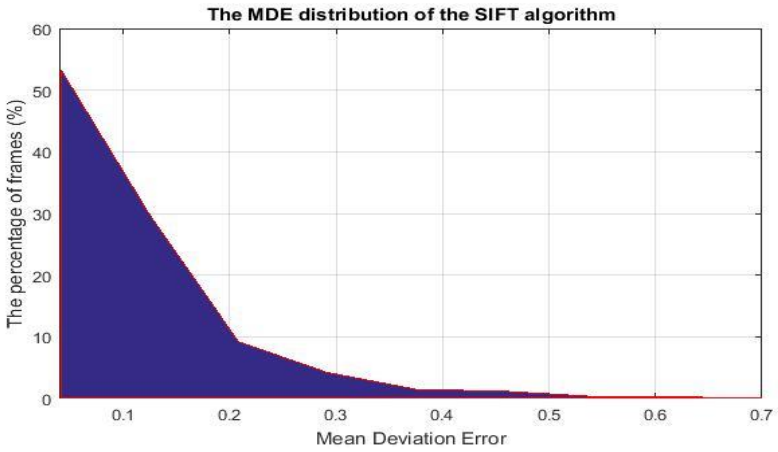
Although the comparison with recent results from other researchers for the evaluation is highly important to evaluate the proposed people counting systems, the comparison between pixel-wise optimization based method with other features regression based methods using the accuracy metrics (MAE, MSE and MDE) is not enough to measure the performance for many reasons: firstly, pixel-wise optimisation based methods can be trained using a small number of frames in comparison to regression based methods (Lempitsky & Zisserman 2010). The proposed pixel-wise optimisation based people counting methods use 100 frames for the training whilst the other state of the art methods use between 400 and 800 frames [50], [81]–[83]. In conclusion, the pixel-wise optimisation based people counting system is more practical

because the set-up time is faster by a factor of at least four (uses 4 times less training frames) compared to regression based methods which lead also to low set-up cost. Secondly, the lower number of training frames required in the training stage of the pixel-wise optimisation based algorithms reduces the potential error being introduced because manually annotation is an error-prone task. The accuracies of people counting systems are significantly affected by errors in the training stage. Thirdly, the pixel-wise optimisation based system is a multipurpose system because it can be used for people counting and also to improve the accuracy of people detection methods (Rodriguez et al. 2011).

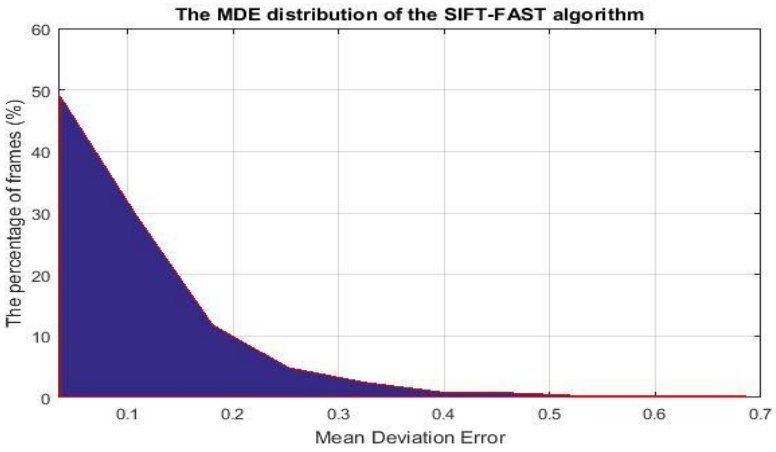
The pixel-wise optimisation based system is compared with other features regression based methods to only show that although this system reduces the training error, speed, cost and can be used to develop more accurate people detection methods, its accuracy is, at least, comparable with the state of the art methods. None of the published results presented in Table 4.1 performs better than the proposed systems based on the metrics used in this thesis. Finally, the MDE of the both proposed systems are less than the acceptable error (0.2) which is meeting the minimum accuracy requirements of system operators (Ryan et al. 2015).

Table 4. 1: Comparison of the Mall dataset results between the proposed systems and the state of the art algorithms

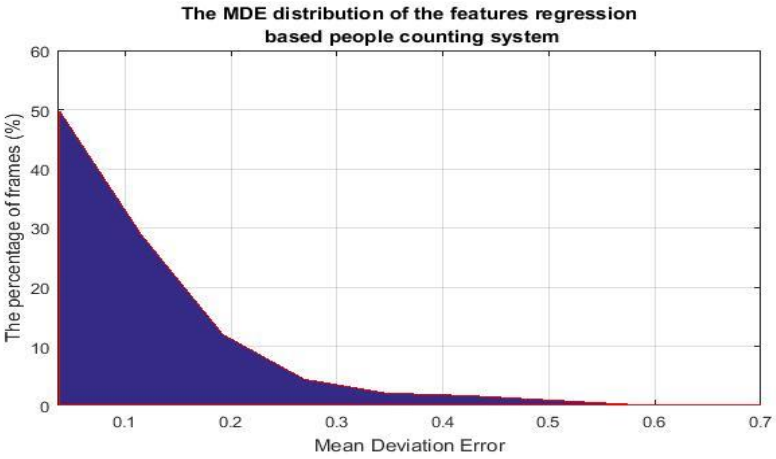
Algorithm/ System	Mall dataset		
	MAE	MSE	MDE
Algorithm 1: SIFT Features Algorithm	**2.96**	**15.30**	**0.096**
Algorithm 2: SIFT-FAST Features Algorithm	**2.83**	**13.92**	**0.092**
Features Regression Based People Counting System	**2.90**	**13.62**	**0.095**
Cost-sensitive Sparse Linear Regression (CS-SLR) (Huang et al. 2016)	3.23	15.77	0.104
Cumulative attribute based model (CA-RR) (Chen et al. 2013)	3.43	17.70	0.105
Gaussian Process Regression (GPR) (Chen et al. 2012; Chen et al. 2013)	3.72	20.10	0.115
Kernel Ridge Regression (KRR) (Chen et al. 2013)	3.51	18.10	0.108
Multi Output Ridge Regression (MORR) (Chen et al. 2012)	3.15	15.70	0.099
Multiple Localised Regression (MLR) (Chen et al. 2012)	3.90	23.90	0.119
Random Forest Regression (RFR) (Chen et al. 2013)	3.91	21.50	0.121
Random Projection Forest (RPF) (Xu & Qiu 2016)	3.22	15.50	-
Ridge regression (RR) (Chen et al. 2012; Chen et al. 2013)	3.59	19.00	0.110
Squares Support Vector Machine Regression (LSSVR) (Chen et al. 2013)	3.51	18.20	0.108
Weighted Ridge Regression (WRR) (Chen & Kamarainen 2014)	3.44	18.00	0.105



(a)

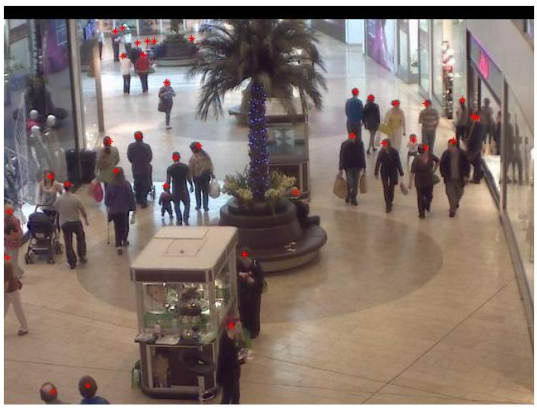


(b)

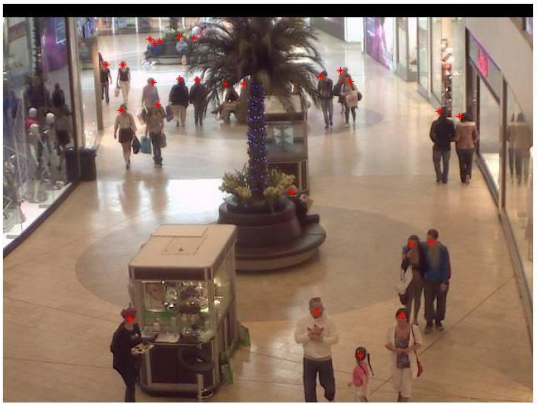


(c)

Figure 4. 1: The MDE distribution of the proposed systems (Mall dataset).



(a) TC= 36, EC1= 38, EC2= 37, EC3= 38



(b) TC= 26, EC1= 29, EC2= 25, EC3=28

(a) TC = 19, EC1= 20, EC2= 20, EC3= 20

(b) TC = 29, EC1= 32, EC2= 28, EC3=31

Figure 4. 2: Examples of the true count (TC) & the estimated count of people using SIFT algorithm (EC1), SIFT-FAST algorithm (EC2) and features regression based crowd counting system (EC3).

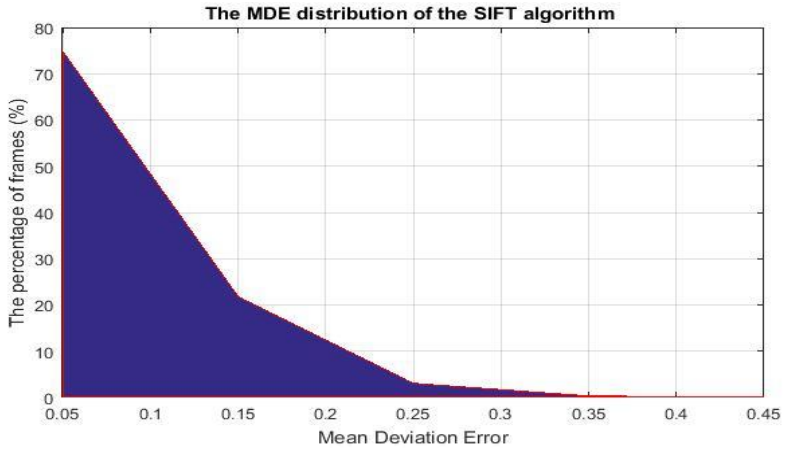
4.2.2 Evaluation of the proposed systems performance using the UCSD dataset

The UCSD dataset represents people moving in two directions along a walkway. As shown in Table 4.2, the MDE of the SIFT features algorithm is 0.065 and 0.064 for the SIFT-FAST features algorithm. The MDE of the features regression based people counting system is 0.066. From the results, it can be seen that the accuracy of SIFT-FAST features algorithm is better than that of SIFT features algorithm and the features regression based people counting system. Figure 4.3 shows the percentage of frames within the MDE distribution of the algorithms. Figure 4.4 shows the true count (TC) of people from sample frames of the UCSD dataset, which is annotated by red dots. In general, the accuracies of the proposed systems with the UCSD dataset are better than the results from the Mall dataset. The potential justification is that the Mall dataset is more complicated in terms of shadows, reflections and crowd size (Saleh et al. 2015; Ryan et al. 2015). In addition, the Mall dataset is collected with more severe perspective distortion than the UCSD dataset. As in the case with the MDE from the Mall dataset, the MDE of this dataset is significantly lower than the acceptable error (0.2) which is meeting the minimum accuracy requirements of system operators (Ryan et al. 2015).

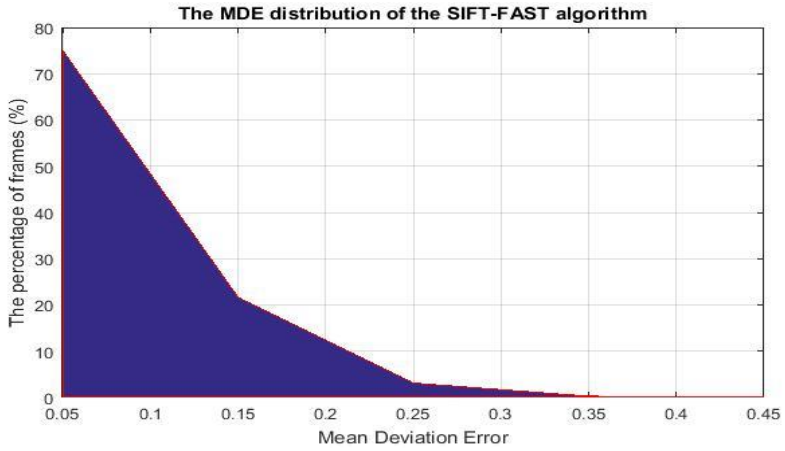
Results of both datasets show that the average accuracy of the SIFT-FAST algorithm is better than the SIFT algorithm. The EC1 and EC2 of the pixel-wise optimisation based algorithms at each frame are correlative because both algorithms use almost the same approach. However, FAST corner points are used with SIFT-FAST features algorithm to improve the accuracy due to the high correlation between the number of people and FAST corner points. The proposed systems give the best or comparable results compared to the published results presented.

Table 4. 2: Comparison of the UCSD dataset results between the proposed systems and the state of the art algorithms.

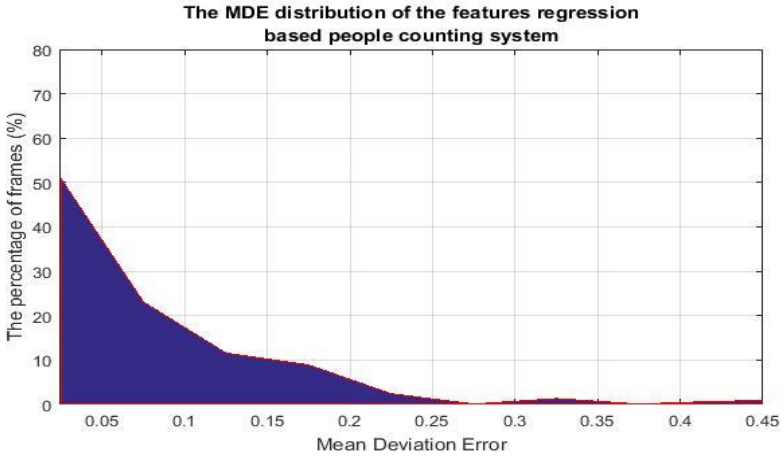
Algorithm/ System	UCSD dataset		
	MAE	MSE	MDE
Algorithm 1: SIFT Features Algorithm	**1.78**	**5.18**	**0.065**
Algorithm 2: SIFT-FAST Features Algorithm	**1.75**	**5.01**	**0.064**
Features Regression Based People Counting System	**1.63**	**4.32**	**0.066**
Improved Iterative Scaling -Label Distribution Learning (IIS-LDL) (Z. Zhang et al. 2015)	2.08	7.25	0.098
Kernel Ridge Regression (KRR) (Z. Zhang et al. 2015)	2.16	7.45	0.107
Random Forest Regression (RFR) (Z. Zhang et al. 2015)	2.42	8.47	0.116
Gaussian Process Regression (GPR) (Z. Zhang et al. 2015; Chen & Kamarainen 2014)	2.30/ 2.24	8.21/ 7.97	0.114/ 0.112
Ridge Regression (RR) (Z. Zhang et al. 2015; Chen & Kamarainen 2014)	2.25	7.82	0.110
Multi Output Ridge Regression (MORR) (Z. Zhang et al. 2015)	2.29	8.08	0.109
Cumulative attribute based model (CA-RR) (Chen et al. 2013; Z. Zhang et al. 2015)	2.07	6.86	0.102
Weighted Ridge Regression (WRR) (Chen & Kamarainen 2014)	2.05	6.75	0.102
Linear regression (LR), Partial Least Squares Regression (PLSR), KRR, LSSVR, GPR and RFR (Loy et al. 2013)	>2.02	>6.67	>0.100
Random Projection Forest (RPF) (Xu & Qiu 2016)	1.90	6.01	-
Cost-sensitive Sparse Linear Regression (CS-SLR) (Huang et al. 2016)	1.83	5.04	0.079
Moving SIFT algorithm (Conte et al. 2013)	3.26	-	0.180



(a)



(b)



(c)

Figure 4. 3: The MDE distribution of the proposed systems (UCSD dataset).

(a) TC = 18, EC1= 19, EC2= 18, EC3= 19

(b) TC = 23, EC1= 22, EC2= 24, EC3=23

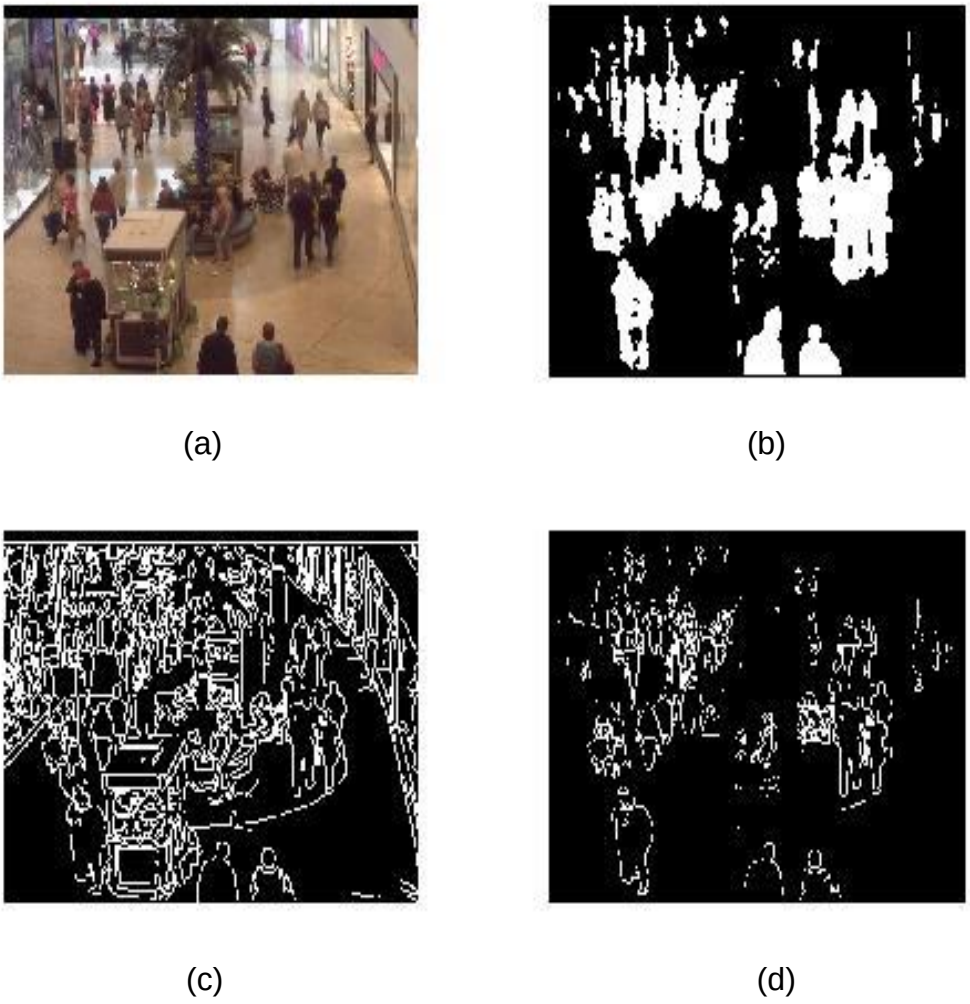
(a) TC = 15, EC1= 15, EC2= 17, EC3= 15

(b) TC = 23, EC1= 21, EC2= 22, EC3=21

Figure 4. 4: Examples of the true count (TC) & the estimated count of people using SIFT algorithm (EC1), SIFT-FAST algorithm (EC2) and features regression based crowd counting system (EC3).

4.2.3 Background subtraction, edge detection and motion edge extraction

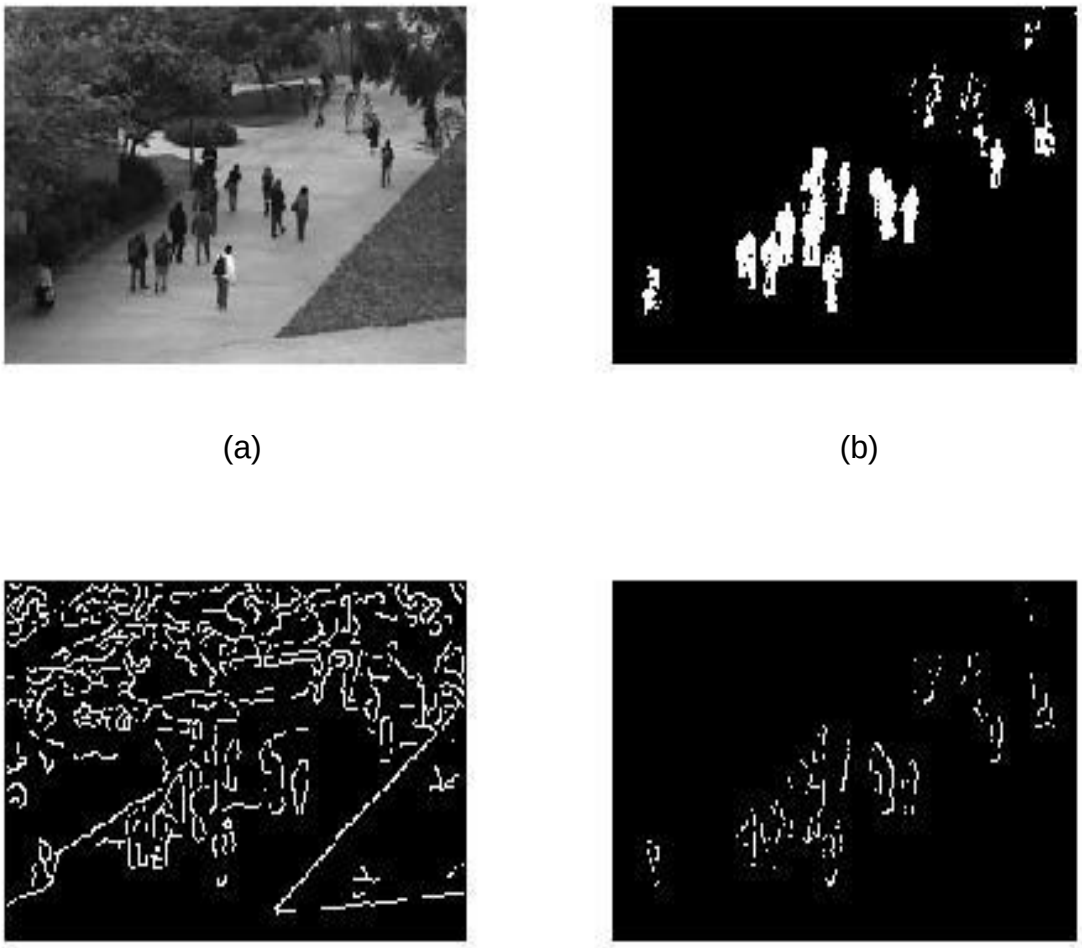
The GMM is used for background subtraction and the Canny edge algorithm is performed to extract the edges of the frames. The logical 'AND' is used to extract motion edge in the pixel-wise optimisation based system. Figures 4.5 and 4.6 show the results of the background subtraction, edge detection and motion edge extraction of two sample frames, one from the Mall dataset and the second from the UCSD dataset.



(a) (b)

(c) (d)

Figure 4. 5: (a) An example of the Mall dataset; (b) foreground, using GMM algorithm; (c) edge using Canny detector; (d) the motion edge, using logical 'AND'.



(a) (b)

(c) (d)

Figure 4. 6: (a) An example of the UCSD dataset; (b) foreground, using GMM algorithm; (c) edge using Canny detector; (d) the motion edge, using logical ‘AND’.

4.2.4 Features selection of the features regression based people counting system

The relationship between the datasets and their optimal combination of features is fuzzy and there are 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 types of features can be used with features regression based people counting systems, e.g. foreground (shape and size), texture, edge and keypoints. Each type of features may be more appropriate for the particular type of environments. 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 (Ryan et al. 2015). Texture features can achieve high performance in environments with high textured backgrounds (Ryan et al. 2015). Some keypoints features are more appropriate for the high perspective distortion environments 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 are 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 (Saleh et al. 2015). Mall dataset has a high complicated background, shadows and reflections than the UCSD dataset (Ryan et al. 2015). In addition, using different kinds of features can help to mitigate the non-linearities that arise from occlusion, segmentation errors and pedestrian configuration (Chan et al. 2008).

4.2.5 Threshold selection of the features regression based people counting system

To study the effect of different choices of threshold on the efficiency of the proposed system, the results of each stage of the proposed multi-stage thresholding method are shown in Table 4.3 and 4.4. 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 the highest accuracy of the stages is selected to be the best threshold. Nine and ten thresholds are shown at the first and second stages, respectively. The best threshold for each dataset is marked in bold.

Table 4. 3: The threshold selection (Mall dataset).

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 4. 4: The threshold selection (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

4.2.6 Performance evaluation in sparse and crowded scenarios

To evaluate the proposed systems with sparse and crowded scenarios, the test set of the Mall dataset is split the same as in [8] into a sparse set which includes all the frames with ground truth (number of people), less than or equal to 30, and crowded set which includes all the frames with ground truth values greater than 30. The test set of the UCSD dataset is also split the same as in [8] into a sparse set which includes all the frames that their ground truth is less than or equal to 23, and crowded set which includes all the frames that their ground truth is greater than 23.

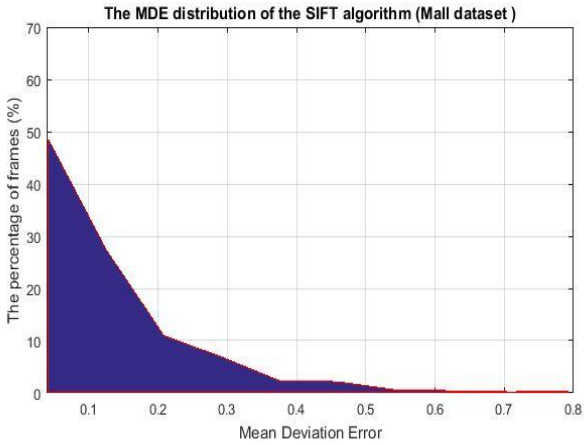
To ensure that the proposed systems are practical and robust, the training set was not been split because the technical definition of the boundary that separates the sparse and crowded frames is not clear (Zeyad Q.H. Al-Zaydi et al. 2016). In addition, partitioning the training set into two sets would be required two training stages. The test sets are processed by the proposed systems jointly and then the results are analysed by splitting them into sparse and crowded sets. In conclusion, the split between sparse and crowded scenarios have mainly been carried out by identifying which frames could be classified into each of the categories. No differential training of the systems has been carried out. Tables 4.5 and 4.6 show the results of both systems with sparse and crowded scenarios. The MDE of both systems in the sparse scenarios are higher than the MDE crowded scenarios. The proposed systems are more applicable for high-density crowds and this can be seen from the achieved good results in crowded scenarios. This opens the door for using the proposed systems in high crowded environments. Figures 4.7 and 4.8 show the percentages of frames within the MDE distribution for the sparse and crowded scenarios based on the Mall and UCSD datasets, respectively.

Table 4. 5: Systems performance with sparse and crowded scenarios (Mall dataset).

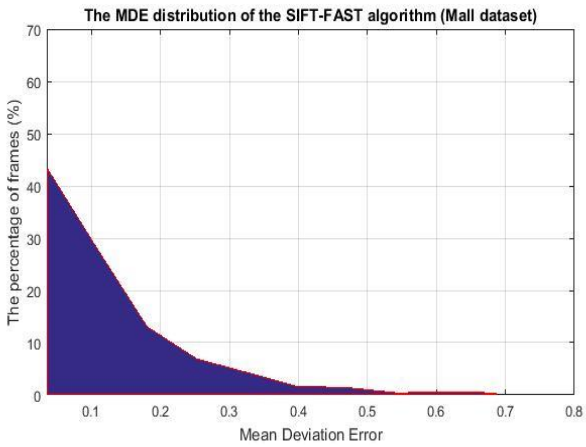
Algorithm	Sparse scenario			Crowded scenario		
	MAE	MSE	MDE	MAE	MSE	MDE
SIFT features Algorithm	3.06	16.90	0.120	2.86	13.62	0.078
SIFT-FAST features Algorithm	2.89	14.98	0.114	2.77	12.81	0.076
Features regression based people counting system	2.93	14.46	0.118	2.86	12.69	0.078

Table 4. 6: Systems performance with sparse and crowded scenarios (UCSD dataset).

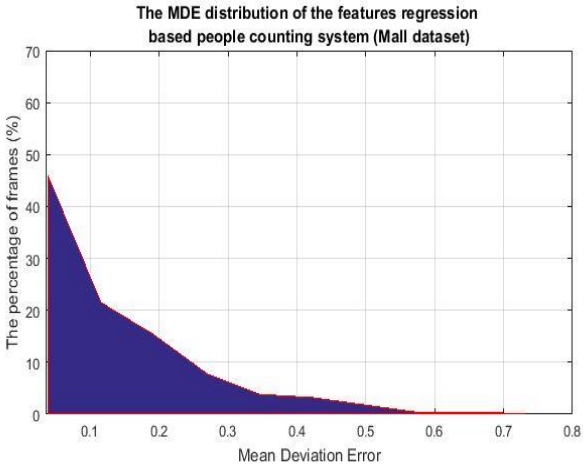
Algorithm	Sparse scenario			Crowded scenario		
	MAE	MSE	MDE	MAE	MSE	MDE
SIFT features Algorithm	1.74	4.95	0.089	1.81	5.35	0.055
SIFT-FAST features Algorithm	1.74	4.99	0.089	1.75	5.02	0.053
Features regression based people counting system	1.65	4.43	0.075	1.52	3.78	0.040



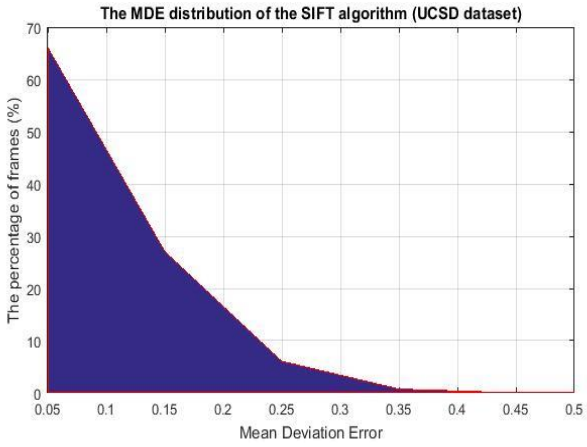
(a)



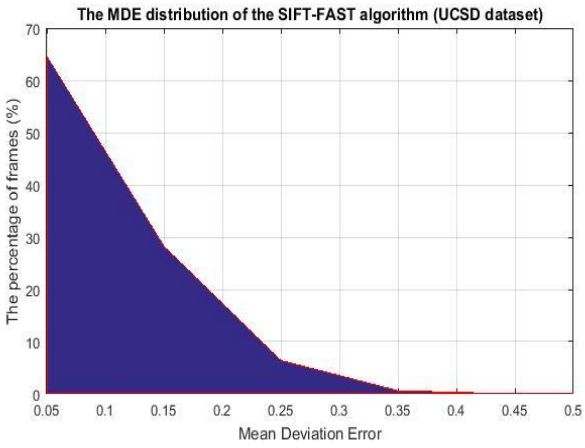
(b)



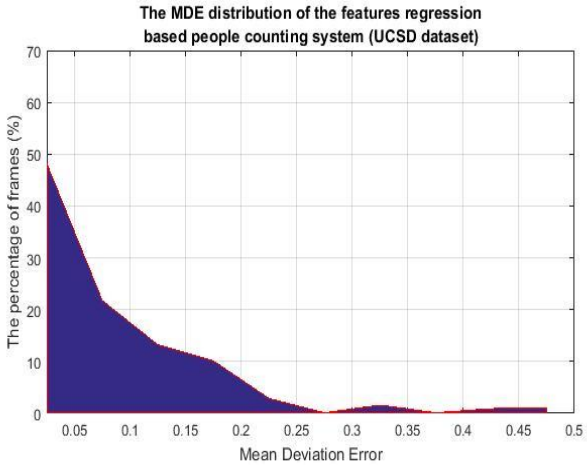
(c)



(d)

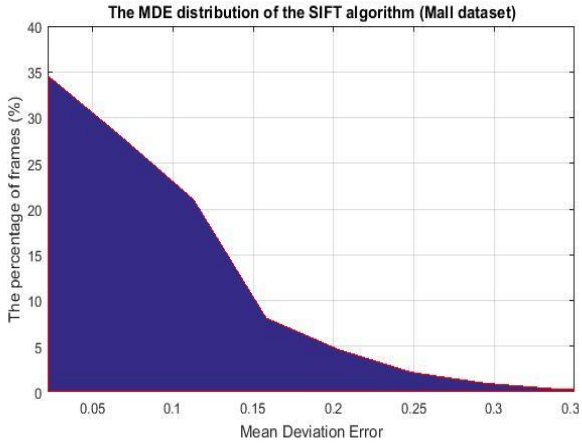


(e)

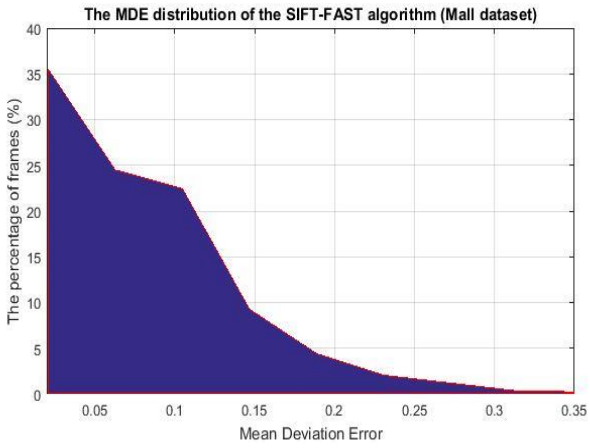


(f)

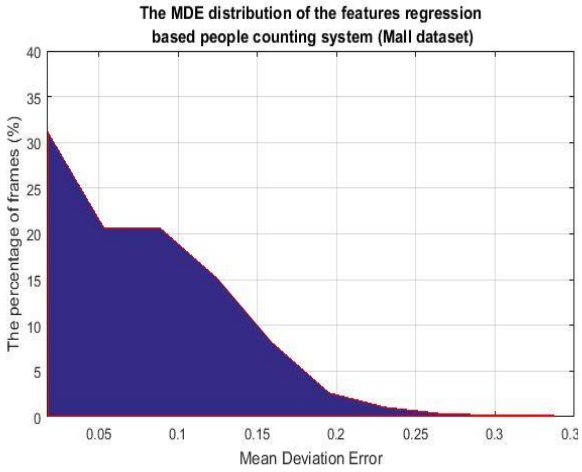
Figure 4. 7: The MDE distribution of the proposed systems in sparse scenarios; (a), (b) and (c) in the mall dataset; (d), (e) and (f) in the UCSD dataset.



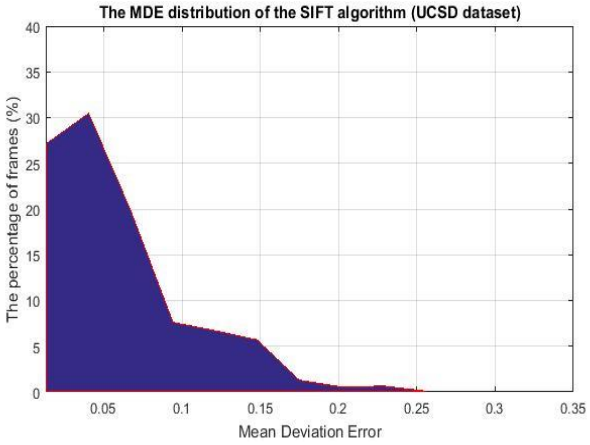
(a)



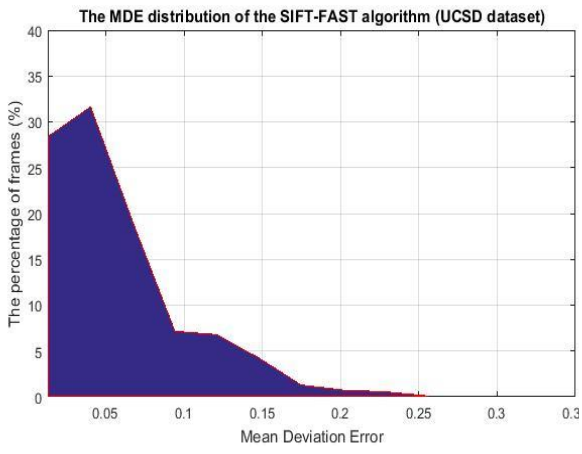
(b)



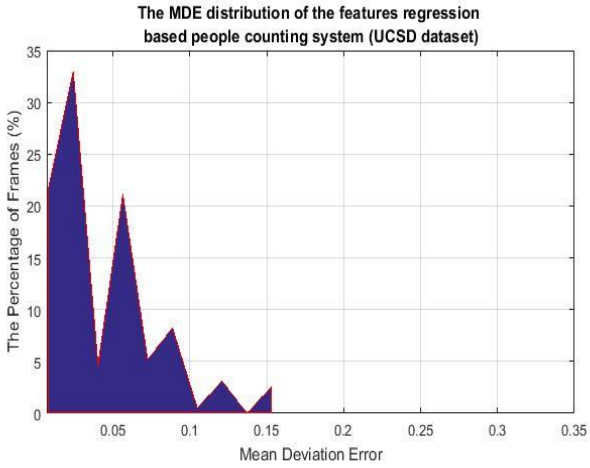
(c)



(d)



(e)



(f)

Figure 4. 8: The MDE distribution of the proposed systems in sparse scenarios; (a), (b) and (c) in the mall dataset; (d), (e) and (f) in the UCSD dataset.

4.2.7 Computation efficiency evaluation

People counting based on ROI method is slow to change over time. Calculating the number of people every second is more than sufficient for the reliable people counting systems (Siva et al. 2016). The proposed systems are implemented using MATLAB software and it is running on a PC with 3.2 GHz core I5 processor and 8 GB memory.

To study the computational efficiency of the proposed features regression based people counting system, a comparison between the classical GPR algorithm and the proposed system is carried on. The classical GPR algorithm is chosen because it uses the same regression model which makes the comparison more reliable. The comparison with the classical GPR algorithm that uses the same regression model will also show the extra computational complexity of the features regression based people counting 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 and software. 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 4.7, the processing speeds of the proposed regression based people counting system are 20.761 fps and 38.471 fps for the MALL and UCSD datasets, respectively. The difference in the processing speeds between the classical GPR method and the proposed regression based people counting system is extremely small.

The processing speeds of the pixel-wise optimisation based people counting system are about 1 fps and 2 fps for the MALL and UCSD datasets, respectively. However, they are significantly lower than the processing speeds of the regression based people counting system, they are sufficient for achieving reliable people counting system.

Table 4. 7: The computation efficiency of the proposed systems.

Algorithm	Processing speed (fps)	
	Mall	UCSD
SIFT features Algorithm	0.928	2.352
SIFT-FAST features Algorithm	0.925	2.337
Features regression based people counting system	20.761	38.471
Classical Gaussian Process Regression (GPR)	20.882	38.701

4.3 Chapter Summary

This chapter presented the experimental results of the proposed systems. The UCSD and Mall datasets have been used to test and evaluate them. The results have shown that the proposed systems achieve good results in heavily occluded environments with perspective distortions. By means of comparisons with other existing low-level features regression methods, our results demonstrate the ability of the proposed systems to outperform the others methods with respect to MAE, MSE and MDE metrics. Experimental results of the proposed systems in sparse and crowded scenarios shows that they perform better in crowded environments.

The computational efficiency results of the proposed systems show that the processing effect of the extra steps of regression based people counting system is extremely small. In addition, it is significantly faster than pixel-wise optimisation based people counting system. The effects of the features and threshold selection on the accuracy of the regression based people counting system have also been presented and discussed.

On the other hand, the proposed pixel-wise optimisation based people counting system is more practical than low-level features regression based methods because it can be trained with a lower number of frames so it is relatively easy to deploy. In addition, it reduces the training error, speed, cost and, opens the door to developing more accurate people detection methods.

Chapter 5: Performance Evaluation in A Challenging Environment

A crowded and complicated environment is used to evaluate the performances of the proposed people counting systems in challenging scenarios. The New York Grand Central Station dataset is selected for that purpose. The results have shown that the proposed systems can achieve good performance in that heavily crowded environment. The MAE, MSE and MDE metrics are used to measure and evaluate the performances of the proposed systems. A comprehensive discussion of the results is also presented to explain the implications of these results.

5.1 Data Description and Experimental Setup

The New York Grand Central dataset was introduced by Zhou (Zhou et al. 2012). It is used for understanding the crowd behaviours by observing the movement trajectories of people. In this work, a region of interest (ROI) is used to specify the same sub-region of all frames, which is only processed for people counting. A binary mask that is the same size as the frames of the dataset is used. In this mask, the pixels that represent the ROI are assigned a value of 1 and a value of 0 for all pixels outside the ROI. Figure 5.1 shows a sample frame from this dataset with the ROI. The red colour mask represents the non-ROI of the sample frame.

Table 5.1 shows the characteristics of the New York Grand Central dataset. It is a video dataset and contains a very large number of frames (46009 frames) and its crowd size (the number of people) changes quite significantly, between 125 and 245 people (Ryan et al. 2015).

The Mall, UCSD and New York Grand Central datasets have been chosen to prove the efficiency and robust of the proposed systems because they cover a wide range of variation of characteristics. They cover a variation of frame rate (fps), resolution, colour, location, shadows, loitering, reflections, crowd size and frame type (Saleh et al. 2015; Ryan et al. 2015).

Figure 5. 1: Sample frame from the New York Grand Central Station dataset with the ROI.

Table 5. 1: The features of the New York Grand Central dataset.

	New York Grand Central dataset
Year	2012
Length (frames)	46009
Frame rate (fps)	23
Resolution	720 x480
Colour	Grey
Location	Indoor
Shadows	No
Reflections	Yes
Loitering	Yes
Crowd size	125-245
Frame type	AVI file

Annotation of all frames in this dataset is not feasible because of the extremely large size of the crowd and the substantial number of frames. A subset of frames has been selected from different locations of this dataset. Five hundred frames are used for the training and testing of the proposed systems. Ten sequences of length 50 frames are selected over a long period of time (33 min) as in (Ryan et al. 2015). Table 5.2 shows the selected subset of frames with the crowd size.

Table 5. 2: The selected subset of frames and the crowd size.

Selected subset	Crowd size
951-1000	132
5951-6000	152
10951-11000	151
15951-16000	160
20951-21000	125
25951-26000	138
30951-31000	141
35951-36000	176
40951-41000	200
45951-46000	245

In the pixel-wise optimisation based people counting system, 100 frames from different locations of New York Grand Central dataset are allocated individually for training and 400 frames for testing. In the features regression based people counting system, a larger number of training frames is required to maintain the accuracy of counting, therefore 300 frames are used for training and 200 frames for testing.

5.2 Experimental Results of Optimisation Based People Counting System

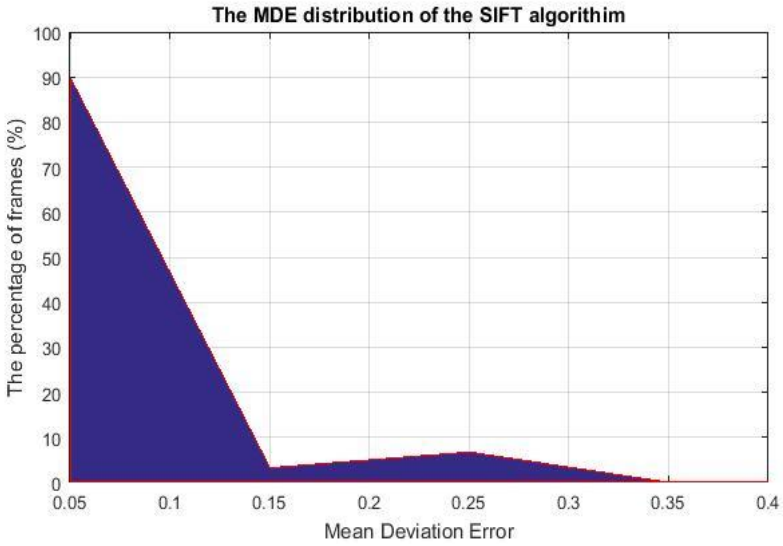
The New York Grand Central dataset is used to test and evaluate the two algorithms of this people counting system. Three metrics are used to measure the performance of the proposed algorithms; MAE, MSE and MDE. The MAE, MSE and MDE of the SIFT features algorithm are 8.74, 167.06 and 0.056, respectively while they are 7.85, 135.32 and 0.051 for the SIFT-FAST features algorithm, respectively. From the results, it can be seen that the accuracy of the SIFT-FAST features algorithm is better than that of SIFT features algorithm. To preserve the practicality by making this system easy to deploy, the same number of the training frames as with the Mall and UCSD datasets is used. One hundred frames are used for the training this system and 400 frames is used for testing.

The results of the proposed system do not compare with the results of the state of the art methods because this dataset is rarely used by the researchers of this field to test and evaluate their people counting systems. In addition, the experimental setup and counting method that are used by those researchers are totally different, making the comparison unreliable. The potential reasons that the researchers avoid using the New York Grand Central dataset are that this dataset is considered the most difficult and challenging dataset due to the large crowd size and the high video resolution. In addition, the reflection and loitering introduce noise that affects the background subtraction and create noise in the extracted features. The camera setting is another factor that increases the complexity of this dataset, because the camera is installed in a high place to cover a large area of monitoring, which decreases the quality of the dataset. This can affect the quality of the extracted features and thereby decrease the accuracy of counting.

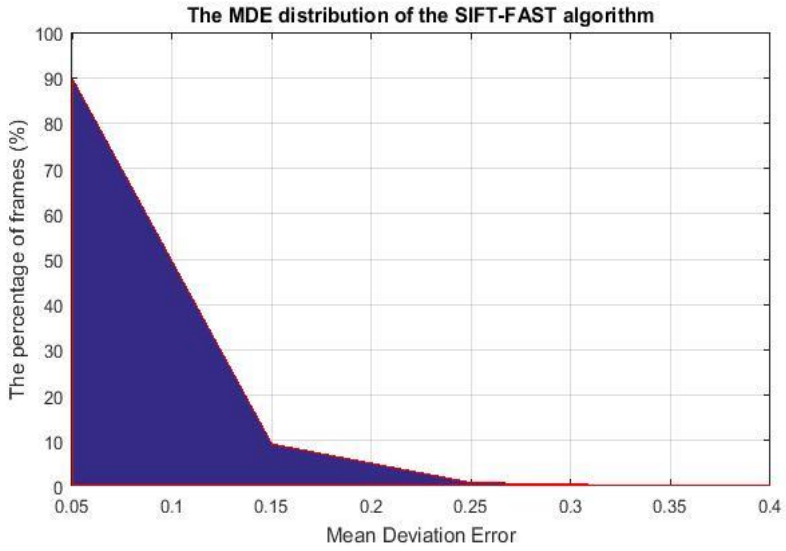
Testing and evaluating any people counting system with this dataset and achieving a high performance can be considered as the ultimate aim to prove

the efficiency of people counting due to the complexity of this dataset. The high performance of the proposed system that achieved using this challenging dataset shows the robust and efficiency of the proposed system. Finally, the MDE of the proposed system is less than the acceptable error (0.2) which is meeting the minimum accuracy requirements of system operators (Ryan et al. 2015).

Figure 5.2 shows the percentage of frames within the MDE distribution of the proposed system. Figure 5.3 shows the true count (TC) of people from sample frames of the New York Grand Central dataset, which is annotated by red dots. EC1 and EC2 represent the estimated number of people using SIFT features and SIFT-FAST features algorithms, respectively.



(a)



(b)

Figure 5. 2: The MDE distribution of the optimisation based people counting system using the New York Grand Central dataset.

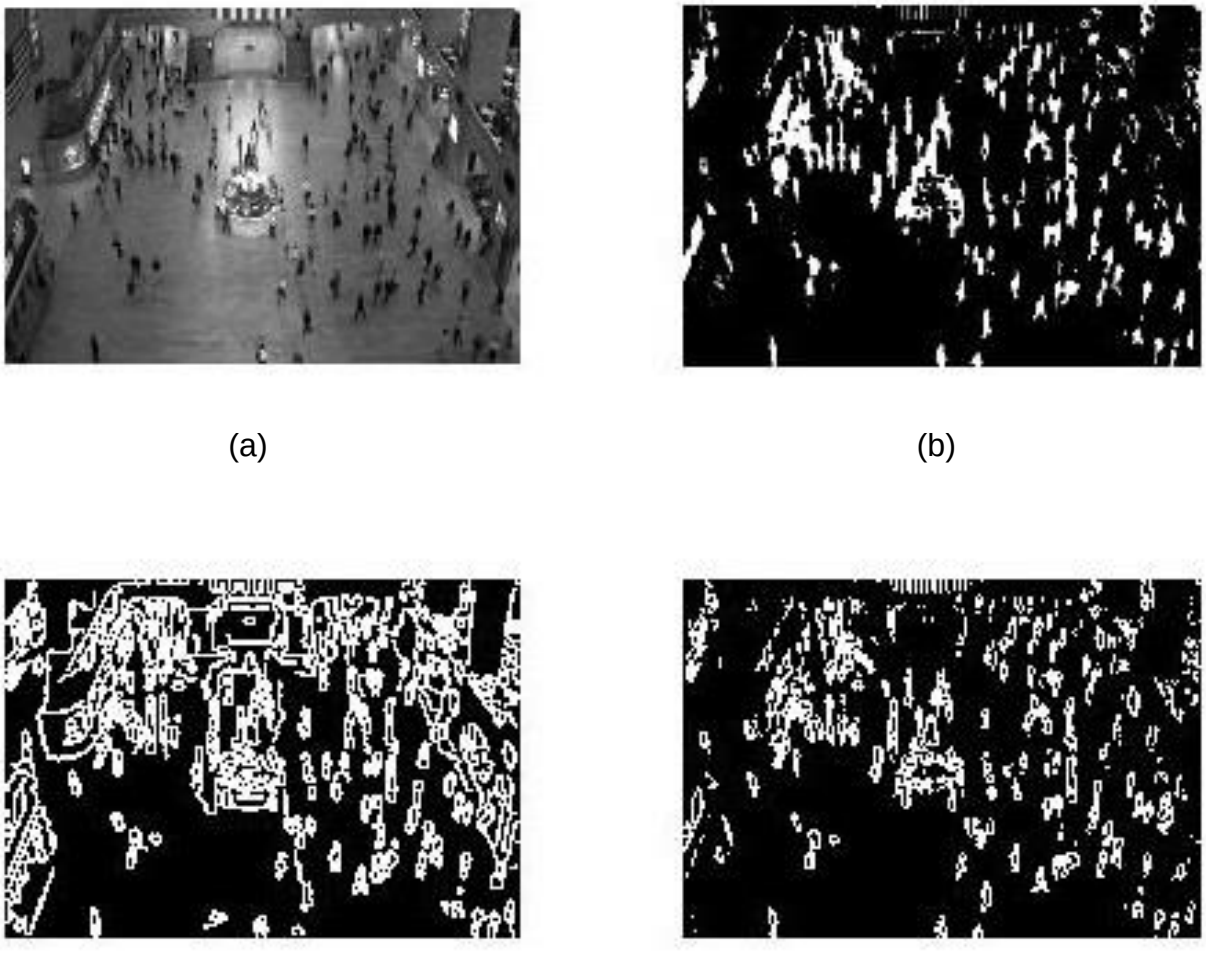
(a) TC = 125, EC1= 125, EC2= 121

(b) TC = 152, EC1= 154, EC2= 152

Figure 5. 3: Examples of the true count (TC) & the estimated count of people using SIFT algorithm (EC1) and SIFT-FAST algorithm (EC2).

5.2.1 Background subtraction, edge detection and motion edge extraction

As with the mall and UCSD datasets, the GMM is used for foreground detection and the Canny edge algorithm is performed to extract the edges of the frames. The motion edges are calculated using the logical 'AND' between the foreground information and detected edges. Figures 5.4 shows the results of the background subtraction, edge detection and motion edge extraction of a sample frame from the New York Grand Central dataset.



(a) (b)

(c) (d)

Figure 5. 4: (a) An example of the New York Grand Central dataset; (b) foreground, using GMM algorithm; (c) edge using Canny detector; (d) the motion edge, using logical 'AND'.

5.3 Experimental Results of Features Regression Based People Counting System

The MAE, MSE and MDE of the features regression based people counting system are 4.41, 25.62 and 0.029, respectively. From the results, it can be seen that the accuracy of this system is better than that of the pixel-wise optimisation based people counting system. Figure 5.5 shows the percentage of frames within the MDE distribution of the proposed system. Figure 5.6 shows some frames from this dataset with their true number of people (TC), which are annotated with red dots, and the estimated number of people (EC) using features regression based people counting system. Finally, the MDE of the proposed system is significantly less than the acceptable error (0.2) which is meeting the minimum accuracy requirements of system operators (Ryan et al. 2015).

The appropriate combination of features with the highest accuracy for this dataset is SPEKT, where S = Size, P = Shape, E = Edges, K = Keypoints and T = Texture. The potential justification for this selection is that using different types of features can help to mitigate the non-linearities that arise from occlusion, segmentation errors and pedestrian configuration (Chan et al. 2008). However, this dataset is high crowded and occluded, it is less complicated than Mall dataset in terms of complicated background and reflections and it does not include shadows. As a consequence, edge features are used with this dataset to achieve the best accuracy while they are not used with the Mall dataset.

Evaluating the performance of the proposed systems with sparse and crowded scenarios is not reasonable because all the frames of this dataset are high crowded. In comparison to the Mall and UCSD datasets, the lower crowded frame in this dataset (125 people) contains a higher number of people than the high crowded frame in the Mall and UCSD datasets (53 people).

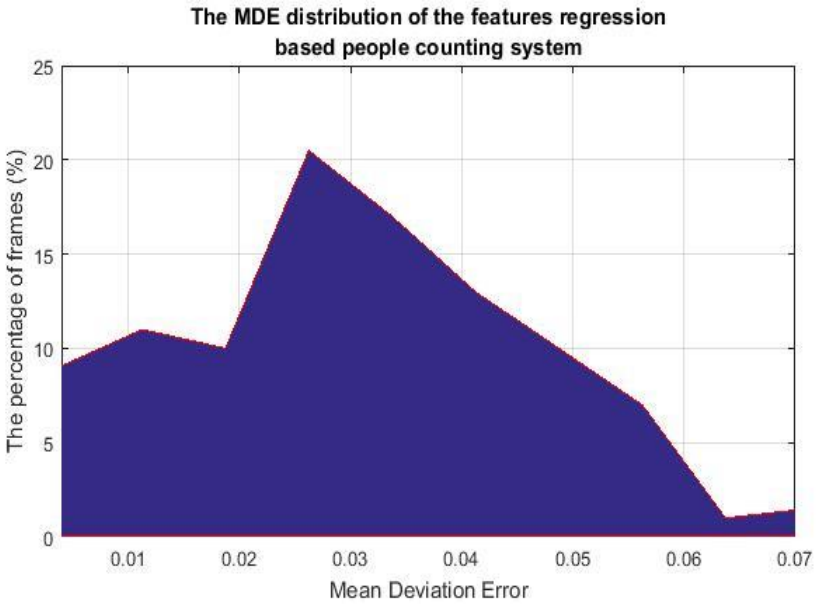


Figure 5. 5: The MDE distribution using the New York Grand Central dataset.

(a) TC = 132, EC= 134

(b) TC = 152, EC= 157

Figure 5. 6: Examples of the true count (TC) & the estimated count of people using features regression based crowd counting system.

5.3.1 Threshold selection

To study the effect of different choices of the threshold on the efficiency of the features regression based people counting system. Table 5.3 shows the results of the New York Grand Central dataset at each stage. The multi-stage thresholding method is also used with this dataset by selecting the best threshold experimentally. 19 potential thresholds are used and the threshold with the highest accuracy of the stages is selected to be the best threshold. Nine and ten thresholds are shown at the first and second stages, respectively. The best threshold for each dataset is marked in bold.

Table 5. 3: The threshold selection of the New York Grand Central dataset.

Threshold	Stage 1			Threshold	Stage 2		
	MAE	MSE	MDE		MAE	MSE	MDE
0.1	6.5100	132.6900	0.0437	0.35	4.7300	30.5500	0.0317
0.2	5.4250	37.6350	0.0364	0.36	4.6300	28.7000	0.0311
0.3	4.9500	33.5900	0.0332	0.37	4.6700	29.3200	0.0313
0.4	4.7600	29.1900	0.0320	0.38	4.7000	30.1300	0.0315
0.5	5.1350	36.4250	0.0345	0.39	4.7650	30.3350	0.0320
0.6	7.0550	68.5550	0.0474	0.41	4.6350	28.1950	0.0311
0.7	6.9150	102.1250	0.0464	0.42	4.5900	27.5800	0.0308
0.8	6.8250	141.3350	0.0458	**0.43**	**4.4150**	**25.6250**	**0.0296**
0.9	6.9850	143.9250	0.0469	0.44	4.8050	32.8750	0.0323
-	-	-	-	0.45	4.6700	30.6800	0.0313

5.4 Computation efficiency evaluation

Counting the number of people every second is more than sufficient for the reliable people counting systems (Siva et al. 2016). In addition, it is also reliable for many applications and environments to count the number of people every few seconds such as shopping malls and grand stations because the estimated number of people using ROI methods is usually slow to change over time. The proposed systems are implemented using MATLAB software and it is running on a PC with 3.2 GHz core I5 processor and 8 GB memory.

As with the Mall and UCSD dataset, a comparison between the classical GPR method and the proposed regression based people counting system is carried on. As shown in Table 5.4, the processing speed of the proposed regression based people counting system is 19.23 fps and 19.79 fps for the classical GPR method. The difference in the processing speeds between them is very small. The processing speed of the pixel-wise optimisation based people counting system is about 0.5 fps. However, it is significantly lower than the processing speed of the regression based people counting system, it is sufficient for achieving a reliable people counting system for grand station environments.

Table 5. 4: The computation efficiency of the proposed systems.

Algorithm	Processing speed (fps)
SIFT features Algorithm	0.529
SIFT-FAST features Algorithm	0.521
Features regression based people counting	19.230
Classical Gaussian Process Regression (GPR)	19.791

5.5 Chapter Summary

This chapter presented the experimental results of the proposed systems using the New York Grand Central dataset. This dataset is one of the most challenging datasets of the people counting datasets. Data is recorded in a highly crowded environment using a high video resolution. In addition, the reflection, loitering, and camera setting increase the complexity.

The results have shown that the proposed systems achieve very good results, especially the features-regression-based people counting system. The computational efficiency results of the proposed systems show that the processing effect of the extra steps of the regression-based people counting system is extremely small and can be neglected. On the other hand, the proposed pixel-wise optimisation-based people counting system is more practical than the features-regression-based people counting system because it has been trained with a small number of frames, so it is relatively easier to deploy. Therefore, it reduces the training error, set-up speed, and cost.

Chapter 6: Conclusion and Future Work

This chapter brings together all the research work presented in this thesis. In particular, it summarised the conclusions, identifies the author's key contributions and proposes future work.

6.1 Conclusions

1- This research focused on people counting aiming to efficiently utilise CCTV cameras for that purpose.

2- Two people counting systems have been presented throughout this thesis: pixel-wise optimisation based and features regression based people counting systems.

3- Each system works independently to count people and may be more appropriate for particular scenarios. The pixel-wise optimisation based people counting system is easier to deploy (better practicality), so it is more appropriate for large distribution surveillance systems. The processing speed of features regression based people counting is higher and it is less practical, so it is more appropriate for small distribution surveillance systems.

4- The results show the efficiency of the proposed systems in comparison to state-of-the-art people counting methods in terms of MAE, MSE, and MDE. The MAE, MSE and MDE of the proposed systems are 2.83, 13.92 and 0.092, respectively, for the Mall dataset; 1.63, 4.32, and 0.066, respectively, for the UCSD dataset; and 4.41, 25.62, and 0.029, respectively, for the New York Grand Central dataset.

5- The processing speeds of the proposed systems depend on the efficiency of hardware and the programming language. The proposed

systems are implemented using MATLAB software and run on a PC with 3.2 GHz core I5 processor and 8 GB memory. The computational efficiency results of the proposed systems are 20.76 fps, 38.47fps and 19.23 fps for the Mall, UCSD, and New York Grand Central datasets, respectively.

6- The accuracy of the proposed systems in the sparse scenarios are lower than the accuracy in the crowded scenarios. The proposed systems are more applicable for high-density crowds and this can be seen from the achieved good results in the crowded scenarios. This opens the door for using the proposed systems in high crowded environments.

7- The importance of this research is not restricted to one application- it is important for many applications e.g. of safety, security, transport, energy management and business intelligence applications.

6.2 Contributions of This Study

6.2.1 Contributions of the proposed pixel-wise optimisation based people counting system

This system used optimisation techniques for people counting. It is a combination of two algorithms and based on the estimation of the density of each pixel in each frame for counting people. SIFT features and clustering were used to represent pixels in the SIFT algorithm whereas FAST corner points with SIFT features were used in the SIFT-FAST algorithm. SIFT and FAST features have been selected due to their high correlation with the number of people. Three datasets have been used to test and evaluate the proposed system. The results have shown that the SIFT-FAST algorithm achieved better results than the SIFT algorithm in all datasets. This proves that the new combination of SIFT and FAST features with pixel-wise technique

improves the performance of people counting. The results show that the both algorithms offer a higher accuracy when compared with some existing low-level features regression based methods. The MAE, MSE and MDE for the proposed algorithms are less than or equal to 2.96, 15.3 and 0.096, respectively, for the Mall dataset and 1.78, 5.18 and 0.065, respectively, for the UCSD dataset. They are less than or equal to 8.74, 167.06 and 0.056, respectively, for the New York Grand Central station dataset.

Edge pixels are used instead of foreground pixels in this system to reduce the number of SIFT descriptors required. However, the number of SIFT descriptors required is decreased, it is difficult for quadratic programming to be used to find the density for all SIFT descriptors (equal to the number of edge motion pixels). To solve this problem, clustering is used to reduce the number of SIFT descriptors to 256 clusters.

A combination of grid map and pixel-wise technique is used to improve the cluster classification in the frames which enables similar clusters in different cells to be assigned different densities depending on their location in the frame. This is used to improve the adaption of the proposed algorithms to high variations in crowd behaviours, distributions, lighting and densities.

This system is more practical than the existing low-level features regression based methods because it can be trained with a small number of frames so it is relatively easy to deploy. In addition, this may reduce the training error, speed, cost and opens the door to developing more accurate people detection methods. This system can be used to estimate crowd densities at specific locations in a scene. This shows significant promise as it can be used to detect localised abnormalities in applications such crowd control, evacuation planning and product displays.

Comparison of the proposed system in sparse and crowded scenarios shows that it performs better in crowded environments in term of the MDE metric. The MDE of the proposed algorithms are less than or equal to 0.078 for the Mall

dataset and 0.055 for the UCSD dataset. It is less than or equal to 0.056 for the all tested frames of the New York Grand Central station dataset.

6.2.2 Contributions of the proposed regression based people counting system

A novel system for people counting based on low-level features regression was presented in this thesis. A multi-Gaussian regression models are used instead of a single model to consider occlusion. In addition, a heterogeneous training data was broken down into linear and non-linear homogeneous data (low-level occlusion frames and high-level occlusion frames). Each homogeneous data was used to train one Gaussian regression model. Training regression models using homogeneous data improved the performance of training because the relationship between the number of people and the extracted features of the homogeneous data is more reliable and accurate. A linear kernel is used with the low-level occlusion regression model while a combination of the linear and RBF kernels is used with high-level occlusion regression model. A comprehensive analysis of the proposed people counting system across three datasets was performed. This system achieved good results under situations of heavy occlusions and perspective distortions. In addition, it outperformed a number of existing approaches. The MAE, MSE and MDE for the proposed system are 2.9, 13.62 and 0.095, respectively, for the Mall dataset and 1.63, 4.32 and 0.066, respectively, for the UCSD dataset. They are 25.62, 168 and 0.029, respectively, for the New York Grand Central station dataset.

The existing low-level features regression methods did not use the level of occlusion to improve the accuracy because there is no equation or formula which can be used to measure it. In this system, a novel equation was used to measure the level of occlusion which is based on the ratio of the number of keypoints to the number of foreground pixels. The measured level of occlusion

of each frame was compared with a predefined threshold to select which regression model should be activated.

A multi-stage thresholding method has been used to determine the predefined threshold. This method consists of two stages and nineteen potential thresholds. The threshold with the highest accuracy of both stages is selected to be the best threshold. This threshold represents the boundary between the low-level and high-level occluded frames. It is selected experimentally because there is no technical definition or clear boundaries between them.

A switch approach between the low-level and high-level occlusion regression models is used to select which regression model should be applied to the input frame. This approach is efficient, fast and less complex than the combination of the regression models that is used by most existing ensemble learning methods. The results of the computation efficiency showed that the difference in speed between the classical Gaussian process regression and this system is very small. This proves that the switch method is fast and efficient to use in the people counting.

The combination of features is selected dynamically depending on the characteristics of each environment. Thirty-one combinations of features were selected to be tested and evaluated with this system and the combination with the highest accuracy has been selected to be the best combination. It is noticed that the use of larger number of different types of features generally improved the performance. It is also noticed that edge features are highly inaccurate in the environments with complicated backgrounds, shadows and reflections.

Comparison of the proposed system in sparse and crowded scenarios shows that it performs better in crowded environments in term of the MDE metric. The MDE of the proposed system is equal to 0.078 for the Mall dataset and 0.075 for the UCSD dataset. It is 0.029 for the all tested frames of the New York Grand Central station dataset.

6.3 Future Work

This thesis has provided a number of contributions to the field of people counting based on computer vision. However, there is no scientific research can be exhaustively investigated due to much uncontrollable constrains such as time, data availability, etc. In this section, avenues for additional future research are briefly proposed.

6.3.1 Evaluation of the proposed systems with other objects

Although the proposed people counting systems achieve high performance, they are only used for one type of object (people). The proposed systems may be modified to count a number of other visual objects, such as cells in a microscopic image, cars or trees.

The optimal set of extracted features should be investigated to select the best combination of features depending on the type of object. In addition, the appropriate regression or optimisation models to use with each type of object should be investigated.

6.3.2 Extension of bidirectional people counting systems

The proposed systems can be extended to consider the direction of people using bidirectional background subtraction methods, such as the mixture of dynamic textures (Chan & Vasconcelos 2008). Bidirectional analysis of people counting may be useful for different applications to monitor the main directions and routes of crowds. The same types of features can be used with all routes of the crowd while individual regression or optimisation models should be used with each one of them. The number of people in each route is counted using these individual regression or optimisation models.

6.3.3 Further study of regression models, optimisation programming and background subtraction methods

In this thesis, the author has chosen one type of regression model (GPR), optimisation programming (quadratic programming) and background subtraction method (GMM) to implement the proposed systems. Although the results of the systems are good and outperform a number of existing people counting methods, the performance of the systems may still be improved by using other types of regression models, optimisation programming and/or background subtraction methods.

The proposed systems are high adaptive because they are not restricted to one type of regression model, optimisation programming and/or background subtraction methods. Different types of these methods should be investigated and a comprehensive analysis of them is required to select the best methods. In addition, high adaptive systems open the door for other researchers to test and evaluate the new developed regression models, optimisation programming and/or background subtraction methods. In conclusions, the performances of the proposed systems can be improved continually.

6.3.4 Improving the processing speed of the processed systems

A Graphics Processing Unit (GPU) is a chip that is used to process any functions relating to what displays on your computer's screen. MATLAB can utilise GPU chips to accelerate the processing speed of the proposed systems so they will be more sophisticated real-time systems. MATLAB software offers the GPU computing technology to clients without having to learn the intricacies of GPU architectures or low-level GPU computing libraries. NVIDIA GPU is only supported by MATLAB while AMD or Intel GPUs are not supported for computation acceleration.

MATLAB software also offers a method to utilise multiple GPUs on a single computer using MATLAB workers in parallel computing toolbox and MATLAB distributed computing server. In conclusion, the processing speed of the proposed systems can be significantly improved by using single or multi NVIDIA GPUs.

References

Adegboye, A., Hancke, G. & Jr, G.H., 2012. Single-pixel approach for fast people counting and direction estimation. In *Southern Africa Telecommunication Networks and Applications*.

Adegboye, A.O., 2013. *Single pixel robust approach for background subtraction for fast people-counting and direction estimation*. Dissertation, University of Pretoria.

Al-zaydi, Z., Ndzi, D. & Sanders, D., 2016. Cascade Method for Image Processing Based People Detection and Counting. In *International Conference on Image Processing, Production and Computer Science*. pp. 30–36.

Al-Zaydi, Z.Q.H. et al., 2016. A robust multimedia surveillance system for people counting. *Multimedia Tools and Applications*.

Al-Zaydi, Z.Q.H. et al., 2016. An adaptive people counting system with dynamic features selection and occlusion handling. *Journal of Visual Communication and Image Representation*, 39, pp.218–225.

Alawi, M.A., Khalifa, O.O. & Rafiqul Islam, M.D., 2013. Performance comparison of background estimation algorithms for detecting moving vehicle. *World Applied Sciences Journal*, 21(SPECIAL ISSUE1), pp.109–114.

Albiol, A. et al., 2009. Video analysis using corner motion statistics. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*. pp. 31–38.

Albiol, A., Albiol, A. & Silla, J., 2009. Statistical video analysis for crowds counting. In *International Conference on Image Processing (ICIP)*. pp. 2569–2572.

Benezeth, Y. et al., 2010. Comparative study of background subtraction algorithms. *Journal of Electronic Imaging*, 19(3), p.33003.

Biadgie, Y. & Sohn, K.A., 2014. Feature detector using adaptive accelerated segment test. In *International Conference on Information Science and Applications*. pp. 1–4.

Biodata Ltd, 2013. Use CCTV to Count People. Available at: http://www.videoturnstile.com/ [Accessed August 1, 2017].

Brostow, G.J. & Cipolla, R., 2006. Unsupervised Bayesian Detection of Independent Motion in Crowds. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Celebi, M.E., Kingravi, H.A. & Vela, P.A., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1), pp.200–210.

Çelik, H., Hanjalić, A. & Hendriks, E. a., 2006. Towards a robust solution to people counting. In *International Conference on Image Processing*. pp. 2401–2404.

Cetinkaya, H.H. & Akcay, M., 2015. People Counting at Campuses. *Procedia - Social*

and Behavioral Sciences, 182, pp.732–736.

Chan, A., Morrow, M. & Vasconcelos, N., 2009. Analysis of Crowded Scenes using Holistic Properties. In *Performance Evaluation of Tracking and Surveillance workshop*. IEEE, pp. 101–108.

Chan, A.B., Liang, Z.S.J. & Vasconcelos, N., 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Chan, A.B. & Vasconcelos, N., 2009. Bayesian poisson regression for crowd counting. In *IEEE International Conference on Computer Vision*. IEEE, pp. 545–551.

Chan, A.B. & Vasconcelos, N., 2012. Counting people with low-level features and bayesian regression. *IEEE Transactions on Image Processing*, 21(4), pp.2160–2177.

Chan, A.B. & Vasconcelos, N., 2008. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5), pp.909–926.

Chen, K. et al., 2013. Cumulative attribute space for age and crowd density estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2467–2474.

Chen, K. et al., 2012. Feature Mining for Localised Crowd Counting. In *British Machine Vision Conference*.

Chen, K. & Kamarainen, J.K., 2014. Learning to Count with Back-Propagated Information. In *International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 4672–4677.

Chen, Y. et al., 2014. An improved image mosaic based on Canny edge and an 18-dimensional descriptor. *Optik*, 125(17), pp.4745–4750.

Cheriyadat, A.M., Bhaduri, B.L. & Radke, R.J., 2008. Detecting multiple moving objects in crowded environments with coherent motion regions. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE.

Cho, S. Y., & Chow, T.W., 1999. A fast neural learning vision system for crowd estimation at underground stations platform. *Neural processing letters*, 10(2), pp.111–120.

Cho, S.Y., Chow, T. & Leung, C. at, 1999. A neural-based crowd estimation by hybrid global learning algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 29(4), pp.535–541.

Chow, T.W.S., Yam, J.Y.F. & Cho, S.Y., 1999. Fast training algorithm for feedforward neural networks: Application to crowd estimation at underground stations. *Artificial Intelligence in Engineering*, 13(3), pp.301–307.

Collberg, B.Y.C. et al., 2016. Repeatability in Computer Systems. *Communications of the ACM*, 59(3), pp.62–69.

Conte, D. et al., 2010a. A method for counting people in crowded scenes. In *IEEE*

International Conference on Advanced Video and Signal Based Surveillance. pp. 225–232.

Conte, D. et al., 2010b. Counting moving people in videos by salient points detection. In *International Conference on Pattern Recognition*. pp. 1743–1746.

Conte, D. et al., 2013. Counting moving persons in crowded scenes. *Machine Vision and Applications*, 24(5), pp.1029–1042.

Cuevas, C., Martínez, R. & García, N., 2016. Detection of stationary foreground objects: A survey. *Computer Vision and Image Understanding*, 152, pp.41–57.

Dalal, N. & Triggs, B., 2005. Histograms of Oriented Gradients for Human Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 886–893.

Davies, A., Yin, J.H. & Velastin, S., 1995. Crowd monitoring using image processing. *Electronics & Communication Engineering Journal*, (February).

DILAX Intelcom, 2015. Public Transport. Available at: https://www.dilax.com/ [Accessed August 1, 2017].

Dollar, P., Belongie, S. & Perona, P., 2010. The Fastest Pedestrian Detector in the West. In *British Machine Vision Conference*.

Domenico, S. Di et al., 2016. Trained-Once Device-Free Crowd Counting and Occupancy Estimation Using WiFi : A Doppler Spectrum Based Approach. In *International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*.

Ershad, S., 2012. Texture Classification Approach Based on Combination of Edge & Co-occurrence and Local Binary Pattern. *arXiv preprint*, pp.626–629.

Felzenszwalb, P.F. et al., 2010. Object Detection with Discriminative Trained Part Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), pp.1627–1645.

Fradi, H. & Dugelay, J., 2012. People counting system in crowded scenes based on feature regression. In *Signal Processing Conference (EUSIPCO)*. IEEE, pp. 136–140.

Fradi, H. & Dugelay, J.L., 2012. Low level crowd analysis using frame-wise normalized feature for people counting. In *International Workshop on Information Forensics and Security*. pp. 246–251.

Fu, M. et al., 2015. Fast crowd density estimation with convolutional neural networks. *Engineering Applications of Artificial Intelligence*, 43, pp.81–88.

Gandhi, V., Čech, J. & Horaud, R., 2012. High-resolution depth maps based on TOF-stereo fusion. In *IEEE International Conference on Robotics and Automation*. pp. 4742–4749.

Gao, C. et al., 2016. People counting based on head detection combining Adaboost and CNN in crowded surveillance environment. *Neurocomputing*, 208, pp.1–9.

Gao, L. et al., 2016. Crowd Pedestrian Counting Considering Network Flow Constraints in Videos. *arXiv preprint*.

Garcia-Bunster, G., Torres-Torriti, M. & Oberli, C., 2012. Crowded pedestrian counting at bus stops from perspective transformations of foreground areas. *IET Computer Vision*, 6(4), p.296.

Ge, W. & Collins, R.T., 2009. Marked point processes for crowd counting. In *Computer Vision and Pattern Recognition Workshops*. IEEE, pp. 2913–2920.

Ghosh, D. & Kaabouch, N., 2016. A survey on image mosaicing techniques. *Journal of Visual Communication and Image Representation*, 34, pp.1–11.

Giveki, D., Soltanshahi, M.A. & Montazer, G.A., 2017. A new image feature descriptor for content based image retrieval using scale invariant feature transform and local derivative pattern. *Optik - International Journal for Light and Electron Optics*, 131, pp.242–254.

Hafeezallah, A. & Abu-Bakar, S., 2016. Crowd counting using statistical features based on curvelet frame change detection. *Multimedia Tools and Applications*.

Han, J. et al., 2013. Enhanced computer vision with Microsoft Kinect sensor: A review. *IEEE Transactions on Cybernetics*, 43(5), pp.1318–1334.

Hashemzadeh, M. & Farajzadeh, N., 2016. Combining keypoint-based and segment-based features for counting people in crowded scenes. *Information Sciences*, 345, pp.199–216.

Hou, Y.L. & Pang, G.K.H., 2011. People counting and human detection in a challenging situation. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, 41(1), pp.24–33.

Hu, X. et al., 2015. Dense crowd counting based on perspective weight model using a fisheye camera. *International Journal for Light and Electron Optics*, 126(1), pp.123–130.

Hu, Y. et al., 2016. Dense crowd counting from still images with convolutional neural networks. *Journal of Visual Communication and Image Representation*, 38, pp.530–539.

Huang, X., Zou, Y. & Wang, Y., 2016. Cost-sensitive sparse linear regression for crowd counting with imbalanced training data. In *IEEE International Conference on Multimedia and Expo (ICME)*.

Ilyas, A., Scuturici, M. & Miguet, S., 2009. Real time foreground-background segmentation using a modified codebook model. In *International Conference on Advanced Video and Signal Based Surveillance*. pp. 454–459.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), pp.651–666.

Jeong, C. Y., Choi, S., & Han, S.W., 2013. A method for counting moving and stationary people by interest point classification. In *IEEE International Conference on Image Processing*. IEEE, pp. 4545–4548.

Jeong, C.Y. & Choi, S., 2016. A Comparison of Keypoint Detectors in the Context of Pedestrian Counting. In *IEEE Information and Communication Technology Convergence (ICTC)*. pp. 1179–1181.

Jin, R. & Liu, H., 2005. A Novel Approach to Model Generation for Heterogeneous Data Classification. In *'Proceedings of the 19th international joint conference on Artificial intelligence (IJCAI)*. pp. 746–751.

Jin, R. & Liu, H., 2004. SWITCH : A Novel Approach to Ensemble Learning for Heterogeneous Data. In *European Conference on Machine Learning*. Springer Berlin Heidelberg, pp. 560–562.

Joshi, N.S. & Choubey, N.S., 2014. Comparison of Traditional Approach for Edge Detection with Soft Computing Approach. *International Journal of Computer Applications*, 96(11), pp.17–23.

Kannan, P.G. et al., 2012. Low cost crowd counting using audio tones. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*. p. 155.

Kaur, G. & Virk, I.S., 2014. Edge Detection through Fuzzy System using Type I Format. *International Journal of Computer Applications*, 102(13), pp.24–27.

Kilambi, P. et al., 2008. Estimating pedestrian counts in groups. *Computer Vision and Image Understanding*, 110(1), pp.43–59.

Kim, K. et al., 2005. Real-time foreground-background segmentation using codebook model. *Real-Time Imaging*, 11(3), pp.172–185.

Kitamura, R., Li, S. & Nakanishi, I., 2015. Spherical FAST Corner Detector. In *IEEE International Conference on Mechatronics and Automation (ICMA)*.

Kong, D., Gray, D. & Tao, H., 2006a. A viewpoint invariant approach for crowd counting. In *International Conference on Pattern Recognition*. pp. 1187–1190.

Kong, D., Gray, D. & Tao, H., 2006b. Counting pedestrians in crowds using viewpoint invariant training. In *International Conference on Pattern Recognition*. pp. 1187–1190.

Leibe, B., Seemann, E. & Schiele, B., 2005. Pedestrian detection in crowded scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 878–885.

Lempitsky, V. & Zisserman, A., 2010. Learning to count objects in images. In *Advances in Neural Information Processing Systems*. pp. 1324–1332.

Lev, L., Brewer, L.J. & Stephenson, G.O., 2008. *Tools for Rapid Market Assessments*,

Li, J., Huang, L. & Liu, C., 2011. Robust people counting in video surveillance: Dataset and system. In *International Conference on Advanced Video and Signal Based Surveillance*. pp. 54–59.

Li, N.-N. et al., 2007. A People-Counting System Based on BP Neural Network. In *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*.

Lin, S., Chen, J. & Chao, H., 2001. Estimation of Number of People in Crowded Scenes Using Perspective Transformation. *IEEE Transactions on Systems, Man, and Cybernetics*, 31(6), pp.645–654.

Lin, T.Y. et al., 2011. Cross camera people counting with perspective estimation and occlusion handling. In *IEEE International Workshop on Information Forensics and Security*. IEEE.

Lin, W.C., Seah, W. & Li, W., 2011. Exploiting radio irregularity in the Internet of Things for automated people counting. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*. pp. 1015–1019.

Lionel, N. et al., 2003. LANDMARC: indoor location sensing using active RFID. *IEEE International Conference on Pervasive Computing and Communications*, pp.407–415.

Longo, S. & Cheng, B., 2015. Privacy preserving crowd estimation for safer cities. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*. pp. 1543–1550.

Lowe, D.G., 2004. Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision*, 60, pp.91–110.

Loy, C. et al., 2013. Crowd counting and profiling: Methodology and evaluation. In *Modeling, Simulation and Visual Analysis of Crowds*. Springer New York, pp. 347–382.

Lumentut, J.S., Gunawan, F.E. & Diana, 2015. Evaluation of Recursive Background Subtraction Algorithms for Real-Time Passenger Counting at Bus Rapid Transit System. *Procedia Computer Science*, 59(Iccsci), pp.445–453.

Luo, J. et al., 2016. Real-time people counting for indoor scenes. *Signal Processing*, 124, pp.27–35.

Ma, H., Zeng, C. & Ling, C.X., 2012. A Reliable People Counting System via Multiple Cameras. *ACM Transactions on Intelligent Systems and Technology*, 3(2), pp.1–22.

Ma, R. et al., 2004. On pixel count based crowd density estimation for visual surveillance. In *IEEE Conference on Cybernetics and Intelligent Systems*. IEEE, pp. 1–3.

Ma, Z. & Chan, A.B., 2013. Crossing the line: Crowd counting by integer programming with local features. In *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2539–2546.

Majumder, G., Bhowmik, K. & Bhatacharjee, D., 2013. Automatic Eye Detection using Fast Corner Detector of North East Indian (NEI) Face Images. *Procedia Technology*, 10(i), pp.646–653.

Marana, A.N. et al., 1999. Estimating crowd density with Minkowski fractal dimension. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. pp. 3521–3524.

Mehmood, M.O., 2016. *People detection methods for intelligent multi-Camera surveillance systems Signal et Images par PRES Universit ´*. Ecole Centrale de Lille.

Mei, J. & Zhao, Y., 2013. An Improved Method of Crowd Counting Based on Regression. In *International Conference on Multimedia Technology*. pp. 143–150.

Merad, D., Aziz, K.E. & Thome, N., 2010. Fast people counting using head detection from skeleton graph. In *Advanced Video and Signal Based Surveillance*. IEEE,

pp. 233–240.

Microsoft, 2011. Kinect. Available at: www.xbox.com/en-US/kinect/ [Accessed August 1, 2016].

Montazer, G.A. & Giveki, D., 2015. Content based image retrieval system using clustered scale invariant feature transforms. *Optik - International Journal for Light and Electron Optics*, 126(18), pp.1695–1699.

Mori, G., Belongie, S. & Malik, J., 2005. Efficient shape matching using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11), pp.1832–1837.

Mu, Y. et al., 2008. Discriminative local binary patterns for human detection in personal album. *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1–8.

Mukherjee, S., 2014. *A Novel Framework for Unique People Count from Monocular Videos*. University of Alberta.

Nakatsuka, M., Iwatani, H. & Katto, J., 2008. A Study on Passive Crowd Density Estimation using Wireless Sensors. In *International Conference on Mobile Computing and Ubiquitous Networking*. pp. 1–6.

Norris, C., Mccahill, M. & Wood, D., 2004. Editorial. The Growth of CCTV: a global perspective on the international diffusion of video surveillance in publicly accessible space. *Surveillance & Society*, 2(2/3), pp.110–135.

Nurhadiyatna, A. et al., 2013. Background subtraction using Gaussian Mixture Model enhanced by Hole Filling Algorithm (GMMHF). *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, pp.4006–4011.

Del Pizzo, L. et al., 2015. Counting people by RGB or depth overhead cameras. *Pattern Recognition Letters*, 81, pp.41–50.

Prathap, K.S.V., Jilani, S.A.K. & Reddy, P.R., 2016. A Real-Time Image Mosaicing using Scale Invariant Feature Transform. *Indian Journal of Science and Technology*, 9(12), pp.7–12.

Rabaud, V. & Belongie, S., 2006. Counting crowded moving objects. In *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 705–711.

Rao, A.S. et al., 2015. Estimation of crowd density by clustering motion cues. *Visual Computer*, 31(11), pp.1533–1552.

Rodriguez, M. et al., 2011. Density-aware person detection and tracking in crowds. In *International Conference on Computer Vision*. IEEE, pp. 2423–2430.

Ryan, D. et al., 2015. An evaluation of crowd counting methods, features and regression models. *Computer Vision and Image Understanding*, 130, pp.1–17.

Ryan, D. et al., 2009. Crowd counting using multiple local features. In *Digital Image Computing: Techniques and Applications*. IEEE, pp. 81–88.

Ryan, D. et al., 2014. Scene invariant multi camera crowd counting. *Pattern*

Recognition Letters, 44, pp.98–112.

Ryan, D.A., 2013. *Crowd Monitoring Using Computer Vision.* Dissertation, Queensland University of Technology.

Sabzmeydani, P. & Mori, G., 2007. Detecting pedestrians by learning shapelet features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Saleh, S.A.M., Suandi, S.A. & Ibrahim, H., 2015. Recent survey on crowd density estimation and counting for visual surveillance. *Engineering Applications of Artificial Intelligence*, 41, pp.103–114.

Sensors, C.R., Fisher, R.B. & Konolige, K., 2008. *Handbook of Robotics Chapter 22 - Range Sensors*,

Shbib, R. et al., 2013. Distributed Monitoring System Based On Weighted Data Fusing Model. *American Journal of Social Issues and Humanities*, 3, pp.53–62.

Shbib, R. et al., 2014. Head Pose Estimation for Car Drivers. *International Journal of u-and e-Service, Science and Technology*, 7(4), pp.359–374.

Shimosaka, M. et al., 2011. Counting pedestrians in crowded scenes with efficient sparse learning. In *Asian Conference on Pattern Recognition (ACPR)*. pp. 27–31.

ShopperTrak, 2013. ShopperTrak Solutions. Available at: http://www.shoppertrak.com/ [Accessed August 1, 2017].

Shrivakshan, G.T. & Chandrasekar, C., 2012. A Comparison of various Edge Detection Techniques used in Image Processing. *International Journal of Computer Science Issues*, 9(5).

Sidla, O. et al., 2006. Pedestrian detection and tracking for counting applications in crowded situations. In *IEEE International Conference on Video and Signal Based Surveillance*.

Siva, P. et al., 2016. Real-time, Embedded Scene Invariant Crowd Counting Using Scale-Normalized Histogram of Moving Gradients (HoMG). In *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 885–892.

Sobral, A. & Vacavant, A., 2014. A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. *Computer Vision and Image Understanding*, 122, pp.4–21.

Stauffer, C. & Grimson, W.E.L., 1999. Adaptive background mixture models for real-time tracking. *Proceedings 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Cat No PR00149*, 2(c), pp.246–252.

Tan, P.N., Steinbach, M. & Kumar, V., 2006. Cluster Analysis: Basic Concepts and Algorithms. In *Introduction to Data Mining*.

Tang, N.C. et al., 2015. Cross-Camera Knowledge Transfer for Multiview People Counting. *IEEE Transactions on Image Processing*, 24(1), pp.80–93.

Technology, A., 2013. Our customers. Available at:

http://www.peoplecounting.co.uk/our-customers [Accessed August 1, 2017].

Tikkanen, T., 2013. *Image-based people detection*. Aalto University.

Tikkanen, T., 2014. *People detection and tracking using a network of low-cost depth cameras*. Aalto University.

Topkaya, I.S., Erdogan, H. & Porikli, F., 2014. Counting People by Clustering Person Detector Outputs. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, pp. 313–318.

Tu, J., Zhang, C. & Hao, P., 2013. Robust real-time attention-based head-shoulder detection for video surveillance. In *IEEE International Conference on Image Processing*. IEEE, pp. 3340–3344.

Tuzel, O., Porikli, F. & Meer, P., 2008. Pedestrian Detection via Classification on Riemannian Manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10).

Valera, M. & Velastion, S.A., 2005. Intelligent distributed surveillance systems: a review. In *Vision, Image and Signal Processing, IEE Proceedings*. pp. 192–204.

Vasco Dantas dos Reis, J., 2014. *Image Descriptors for Counting People with Uncalibrated Cameras*. University of Porto.

Viola, P. & Jones, M., 2004. Robust real-time face detection. *International Journal of Computer Vision*, 57(2), pp.137–154.

Viola, P., Jones, M.J. & Snow, D., 2005. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2), pp.153–161.

Wang, B. et al., 2013. Crowd Density Estimation Based on Texture Feature Extraction. *Journal of Multimedia*, 8(4), pp.331–337.

Wang, J. et al., 2014. Spatiotemporal Group Context for Pedestrian Counting. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(9), pp.1620–1630.

Wang, M., 2014. *Data Assimilation for Agent-Based Simulation of Smart Environment*. Dissertation, Georgia State University.

Weppner, J. & Lukowicz, P., 2011. Collaborative Crowd Density Estimation with Mobile Phones. In *Proceedings of ACM PhoneSense*.

Williams, C.K.I. & Rasmussen, C.E., 2008. *Gaussian processes for regressionn: A Quick Introduction*,

Wu, B. & Nevatia, R., 2007. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2), pp.247–266.

Xi, W. et al., 2014. Electronic frog eye: Counting crowd using WiFi. *Proceedings - IEEE INFOCOM*, pp.361–369.

Xiaohua, L., Lansun, S. & Huanqin, L., 2006. Estimation of Crowd Density Based on

Wavelet and Support Vector Machine. *Transactions of the Institute of Measurement and Control*, 28(3), pp.299–308.

Xing, X., Wang, K. & Lv, Z., 2015. Fusion of Gait and Facial Features using Coupled Projections for People Identification at a Distance. *Signal Processing Letters*, 22(12), pp.2349–2353.

Xu, B. & Qiu, G., 2016. Crowd Density Estimation based on Rich Features and Random Projection Forest. In *IEEE Winter Applications of Computer Vision*. pp. 1–8.

Xu, T. et al., 2016. Crowd Counting Using Accumulated HOG. In *International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*. pp. 1877–1881.

Yoshinaga, S., Shimada, A. & Taniguchi, R.I., 2010. Real-time people counting using blob descriptor. *Procedia - Social and Behavioral Sciences*, 2(1), pp.143–152.

Yuan, Y. et al., 2011. Crowd Density Estimation Using Wireless Sensor Networks. In *Seventh International Conference on Mobile Ad-hoc and Sensor Networks*. IEEE, pp. 138–145.

Yuan, Y. et al., 2013. Estimating crowd density in an RF-based dynamic environment. *IEEE Sensors Journal*, 13(10), pp.3837–3845.

Yuk, J.S.C. et al., 2006. Real-time multiple head shape detection and tracking system with decentralized trackers. In *International Conference on Intelligent Systems Design and Applications*. pp. 384–389.

Zeng, C. & Ma, H., 2010. Robust head-shoulder detection by PCA-based multilevel HOG-LBP detector for people counting. *Proceedings - International Conference on Pattern Recognition*, pp.2069–2072.

Zhang, C., Li, H. & Wang, X., 2015. Cross-scene Crowd Counting via Deep Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, J. et al., 2011. Predicting pedestrian counts in crowded scenes with rich and high-dimensional features. *IEEE Transactions on Intelligent Transportation Systems*, 12(4), pp.1037–1046.

Zhang, X. et al., 2012. Water filling: Unsupervised people counting via vertical kinect sensor. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*. IEEE, pp. 215–220.

Zhang, Y. et al., 2016. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 589–597.

Zhang, Z., Wang, M. & Geng, X., 2015. Crowd counting in public video surveillance by label distribution learning. *Neurocomputing*, 166, pp.151–163.

Zhong, S.H., Liu, Y. & Chen, Q.C., 2015. Visual orientation inhomogeneity based scale-invariant feature transform. *Expert Systems with Applications*, 42(13), pp.5658–5667.

Zhou, B., Wang, X. & Tang, X., 2012. Understanding collective crowd behaviors: Learning a Mixture model of Dynamic pedestrian-Agents. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.2871–2878.

Zhu, F.Z.F. et al., 2009. A New Method for People-Counting Based on Support Vector Machine. *International Conference on Intelligent Networks and Intelligent Systems*, 1(5), pp.5–8.

Appendices

Appendix A: List of Features

The following table lists the features that used with the features regression based crowd counting system. Four categories were used with this system and as in Section 3.5.6.

Features	Description
Foreground segment	segment area segment perimeter perimeter orientation histogram (90 degrees) perimeter orientation histogram (120 degrees) perimeter orientation histogram (150 degrees) perimeter orientation histogram (0 degrees) perimeter orientation histogram (30 degrees) perimeter orientation histogram (60 degrees) perimeter-area ratio Blob count
Edge	internal edge length internal edge orientation histogram (90 degrees) internal edge orientation histogram (120 degrees) internal edge orientation histogram (150 degrees) internal edge orientation histogram (0 degrees) internal edge orientation histogram (30 degrees) internal edge orientation histogram (60 degrees)
Texture	GLCM energy (0 degrees) GLCM homogeneity (0 degrees) GLCM entropy (0 degrees) GLCM energy (45 degrees) GLCM homogeneity (45 degrees) GLCM entropy (45 degrees) GLCM energy (90 degrees) GLCM homogeneity (90 degrees) GLCM entropy (90 degrees) GLCM energy (135 degrees) GLCM homogeneity (135 degrees) GLCM entropy (135 degrees)
Keypoints	SIFT

Appendix B: List of Publications

Journal Papers:

- Al-Zaydi, Z. Q., Ndzi, D. L., Yang, Y., & Kamarudin, M. L. (2016). An adaptive people counting system with dynamic features selection and occlusion handling. *Journal of Visual Communication and Image Representation*, *39*, 218-225 (impact factor 2.16).

- Al-Zaydi, Z. Q., Ndzi, D. L., Kamarudin, M. L., Zakaria, A., & Shakaff, A. Y. (2016). A robust multimedia surveillance system for people counting. *Multimedia Tools and Applications*, 1-28 (impact factor 1.53).

Conference Proceedings:

- Al-Zaydi, Z. Q., Ndzi, D., & Sanders, D. (2016). Cascade method for image processing based people detection and counting. *Proceedings of 2016 International Conference on Image Processing, Production and Computer Science (ICIPCS'2016),* 30-36.

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

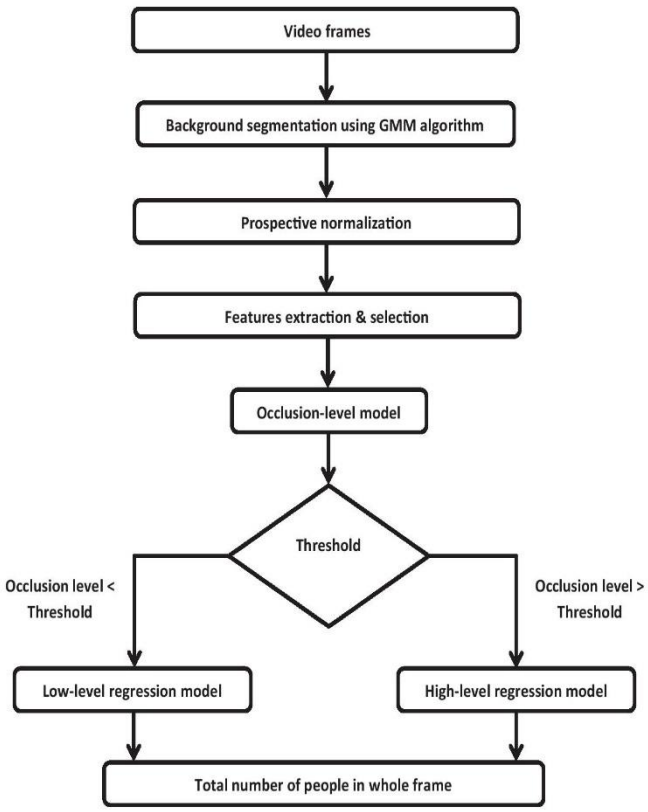


Fig. 1. Flow diagram of the proposed framework.

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]:

$$Level\ of\ crowd = \frac{No\ of\ keypoints}{No\ of\ keypoints\ per\ person} \tag{6}$$

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

$$Level\ of\ occlusion = \frac{No\ of\ keypoints}{No\ of\ foreground\ pixels} \tag{7}$$

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 s 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 normalisation, 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]. Grey-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.

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. Fig. 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]:

$$weight_{line} = \frac{h_{ab} w_{ab}}{h_{line} w_{line}} \tag{8}$$

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.

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 2.5 GHz core I5 processor and 4 GB memory. The mean deviation error is given as:

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

$$\mathrm{MDE} = \frac{1}{N}\sum_{n=1}^{N}\frac{|y_n - \hat{y}_n|}{y_n} \tag{9}$$

The mean absolute error is given as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n| \tag{10}$$

The mean squared error is given as:

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

where N is the total number of the test frames, y_n is the actual count, and $\hat{y}_n$ is the estimated count (Table 1).

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	

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

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. Fig. 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.

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. Fig. 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.

3.3. Computation efficiency

To study the computational efficiency of the proposed method, comparisons between the state of the art GPR algorithm and the

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

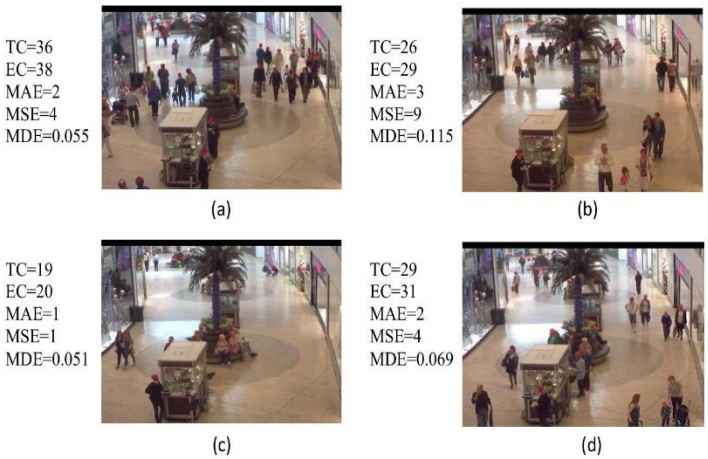


Fig. 3. Examples of the Mall dataset frames and their results.

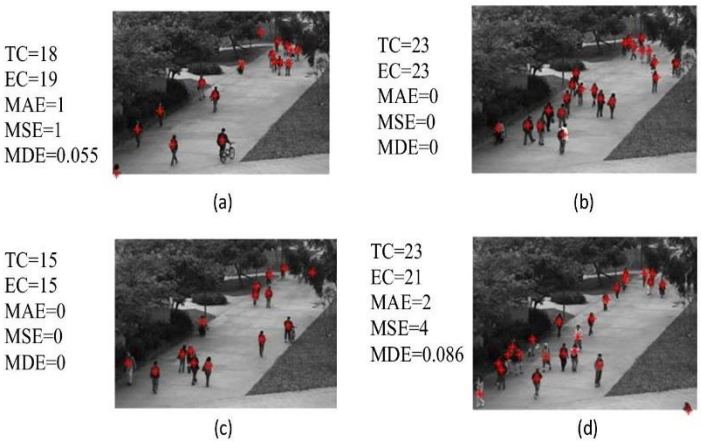


Fig. 4. Examples of the UCSD dataset frames and their results.

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

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.

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 Tables 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. Tables 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.

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

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

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 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

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, S. Sridharan, Scene invariant multi camera crowd counting, Pattern Recognit. Lett. 44 (2014) 98–112.
[2] A. Technology, Our customers available: <www.peoplecounting.co.uk/our-customers>2013 (accessed: 23.03.15).
[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, T. Xiang, Crowd counting and profiling: methodology and evaluation, 2013.
[6] ShopperTrak, ShopperTrak Solutions available: <www.shoppertrak.com/products>2013 (accessed: 23.03.15).
[7] Biodata Ltd, Use CCTV to Count People available: <www.videoturnstile.com>2013 (accessed: 23.03.15).
[8] DILAX Intelcom, Public Transport available: <www.dilax.net/electronic-people-counter-passenger-counter-customer-counter/passenger-counting-systems-for-public-transportation>2015 (accessed: 23.03.15).
[9] F.Z.F. Zhu, X.Y.X. Yang, J.G.J. Gu, 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, 2009, pp. 5–8. no. 5.
[10] N.-N. Li, J. Song, R.-Y. Zhou, J.-H. Gu, A people-counting system based on BP neural network, Fourth Int. Conf. Fuzzy Syst. Knowl. Discov. (FSKD 2007), vol. 3, 2007. no. Fskd.
[11] M. Nakatsuka, H. Iwatani, J. Katto, A study on passive crowd density estimation using wireless sensors, 4th Intl. Conf. on Mob. Comput. Ubiquitous Netw. (ICMU 2008), 2008, pp. 1–6. no. 2.
[12] Z. Ma, A.B. Chan, Crossing the line: crowd counting by integer programming with local features, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2013, pp. 2539–2546.
[13] Y. Yuan, C. Qiu, W. Xi, J. Zhao, Crowd density estimation using wireless sensor networks, 2011 Seventh Int. Conf. Mob. Ad-hoc Sens. Networks, 2011, pp. 138–145.
[14] R. Shbib, S. Zhou, D. Ndzi, K. Al-kadhimi, Distributed monitoring system based on weighted data fusing model, Am. J. Soc. Issues Humanit. 3 (March) (2013) 53–62.
[15] W. Xi, J. Zhao, X.Y. Li, K. Zhao, S. Tang, X. Liu, Z. Jiang, Electronic frog eye: counting crowd using WiFi, Proc. – IEEE INFOCOM, 2014, pp. 361–369.
[16] W.-C. Lin, W. Seah, W. Li, Exploiting radio irregularity in the internet of things for automated people counting, IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC, 2011, pp. 1015–1019.
[17] V. Gandhi, J. Čech, R. Horaud, High-resolution depth maps based on TOF-stereo fusion, in: Proc. – IEEE Int. Conf. Robot. Autom., 2012, pp. 4742–4749.
[18] Y.L. Hou, G.K.H. Pang, People counting and human detection in a challenging situation, IEEE Trans. Syst. Man, Cybern. Part A: Syst. Hum. 41 (1) (2011) 24–33.
[19] T. Tikkanen, People detection and tracking using a network of low-cost depth cameras, 2014.
[20] J. Li, L. Huang, C. Liu, Robust people counting in video surveillance: dataset and system, in: 2011 8th IEEE Int. Conf. Adv. Video Signal based Surveillance, AVSS 2011, 2011, pp. 54–59.
[21] J. Tu, C. Zhang, P. Hao, Robust real-time attention-based head-shoulder detection for video surveillance, Image Process. (2013) 3340–3344.
[22] A. Adegboye, G. Hancke, G. Hancke Jr., Single-pixel approach for fast people counting and direction estimation, Satnac. Org.Za.
[23] C. Norris, M. Mccahill, D. Wood, The growth of CCTV: a global perspective on the international diffusion of video surveillance in publicly accessible space, Surveill. Soc. 2 (2/3) (2004) 110–135.
[24] S.A.M. Saleh, S.A. Suandi, H. Ibrahim, Recent survey on crowd density estimation and counting for visual surveillance, Eng. Appl. Artif. Intell. 41 (2015) 103–114.
[25] K. Chen, J.-K. Kamarainen, Learning to count with back-propagated information, Pattern Recognit. (2014) 4672–4677.
[26] K. Chen, C.C. Loy, S. Gong, T. Xiang, Feature mining for localised crowd counting, Proceedings Br. Mach. Vis. Conf. 2012, 2012, pp. 21.1–21.11.
[27] H. Foroughi, N. Ray, H. Zhang, People counting with image retrieval using compressed sensing, in: Acoust. Speech Signal Process. (ICASSP), IEEE, 2014, pp. 4354–4358.
[28] O. Sidla, Y. Lypetskyy, N. Brändle, 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, C.X. Ling, A reliable people counting system via multiple cameras, ACM Trans. Intell. Syst. Technol. 3 (2) (2012) 1–22.
[30] M. Rodriguez, E.N. Superieure, I. Laptev, J. Sivic, J.-Y. Audibert, Density-aware person detection and tracking in crowds, in: Comput. Vis. (ICCV), IEEE, 2011, pp. 2423–2430.
[31] P. Viola, M. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.

[32] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR '05 Proc. 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 1, 2005, pp. 886–893.
[33] D. Gerónimo, A.M. López, A.D. Sappa, T. Graf, Survey of pedestrian detection for advanced driver assistance systems, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1239–1258.
[34] A.B. Chan, Z.S.J. Liang, N. Vasconcelos, Privacy preserving crowd monitoring: counting people without people models or tracking, in: 26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR, 2008.
[35] J. Zhang, B. Tan, F. Sha, L. He, Predicting pedestrian counts in crowded scenes with rich and high-dimensional features, IEEE Trans. Intell. Transp. Syst. 12 (4) (2011) 1037–1046.
[36] T.W.S. Chow, J.Y.F. Yam, S.Y. Cho, Fast training algorithm for feedforward neural networks: application to crowd estimation at underground stations, Artif. Intell. Eng. 13 (3) (1999) 301–307.
[37] A.B. Chan, N. Vasconcelos, Counting people with low-level features and bayesian regression, IEEE Trans. Image Process. 21 (4) (2012) 2160–2177.
[38] A.B. Chan, N. Vasconcelos, Bayesian poisson regression for crowd counting, in: Proc. IEEE Int. Conf. Comput. Vis., no. ICCV, 2009, pp. 545–551.
[39] A. Chan, M. Morrow, N. Vasconcelos, Analysis of crowded scenes using holistic properties, in: 9th Int. Symp. PETS, no. Pets, 2009.
[40] S.Y. Cho, T.W. Chow, C. Leung, A neural-based crowd estimation by hybrid global learning algorithm, IEEE Trans. Syst. Man Cybern. B Cybern. 29 (4) (1999) 535–541.
[41] S.Y. Cho, T.W. Chow, A fast neural learning vision system for crowd estimation at underground stations platform, Neural Process. Lett. 10 (2) (1999) 111–120.
[42] D. Ryan, S. Denman, C. Fookes, S. Sridharan, Crowd counting using multiple local features, DICTA 2009 – Digit. Image Comput. Tech. Appl., 2009, pp. 81–88.
[43] I.S. Topkaya, H. Erdogan, F. Porikli, Counting people by clustering person detector outputs, 2014, pp. 313–318.
[44] R. Ma, L. Li, W. Huang, Q. Tian, On pixel count based crowd density estimation for visual surveillance, IEEE Conf. Cybern. Intell. Syst. 2004., vol. 1, 2004, pp. 1–3.
[45] A. Davies, J.H. Yin, S. Velastin, Crowd monitoring using image processing, no. February, 1995.
[46] G.J. Brostow, R. Cipolla, Unsupervised Bayesian detection of independent motion in crowds, Comput. Vis. Pattern Recognit. (2006).
[47] V. Rabaud, S. Belongie, Counting crowded moving objects, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 1, 2006, pp. 705–711.
[48] A.M. Cheriyadat, B.L. Bhaduri, R.J. Radke, Detecting multiple moving objects in crowded environments with coherent motion regions, in: 2008 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. CVPR Work., 2008.
[49] D. Merad, K.E. Aziz, N. Thome, Fast people counting using head detection from skeleton graph, in: Proc. – IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2010, 2010, pp. 233–240.
[50] C.K.I. Williams, 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, H.Y.M. Liao, Cross camera people counting with perspective estimation and occlusion handling, in: 2011 IEEE Int. Work. Inf. Forensics Secur. WIFS 2011, 2011.
[52] J. Mei, Y. Zhao, An improved method of crowd counting based on regression, 2013, pp. 143–150.
[53] R. Jin, H. Liu, SWITCH : a novel approach to ensemble learning for heterogeneous data, in: Mach. Learn. ECML 2004, 2004, pp. 560–562.
[54] R. Jin, H. Liu, A novel approach to model generation for heterogeneous data classification, in: IJCAI'05 Proc. 19th Int. Jt. Conf. Artif. Intell., 2005, pp. 746–751.
[55] A. Albiol, M.J. Silla, J.M. Mossi, A. Albiol, Video analysis using corner motion statistics, in: 11th IEEE Int. Work. Perform. Eval. Track. Surveill. (PETS 2009), 2009, pp. 31–38.
[56] A. Albiol, A. Albiol, J. Silla, Statistical video analysis for crowds counting, in: Proc. – Int. Conf. Image Process. ICIP, 2009, pp. 2569–2572.
[57] D. Conte, P. Foggia, G. Percannella, F. Tufano, M. Vento, Counting moving people in videos by salient points detection, in: Proc. – Int. Conf. Pattern Recognit., 2010, pp. 1743–1746.
[58] S.W. Jeong, C.Y. Choi, S. Han, A method for counting moving and stationary people by interest point classification, Image Process. (2013) 4545–4548.
[59] H. Fradi, J. Dugelay, People counting system in crowded scenes based on feature regression, in: Signal Process. Conf., no. Eusipco, 2012, pp. 136–140.
[60] D. Ryan, S. Denman, S. Sridharan, C. Fookes, An evaluation of crowd counting methods, features and regression models, Comput. Vis. Image Underst. 130 (2015) 1–17.
[61] A.N. Marana, L.D.F. Costa, R.A. Lotufo, S.A. Velastin, Estimating crowd density with Minkowski fractal dimension, in: 1999 IEEE Int. Conf. Acoust. Speech, Signal Process. Proceedings. ICASSP99 (Cat. No. 99CH36258), vol. 6, 1999, pp. 3521–3524.
[62] S. Ershad, Texture classification approach based on combination of edge & co-occurrence and local binary pattern, arxiv Prepr. arxiv 1203.4855, 2012, pp. 626–629.
[63] B. Wang, H. Bao, S. Yang, H. Lou, Crowd density estimation based on texture feature extraction, J. Multimedia 8 (4) (2013) 331–337.
[64] K. Chen, S. Gong, T. Xiang, C.C. Loy, Cumulative attribute space for age and crowd density estimation, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2, 2013, pp. 2467–2474.
[65] Z. Zhang, M. Wang, X. Geng, Crowd counting in public video surveillance by label distribution learning, Neurocomputing 166 (2015) 151–163.

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 Ph.D. 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 B.Sc. (Joint Honours) in Electronics and Mathematics from Keele University in 1994, and a Ph.D. 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.

Multimed Tools Appl
DOI 10.1007/s11042-016-4156-x

A robust multimedia surveillance system for people counting

Zeyad Q. H. Al-Zaydi[1,2] · David L. Ndzi[1] · Munirah L. Kamarudin[3] · Ammar Zakaria[4] · Ali Y. M. Shakaff[4]

Received: 12 May 2016 / Revised: 7 October 2016 / Accepted: 14 November 2016

Abstract Closed circuit television cameras (CCTV) are widely used in monitoring. This paper presents an intelligent CCTV crowd counting system based on two algorithms that estimate the density of each pixel in each frame and use it as a basis for counting people. One algorithm uses scale-invariant feature transform (SIFT) features and clustering to represent pixels of frames (SIFT algorithm) and the other uses features from accelerated segment test (FAST) corner points with SIFT features (SIFT-FAST algorithm). Each algorithm is designed using a novel combination of pixel-wise, motion-region, grid map, background segmentation using Gaussian mixture model (GMM) and edge detection. A fusion technique is proposed and used to validate the accuracy by

Highlights
- Two people counting algorithms based on CCTV cameras are proposed.
- Training error, set-up time and cost have been reduced by the proposed system.
- Motion edges, grid map, pixel-wise and fusion techniques are used in the proposed algorithms.
- Two indoor and outdoor datasets are used for evaluation.
- The accuracy of the proposed system is, at least, comparable with the state of the art methods.

✉ Zeyad Q. H. Al-Zaydi
zeyad.al-zaydi@port.ac.uk

David L. Ndzi
david.ndzi@port.ac.uk

Munirah L. Kamarudin
latifahmunirah@unimap.edu.my

Ammar Zakaria
ammarzakaria@unimap.edu.my

Ali Y. M. Shakaff
aliyeon@unimap.edu.my

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

[2] Computer Centre, University of Technology, Baghdad, Iraq

[3] School of Computer and Communication Engineering, University Malaysia Perlis, Perlis, Malaysia

[4] School of Mechatronic Engineering, University Malaysia Perlis, Perlis, Malaysia

Published online: 23 November 2016

combining the result of the algorithms at frame level. The proposed system is more practical than the state of the art regression methods because it is trained with a small number of frames so it is relatively easy to deploy. In addition, it reduces the training error, set-up time, cost and open the door to develop more accurate people detection methods. The University of California (UCSD) and Mall datasets have been used to test the proposed algorithms. The mean deviation error, mean squared error and the mean absolute error of the proposed system are less than 0.1, 16.5 and 3.1, respectively, for the Mall dataset and less than 0.07, 5.5 and 1.9, respectively, for UCSD dataset.

Keywords Crowd counting systems · Monitoring · CCTV cameras · Background segmentation

1 Introduction

Closed circuit television cameras (CCTV) have already become ubiquitous and their use is growing exponentially. For instance, 4.2 million CCTV cameras were used in the United Kingdom [55] in 2004 and an estimated up to around 5.9 million in 2015 [79]. People counting systems are one of the most challenging systems in computer vision to implement [8, 35, 36, 50, 59, 63, 76]. People counting is a useful task for safety, security and operational purposes and can be important for improving awareness [50, 51, 59, 65, 70]. The number of people in a given space can be used to develop business intelligence, such as improving location of products within a shop and finding the number of visitors [51, 65, 70]. Crowd management [70], transport [39] and staff planning applications can be improved by using this kind of information. Heating, lighting and air conditioning can also be optimised using people counting and distribution information to enhance energy management [34, 74], or to improve emergency evacuation plan [74].

A significant amount of research has been carried out to find an accurate computer vision solution but there are still many challenges that need to be resolved. These include occlusions, varying lighting, long processing time and improving the accuracy in image processing [2, 36, 50, 63].

This work distinguishes itself with the following four main contributions. First, a new combination of SIFT and FAST features with pixel-wise technique is used to improve the accuracy. Second, motion edge pixels are used instead of foreground pixels to reduce the number of SIFT descriptors required. Third, a combination of grid map and pixel-wise technique is used to improve the cluster classification in frames which enables similar clusters in different cells to be assigned different densities depending on their location in the frame. Fourth, the algorithms are comprehensively tested and validated using two datasets, the University of California, San Diego (UCSD) and Mall datasets [13, 16].

2 Related work

Crowd counting can be classified into four categories; crowd counting based on detection, clustering, regression and optimisation.

2.1 People detection based algorithms

Detection based algorithms start by detecting people individually and then counting them [2]. The detection process depends either on the person's entire body or parts of the body such as face, head or head-shoulder [2]. The main advantages of these algorithms are that they count people and find

their locations as well, therefore they are useful in people tracking [35]. The main disadvantages of these algorithms are that they are severely affected by varying lighting, occlusion and have long processing times [72]. They achieve good results in sparsely populated scenarios, whereas in crowded scenarios, the accuracy decreases significantly [35]. In addition, a high resolution camera is required to obtain a good accuracy [35]. Triggs and Dalal proposed histogram of oriented gradients (HOG) method thereby creating a basis for the development of a fast appearance-based detection algorithm [25]. Many improvements of the HOG technique have been proposed. One of the most promising variants is the fastest pedestrian detection in the west (FPDW) which has significantly increased the speed of detection [27].

Pedestrian detection is constrained to horizontal or vertical camera angles. In people counting, horizontal camera angles can be used but a vertical or downward facing angle is often preferred to minimise occlusions [67]. A majority of commercial people counting products are cameras that are placed on the ceiling pointing downwards to get the best view. However, this is not an optimal set-up if the detection area needs to be maximized.

People detection based algorithms can be classified into six categories: full body detection [25, 46, 73]; part body detection [28, 49, 77]; 3D camera detection [33]; shape matching detection, where ellipse and Bernoulli shapes are used to identify and count people in each blob [31, 48]; multi-camera detection, which is used to avoid occlusion [53]; and density-aware detection, which is used to reduce the false positive per image (FPPI) in low crowd density locations and decreases the miss rate in high crowd density locations in the frames [58].

2.2 Features trajectories clustering based algorithms

Clustering based algorithms track visual features over time and the feature trajectories are then clustered into unique tracks using temporal, spatial and other factors [9, 18, 56, 71]. The number of clusters is the estimated number of people [54]. Different approaches have been used to study the similarities between trajectories such as Dynamic Time Warping (DTW). DTW is a time series method that is widely used to measure similarities between two temporal sequences [4, 6]. Kanade–Lucas–Tomasi (KLT) feature-matching algorithm is sometimes used to find the trajectories of features [56]. The advantages of clustering based algorithms are that they can decrease the occlusion and the angle of the camera effects [75]. However, their accuracies significantly decrease in highly crowded environments with cluttered background and heavy occlusion. A complicated trajectory management technique is required to assess the similarities of trajectories with different lengths, which is another limitation of these algorithms [64]. In addition, errors in the number of people due to the cohesiveness of features that belong to different people also affect their accuracies [64]. These algorithms also require high video frame rate to work well because motion information can reliably be extracted [16]. Features trajectory clustering algorithms can be used to count people but it is difficult to use them in a real-time environments due to their long processing times.

2.3 Low-level features regression based algorithms

Regression based algorithms usually consist of three steps, starting with a background segmentation that is used on a frame by frame basis to detect the foreground information. Low-level features are then extracted from the foreground such as edge features [16, 19–21, 60], segment features [11–14, 19–21, 35, 81], texture features [15, 17, 19, 78] and keypoints [24]. A regression function is then trained using these features to find the relationship between the number of people and the extracted features which is then used to estimate the number of people [71]. Various types of regression

functions have been used such as support vector machine tree [23, 78], linear [26, 52], neural networks [19–21, 60] and Gaussian process algorithms [11–14, 54]. A significant amount of research has been carried out to improve these algorithms by varying the number of features. Some other researchers have tried to improve them by using more than one regression function and then choose the best fitting features [29]. The main advantages of these algorithms are that the accuracy is higher than feature trajectory clustering and detection based algorithm in crowded scenarios, and the computational time is shorter [16, 29, 72]. Their main disadvantage is that different training datasets are required with different environments or camera set-ups [71]. Some new contributions have also been presented to improve their accuracies, handling occlusions and adapt to new environments. Recent technique in crowd counting has been tested using static pictures from crowded environments [37]. A deep-learning approach that uses convolutional neural networks to predict the number of people has been proposed in that technique. Occlussion is another problem that a new proposed technique tries to minimise [3]. Research in [3] takes occlusion into account by using two regression functions, one for the low occlusion frames and the second for the high occlusion frames. In addition, adaptive combination of features is used in each environment according to their nature. Statistical features have been used by Hafeezallah et al. [32] to train a neural network to develope a highly accurate crowd counting algorithm. The differences of the sequential frames with curvelet transform has been proposed by [32] to improve the accuracy. A random projection forest, as a regression function, has also been proposed by other researchres to increase the maximum number of features that is used for training [80]. A small number of features can be handled by traditional regression functions which can negatively affect the performances of crowd counting systems. Aravinda et al. [57] have proposed a combination of optical flow for motion cues and hierarchical clustering to estimate the crowd density. Hierarchical clustering have been in [57] used to isolate distinct pixels that correspond to different people in the frame. Multi-cameras knowledge transfer technique has been used by Nick et al. [69] to provide different views of the crowd which are used to minimise occlusion and improve performance. The main disadvantage of the technique is the long set-up time required and the high cost of the hardware. Finally, a quadratic programming technique is used with a regression function and network flow constraints to improve the accuracy of estimating the number of people [30]**.** They take into account the temporal domain of a series of frames to improve the accuracy. Regression based algorithms are classified into three categories; holistic, histograms (intermediate) and local algorithms [62].

Holistic algorithms use global image features and one regression function for the whole frame [11–14, 19–21]. The types of features that used by these algorithms include foreground, edge , keypoints and texture features. A limitation of these algorithms is that they apply one global regression function over the whole image thereby not taking into account the high variability of crowd distribution, behaviour and density in different regions of the image [62].

Histogram features are used by histograms algorithms such as, edge orientation histogram, blob size histogram and histogram of oriented gradients (HOG) [44, 45]. One global regression function is trained by these features to find the estimated number of people. These algorithms use histogram bin magnitude and edge direction to avoid noise and to distinguish people, respectively. Histograms algorithms also ignore high variations in crowd behaviour, distribution and density in different regions of the image [62].

Local algorithms count the number of people by partitioning the frame into several regions and one local regression function is trained for each region to count the total number of people in the whole frame. The regions can be cells having regular or irregular sizes [16] or the regions can be foreground blobs and the total number of people is counted by summing the numbers in all regions [10, 22, 40, 43, 60].

2.4 Pixel-wise optimisation based algorithms

Some researchers use pixel-wise techniques to estimate the number of people [47]. In this approach, the density of each pixel is found and then integrated over the whole frame to estimate the total number of people [47]. Optimisation is used instead of regression to train crowd counting systems. This approach can be used to improve people detection algorithms by combining it with full or part body detection based algorithms [58]. Full body, head and head-shoulder detection based algorithms can be improved and the accuracy can be increased by using the density of pixels [58]. The aim of this combination is to reduce the false positive per image (FPPI) in low crowd density locations in the frames which happens when it inaccurately detects the presence of people when there is actually nobody. In addition, this approach decreases the miss rate in high crowd density locations in the frames.

Pixel-wise optimisation based algorithms can be trained using a small number of frames in comparison to regression based algorithms [47]. As a consequence, the set-up time of the system can be reduced by more than 25% in comparison to regression based algorithms which lead also to low set-up cost. Using a large number of training frames can negatively affect the accuracy of the training because manually annotation is an error-prone task.

3 System design

In this paper, the proposed system depends on supervised learning to estimate the number of people. The training frames are annotated and Gaussian representation is used to represent people. Quadratic programming is used for learning and maximum excess over subarrays distance (D_{MESA}) is used to measure the difference between the true and predicted count which represent the loss function as given by Eq. (2).

The proposed system assumes that each pixel (p) in a frame is represented by a SIFT or SIFT-FAST feature vector. The density function of each pixel is represented as a linear transformation of the pixel representation (x_p) as given by Eq. (1);

$$F(p) = \mathrm{w}^{\mathrm{T}}\mathrm{x}_{\mathrm{p}} \tag{1}$$

Where w^T is the weight of each pixel in the frame. At the learning stage, a training frames set with their ground truth (true count) are used to find the correct weight (w^T) of each pixel. Then the densities of all pixels in the frame are summed to find the predicted count. D_{MESA} is used to compare between the predicted count and true count as a loss function. D_{MESA} is defined as [58];

$$D_{\mathrm{MESA}}(F1, F2) = max\left|\sum\nolimits_{p\in B} F1(p) - \sum\nolimits_{p\in B} F2(p)\right| \tag{2}$$

Where $F1(p)$ and $F2(p)$ are the predicted count and true count of people in a frame. D_{MESA} is chosen for the proposed system because it is not significantly affected by jitter and noise but it has a strong relationship with the number and positions of people [47]. The ultimate goal of the learning stage is to find the best weight for each pixel that minimises the sum of the errors between the true counts and the predicted counts (the loss function) [47];

$$w = \mathrm{argmin}_w\left(w^T w + \gamma\sum\nolimits_{i=1}^{N} D_{\mathrm{MESA}}\right) \tag{3}$$

Where γ is a scalar parameter to control the regularization strength, $argmin_w$ represents the best weight that minimises the D_{MESA}. Quadratic programming can be used to solve Eq. (3) by using;

$$\min_{w,\xi_1,.....,\xi_N}\left(w^T w+\gamma\sum\nolimits_{i=1}^{N}\xi_i\right) \tag{4}$$

Subject to;

$$\xi_i \geq \sum\nolimits_{p\in B}(F1(p)-F2(p)), \quad \xi_i \geq \sum\nolimits_{p\in B}(F2(p)-F1(p)) \tag{5}$$

Where ξ_i are the auxiliary variables of training frames. Quadratic programming uses iterations to optimise the results and find the best weight (w^T) of each pixel. The iterations terminate when the right side of equation (5) is within ($\xi_i + \beta$) factor. β is a small constant ($\beta << 1$). It uses to decrease the number of iterations and faster convergence. Choosing β equal to 0 solves the equations (4) and (5) exactly. However, the convergence will finish faster if β is chosen to a very small value and that will not affect the performance of training [47]. In the experiments of the proposed system, β has been chosen to be equal to 0.001. The flow diagram of the proposed system is illustrated in the Fig. 1. It consists of two counting algorithms, one video source and one fusion model.

3.1 Algorithm 1: SIFT features algorithm

This algorithm combines the following techniques to count the number of people; motion edges, SIFT descriptors, gird map and pixel-wise techniques. This combination that is used to find the density of each pixel, is novel. Edge pixels are used because their number is less than foreground pixels. As a consequence, the required time to find the SIFT descriptors and cluster them in a frame will be significantly reduced which makes the proposed system faster than other people counting techniques based on D_{MESA} optimisation. There is a high correlation between SIFT descriptors and the number of people. This is difficult for quadratic programming to be used to find the density for a large number of SIFT descriptors (equal to the number of edge motion pixels). To solve this problem, clustering is used to reduce the number of SIFT descriptors to 256 clusters. The main disadvantage of using clustering is that many SIFT descriptors can be grouped into one cluster to reduce the problem space but they represent different densities. Grid map is used to improve the cluster classification in the frames which enables similar clusters in different cells to be assigned different densities depending on their location in the frame. The proposed algorithm can better adapt to high variations in crowd behaviours, distributions and densities. As a result, the accuracy is improved. Figure 2 shows the flow diagram of this algorithm. The procedure of the algorithm is illustrated in the following steps;

1- Implement Gaussian mixture model (GMM) to find the foreground information of the frame.

$$F_{GMM} = GMM(i,j) \tag{6}$$

Where F_{GMM} is the foreground pixels of the frame and $GMM(i,j)$ is the Gaussian mixture model of each pixel of the frame.

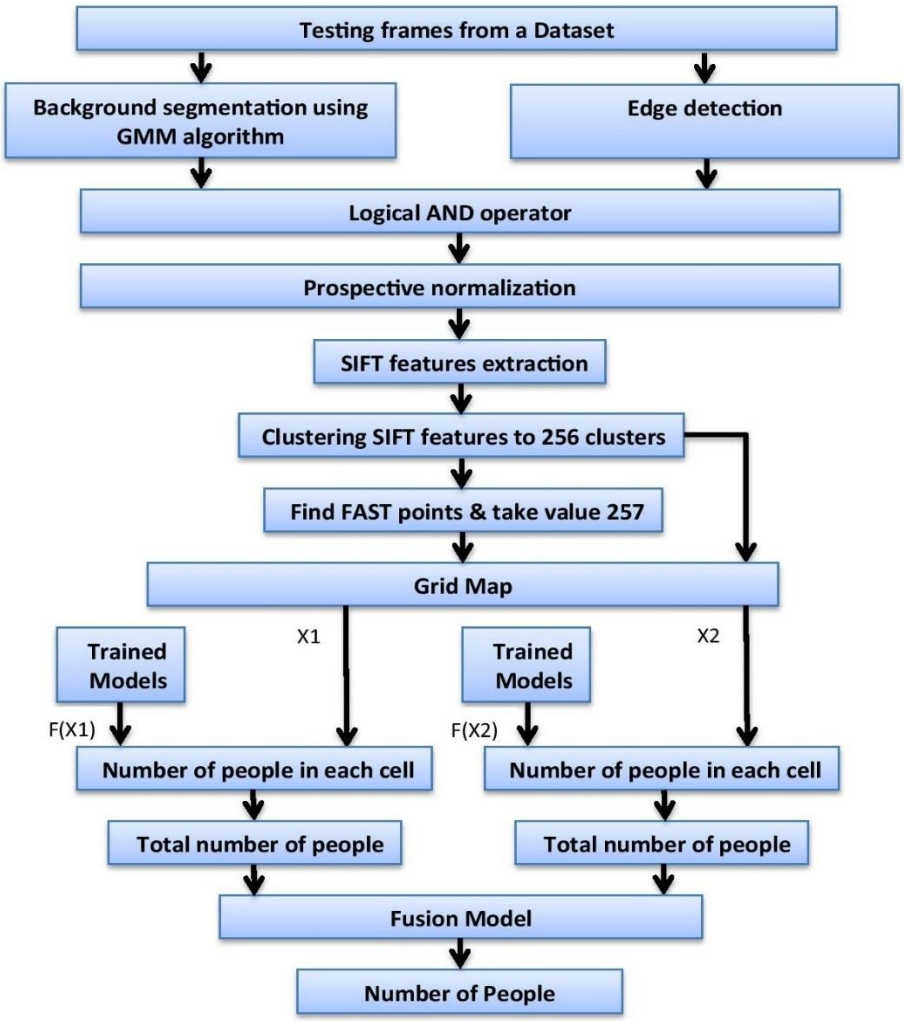


Fig. 1 Flow diagram of the proposed system

2- Implement edge detection to find the edges of the frame.

$$F_{Edge} = E(i,j) \tag{7}$$

Where F_{Edge} is the edge of the frame and $E(i,j)$ is the detected edge of each pixel of the frame.

3- Perform logical (AND) operation between the foreground pixels of the frame and the detected edge to find the motion edge of the frame.

$$F_{motion\ edge} = F_{GMM}\ (i,j)\&\&F_{\mathrm{Edge}}\ (i,j) \tag{8}$$

Where $F_{motion\ edge}$ is the motion edge for the frame.

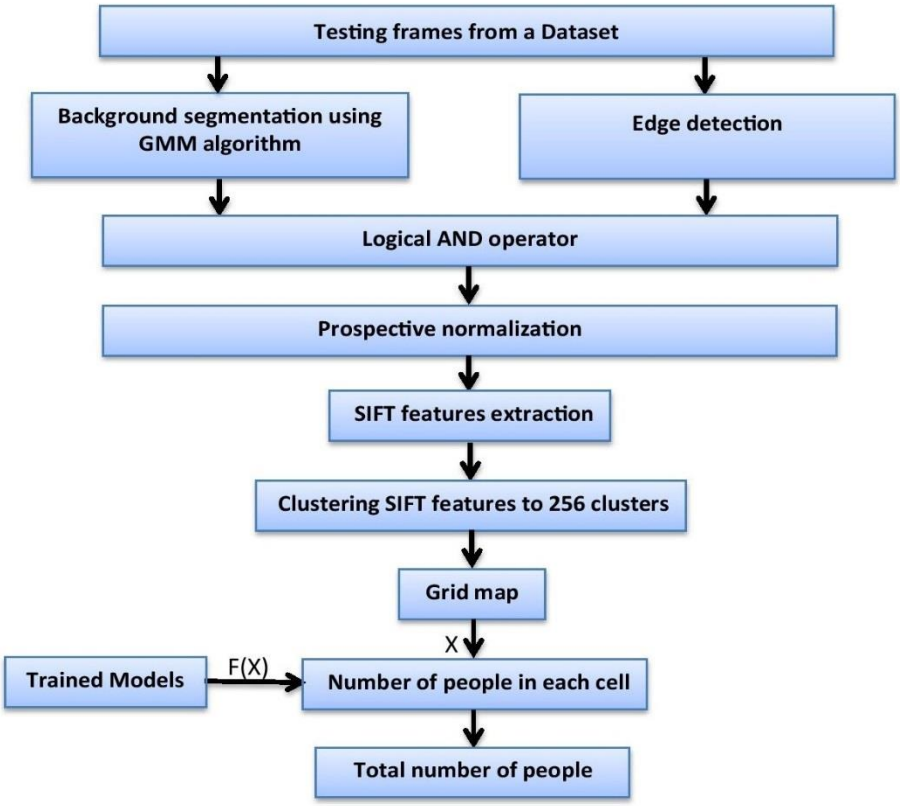


Fig. 2 Flow diagram of the SIFT Features algorithm

4- The pixels in each line of the frame are assigned different weight as a perspective normalization.
5- Find the SIFT descriptor for each motion edge pixel. Then, cluster the SIFT descriptors to 256 clusters. The centres of SIFT features are used as criteria for clustering them.

$$F_{SIFT} = SIFT(i,j) \qquad (i,j) \in \text{motion edge} \tag{9}$$

$$F_{Cluster} = Cluster(F_{SIFT}) \qquad (i,j) \in \text{motionedge} \tag{10}$$

Where F_{SIFT} is the SIFT descriptors of the frame and $F_{Cluster}$ is the SIFT descriptors clustering.

6- Divided the frames into cells (as a grid map) and count the number of people in each cell.

$$F_{Grid} = \sum\nolimits_{n} C_n \tag{11}$$

Where F_{Grid} is the grid map of each frame, C is a cell in the grid map and n is the number of cells in the grid map. Four cells configuration has been used in the proposed system which gives the best accuracies experimentally.

7- Use a quadratic programming (Interior-point-convex algorithm) to find the density of each cluster in each cell.
8- Integrate the densities of pixels over each cell to find the number of people in each cell.

$$N_{cell} = \sum_{(i,j)\in B_n} P_{density}(i,j) \tag{12}$$

Where N_{cell} is the number of people in each cell and $P_{density}(i,j)$ is the density of each pixel that belongs to this cell.

9- The summation of the number of people in all cells represents the total number of people in the frame.

$$N_{total} = \sum\nolimits_n N_{cell} \tag{13}$$

Where N_{total} is the total number of people in a frame and n is the number of cells.

3.2 Algorithm 2: SIFT-FAST features algorithm

This algorithm uses two features; FAST and SIFT. This algorithm combines the following techniques to count the number of people; motion edges, grid map, SIFT & FAST features and pixel-wise techniques. Edge pixels are used because their number is less than those of foreground pixels. The same approach as for SIFT feature algorithm described in Section 3.1 is used. However, FAST corner points are used to improve the accuracy due to the high correlation between the number of people and FAST corner points. The algorithm can also better adapt to high variations due to crowd behaviours, distribution and density. Figure 3 shows the flow diagram of the algorithm. Steps 1 to 5 are the same as for SIFT feature algorithm and descriptions from step 6 are as follows:

6- Find FAST points in each frame within the motion region.

$$F_{FAST} = F(i,j) \qquad (i,j)\in \text{motion regions} \tag{14}$$

Where F_{FAST} is the FAST corner points of a frame.

7- All pixels that are FAST corner points are assigned the value 257 so that quadratic programming can be used to find 257 density values instead of 256.
8- Divide the frame into cells (as a grid map) and the number of people in each cell is counted individually.

$$F_{Grid} = \sum\nolimits_n C_n \tag{15}$$

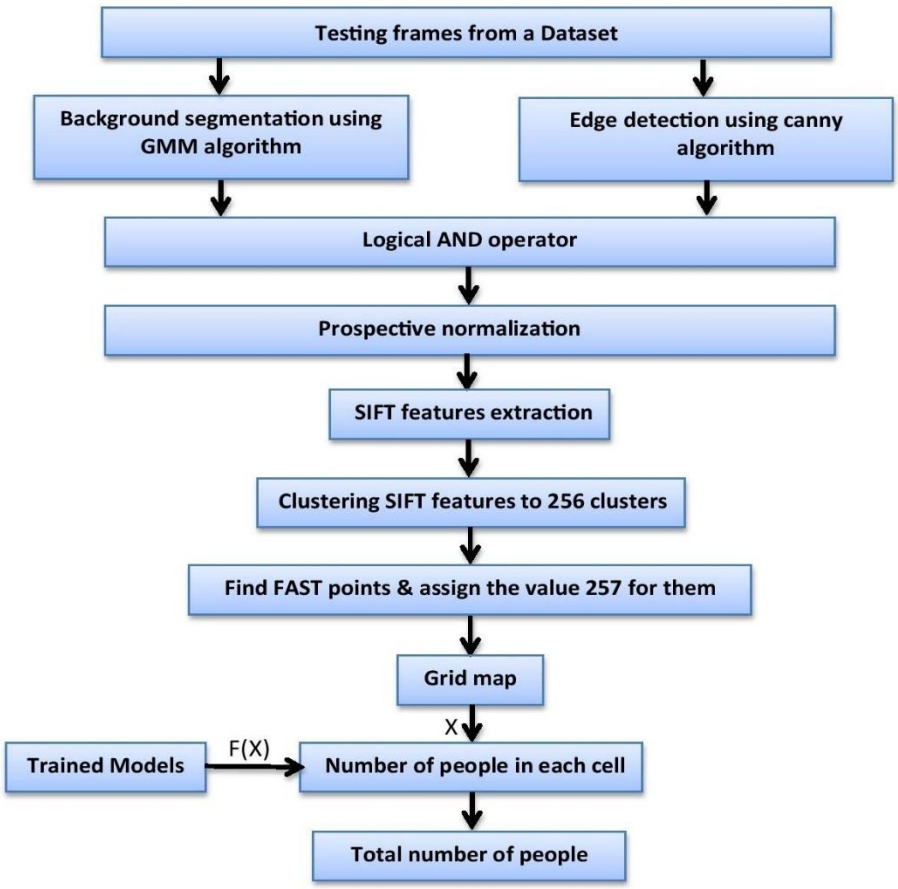


Fig. 3 Flow diagram of the SIFT-FAST features algorithm

Where F_{Grid} is the grid map of the frames, C is a cell in the grid map and n is the number of cells in the grid map.

9- Use a quadratic programming (Interior-point-convex algorithm) to find the density value of each cluster.

10- Integrate the densities of pixels over each cell to find the number of people in each cell.

$$N_{cell} = \sum_{(i,j)\in B_n} P_{density}(i,j) \tag{16}$$

Where N_{cell} is the number of people in each cell and $P_{density}(i,j)$ is the density of each pixel that belongs to the cell.

11- The summation of the number of people in all cells represents the total number of people in each frame.

$$N_{total} = \sum\nolimits_{n} N_{cell} \tag{17}$$

Where N_{total} is the total number of people in a frame, n is the number of cells.

3.3 Fusion technique

The fusion model is updated periodically using the results of all the algorithms. Each algorithm works independently to count the number of people and then they update the fusion model. Fusion is used to improve accuracy by determining the average error for each frame and to increase the confidence of the proposed system because the result of one algorithm is confirmed by that of another. This produces a cooperative paradigm and improves the confidence level in the results.

3.4 Geometric correction

At long distances, people appear smaller than those closer to the camera. Therefore, the extracted features of the same person at different locations in the scene are significantly different.

Re-scaling the pixels of the frames is implemented by assigning different weights to solve this problem. Fig. 4 shows the different sizes of the same pedestrian at different depths. Line (*ab*) is the reference line so the pixel's weight on that line is 1, the pixels of other lines are scaled and weighted using equation (18) [50];

$$weight_{line} = \frac{h_{ab} w_{ab}}{h_{\text{line}} w_{\text{line}}} \tag{18}$$

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

3.5 Background segmentation

Background segmentation is a process of extracting foreground information on a frame by frame basis. Background segmentation algorithms usually consist of three steps; background initialization, foreground detection and background maintenance [68]. In the background initialization, various techniques such as statistical, fuzzy and neuro-inspired techniques are used to build a background model. In foreground detection, a comparison is implemented between the current frame and the background model. Updating a background model according to changes in the environment is processed in the background maintenance step. Background segmentation

Fig. 4 The change of size of the same person at different locations

methods can be classified into recursive and non-recursive algorithms [2]. In non-recursive algorithms, the background model is considered to be static and does not update, whereas in recursive algorithm, it is a dynamic and changes depending on the change of environment [2]. Figure 5 shows the general block diagram of background segmentation algorithms.

GMM is one of the most widely used algorithms for background segmentation. This algorithm is a robust in light varying conditions and in environments with animated textures such as waves on the surface of water or trees being blown by wind [1]. Each pixel in a background model is formed using a mixture of Gaussian distributions (normally from three to five distributions) rather than one Gaussian distribution [1, 5].

$$p(X_t) = \sum_{i=1}^{K} w_{i,t} * f\left(x_t | \mu_{i,t}, \Sigma_{i,t}\right) \tag{19}$$

Where K is the number of Gaussian distributions and $w_{i,t}$ is the weight of the i^{th} distribution at time t. Each Gaussian distribution can be found using the probability density function;

$$f(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right) \tag{20}$$

Where μ is the mean and Σ is the covariance matrix. The background model is updated using an adaptive filter;

$$\mu_t = \alpha X_t + (1-\alpha)\mu_{t-1} \tag{21}$$

Where;

- μ_t denotes the spatial mean of the pixels at time t,
- μ_{t-1} denotes the previous spatial mean of the pixels at time t − 1,
- α is an empirical weight and.
- X_t is the current pixels values.

3.6 Edge detection

They refer to the process of localising pixel intensity transitions [61]. There is a strong relationship between the complexity of crowds and the number of people because crowded

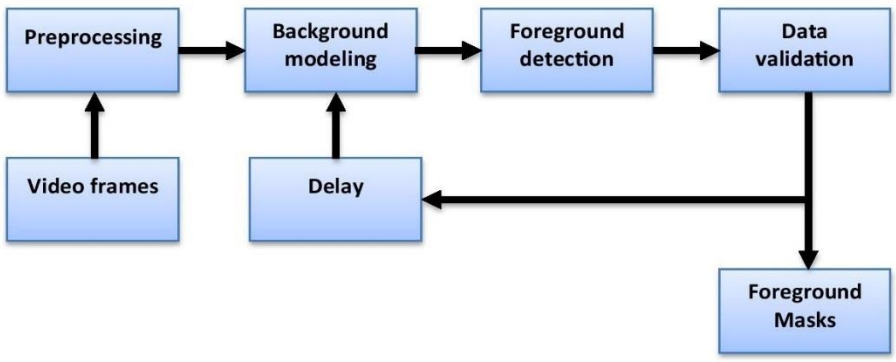


Fig. 5 General block diagram of background segmentation algorithms

environments tend to produce complex edges, while sparse environments tend to produce coarse edges [50]. Edges can be extracted using different algorithms such as Sobel, Canny, Prewitt, Roberts and Fuzzy logic algorithms [41, 42]. Canny edge detection is used in the proposed system. The following steps explain the procedure of canny edge algorithm [66]:

1- Smooth the image using a Gaussian filter to minimise noise.

$$S(i,j) = G(i,j,\Sigma) * I(i,j) \tag{22}$$

Where $G(i,j,\Sigma)$ is a Gaussian filter and $I(i,j)$ is a pixel.

2- Use derivative approximation by finite differences to find gradient magnitude and orientation. Firstly, partial derivatives $X(i,j)$ and $Y(i,j)$ is found by using the smoothed array $S(i,j)$:

$$X(i,j) \approx (S(i,j+1) - S(i,j) + S(i+1,j+1) - S(i+1,j))/2 \tag{23}$$

$$Y(i,j) \approx (S(i,j) - S(i+1,j) + S(i,j+1) - S(i+1,j+1))/2 \tag{24}$$

The partial derivatives $X(i,j)$ and $Y(i,j)$ are then used to find the magnitude and orientation of the gradient:

$$M(i,j) = \sqrt{X(i,j)^2 + Y(i,j)^2} \tag{25}$$

$$\theta(i,j) = \arctan(X(i,j), Y(i,j)) \tag{26}$$

3- Non-Maximal Suppression algorithm (NMS) is performed to thin out the edges. The edges are then detected using the double thresholding algorithm.

3.7 Clustering

Clustering is used in the proposed system to reduce the number of different descriptors (hundreds of thousands for 640 × 480 frame size) into a reasonable number of clusters (256 clusters in the SIFT features algorithm and 257 clusters in the SIFT-FAST features algorithm) that can be used with quadratic programming. K-means clustering is a method of vector quantisation and aims to partition n observations into $k \leq n$ clusters such that each observation belongs to the cluster with the nearest mean [7]. In other words, it aims to find:

$$argmin_S \sum_{i=1}^{K} \sum_{X \in S_i} \|X - \mu_i\|^2 \tag{27}$$

Where X is the observation, S_i is the i^{th} cluster and μ_i is the mean of cluster S_i. In the proposed system, the k-means algorithm is used to cluster the SIFT descriptors of the datasets frames and produce a codebook of 256 entries. The codebook is constructed using only the descriptors of the training frames and then the descriptors of the testing frames are clustered by the K-means algorithm and the codebook. The SIFT descriptor of each pixel is represented by one value between 1 and 256. A vector of length 256 is used to convert each pixel and quantise it by comparing them with the centroids in the codebook.

4 Results and discussion

4.1 Benchmark datasets

The pedestrian dataset from University of California, San Diego (UCSD) and the Mall datasets have been used to evaluate the proposed system [13, 16]. UCSD dataset has been widely used for testing and validating people counting methods [82]. Mall dataset is a newer and more comprehensive dataset to use in that it covers a different range of crowd densities, different activity patterns (static and moving crowds), collected under a large range of illumination conditions at different times of the day with more severe perspective distortion. Thus individual objects may exhibit larger variations in size and appearance at different depths of the scene [50]. The Mall dataset was introduced by Chen [16]. It has been collected inside a cluttered indoor and includes 2000 annotated frames. The two datasets have the same length (2000 frames) but they have different features in terms of the frame rate (fps), resolution, colour, location, shadows, reflections, crowd size and frame type [62, 63]. Table 1 shows the features of each dataset.

For the Mall and UCSD datasets, the datasets are partitioned into a training set, for learning the proposed system, and a test set, for validation. 100 frames from different locations of each dataset are allocated indivdually for training and 1900 frames for testing.

4.2 Evaluation metrics

Three metrics have been used as performance indicators for crowd counting; Mean deviation Error (MDE), mean absolute error (MAE) and mean squared error (MSE) [50]. The MDE is defined as;

$$MDE = \frac{1}{N}\sum_{n=1}^{N}\frac{\left|y_n - \hat{y}_n\right|}{y_n} \tag{28}$$

Table 1 The features of the benchmark datasets

	Mall dataset	UCSD dataset
Year	2012	2008
Length (frames)	2000	2000
Frame rate (fps)	<2	10
Resolution	640 × 480	238 × 158
Colour	RGB	Grey
Location	Indoor	Outdoor
Shadows	Yes	No
Reflections	Yes	No
Loitering	Yes	No
Crowd size	11–45	13–53
Frame type	.jpeg	.png

The MAE is defined as;

$$MAE = \frac{1}{N}\sum_{n=1}^{N}\left|y_n - \hat{y}_n\right| \tag{29}$$

The MSE is given as;

$$MSE = \frac{1}{N}\sum_{n=1}^{N}\left(y_n - \hat{y}_n\right)^2 \tag{30}$$

Where N is the total number of the test frames, y_n is the actual count, and y_n is the estimated count of n_{th} frames. MAE and MSE are indicative quantities of the error of the estimated crowd count but they contain no information about how crowded the environment is [50]. MDE takes into account the crowdedness and gives an indication of how good a measurement is relative to the actual count [32].

4.3 Background segmentation, edge detection and motion edge extraction

The GMM is used for background segmentation and the Canny edge algorithm is performed to extract the edges of the frames. The logical 'AND' is used to extract motion edge. Figs. 6 and 7 show the results of the background segmentation, edge detection and motion edge extraction of two sample frames, one from the Mall dataset and the second from the UCSD dataset.

4.4 Performance evaluation of the proposed system using the mall dataset

As shown in Table 2, the mean deviation error (MDE) of the SIFT features algorithm is 0.099 and 0.094 for SIFT-FAST features algorithm. The results are compared with results presented by other researchers for the same dataset as a measure of accuracy of the proposed system. From the results, we can see that the accuracy of the SIFT-FAST features algorithm is slightly better than that of SIFT features algorithm. It shows that there is a reasonable improvement in the accuracies of the implemented algorithms when compared to those published by other researchers. Figure 8 shows the percentage of frames within the MDE distribution of the algorithms. Figure 9 shows the true count (TC) of people from sample frames of the Mall dataset, which is annotated by red dots. EC1 and EC2 represent the estimated number of people using SIFT and SIFT-FAST features algorithms, respectively.

The performance of crowd systems is measured using the accuracy (MAE, MSE and MDE) and practicality. Practicality is measured by the percentage of the training frames minimisation [60]. Crowd counting systems are practical if they are easy to deploy. In the real world, crowd systems are deployed in different environments which means they are individually trained for the location. Therefore, it is very important to reduce the number of the training frames required. The ground truth (the actual number of people) for each training frame is required when training crowd counting systems. Each environment needs several hundreds of frames (usually 400–800 training frames) for the training [15, 23, 24, 83], so the training process becomes time-consuming.

The results of the proposed system have been compared with recent results from other researchers for the evaluation. The comparison with other methods based on the accuracy metrics (MAE, MSE and MDE) is not enough to measure the performance for many reasons:

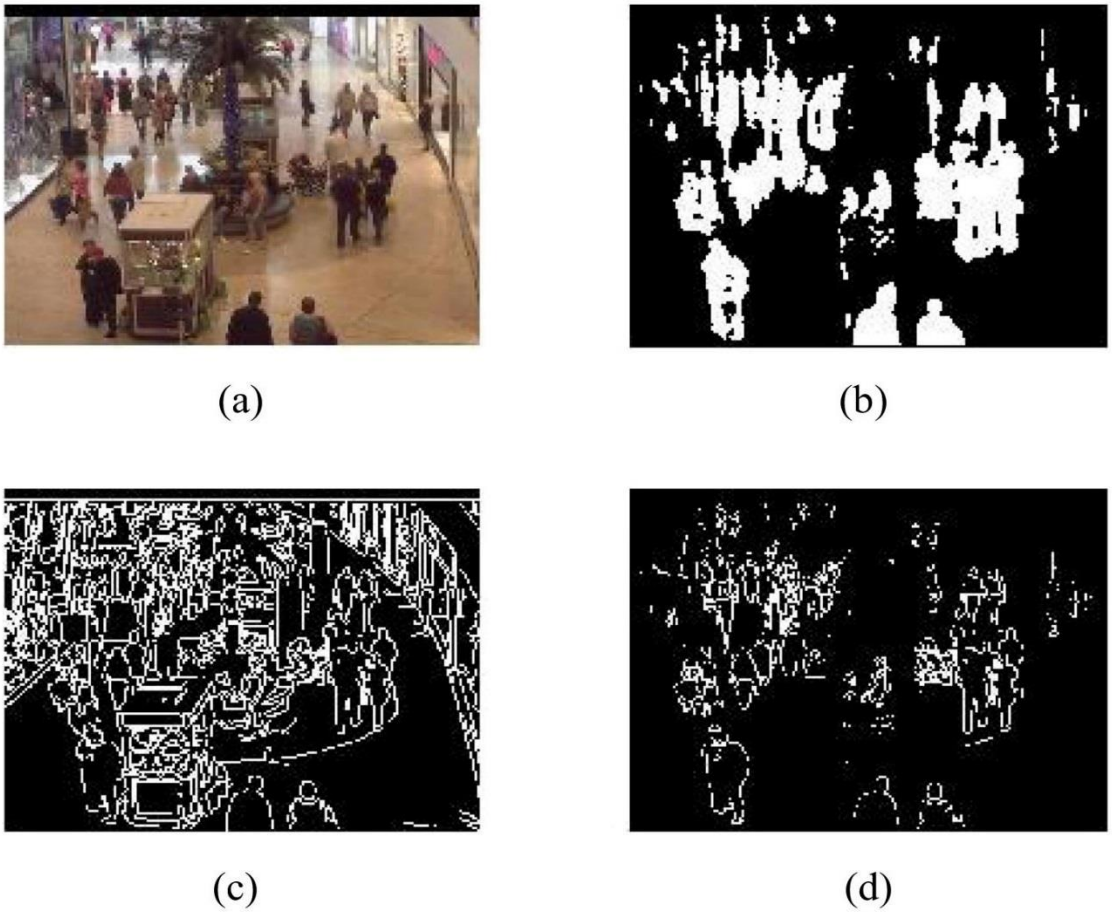


Fig. 6 **a** An example of the Mall dataset form; **b** foreground, using GMM algorithm; **c** edge using Canny detector; **d** the motion edge, using logical 'AND'

firstly, pixel-wise optimisation based algorithms can be trained using a small number of frames in comparison to regression based algorithms [47]. The proposed system uses 100 frames for the training whilst the other state of the art methods use between 400 and 800 frames [15, 23, 24, 83]. In conclusion, the proposed system is more practical because the set-up time is faster by a factor of at least four (uses 4 times less training frames) compared to regression based algorithms which lead also to low set-up cost. Secondly, the lower number of training frames required in the training stage reduces the potential for error being introduced because manually annotation is an error-prone task. The accuracy of crowd counting systems are significantly affected by errors in the training stage. Thirdly, the proposed system is a multipurpose system because it can be used for crowd counting and also in people detection [58].

The comparison is only used to show that although the proposed system reduces the training error, speed, cost and can be used to develop more accurate people detection methods, its accuracy is, at least, comparable with the state of the art methods. None of the published results presented in Table 2 performs better than SIFT-FAST features algorithm based on the metrics used in this paper. In terms of MSE, only the algorithms presented in [16, 80] and [38] produced slightly better results but not in terms of MAE and MDE metrics. Finally, the MDE of the proposed system is less than the acceptable error (0.2) which is meeting the minimum accuracy requirements of system operators [62].

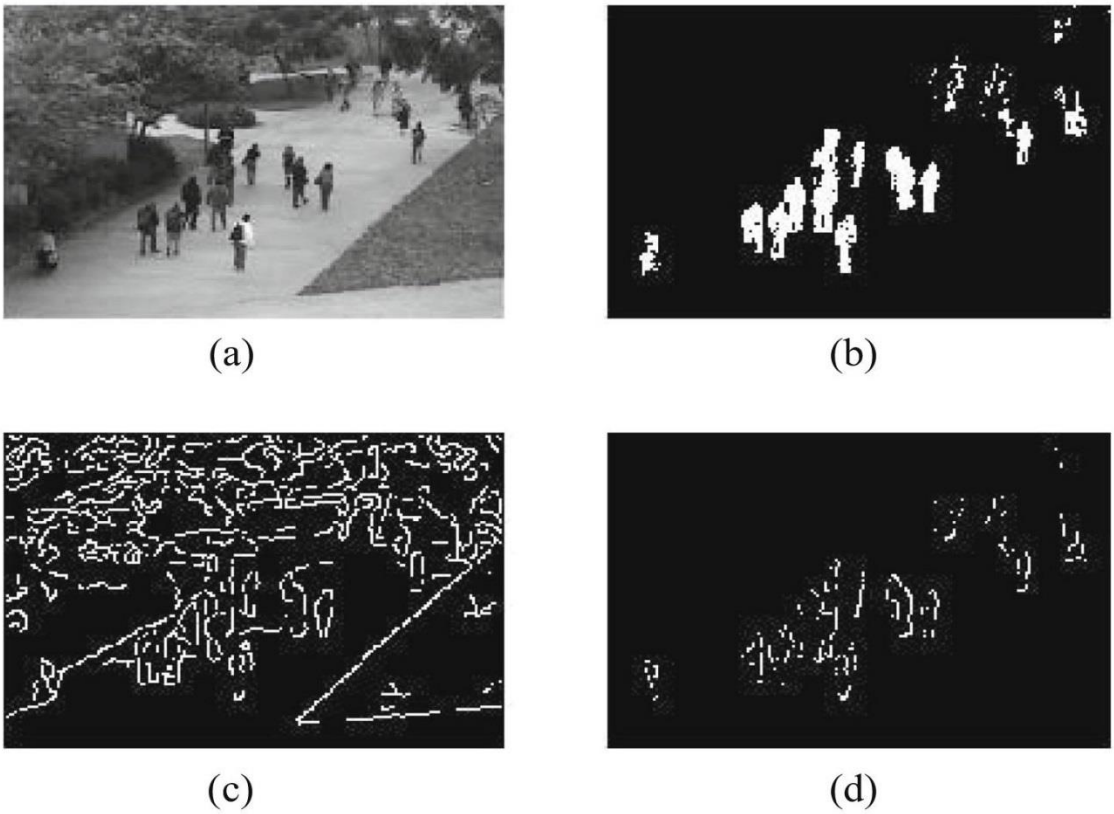


Fig. 7 **a** An example of the UCSD dataset form; **b** foreground, using GMM algorithm; **c** edge using Canny detector; **d** the motion edge, using logical 'AND'

Table 2 Comparison for the Mall dataset results between the proposed system and the state of the art algorithms

Algorithm	Mall dataset		
	MAE	MSE	MDE
Algorithm 1: SIFT Features Algorithm	3.08	16.31	0.099
Algorithm 2: SIFT-FAST Features Algorithm	2.94	14.64	0.094
Cumulative attribute based model (CA-RR) [17]	3.43	17.70	0.105
Squares Support Vector Machine Regression (LSSVR) [17]	3.51	18.20	0.108
Kernel Ridge Regression (KRR) [17]	3.51	18.10	0.108
Random Forest Regression (RFR) [17]	3.91	21.50	0.121
Gaussian Process Regression (GPR) [16, 17]	3.72	20.1	0.115
Ridge regression (RR) [16, 17]	3.59	19.00	0.110
Multi Output Ridge Regression (MORR) [16]	3.15	15.70	0.099
Multiple Localised Regression (MLR) [16]	3.90	23.90	0.119
Weighted Ridge Regression (WRR) [15]	3.44	18.00	0.105
Random Projection Forest (RPF) [80]	3.22	15.50	-
Cost-sensitive Sparse Linear Regression (CS-SLR) [38]	3.23	15.77	0.104

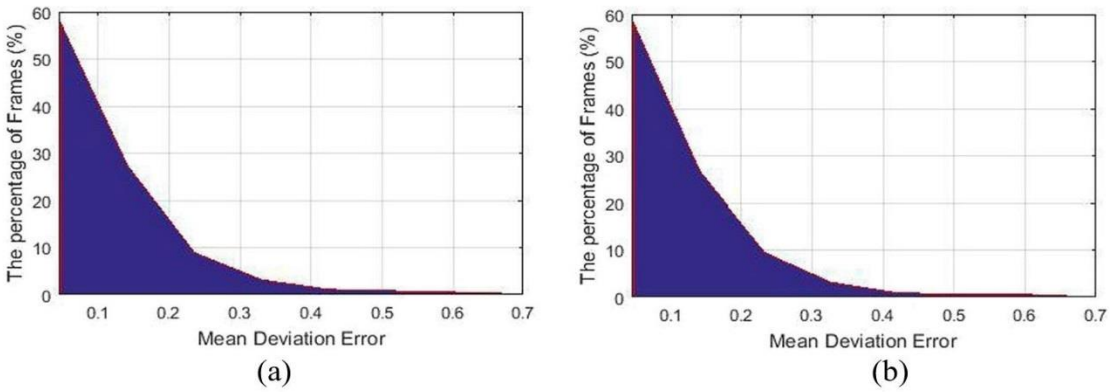


Fig. 8 **a** The MDE of SIFT algorithm **b** the MDE of SIFT-FAST algorithm

4.5 Performance evaluation of the proposed system using the UCSD dataset

The UCSD dataset represents people moving in two directions along a walkway. As shown in Table 3, the MDE of SIFT features algorithm is 0.066 and 0.064 for SIFT-FAST features algorithm. From the results, it can be seen that the accuracy of SIFT-FAST features algorithm is better than that of SIFT features algorithm. Figure 10 shows the percentage of frames within

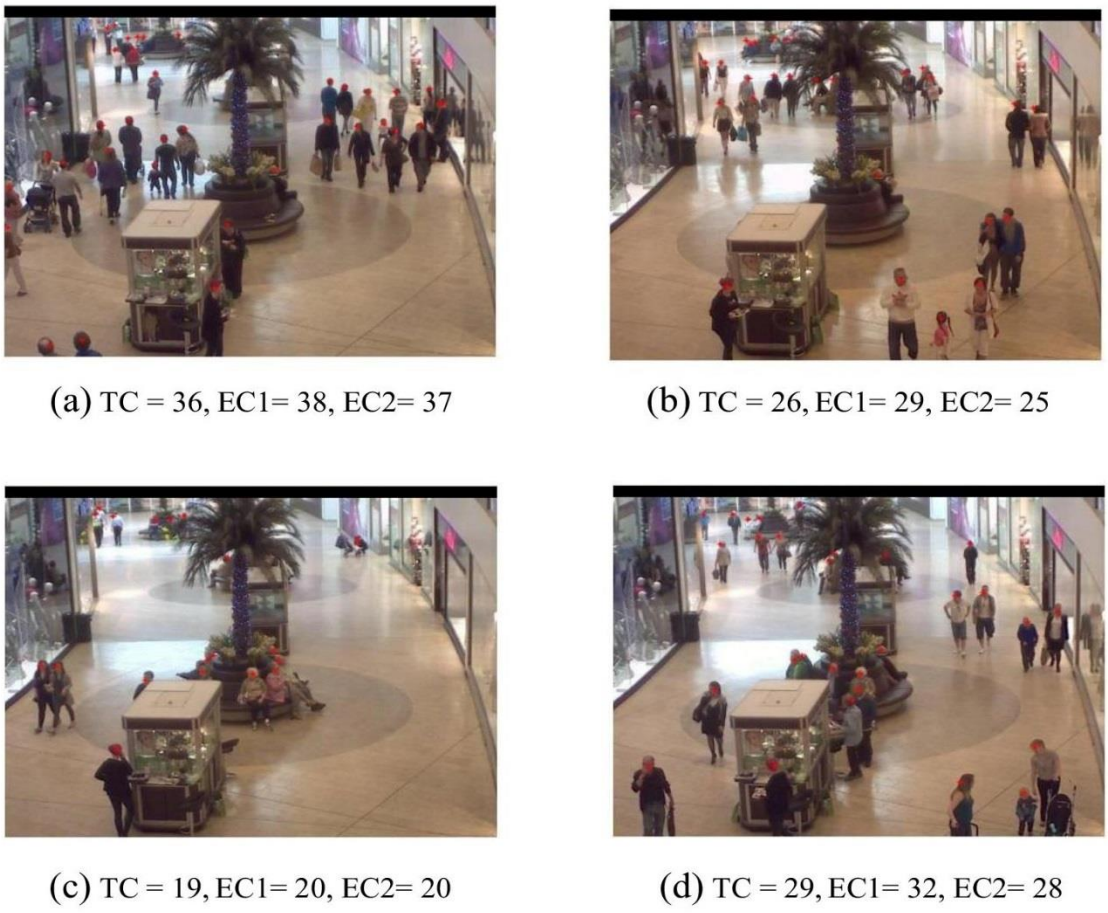


Fig. 9 Examples of the true count (TC) & the estimated count of people using SIFT (EC1) and SIFT-FAST (EC2) algorithms

Table 3 Comparison for the UCSD dataset results between the proposed system and the state of the art algorithms

Algorithm	UCSD dataset		
	MAE	MSE	MDE
Algorithm 1: SIFT Features Algorithm	1.82	5.24	0.066
Algorithm 2: SIFT-FAST Features Algorithm	1.76	4.93	0.064
Improved Iterative Scaling -Label Distribution Learning (IIS-LDL) [83]	2.08	7.25	0.098
Kernel Ridge Regression (KRR) [83]	2.16	7.45	0.107
Random Forest Regression (RFR) [83]	2.42	8.47	0.116
Gaussian Process Regression (GPR) [15, 83]	2.24	7.97	0.112
Ridge Regression (RR) [15, 83]	2.25	7.82	0.110
Multi Output Ridge Regression (MORR) [83]	2.29	8.08	0.109
Cumulative attribute based model (CA-RR) [17, 83]	2.07	6.86	0.102
Weighted Ridge Regression (WRR) [15]	2.05	6.75	0.102
Linear regression (LR), Partial Least Squares Regression (PLSR), KRR, LSSVR, GPR and RFR [50]	>2.02	>6.67	>0.100
Random Projection Forest (RPF) [80]	1.90	6.01	-
Cost-sensitive Sparse Linear Regression (CS-SLR) [38]	1.83	5.04	0.079
Moving SIFT algorithm [24]	3.26	-	0.180

the MDE distribution of the algorithms. Figure 11 shows the true count (TC) of people from sample frames of the UCSD dataset, which is annotated by red dots. In general, the accuracies of the proposed system with the UCSD dataset are better than the results from the Mall dataset. The potential justification is that the Mall dataset is more complicated in terms of shadows, reflections and crowd size [62, 63]. In addition, the Mall dataset is collected with more severe perspective distortion than the UCSD dataset. As is the case with the MDE from the Mall dataset, the MDE of this dataset is significantly lower than the acceptable error (0.2) which is meeting the minimum accuracy requirements of system operators [62]. Results of both datasets show that their average accuracies for each dataset are almost similar but their accuracies at frame level are different. The difference of estimation between the SIFT and SIFT-FAST algorithms for each frame is usually between 0 and 4. EC1 and EC2 at each frame are

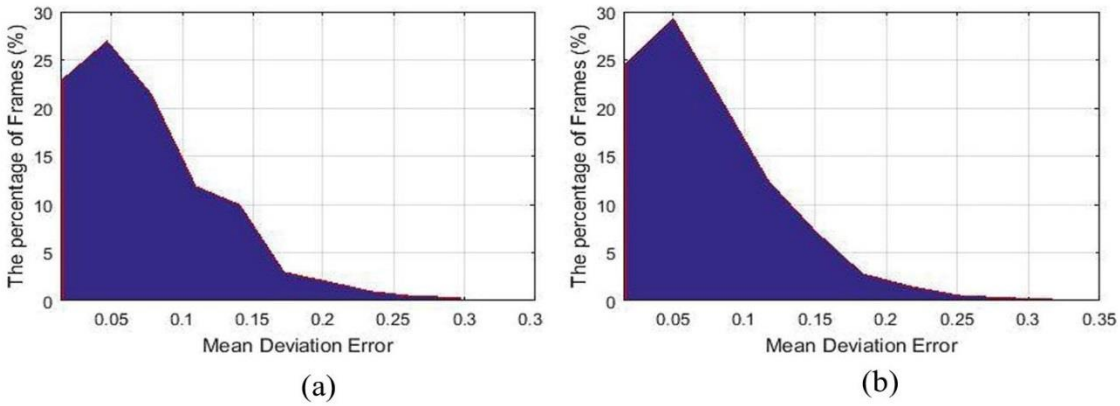


Fig. 10 **a** MDE of SIFT algorithm **b** MDE of SIFT-FAST algorithm

(a) TC = 18, EC1= 19, EC2= 18

(b) TC = 23, EC1= 22, EC2=24

(c) TC = 15, EC1= 15, EC2= 17

(d) TC = 23, EC1= 21, EC2= 22

Fig. 11 Examples of the true count (TC) & the estimated count of people using SIFT (EC1) and SIFT-FAST (EC2) algorithms

correlative because both algorithms use almost the same approach. However, FAST corner points are used with SIFT-FAST features algorithm to improve the accuracy due to the high correlation between the number of people and FAST corner points. SIFT-FAST features algorithm gives the best results compared to all published results presented in Table 3. Only results presented in [38] gives a comparable results to the SIFT features algorithgm.

4.6 Performance evaluation of the proposed system in sparse and crowded scenarios

To evaluate the proposed system with sparse and crowded scenarios, the test set of the Mall dataset is split the same as in [50] into a sparse set which includes all the frames with ground truth (number of people), less than or equal to 30, and crowded set which includes all the frames with ground truth

Table 4 System performance with sparse and crowded scenarios (Mall dataset)

Algorithm	Sparse scenario			Crowded scenario		
	MAE	MSE	MDE	MAE	MSE	MDE
SIFT features Algorithm	3.20	18.39	0.126	2.96	14.27	0.081
SIFT-FAST features Algorithm	3.15	17.21	0.124	2.73	12.11	0.075

Table 5 System performance with sparse and crowded scenarios (UCSD dataset)

Algorithm	Sparse scenario			Crowded scenario		
	MAE	MSE	MDE	MAE	MSE	MDE
SIFT features Algorithm	1.67	4.41	0.093	1.93	5.84	0.055
SIFT-FAST features Algorithm	1.65	4.29	0.084	1.84	5.41	0.056

values greater than 30. The test set of the UCSD dataset is also split the same as in [50] into a sparse set which includes all the frames that their ground truth is less than or equal to 23, and crowded set which includes all the frames that their ground truth is greater than 23.

To ensure that the proposed system is practical and robust, the training set was not been split because the technical definition of the boundary that separates the sparse and crowded frames is not clear [3]. In addition, partitioning the training set into two sets would required two training stages. The test sets are processed by the proposed system jointly and then the results are analysed by splitting them into sparse and crowded sets. In conclusion, the split between sparse and crowded scenarios have mainly been carried out by identifying which frames could be classified into each of the categories. No differential training of the system has been carried out. Tables 4 and 5 show the results of both algorithms with sparse and crowded scenarios. The MDE of both algorithms in the sparse scenarios is higher than the MDE crowded scenarios. The proposed system is more applicable for high density crowds and this can be seen from the achieved good results in crowded

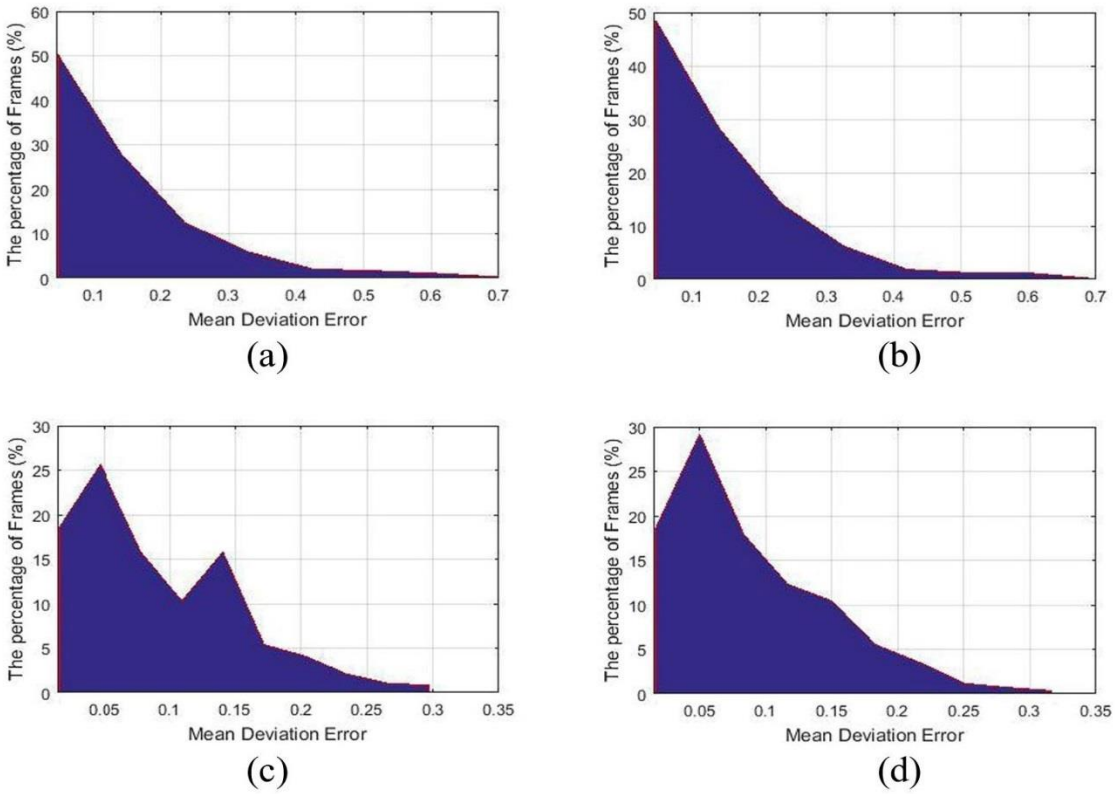


Fig. 12 System performance with sparse scenarios. (**a**) and (**b**) are the MDE of the of SIFT algorithm and SIFT-FAST algorithm on Mall dataset, respectively; (**c**) and (**d**) are the MDE of the of SIFT algorithm and SIFT-FAST algorithm on UCSD dataset, respectively

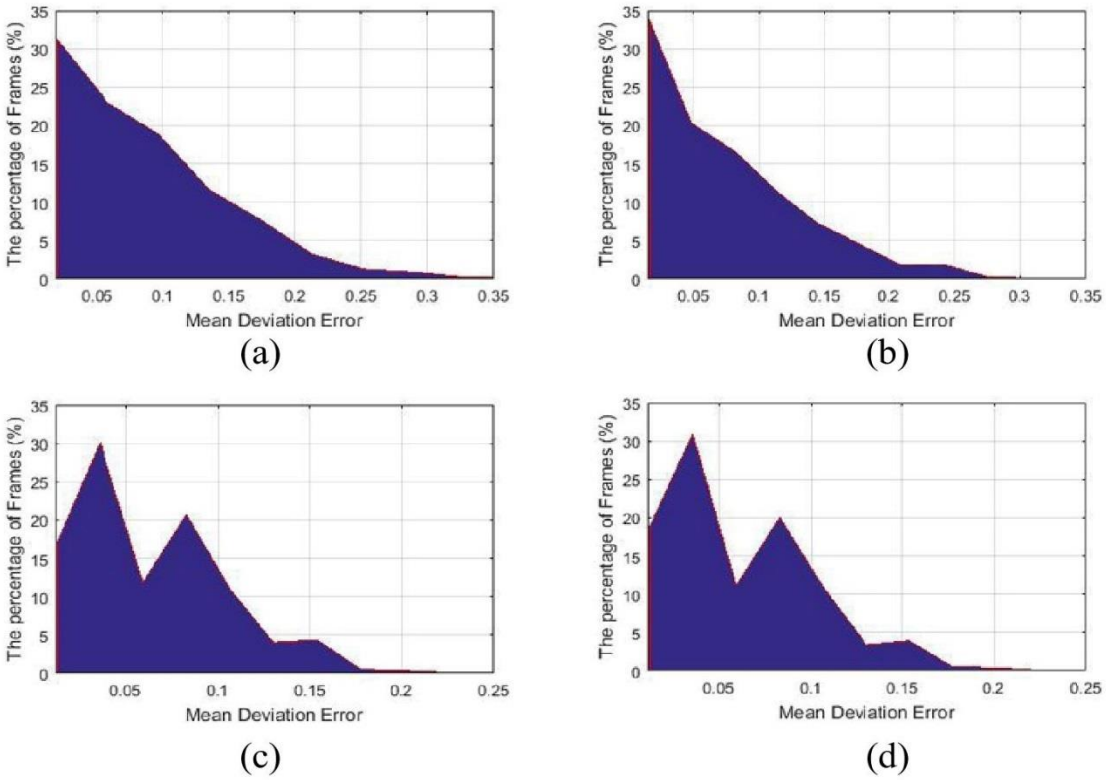


Fig. 13 System performance with crowded scenarios. (**a**) and (**b**) are the MDE of the of SIFT algorithm and SIFT-FAST algorithm on Mall dataset, respectively; (**c**) and (**d**) are the MDE of the of SIFT algorithm and SIFT-FAST algorithm on UCSD dataset, respectively

scenarios. This opens the door for using the proposed system in a high crowded environments. Figures 12 and 13 show the percentages of frames within the MDE distribution for the sparse and crowded scenarios based on the Mall and UCSD datasets, respectively.

5 Conclusions

CCTV cameras are already widely used, the objective of the research presented in this paper was to develop a system that can be incorporated with existing CCTV cameras to provide the number of people in a given space. Two algorithms have been proposed and implemented using a novel combination of four techniques; motion edges, grid map, SIFT & FAST features and pixel-wise techniques. The use of edge pixels for which their number is small compared to foreground pixels significantly reduces the run time of the algorithms. SIFT and FAST features have been chosen due to their high correlation with the number of people. In addition, a grid map approach has been proposed and used to allow similar clusters in different cells to be assigned different densities depending on their location in the frame. This is used to improve the adaption of the proposed algorithms to high variations in crowd behaviours, distributions, lighting and densities.

The UCSD and Mall datasets have been used to evaluate the proposed system. The results have shown that the proposed algorithms achieve good results in heavily occluded environment with perspective distortions. Comparisons with the low-level features regression based methods published in literature show that the proposed algorithms improve the accuracies based on MDE, MSE and MAE metrics (less than 0.1, 16.5 and 3.1, respectively, for the Mall dataset and less than 0.07,

5.5 and 1.9, respectively, for UCSD dataset). The proposed system is more practical than low-level features regression based methods because it can be trained with a lower number of frames so it is relatively easy to deploy. In addition, it reduces the training error, speed, cost and, opens the door to developing more accurate people detection methods. The proposed algorithms can also be used to estimate crowd densities at specific locations in a scene. This shows significant promise as it can be used to detect localised abnormalities in applications such crowd control, evacuation planning and product displays. Comparison of the proposed system in sparse and crowded scenarios shows that it performs better in crowded environments.

References

1. Adegboye AO (2013) Single pixel robust approach for background subtraction for fast people-counting and direction estimation. University of Pretoria, Dissertation
2. Adegboye A, Hancke G, Jr GH (2012) Single-pixel approach for fast people counting and direction estimation. South. Africa Telecommun, Networks Appl
3. Al-Zaydi ZQH, Ndzi DL, Yang Y, Kamarudin ML (2016) An adaptive people counting system with dynamic features selection and occlusion handling. J Vis Commun Image Represent 39:218–225. doi:10.1016/j.jvcir.2016.05.018
4. Antonini G, Thiran JP (2004) Trajectories clustering in ICA space an application to automatic counting of pedestrians in video sequences. Adv. Concepts Intell. Vis, Syst
5. Benezeth Y, Jodoin P-M, Emile B et al (2010) Comparative study of background subtraction algorithms. J Electron Imaging 19:33003. doi:10.1117/1.3456695
6. Berndt D, Clifford J (1994) Using dynamic time warping to find patterns in time series. Report, AAAI
7. Bottesch T, Markus K, Kaechele M, Ulm U (2016) Speeding up k -means by approximating Euclidean distances via block vectors. Int. Conf. Mach. Learn, In, pp. 2578–2586
8. Bouwmans T, El Baf F, Vachon B (2008) Background modeling using mixture of Gaussians for foreground detection - a survey. Recent Patents Comput Sci 1:219–237. doi:10.2174/2213275910801030219
9. Brostow GJ, Cipolla R (2006) Unsupervised Bayesian detection of independent motion in crowds. IEEE Conf Comput Vis Pattern Recognit. doi:10.1109/CVPR.2006.320
10. Çelik H, Hanjalić A, Hendriks EA (2006) Towards a robust solution to people counting. Int. Conf. Image Process, In, pp. 2401–2404
11. Chan AB, Vasconcelos N (2009) Bayesian poisson regression for crowd counting. In: IEEE Int. Conf, Comput. Vis. IEEE, pp. 545–551
12. Chan AB, Vasconcelos N (2012) Counting people with low-level features and bayesian regression. IEEE Trans Image Process 21:2160–2177. doi:10.1109/TIP.2011.2172800
13. Chan AB, Liang ZSJ, Vasconcelos N (2008) Privacy preserving crowd monitoring: counting people without people models or tracking. IEEE Conf Comput Vis Pattern Recognit. doi:10.1109/CVPR.2008.4587569
14. Chan A, Morrow M, Vasconcelos N (2009) Analysis of crowded scenes using holistic properties. In: Perform. Eval, Track. Surveill. Work. IEEE, pp. 101–108
15. Chen K, Kamarainen J-K (2014) Learning to count with back-propagated information. Int. Conf. Pattern Recognit. IEEE, In, pp. 4672–4677
16. Chen K, Loy CC, Gong S, Xiang T (2012) Feature Mining for Localised Crowd Counting. Br Mach Vis Conf. doi:10.5244/C.26.21
17. Chen K, Gong S, Xiang T, Loy CC (2013) Cumulative attribute space for age and crowd density estimation. In: IEEE Conf. Comput, Vis. Pattern Recognit, pp. 2467–2474
18. Cheriyadat AM, Bhaduri BL, Radke RJ (2008) Detecting multiple moving objects in crowded environments with coherent motion regions. IEEE Conf Comput Vis Pattern Recognit Work. doi:10.1109/CVPRW.2008.4562983
19. Cho SY, Chow TW (1999) A fast neural learning vision system for crowd estimation at underground stations platform. Neural Process Lett 10(2):111–120
20. Cho S-Y, Chow T, Leung C (1999) A neural-based crowd estimation by hybrid global learning algorithm. IEEE Trans Syst Man, Cybern Part B Cybern 29:535–541. doi:10.1109/3477.775269
21. Chow TWS, Yam JYF, Cho SY (1999) Fast training algorithm for feedforward neural networks: application to crowd estimation at underground stations. Artif Intell Eng 13:301–307. doi:10.1016/S0954-1810(99)00016-3
22. Conte D, Foggia P, Percannella G et al (1743–1746) (2010) counting moving people in videos by salient points detection. Int. Conf. Pattern Recognit. pp, In

23. Conte D, Foggia P, Percannella G et al (2010) A method for counting people in crowded scenes. In: IEEE Int. Conf, Adv. Video Signal Based Surveill, pp. 225–232
24. Conte D, Foggia P, Percannella G, Vento M (2013) Counting moving persons in crowded scenes. Mach Vis Appl 24:1029–1042. doi:10.1007/s00138-013-0491-3
25. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE Conf. Comput, Vis. Pattern Recognit, pp. 886–893
26. Davies A, Yin JH, Velastin S (1995) Crowd monitoring using image processing. Electron Commun Eng J. doi:10.1049/ecej:19950106
27. Dollar P, Belongie S, Perona P (2010) The fastest pedestrian detector in the west. Br Mach Vis Conf. doi:10.5244/C.24.68
28. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminative trained part based models. IEEE Trans Pattern Anal Mach Intell 32:1627–1645
29. Fradi H, Dugelay JL (2012) Low level crowd analysis using frame-wise normalized feature for people counting. Int. Work. Inf. Forensics Secur, In, pp. 246–251
30. Gao L, Wang Y, Ye X, Wang J (2016) Crowd Pedestrian Counting Considering Network Flow Constraints in Videos. arXiv Prepr
31. Ge W, Collins RT (2009) Marked point processes for crowd counting. In: Comput. Vis, Pattern Recognit. Work. IEEE, pp. 2913–2920
32. Hafeezallah A, Abu-Bakar S (2016) Crowd counting using statistical features based on curvelet frame change detection. Multimed Tools Appl. doi:10.1007/s11042-016-3869-1
33. Harville M (2002) Stereo person tracking with adaptive plan-view statistical templates. Proc. ECCV Work. Stat. Methods Video Process, In, pp. 67–72
34. Hashimoto K, Morinaka K, Yoshiike N et al (1997) People count system using multi-sensing application. Int. Solid State Sensors Actuators Conf, In, pp. 1291–1294
35. Hou YL, Pang GKH (2011) People counting and human detection in a challenging situation. IEEE Trans Syst Man, Cybern Part ASystems Humans 41:24–33. doi:10.1109/TSMCA.2010.2064299
36. Hu X, Zheng H, Chen Y, Chen L (2015) Dense crowd counting based on perspective weight model using a fisheye camera. Int J Light Electron Opt 126:123–130. doi:10.1016/j.ijleo.2014.08.132
37. Hu Y, Chang H, Nian F et al (2016) Dense crowd counting from still images with convolutional neural networks. J Vis Commun Image Represent 38:530–539. doi:10.1016/j.jvcir.2016.03.021
38. Huang X, Zou Y, Wang Y (2016) Cost-sensitive sparse linear regression for crowd counting with imbalanced training data. IEEE Int. Conf. Multimed, Expo
39. Intelcom DILAX (2015) Public Transport https://www.dilax.com/. Accessed 1 Oct 2016
40. Jeong CY, Choi S, Han SW (2013) A method for counting moving and stationary people by interest point classification. In: IEEE Int. Conf, Image Process. IEEE, pp. 4545–4548
41. Joshi NS, Choubey NS (2014) Comparison of traditional approach for edge detection with soft computing approach. Int J Comput Appl 96:17–23
42. Kaur G, Virk IS (2014) Edge detection through fuzzy system using type I format. Int J Comput Appl 102:24–27
43. Kilambi P, Ribnick E, Joshi AJ et al (2008) Estimating pedestrian counts in groups. Comput Vis Image Underst 110:43–59. doi:10.1016/j.cviu.2007.02.003
44. Kong D, Gray D, Tao H (2005) Counting pedestrians in crowds using viewpoint invariant training. Procedings Br Mach Vis Conf. doi:10.5244/C.19.63
45. Kong D, Gray D, Tao H (2006) A viewpoint invariant approach for crowd counting. Int. Conf. Pattern Recognit, In, pp. 1187–1190
46. Leibe B, Seemann E, Schiele B (2005) Pedestrian detection in crowded scenes. In: IEEE Conf. Comput, Vis. Pattern Recognit, pp. 878–885
47. Lempitsky V, Zisserman A (2010) Learning to count objects in images. Adv. Neural Inf. Process. Syst, In, pp. 1324–1332
48. Li J, Huang L, Liu C (2011) Robust people counting in video surveillance: dataset and system. Int. Conf. Adv. Video Signal Based Surveill, In, pp. 54–59
49. Lin S, Chen J, Chao H (2001) Estimation of number of people in crowded scenes using perspective transformation. IEEE Trans Syst Man Cybern 31:645–654
50. Loy C, Chen K, Gong S, Xiang T (2013) Crowd counting and profiling: methodology and evaluation. Model. Simul. Vis. Anal. Crowds. Springer New York, In, pp. 347–382
51. Ltd B (2013) Use CCTV to Count People http://www.videoturnstile.com/. Accessed 1 Oct 2016
52. Ma R, Li L, Huang W, Tian Q (2004) On pixel count based crowd density estimation for visual surveillance. In: IEEE Conf. Cybern, Intell. Syst. IEEE, pp. 1–3
53. Ma H, Zeng C, Ling CX (2012) A reliable people counting system via multiple cameras. ACM Trans Intell Syst Technol 3:1–22. doi:10.1145/2089094.2089107

54. Merad D, Aziz KE, Thome N (2010) Fast people counting using head detection from skeleton graph. Adv. Video Signal Based Surveill. IEEE, In, pp. 233–240
55. Norris C, Mccahill M, Wood D (2004) Editorial. The growth of CCTV: a global perspective on the international diffusion of video surveillance in publicly accessible space. Surveill Soc 2:110–135
56. Rabaud V, Belongie S (2006) Counting crowded moving objects. In: IEEE Conf. Comput, Vis. Pattern Recognit, pp. 705–711
57. Rao AS, Gubbi J, Marusic S, Palaniswami M (2015) Estimation of crowd density by clustering motion cues. Vis Comput 31:1533–1552. doi:10.1007/s00371-014-1032-4
58. Rodriguez M, Superieure EN, Laptev I et al (2011) Density-aware person detection and tracking in crowds. Int. Conf. Comput. Vis. IEEE, In, pp. 2423–2430
59. Ryan DA (2013) Crowd monitoring using computer vision. Queensland University of Technology, Dissertation
60. Ryan D, Denman S, Fookes C, Sridharan S (2009) Crowd counting using multiple local features. Digit. Image Comput. Tech. Appl. IEEE, In, pp. 81–88
61. Ryan D, Denman S, Fookes C, Sridharan S (2014) Scene invariant multi camera crowd counting. Pattern Recogn Lett 44:98–112. doi:10.1016/j.patrec.2013.10.002
62. Ryan D, Denman S, Sridharan S, Fookes C (2015) An evaluation of crowd counting methods, features and regression models. Comput Vis Image Underst 130:1–17. doi:10.1016/j.cviu.2014.07.008
63. Saleh SAM, Suandi SA, Ibrahim H (2015) Recent survey on crowd density estimation and counting for visual surveillance. Eng Appl Artif Intell 41:103–114. doi:10.1016/j.engappai.2015.01.007
64. Shbib R, Zhou S, Ndzi D, Al-kadhimi K (2013) Distributed monitoring system based on weighted data fusing model. Am J Soc Issues Humanit 3:53–62
65. ShopperTrak (2013) ShopperTrak Solutions http://www.shoppertrak.com/. Accessed 1 Oct 2016
66. Shrivakshan GT, Chandrasekar C (2012) A Comparison of various Edge Detection Techniques used in Image Processing Int J Comput Sci Issues:9
67. Sidla O, Lypetskyy Y, Brändle N, Seer S (2006) Pedestrian detection and tracking for counting applications in crowded situations. IEEE Int Conf Video Signal Based Surveill. doi:10.1109/AVSS.2006.91
68. Sobral A, Vacavant A (2014) A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. Comput Vis Image Underst 122:4–21. doi:10.1016/j.cviu.2013.12.005
69. Tang NC, Lin Y-Y, Weng M, Liao HM (2015) Cross-camera knowledge transfer for Multiview people counting. IEEE Trans Image Process 24:80–93. doi:10.1109/TIP.2014.2363445
70. Technology A (2013) Our customers http://www.peoplecounting.co.uk/our-customers. Accessed 1 Oct 2016
71. Topkaya IS, Erdogan H, Porikli F (2014) Counting people by clustering person detector outputs. In: IEEE Int. Conf, Adv. Video Signal Based Surveill. IEEE, pp. 313–318
72. Tu J, Zhang C, Hao P (2013) Robust real-time attention-based head-shoulder detection for video surveillance. In: IEEE Int. Conf, Image Process. IEEE, pp. 3340–3344
73. Tuzel O, Porikli F, Meer P (2008) Pedestrian detection via classification on Riemannian manifolds. IEEE Trans Pattern Anal Mach Intell. doi:10.1109/TPAMI.2008.75
74. Wang M (2014) Data assimilation for agent-based simulation of smart environment. Georgia State University, Dissertation
75. Wang M, Wang X (2011) Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In: IEEE Conf. Comput, Vis. Pattern Recognit, pp. 3401–3408
76. Wang J, Fu W, Liu J et al (2014) Spatiotemporal group context for pedestrian counting. IEEE Trans Circuits Syst Video Technol 24:1620–1630
77. Wu B, Nevatia R (2007) Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. Int J Comput Vis 75:247–266. doi:10.1007/s11263-006-0027-7
78. Xiaohua L, Lansun S, Huanqin L (2006) Estimation of crowd density based on wavelet and support vector machine. Trans Inst Meas Control 28:299–308. doi:10.1191/0142331206tim178oa
79. Xing X, Wang K, Lv Z (2015) Fusion of gait and facial features using coupled projections for people identification at a distance. Signal Process Lett 22:2349–2353
80. Xu B, Qiu G (2016) Crowd density estimation based on rich features and random projection Forest. IEEE Winter Appl. Comput. Vis, In, pp. 1–8
81. Zhang J, Tan B, Sha F, He L (2011) Predicting pedestrian counts in crowded scenes with rich and high-dimensional features. IEEE Trans Intell Transp Syst 12:1037–1046. doi:10.1109/TITS.2011.2132759
82. Zhang C, Li H, Wang X (2015a) Cross-scene crowd counting via deep convolutional neural networks. Proc IEEE Conf Comput Vis Pattern Recognit. doi:10.1109/CVPR.2015.7298684
83. Zhang Z, Wang M, Geng X (2015b) Crowd counting in public video surveillance by label distribution learning. Neurocomputing 166:151–163. doi:10.1016/j.neucom.2015.03.083

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 multimedia surveillance systems, 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 video and image processing, wireless sensor networks and mesh networks for applications in precision agriculture, environmental monitoring, behavioural economics, security, building control and energy management, etc.

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.

Dr. Ammar Zakaria graduated with B. Eng. (Hons.) in Electronic and Computer Engineering from the University of Portsmouth, UK. in 2008. He received his a PhD in Mechatronic Engineering from Universitiy Malaysia Perlis, 2013. His research interest include sensor technology, robotics, wireless sensor and actuator network and Image processing.

Prof. Ali Y.M. Shakaff is currently a Professor at the School of Mechatronic Engineering, Universiti Malaysia Perlis. His academic career has spanned over a period of 23 years, throughout which he has been thoroughly involved in teaching and research in various areas of electronic engineering. He has also spent a considerable time in academic management, notably in the development of new engineering schools for Universiti Malaysia Perlis (UniMAP) and before that, Universiti Sains Malaysia (USM). Prior to joining UniMAP in 2002 (as part of the founding team), he was the Dean for the School of Electrical & Electronic Engineering, USM (1996-2002). He was formerly UniMAP's Deputy Vice-Chancellor for Academic and International Affairs (2002-2009).

ISBN 978-93-84422-71-4
Proceedings of 2016 International Conference on Image Processing, Production and Computer Science (ICIPCS'2016)
London (UK), March 26-27, 2016 pp.30-36

Cascade Method for Image Processing Based People Detection and Counting

Zeyad Al-Zaydi, David Ndzi and David Sanders

School of Engineering, University of Portsmouth, Anglesea Road, Portsmouth PO1 3DJ, UK.
E-mail address: {zeyad.al-zaydi, david.ndzi, david.sanders}@port.ac.uk

Abstract: *People detection is of great importance in video surveillance. Different approaches have been proposed to achieve accurate detection system. The main problem in people detection systems is that it must maintain a balance between the number of false detections and the number of missing people which limits the global detection results. In order to solve this problem and add robustness to detection, we propose a multiplexor and collector model composed of multiple independent detectors. This model is used to keep the true positive detections provided by a number of detectors and reduce the miss rate. In addition, a fusion model is proposed to check the robustness of the cascaded detection system. A pipeline technique will also be used to avoid the increasing of detection time.*

Keywords: *people detection, counting, surveillance systems, image processing, computer vision.*

1. Introduction

People detection is one of the most challenging task in computer vision [1]. Although significant research has been carried out to find an accurate solution for this task, there are still many challenges that need to be resolved. These include variability in appearances, crowded scenarios, handling complex backgrounds and occlusion which lead to high false detection and miss rate. The trade-off between false detection and miss rate renders most methods ineffective.

People detection is fundamental in intelligent video surveillance systems as it provides important information for establishing awareness. People detection can be used for people counting and tracking. An efficient and accurate people counting, tracking and distribution system would be beneficial and fundamental to a lot of applications such as in

- safety applications; e.g. an indicator of over-crowded situations, for possible emergency evacuation processes and crowd management [2], [3];
- security applications; e.g. an indicator of fighting, rioting, violent protest, mass panic and excitement [4], [5];
- business intelligence and behavioural economics applications; e.g. the distribution of costumers may be used for product placement, floor planning and staff management [6]. In addition, the overall crowd in a retail store may be monitored to assess store performance over time [7];
- Transport applications; e.g. to improve the distribution of buses over different routes which, is fundamental to the optimisation of transport network and for scheduling public transport [8], [9]; and
- energy management applications such as optimising air conditioning, lighting and heating in buildings according to occupancy density and distributions [3], [10].

This paper proposes firstly, a multi stage independent detection system employed to minimise miss rate; secondly a fusion technique that can be used to compensate the limitations of each independent detector and finally, a pipeline technique to that minimises processing, and hence detection, time.

The remainder of this paper is structured as follows: Section 2 describes the related work; Sections 3 describes the system design; Finally, Section 4 summarizes the main conclusions and future work.

2. Related Work

There is an extensive literature on people detection approaches. They can be generally grouped into six paradigms. In this section, we provide an overview on each of the paradigms.

2.1.Full Body Detection Based Algorithms

They are direct approaches to count the number of people in a scene through detection. The algorithms are trained using the full body appearance of a set of people [11]–[13]. They suffer from large pose variations and partial occlusion as the number of people increases [14]. Different features are used to represent the full body appearance such as Haar like features [15] and histogram of oriented gradient (HOG) features [13]. Different linear and nonlinear classifiers are also used to find the relationship between the features and the number of people such as linear support vector machine (SVM), neural network (NN) and Gaussian process regression (GPR) [16], [17]. The accuracy of full body detectors is acceptable in sparse environments but the accuracy decreases significantly in crowded environments.

2.2.Part Body Detection Based Algorithms

A significant amount of research has been carried out to mitigate partial occlusion by detecting only part of the body such as heads, faces, eyes and head-shoulders [18]–[20]. The shape of people's heads changes or differ with hair styles and head coverings. Hence head based detection is not robust enough for counting people [14]. On the other hand, a head-shoulder region occupies a larger proportion of a human body image than a head alone and they are more likely to be detected [14]. Faces and eyes are rarely used to count the number of people because a lot of people do not look at the cameras when passing and faces and eyes are easily occluded.

2.3.Shape Matching Detection Based Algorithms

Ellipses are used by some researchers to count the number of people [21]. In this approach, the background subtraction method is applied to segment the foreground blobs [22], [23] and ellipse detection is applied to identify the number of people in each blob. Other shapes such as Bernoulli shapes have been used by some researchers to count the number of people [24]. The accuracy of shape matching detectors is acceptable in sparsely occupied environments but the accuracy decreases significantly in crowded environments.

2.4.Multi Camera Detection Based Algorithms

Many research studies have focused on counting the number of people using a single camera which can fail in crowded environments (i.e., heavy occlusion occurs). Some researchers have used multiple cameras to count people to avoid occlusion [25]. The cost of hardware and the multi-camera set-up is the main disadvantages of this approach.

2.5.3D camera detection based algorithms

Depth information has great potential to improve people counting for many reasons e. g. [26], it can be used to;

- Improve foreground segmentation; and
- Handle occlusions more efficiently.

This third dimension allows time-of-flight (TOF) and stereo vision features to be used to obtain image depth [27]. For instance, Microsoft Kinect devices could be used to obtain image depth, which provides high quality images at a lower price compared to previous technologies [28]. The performance of 3D camera based detection techniques is affected by changing illumination and when monitoring a large area of similar colors and little edges, because it may be difficult to find features [29]. In addition, developing a people counting algorithm based on a depth sensing system is more complex and would therefore require a significant amount of computational time [30].

2.6.Density-Aware Detection Based Algorithms

This approach combines full or part body detection based algorithms and crowd density estimation [31]. Full body, head and head-shoulder detection based algorithm can be improved and the accuracy can be increased by using density-aware information [31]. The aim of this approach is to reduce the false positive per image (FPPI) in low crowd density locations in the frame. This occurs when it incorrectly detects the presence of a person, when there is actually nobody. In addition, this approach reduces the miss rate in high crowd density locations in the frame.

3. System Design

The main aim of classical cascade classifiers is to reduce the FPPI. In this approach FPPI will be improved, obviously, but the miss rate of detection (FNR) will increase rapidly. The FNR is measured using [13]:

$$FNR = \frac{FN}{FN+TP} \quad (1)$$

Where FN is the number of times it incorrectly indicates that there is nobody, TP is the number of correct detections. In addition, a classical cascade classifier usually consists of multiple stages of weak classifiers of the same kind but do not use multiple independent detectors. As shown in Figure 1, all detected windows of Non-Person (NP) are rejected directly at each classifier while the windows of Candidate Person (CP) are passed to the next stage for more checking by the cascaded classifiers.

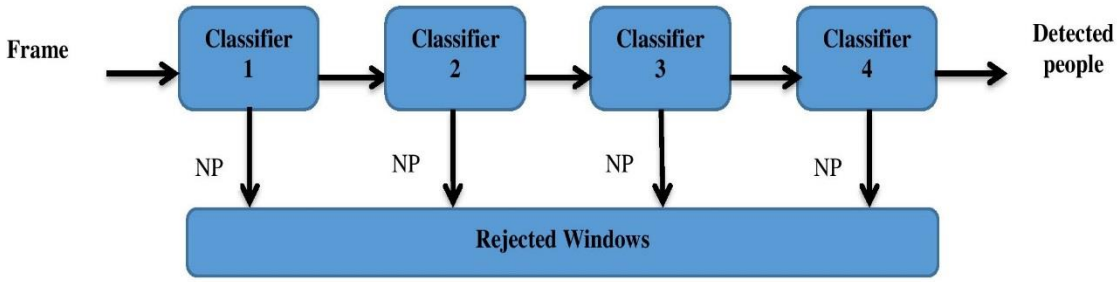


Fig. 1: Flow diagram of a cascade classifier.

In this paper, multiple independent detectors and a novel multiplexer cascade is proposed. Different detectors are used to detect people. Each one has some advantages and disadvantages mainly because each technique is based on different features extraction, learning method and person models. Frame-by-frame, different independent detectors may produce different results so a novel cascade of different independent detectors can be implemented to improve the true positive detections provided by a number of detectors. That will reduce the FNR especially when the advantages of each detector is exploited in this cascaded model. In addition, the rejected windows from all detectors can be fused and compared with a predefined threshold for detection purposes. Figure 2 shows the block diagram of the proposed method which consists of three independent detectors and one fusion model. Where

CP is the windows of candidate person, LCP is the windows of low-level candidate person, DP is the windows of detected people and RP is the windows of rejected people. Two fusion models are developed in this paper.

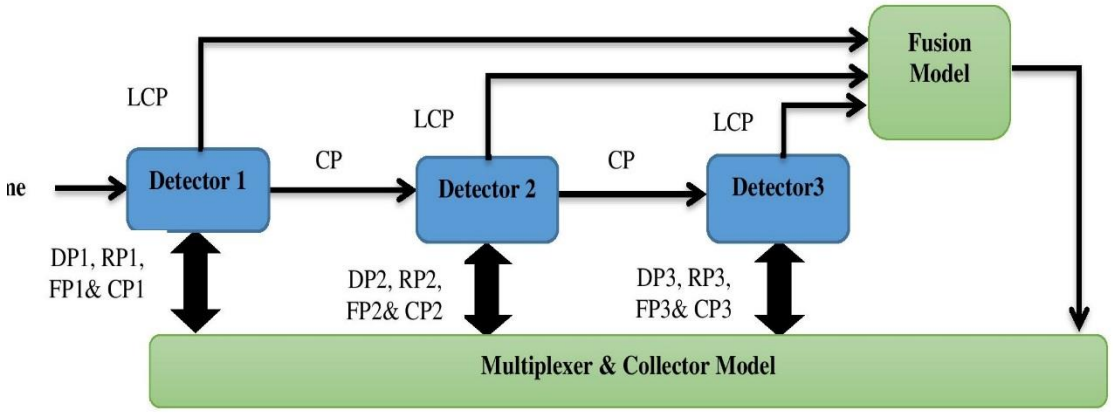


Fig. 2: Flow diagram of a novel multiplexer cascade detector.

The confidence level of the detectors will be used to classify windows into detected person windows, rejected person windows and candidate person windows. In addition, three predefined thresholds will be used in this classification; high-quality, medium-quality and low-quality thresholds. The multiplexer and collector model will use the following rules to classify the windows;

$$\boldsymbol{if}\ (CFW > High\ quality\ threshold)\ \boldsymbol{Then}\ DP$$

$$\boldsymbol{if}\ (CFW < Low\ quality\ threshold)\ \boldsymbol{Then}\ RP$$

$$\boldsymbol{if}(CFW < High\ quality\ threshold\ \boldsymbol{and}\ > Low\ quality\ threshold)\ \boldsymbol{Then}\ CP$$

$$\boldsymbol{if}(CFW < High\ quality\ threshold\ \boldsymbol{and}\ > miduim\ quality\ threshold)\ \boldsymbol{Then}\ LCP$$

where CFW is the confidence level of a window. The multiplexer and detection windows model will use the following equation to collect the results of all detectors and the fusion model;

$$Total\ detected\ people = DP1 + DP2 + DP3 + DP4 \quad (2)$$

The fusion model measures the robustness of the LCP windows by comparing them with the medium-quality threshold. All LCP windows that get more than this threshold at all detectors will be considered as DP windows. The fusion model will use the following rule;

$$\boldsymbol{if}\ (LCP1\ \boldsymbol{and}\ LCP2\ \boldsymbol{and}\ LCP3 > miduim\ quality\ threshold\ )\ \boldsymbol{Then}\ DP$$

The proposed method will use pipeline techniques to avoid increasing the detection or processing time. It is an implementation technique in which multiple frames are overlapped during execution. Three frames are processed simultaneously using different detectors. In this case, the processing time is not increased.

4. Experimental Results

Matlab software is used to implement the proposed system. Preliminary results of the proposed method have been obtained. The proposed approach is implemented and tested using pictures. Two detectors are used in to produce the results presented in this section; the Haar-like detector, which is very famous and widely used in detection systems and the second one is a full body detector. As shown in Figure 3, the miss rate of the proposed approach is 0.3 and 0 for the first and second pictures. From the results, it can be seen that the miss rate is lower than each detector individually.

(a) Detected people using the first detector

(b) People not detected using the first detector due to low-level of confidence.

(c) The detected people by the Haar-like detector detector.

(d) The detected people by the proposed approach.

Fig 3: The performance of the proposed approach.

5. Conclusion

This paper has outlined the general principles of a new approach for people detection and counting using video surveillance. This approach combines multiple independent detectors using a multiplexer & collector model, fusion model and a pipeline technique in order to keep the correct positive detections by a number of detectors and, at the same time, reduce the miss rate. This integration technique performs better, provides more accurate results and enhances the detection rate. The proposed approach has been implemented and tested using frames as well as pictures. Initial results are very promising and shows that the proposed approach reduces the miss detection rate and hence, improves accuracy. Further development and evaluation of the proposed technique using different videos in different environments with different crowd densities in real-time environments is currently being carried out.

6. References

[1] S. Mukherjee and K. Das, "Omega Model for Human Detection and Counting for application in Smart Surveillance System," *arXiv Prepr. arXiv1303.0633*, vol. 4, no. 2, pp. 167–172, 2013.

[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. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Scene invariant multi camera crowd counting," *Pattern Recognit. Lett.*, vol. 44, pp. 98–112, 2014.

[5] Z. Zhang, M. Wang, and X. Geng, "Crowd counting in public video surveillance by label distribution learning," *Neurocomputing*, vol. 166, pp. 151–163, 2015.

[6] C. Loy, K. Chen, S. Gong, and T. Xiang, *Crowd counting and profiling: Methodology and evaluation*. 2013.

[7] D. A. Ryan, "Crowd Monitoring Using Computer Vision," Queensland University of Technology, 2013.

[8] 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.

[9] A. Bartolini, F., Cappellini, V., & Mecocci, "Counting people getting in and out of a bus by real-time image-sequence processing," *Image Vis. Comput.*, pp. 36–41, 1994.

[10] K. Hashimoto, K. Morinaka, N. Yoshiike, C. Kawaguchi, and S. Matsueda, "People count system using multi-sensing application," *Proc. Int. Solid State Sensors Actuators Conf. (Transducers '97)*, vol. 2, pp. 1291–1294, 1997.

[11] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian Detection via Classification on Riemannian Manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, 2008.

[12] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," *Comput. Vis. Pattern Recognition, 2005. CVPR 2005. IEEE Comput. Soc. Conf.*, vol. 1, pp. 878–885, 2005.

[13] 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.

[14] J. S. C. Yuk, K. Y. K. Wong, R. H. Y. Chung, F. Y. L. Chin, and K. P. Chow, "Real-time multiple head shape detection and tracking system with decentralized trackers," *Proc. - ISDA 2006 Sixth Int. Conf. Intell. Syst. Des. Appl.*, vol. 2, pp. 384–389, 2006.

[15] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[16] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, 2005.

[17] 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.

[18] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminative Trained Part Based Models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[19] S. Lin, J. Chen, and H. Chao, "Estimation of Number of People in Crowded Scenes Using Perspective Transformation," *IEEE Trans. Syst. Man. Cybern.*, vol. 31, no. 6, pp. 645–654, 2001.

[20] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, 2007.

[21] 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.

[22] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[23] A. Ilyas, M. Scuturici, and S. Miguet, "Real time foreground-background segmentation using a modified codebook model," *6th IEEE Int. Conf. Adv. Video Signal Based Surveillance, AVSS 2009*, pp. 454–459, 2009.

[24] W. Ge and R. T. Collins, "Marked point processes for crowd counting," *2009 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. CVPR Work. 2009*, pp. 2913–2920, 2009.

[25] 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.

[26] M. Harville, "Stereo person tracking with adaptive plan-view statistical templates," *Proc. ECCV Work. Stat. Methods Video Process.*, pp. 67–72, 2002.

[27] 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.

[28] Microsoft, "Kinect," 2011. [Online]. Available: www.xbox.com/en-US/kinect/. [Accessed: 23-Mar-2015].

[29] C. R. Sensors, R. B. Fisher, and K. Konolige, "Handbook of Robotics Chapter 22 - Range Sensors," 2008.

[30] T. Tikkanen, "People detection and tracking using a network of low-cost depth cameras," 2014.

[31] 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.

Appendix C: Ethical Approval and Form UPR16 (Research Ethics Review Checklist)

Ethical approval was confirmed by email from the Faculty of Technology Ethics Committee at the University of Portsmouth, dated 19 June 2017.

FORM UPR16

Research Ethics Review Checklist

Please include this completed form as an appendix to your thesis (see the Postgraduate Research Student Handbook for more information

Postgraduate Research Student (PGRS) Information				Student ID:	UP714763
PGRS Name:	Zeyad Qasim Habeeb Al-zaydi				
Department:	Engineering	**First Supervisor:**	Dr Branislav Vuksanovic		
Start Date: (or progression date for Prof Doc students)	01/10/2014				
Study Mode and Route:	Part-time ☐	MPhil ☐	MD ☐		
	Full-time ☒	PhD ☒	Professional Doctorate ☐		

Title of Thesis:	Image Processing Based Ambint Context-aware Information Fusion
Thesis Word Count: (excluding ancillary data)	30088

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University's Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist:
(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: http://www.ukrio.org/what-we-do/code-of-practice-for-research/)

a)	Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?	YES ☒ NO ☐
b)	Have all contributions to knowledge been acknowledged?	YES ☒ NO ☐
c)	Have you complied with all agreements relating to intellectual property, publication and authorship?	YES ☒ NO ☐
d)	Has your research data been retained in a secure and accessible form and will it remain so for the required duration?	YES ☒ NO ☐
e)	Does your research comply with all legal, ethical, and contractual requirements?	YES ☒ NO ☐

Candidate Statement:

I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)

Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):	1B6D-6FAC-1170-B151-5D1E-B69C-CAD7-773D

If you have *not* submitted your work for ethical review, and/or you have answered 'No' to one or more of questions a) to e), please explain below why this is so:

Signed *(PGRS)*:	Zeyad	**Date:** 29/08/2017

UPR16 – August 2015