PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

VISION-BASED PEDESTRIAN COUNTING AT BUS STOPS

GUILLERMO ARTURO GARCÍA BUNSTER

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Advisor:
MIGUEL TORRES TORRITI

Santiago de Chile, December 2009

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

VISION-BASED PEDESTRIAN COUNTING AT BUS STOPS

GUILLERMO ARTURO GARCÍA BUNSTER

Members of the Committee:
MIGUEL TORRES TORRITI
DOMINGO MERY QUIROZ
PABLO ZEGERS FERNÁNDEZ
JUAN COEYMANS AVARIA

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Santiago de Chile, December 2009

Gratefully to Bárbara.

ACKNOWLEDGEMENTS

This work was supported by CONICYT of Chile under PBCT Grant ACT-32 Dept. of Electrical Engineering, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Casilla 306-22, Santiago.

Special thanks to Miguel Torres for all his support, guidance, advices and remarks over the development of this work. Thanks to the members of the Robotics and Automation Laboratory that helped in the collecting of samples. Also thanks to the members of the committee for the remarks on this thesis.

To my family for all the hard work in raising me and granting me a proper education and to Bárbara for her constant love and support over the past years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Accurately counting people waiting at bus stops is essential for automated bus fleet scheduling and dispatch. Estimating the passenger demand in regular open bus stops by means of image processing is a nontrivial problem because of the varying conditions, such as illumination, crowdedness, people poses, to name a few. This paper presents a simple, but very effective approach to estimate the passenger count using people density estimates. People density is obtained from foreground segmentation using a Gaussian mixture background model. The final people count estimates is obtained using a classifier based on linear discriminant analysis. The approach is compared to the well-known Viola-Jones detector and shown to yield better people count estimates despite its simplicity, because it is more robust to occlusions, pose changes, and due to the fact that it does not attempt to find body parts. Additionally the algorithm shows promising results when applied to images captured using a thermal camera, as well as a omnidirectional panoramic camera. The proposed method is general and it can be employed to count people in other public spaces, such as crosswalks and buildings.

Keywords: Pedestrian detection, pedestrian counting, background subtraction, expectation maximization, Haar-features, density-based demand estimation.

RESUMEN

Con el fin de automatizar y optimizar el despacho y control de flota en sistemas modernos de transporte público, es necesario medir con exactitud y en tiempo real la cantidad de peatones esperando locomoción colectiva en paraderos de buses. Realizar esta medición en lugares abiertos mediante procesamiento de imágenes es un problema complejo debido a las adversas condiciones en que las imágenes son capturadas, donde influyen principalmente la iluminación variante e irregular en la escena, la aglomeración de personas en un solo lugar y las distintitas posiciones que estas asumen. Este trabajo presenta un método simple y efectivo para estimar la cantidad de pasajeros en paraderos midiendo la densidad de objetos en la imagen que no pertenecen al plano de fondo de la escena. Estos objetos son identificados mediante un proceso de segmentación de fondo en la imagen utilizando un modelo probabilístico de distribuciones normales multimodales. Finalmente, considerando los efectos de la perspectiva en la imagen capturada, se construye un modelo fenomenológico para estimar la cantidad total de peatones a partir de la superficie de los objetos que no pertenecen a la escena. Este enfoque es comparado con el popular algoritmo de Viola y Jones para detección de objetos aplicado al problema de detección de caras y personas en imágenes, y se demuestra que mediante este enfoque se obtienen mejores medidas de error. Esto se debe a que el método propuesto es más robusto ante oclusiones y es independiente a la posición y orientación del individuo, dificultades que los enfoques tradicionales utilizados en la detección de personas no han logrado superar. Además, este algoritmo demuestra mantener su efectividad al utilizar imágenes capturadas con cámaras infrarrojas y omnidireccionales. El método propuesto es general y se puede utilizar para contar personas en distintas aplicaciones y en cualquier tipo de ambientes públicos, como vestíbulos de edificios o cruces peatonales.

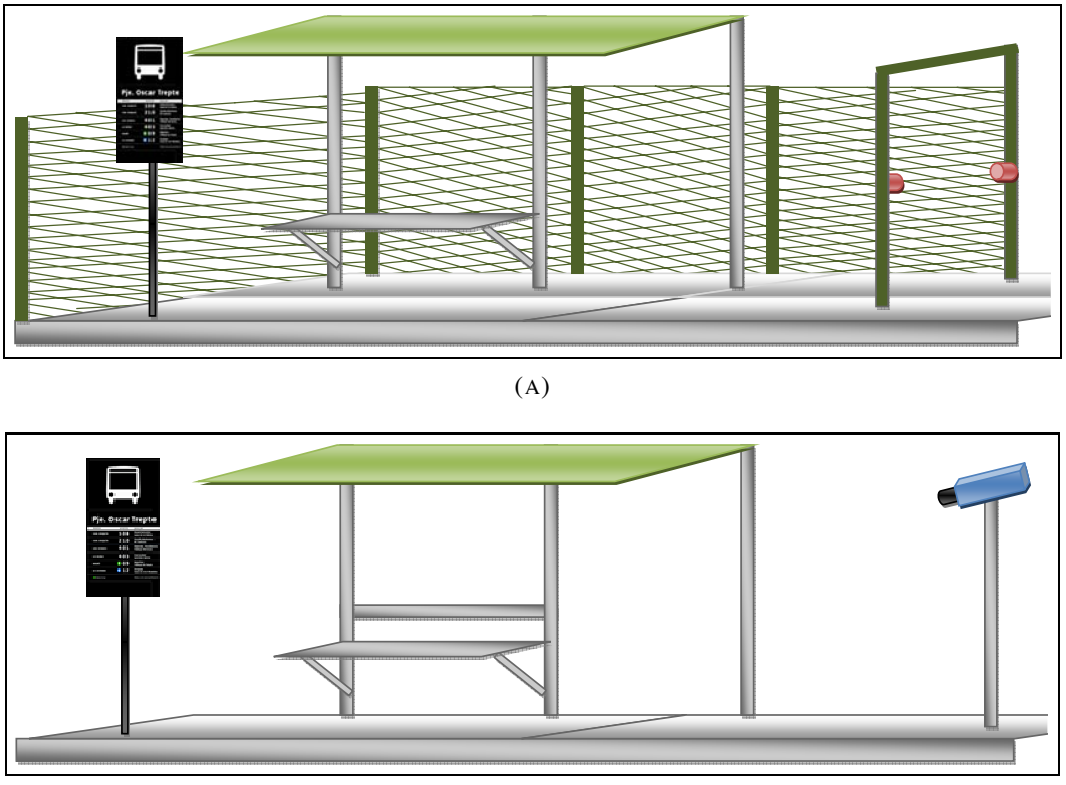
Palabras Claves: Detección de peatones, conteo de peatones, sustracción de fondo, maximización de la esperanza en distribuciones gaussianas, características de Haar, estimación de demanda basada en densidad.

1. INTRODUCTION

Modern public transportation systems require accurate real-time information of route conditions and demand for optimal fleet scheduling and control (Cortés, Sáez, Sáez, Núñez, & Tirachini, 2007; Núñez, Cortés, Sáez, & Riquelme, 2008). Technologies for traffic flow monitoring are well established, in contrast to the state of the art in real-time sensing of passenger demand at bus stops. Traditionally demand information has been obtained using statistical models built off-line from manually collected data. Hence, while planning, transport system operators are constrained to assume a fixed demand in each bus stop, disregarding short term demand fluctuations that are not considered in the statistical model. On the other hand, modern transport control systems require real time demand information in order to optimize short term performances regarding passengers awaiting time and travel time, therefore, developing automated methods to reliably count passengers at bus stops and within the buses is necessary. Current demand sensing technologies rely on passengers to access the bus stop through a turnstile or a movement sensing threshold, making these technologies very invasive in an urban sense. Vision-based solutions, on the other hand, don't require considerable infrastructure, making them an elegant and urbanistically passive solution to the demand estimation problem. Moreover, this particular application hasn't been studied widely in the literature, making this problem interesting for further research.

Based on the sensing technique, current pedestrian detection technologies can be divided into two categories: sensors which detect an individual person at a time, and sensors that detect multiple pedestrians at once. The first type of sensor requires the person to be in contact or very close to the sensing device, such as turnstiles, ultrasonic/infrared beam based sensors or proximity RDIF card readers. The main disadvantage of these solutions is that the bus stop area must be physically enclosed, so that passengers walk through passageways or portals equipped with any of the sensors, thus making their implementation more expensive and very invasive as illustrated in figure 1.1a. This constraint makes the implementation of these solutions very invasive in an urbanistic sense and carries further constructing costs. Other variants of this solution consider enclosing the bus stop area with

multiple portals but without any improvements in either sensing capabilities or infrastructure costs.



(A)

(B)

FIGURE 1.1. Different bus stops configurations for demand estimation: (1.1a) using conventional technologies and (1.1b) using a wide-view sensor.

On the other hand, the second type of sensors, such as laser scanners (Fuerstenberg, Dietmayer, & Willhoeft, 2002), long range RFID (Polivka, Svanda, Hudec, & Zvanovec, 2009), thermal (Bi, Tsimhoni, & Liu, 2009) and conventional cameras (Bu & Chan, 2005), have a wide field of view and do not need invasive constraints on the bus stop structure since they cover a large area. The hardware cost of most of these technologies is high compared to the first category, but a considerable amount of research has been carried out in their application to pedestrian detection, particularly in collision avoidance for intelligent vehicles (Gavrila, 2000; Bu & Chan, 2005; Gandhi & Trivedi, 2006) and surveillance applications (Haritaoglu, Harwood, & Davis, 2000; Heikkilä & Silvén, 2004). Laser scanners

are very expensive and in most cases limited to scanning along a single plane thus dificulting the people counting process. Vision based pedestrian detection using conventional (fig. 1.2a), omnidirectional (fig. 1.2c) or thermal cameras (fig. 1.2b), provides rich information about the surrounding environment compared to the other sensor technologies and can be employed simultaneously for counting and surveillance purposes. Furthermore, standard and thermal cameras are becoming more accessible due to lowering prices. Considering these advantages, this paper proposes a vision-based approach to pedestrian counting at bus stops using standard digital video cameras.

1.1. Objectives

The main objective of our research is to estimate, with reasonable accuracy and precision, the total amount of pedestrians awaiting at bus stops using a static video camera by means of computer vision and image processing techniques. Additionally the proposed solution must not consider the use of passageways or portals for controlling the access to bus stops. Within the proposed solution lie the following secondary objectives:

- Implement feature extraction and registration methods to develop a probabilistic model of the background.
- Develop methods for extracting foreground density features among predetermined regions of the image.
- Develop, test and validate a set inference mechanisms based on pattern recognition for estimating the total amount of pedestrians in the scene using the density features.

1.2. Hypothesis

The main hypothesis is that foreground (non-static) regions in each image are proportional to the visible surface of people in the scene, and that this surface is correlated to the number of pedestrians. Hence the foreground information is suitable to estimate the total amount of people in the scene. It will be shown that this first hypothesis is valid

(A) (B)

(C)

FIGURE 1.2. (a) Conventional, (b) infrared and (c) omnidirectional image of people standing at a bus stop.

because of the projective geometry of the camera system. Another assumption is that pedestrians stand still for a short period of time, therefore it is possible to obtain accurate foreground/background statistics by looking at a group of consecutive frames at a time. Considering these assumptions the method should prove more accurate than color or edge based approaches, particularily under crowded situations and scenes with irregular lighting conditions.

1.3. Review

The first attempts to detect pedestrians in images considered the human body as a whole and employed shape models or edge templates. Papageorgiou and Poggio (1999) employed Haar wavelets and a trained Support Vector Machine (SVM) for deciding if a region of interest contains a human or not. Gavrila (2000); Gavrila and Giebel (2002); Gavrila (2007), studied the performance of the Chamfer System for detecting pedestrians, which consists in correlating the Distance Transform of the source image with edge templates of humans in different poses. The final classification from the correlation image is done by SVMs or Neural Networks (NN). On a similar manner, Felzenszwalb (2001) compares edge templates with the source image using the Hausdorff distance. These approaches have proved quite useful for detecting non-occluded objects, such as pedestrians, faces and cars. However, detection of the whole body using holistic models often yields poor results in practice due to the many different poses pedestrians can assume. In addition to the lack of pose invariance, the detection of edges for proper segmentation is further complicated by occlusions and clothing colors.

Viola and Jones (2001) proposed a different framework for object detection based on the selection of discriminative Haar features to train a cascade of AdaBoost detectors (Freund & Schapire, 1996). Their approach proved very efficient in detecting invariant objects, particularly in the face detection problem (Viola & Jones, 2001; Viola, Jones, & Snow, 2005) reporting detection rates of about 95% and false positives rates below 1%. The main advantage of their classification scheme is the detection of objects in real time, hence making it very effective in practical applications. This is achieved extracting large amounts of Haar features that are fast to compute, and using a set of simple classifiers in cascade, rather than a single complex and time consuming classifier. However, this approach exhibits a low performance when applied to the detection of deformable objects, such as walking pedestrians. In order to overcome this problem, Viola and Jones added features

from motion patterns (Viola et al., 2005). This technology, applied to face detection, is currently incorporated in commercial digital cameras, enabling them to detect faces in a picture and automatically adjust the camera's focus. Regarding the people detection problem, other researchers have decided to divide the whole body into some invariant segments, like head-shoulders, arms, and legs, considering frontal, lateral and profile views of those parts (Mohan, Papageorgiou, and Poggio (2001), Mikolajczyk, Schmid, and Zisserman (2004), Wu and Nevatia (2007)). Each of these part detectors consider features built from histograms of oriented gradients and AdaBoost or SVM classifiers. These authors conclude that detecting humans by their components allows finding partially occluded pedestrians and increases detection rates.

In many computer vision applications, the camera employed is fixed at a specific position. This allows the identification of a scene background model which allows to identify objects that do not belong to the scene. In the case of pedestrian counting at bus stops, background subtraction is a useful tool since the scene is mostly invariant and the objects that do not belong to the scene are very likely to be people.

The simplest way of modeling the scene is constructing a background image by averaging a sequence of frames, subsequently the foreground objects can be identified by thresholding the difference between the background and the input image. The main disadvantage of this solution is that it does not handle properly multimodal color distributions of each pixel in the image and carries the difficulty of assigning the appropriate threshold for each pixel (Rosin, 1998).

Wren, Azarbayejani, Darrell, and Pentland (1996) considered that each pixel color values have a multivariate Gaussian distribution and each threshold is determined by the Mahalanaobis distance to the pixel mean. The mean of the Gaussian distribution is estimated by a running average (2.3) and the covariance matrix is determined by a running scatter matrix. This approach proved successful results in controlled indoor environments, but in outdoor situations, due to varying lighting conditions or dynamic background objects, the system's performance diminishes. To cope with changing illumination conditions

and dynamic background objects Friedman and Russell (1997) and Stauffer and Grimson (1999) introduced a Multivariate Mixture of Gaussians (MMoG) background model, where each parameter of the MMoG can be determined using K-Means Clustering or employing the Expectation Maximization algorithm (Dempster, Laird, & Rubin, 1977). The main advantage of this method is that each of the parameters from the mixture are adapted through time, evolving according to the lighting conditions.

Other authors embrace the concept of Eigen-Backgrounds (Oliver, Rosario, & Pentland, 2000; Rymel, Renno, Greenhill, Orwell, & Jones, 2004; Zhang & Zhuang, 2007), that consists in building a representative subspace of the image space by Principal Component Analysis (PCA). This method consists in calculating the eigenvalues and eigenvectors of the covariance matrix of a sequence of sample images and building a transformation matrix with a subset of eigenvectors (or eigen-backgrounds) with higher eigenvalues. Finally the foreground area is detected by comparing the captured image to a generated eigen-background. Since each sample is conformed by the intensity values of the pixels from the whole image, the final classification of each individual pixel is done with information from the entire image rather than with information from the pixel alone. In the foreground segmentation task this method shows good results but, because of the high dimensionality of the feature space and the amount of samples needed, the memory requirements for PCA in this application are very high making this method often impracticable. However, once the eigen-backgrounds are obtained the segmentation processing time is smaller than that of MMoG according to Oliver et al. (2000). Another drawback of this approach is that the eigenspace must be obtained from a representative set of samples (Rymel et al., 2004), or calculated incrementally (Zhang & Zhuang, 2007), such that it can represent different backgrounds over time. In either case, the latter adds complications to the practical implementation.

Other methods consider building a non-parametric model formed by kernels built from a set of samples that are most likely to be background (Elgammal, Harwood, & Davis, 2000). The main disadvantage is that there is no certainty on which samples should be added to the model, but using simple rules, e.g. low variance over time, some certainty

is gained while considering a given sample. Authors also consider the use of short and long term models, the first for sensitive detection and the second to gain a more stable background representation. Other authors consider optical flow components in order to identify still objects that can be regarded as background (Mittal & Paragios, 2004).

Another line of work considers modeling each pixel according to color and/or gradient histograms in a bounding region (Noriega & Bernier, 2006). The main drawback of this solution is that the feature space must be reduced in order to restrict the size of the histograms, hence diminishing the sensibility of the classification scheme, but the approach shows good results in cases where there is significant contrast between foreground and background (Panahi, Sheikhi, Hadadan, & Gheissari, 2008). Other authors propose the use of energy or entropy functions in terms of color and contrast information in a region (Sun, Zhang, Tang, & Shum, 2006) to model the background, but relying on a probabilistic model, such as MMoG, to calculate entropies.

In general, all foreground segmentation algorithms have different advantages depending on the application and on the background properties. For example, some backgrounds have considerable noise, adding complexity to the segmentation algorithms, while other backgrounds are stable over time and only require a simple classifier. A recent evaluation of the most popular background subtraction algorithms under different type of scenarios can be found in Panahi et al. (2008).

1.4. Contribution

This work proposes an approach based on computer vision techniques to count passengers waiting at regular bus stops without any sort of passageway sensors. This method can be applied to a sequence of images captured from either conventional perspective cameras (fig.1.2a), infrared cameras (fig.1.2b) or omnidirectional cameras (fig.1.2c). Infrared cameras have the advantage of being able to detect pedestrians on nighttime, since it detects the thermal emissions of objects, rather than visible light. This makes them the appropriate sensor when there are no adequate lighting conditions for using a normal camera or when a

considerable amount of passengers are awaiting in a bus stop during nighttime. Omnidirectional cameras, on the other hand, have a larger field of view compared to other cameras, allowing a $360^\circ \times 80^\circ$ field of view. This permits a panoramic view of the scene and lays aside the need of mounting the camera far from the bus stop, as illustrated in figure 3.2.

As described extensively in the next chapter, the novelty of this approach lies in the fact that the people counting process is done using foreground pixel density estimates. This approach will be referred to as PDM (People Density Method). It is shown that despite PDM's simplicity, it is more accurate and reliable than an alternative solution based on the well-known Viola-Jones (VJ) detection scheme applied to foreground regions. This is mainly because PDM does not require to solve a recognition problem with strong assumptions of object shape-invariance (non-deformability), as is implicit in the VJ approach. On the other hand, shape extracted from contours is often lost in crowded scenes. The advantage of PDM is that no special assumptions on the pose or motion of people is required. This method can be applied in either conventional perspective cameras, infrared cameras or omnidirectional cameras. The only two basic assumptions required are i) that the background can be modeled with reasonable accuracy, and ii) that all foreground pixels correspond to people. Both of these assumptions are fulfilled most of the time because background subtraction can effectively be solved using the approach presented in section 2.2. On the other hand, most of the time foreground pixels are generated by pedestrians, and even if sometimes small pets or people carrying large objects may produce larger foreground areas, the additional foreground pixels introduce negligible errors.

The proposed approach consists on a two stage process, as described in figure 1.3. The first stage implements a foreground segmentation algorithm that uses a MMoG distribution to model the background using the color and gradient information of each pixel. The second stage consists in testing pattern recognition algorithms in the task of estimating the total amount of pedestrians in the scene using only foreground information.

The contribution of this work can be summarized in a simple and novel method for pedestrian counting at bus stops (or any open environment) based on area computations.

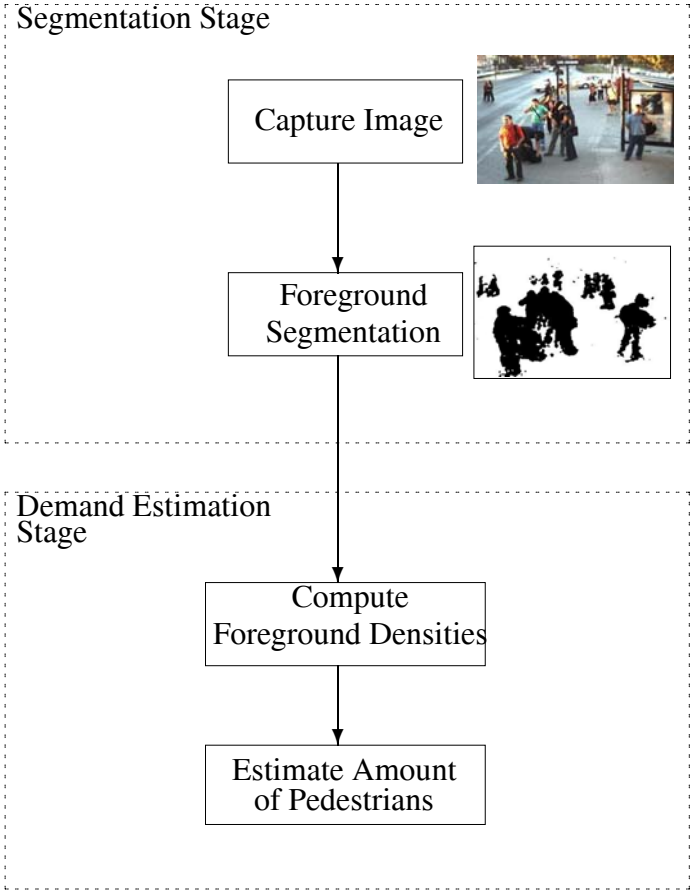


FIGURE 1.3. Outline of the proposed method.

The approach does not require people to walk through a portal. The thesis shows that the proposed is more accurate than the method based on a combined action of a background subtraction process and Viola-Jones approach. Both approaches are tested using image sequences from a real bus stop.

2. PROPOSED APPROACH

2.1. System Overview

The proposed approach and the reference method are implemented as a two-stage process (see fig. 2.1). First the foreground is extracted as the complement to the background, which is modeled as a mixture of Gaussians. In the second stage, foreground pixel densities are employed to estimate the people count. To make the comparison with the Viola-Jones approach valid, the body-parts detected with this method are selected only if they belong to the foreground in order to remove false detections.

2.2. Background Identification

Background identification is an essential preliminary step whose purpose is to reduce the search for pedestrian contours or their extremities to non-background regions. This reduced search area does not only saves valuable processing time, but also allows to increase detection certainty because it should be easier to identify mostly static backgrounds than trying to find directly the many different dynamic objects there might be in the foreground. The background identification can be achieved using different methods, such as texture analysis, edge extraction, color and intensity filters. Texture and edge based methods are more robust to brightness changes. However, texture segmentation requires computationally demanding calculations, while edges alone may provide insufficient information to bound the background regions. On the other hand, color and intensity filters are more vulnerable to brightness variations or confusion in the presence of foreground objects with colors that closely resemble those of the background.

In this work background segmentation is performed using a Multivariate Mixture of Gaussians (MMoG) probability distribution in the color & gradient feature space. The feature vector is defined as a five component vector $\mathbf{f} = [r, g, b, g_o, g_m]^T$ containing the red, green and blue color values, gradient orientation and gradient magnitude at a given pixel.

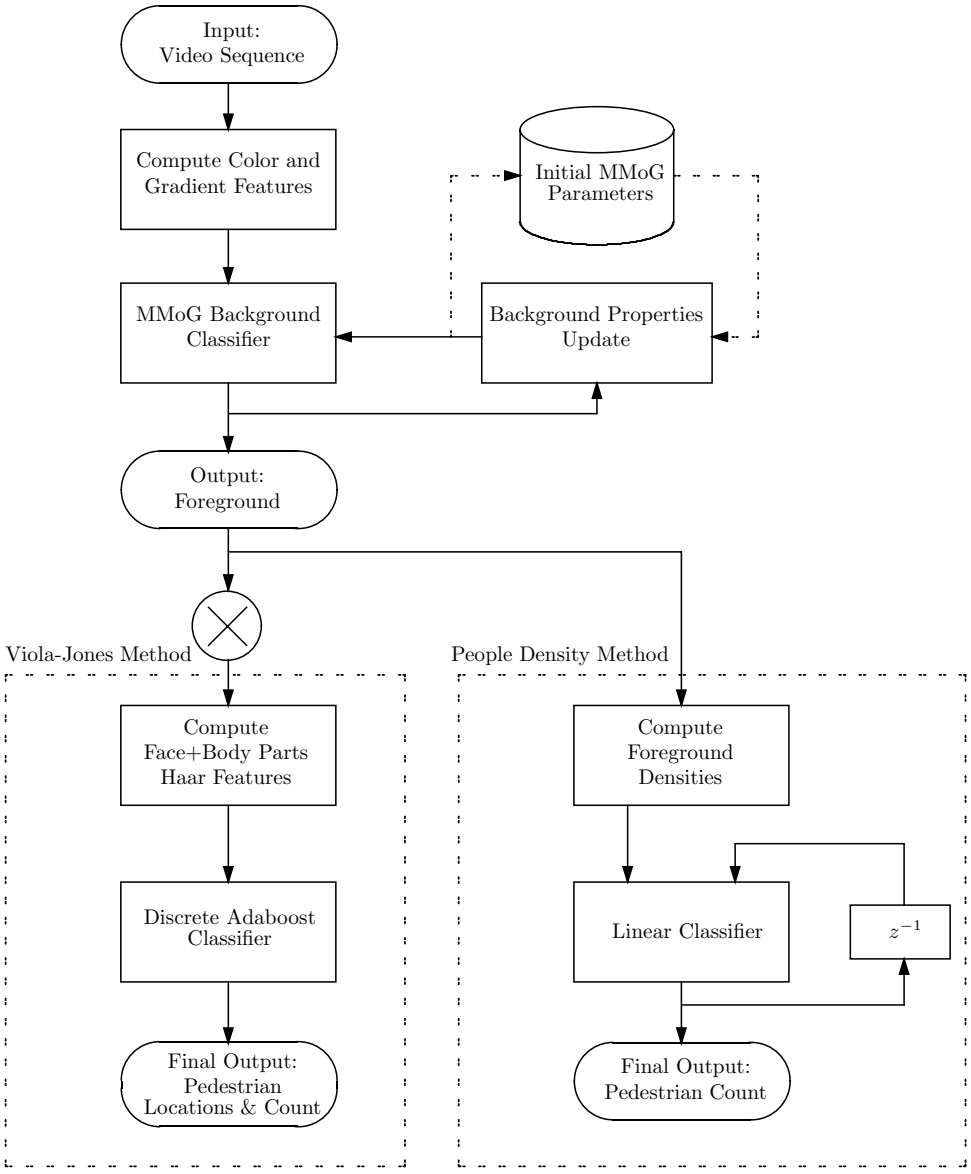


FIGURE 2.1. Detection scheme.

The gradient magnitude and orientation at pixel (u, v) is calculated as

$$\begin{aligned} g_m &= \sqrt{g_u^2 + g_v^2} \\ g_o &= \arctan(g_u, g_v) \end{aligned} \tag{2.1}$$

where

$$\begin{aligned} g_u &= I_{u+1,v} - I_{u-1,v} \\ g_v &= I_{u,v+1} - I_{u,v-1} \end{aligned} \tag{2.2}$$

and $I_{u,v}$ denotes the image intensity value at pixel (u, v).

The background subtraction approach provides one of the simplest techniques to detect objects of interest in urban scenes, such as pedestrians and cars, using static cameras. This approach relies on the assumption that the background does not change significantly from one video frame to the other, and hence, computing the difference between frames would yield the moving foreground objects. If the foreground objects are also static, the subtraction of contiguous frames will evidently not reveal the foreground objects. This problem can be avoided if the difference is computed between the current frame and a reference background (free from foreground elements). However, obtaining a reference background is a challenging problem because real backgrounds are not constant due to illumination changes during the day, sudden appearance of clouds, presence of reflecting objects, camera oscillations, high-frequency background variations like rain or moving leaves.

There are many approaches to model the background for background subtraction purposes (Piccardi, 2004). The simplest method is to employ a first order IIR running average filter to compute the background model B_k at instant k in terms of the previous background model B_{k-1} and the new image frame I_k as:

$$B_k = (1 - \alpha)B_{k-1} + \alpha I_k \tag{2.3}$$

where α is the filter update or *learning* rate. This method requires a small amount of memory and is fast to compute, however it does not consider the fact that each pixel may have different intensity variations and that background colors can have multimodal distributions over time.

An approach that copes with multimodal distributions is the method by Stauffer and Grimson (1999), which models each pixel as a mixture of L Gaussians with parameters

$\mu_k^l \in \mathbb{R}^n$, $\mathbf{\Sigma}_k^l \in \mathbb{R}^{n \times n}$, $l = 1, 2, \ldots, L$ at time k. More specifically, the probability of observing a feature vector $\mathbf{f}_k$ at a given pixel is given by:

$$P(\mathbf{f}_k) = \sum_{l=1}^{L} \omega^l P(\mathbf{f}_k | \mu^l, \mathbf{\Sigma}^l) \tag{2.4}$$

where $\omega^l = P(l)$ are the priors weighting the Gaussian distributions:

$$P(\mathbf{f}_k | \mu^l, \mathbf{\Sigma}^l) = \frac{1}{(2\pi)^{n/2} |\mathbf{\Sigma}^l|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{f}_k - \mu^l)^T \mathbf{\Sigma}^{l\,-1} (\mathbf{f}_k - \mu^l)} \tag{2.5}$$

that form the mixture.

It is to be noted that the mixture models both the foreground and background without distinction. Hence, L is not the number of background classes, but the number of all possible pixel distributions. Because of this, the choice of L should be $L \geq B + 1$, if there are B background classes. In practice, $B \geq 2$ to model at least two background classes, hence $L \geq 3$. The current literature reports values of $L \leq 7$, however, significant improvements are unlikely for values beyond $L = 5$. Following Stauffer and Grimson (1999) and Cheng, Yang, Zhou, and Cui (2006), B is chosen as the smallest number of modes with representative weight, i.e.

$$B = \underset{b}{\operatorname{argmin}} \left(\sum_{l=1}^{b} \omega^l > \delta_b \right) \tag{2.6}$$

where each Gaussian is sorted in decreasing order according to $\omega_k^l / \left\| \mathbf{\Sigma}_k^l \right\|$, and δ_b is a thresholding parameter related to the overall background probability.

A pixel is declared to match one of the L Gaussian distributions, if

$$\sqrt{(\mathbf{f}_k - \mu_k^l)^T \mathbf{\Sigma}_k^{l\,-1} (\mathbf{f}_k - \mu_k^l)} < \lambda \tag{2.7}$$

for some l, where λ represents the number of standard deviations from the mean that defines the matching threshold ($\lambda = 2.5$ in Stauffer and Grimson (1999)). The pixel is declared to belong to the background whenever it matches one of the first B distributions.

If a pixel matches one of the L possible distributions, the parameters of the corresponding probability density function are updated using an incremental form of the Expectation Maximization algorithm (Dempster et al., 1977) applied to MMoG:

$$\mu_k^l = \frac{S_{k-1}}{S_k}\mu_{k-1}^l + \frac{1}{S_k}\tilde{\omega}^l \tag{2.8}$$

$$\mathbf{\Sigma}_k^l = \frac{S_{k-1}}{S_k}\mathbf{\Sigma}_{k-1}^l + \frac{1}{S_k}\tilde{\omega}^l \left(\mathbf{f}_k - \mu_k^l\right)\left(\mathbf{f}_k - \mu_k^l\right)^T \tag{2.9}$$

$$\omega_k^l = \frac{1}{k}S_{k-1} + \frac{1}{k}\tilde{\omega}_k^l \tag{2.10}$$

where

$$S_k = \min\left\{S_{k-1} + \tilde{\omega}_k^l, \alpha\right\} \tag{2.11}$$

$$\tilde{\omega}_k^l = \frac{P\left(\mathbf{f}_k \left| \mu^l, \mathbf{\Sigma}^l\right.\right)\omega_k^l}{\sum_{c=1}^{L} P\left(\mathbf{f}_k \left| \mu^c, \mathbf{\Sigma}^c\right.\right)\omega_k^c} \tag{2.12}$$

where α is the fixed learning rate used upon convergence. Since S_k increases over time, each class parameter becomes stiff, insensitive to incoming data, hence the upper bound α guaranties flexibility once the parameters of the model converged. If a pixel does not match any of the L distributions, the least probable distribution (i.e. the one with lowest ω^l) is replaced by a distribution with the current pixel value as its mean, an initially high variance and a low prior ω^l. This step allows to update the background model without degrading the model as in the case of the unimodal distribution, because when some new object appears in the scene, the background parameters are not lost until one of them becomes the L least likely distribution, i.e. when the background class weight w^l becomes the smallest of the L weights. The algorithm is initialized with a predetermined number of classes L, each with random class means, all with equal covariances and equal class probabilities. Once the MMoG parameters have converged, those modes that do not satisfy $\omega^l > \omega_{min}$ are eliminated from the set, and each remaining ω^l is renormalized such that $\sum_{l=1}^{L_{new}} \omega^l = 1$. The parameter ω_{min} should be near zero, so that only the modes that practically never occur are eliminated. This allows the algorithm to automatically choose the number of modes in each of the pixels, thus reducing the computational load and processing time.

2.3. Pedestrian Count Using Density Estimates

For clarity of exposition the basic notation will be introduced first. The pinhole camera model (Hartley & Zisserman, 2004) relates a point in space $\mathbf{x}$ with homogeneous coordinates $[\ x \quad y \quad z \quad 1\ ]^T$ to its projection in the optical plane (e.g. CCD, CMOS array) $\mathbf{x}_{cam}$ with homogeneous coordinates $[\ u \quad v \quad 1\ ]^T$ according to:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f\rho_x & 0 & u_0 & 0 \\ 0 & f\rho_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & -h_{cam} \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{2.13}$$

where f is the focal length of the camera, ρ_x and ρ_y are the scale factors in each of the optical plane axes, u_0 and v_0 are the vertical and horizontal offsets that relate the origin of the global coordinate frame to the image plane origin, and h_{cam} is the height of the camera's stand-pole. The global and image plane coordinate systems are disposed as shown in figure 2.2, where x is the distance from the stand-pole, y and z are the horizontal and vertical distances respectively, u and v are the vertical and horizontal coordinates in the image plane which are coplanar with the plane formed by y and z. The 4×4 matrix in

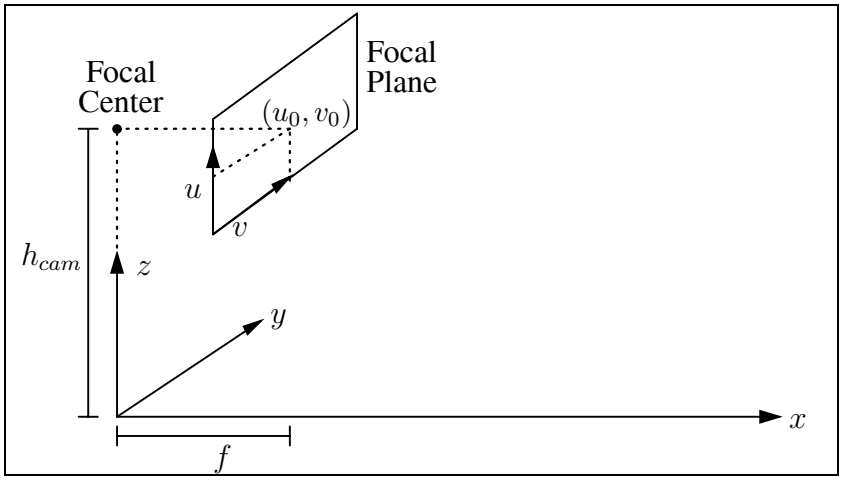


FIGURE 2.2. Coordinates systems in the global frame and in the image plane.

equation (2.13) is a rotation and translation operator that transforms global coordinates to

the cameras local coordinates and the 3×4 matrix is the pinhole projection matrix that maps local coordinates to focal plane pixel coordinates. Combining the two operations, the transformation can be stated as:

$$\lambda \mathbf{x}_{cam} \quad = \quad H\mathbf{x} \tag{2.14}$$

where H is a 3×4 projection matrix.

It is possible to estimate the total amount of pedestrians in the scene by measuring the foreground area that each of these individuals produce considering their relative position from the camera. Assuming that people on the scene are standing in a upward position,

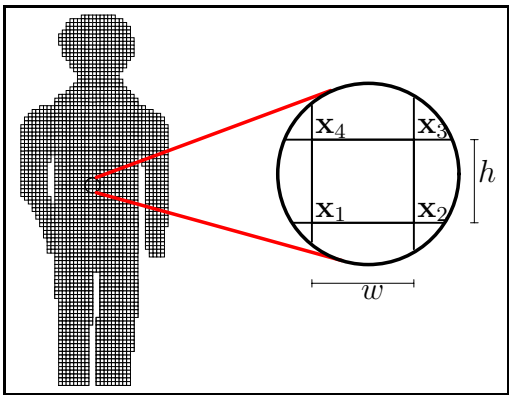


FIGURE 2.3. Foreground elements divided into rectangular sections.

the projected surface of each individual can be approximated by the sum of small vertical rectangular surfaces as illustrated in figure 2.3. The relationship between the area of each rectangle and the foreground area in the image plane can be formulated considering that each of these rectangles is at a distance x from the stand-pole and is defined by its four corners in homogeneous coordinates:

$$\begin{aligned}
\mathbf{x}_1 &= \begin{bmatrix} x & y_0 & z_0 & 1 \end{bmatrix}^T \\
\mathbf{x}_2 &= \begin{bmatrix} x & y_0 + w & z_0 & 1 \end{bmatrix}^T \\
\mathbf{x}_3 &= \begin{bmatrix} x & y_0 + w & z_0 + h & 1 \end{bmatrix}^T \\
\mathbf{x}_4 &= \begin{bmatrix} x & y_0 & z_0 + h & 1 \end{bmatrix}^T
\end{aligned} \tag{2.15}$$

where h and w are the rectangle's height and width respectively and y_0 and z_0 is the horizontal and vertical distances from the global coordinate system origin. Each point is projected on the focal plane at $\lambda_i \mathbf{x}_i^{cam} = H\mathbf{x_i}$, $i = 1 \ldots 4$.
In the optical plane, the area of the projected quadrilateral can be calculated as (Beyer, 1987):

$$\Delta A_p \;=\; \frac{1}{2}\sum_{i=1}^{4} u_i v_{i+1} - u_{i+1} v_i \quad [pixels^2] \tag{2.16}$$

Replacing (2.14) and (2.15) in equation (2.16), the area of a projected quadrilateral can be stated as:

$$\Delta A_p \;=\; \rho_x \rho_y f^2 \frac{\Delta A}{x^2} \quad [pixels^2] \tag{2.17}$$

where $\Delta A = hw$ is the original area of the rectangle. It follows from (2.17) that the projected area of each rectangle is proportional to the original surface and inversely proportional to the squared distance from the stand-pole. Hence the area projected on the image plane by a pedestrian is proportional to its real visible surface.

The vertical position in the image plane of an object of height l measured from the ground is given by

$$u \;=\; u_0 + \frac{1}{2}\frac{f\rho_x(l - 2h_{cam})}{x} \quad [pixels] \tag{2.18}$$

In the context of our work, x represents the distance at which pedestrians stand from the camera, whereas l represents the height of the pedestrians. Since the variations in x are much larger than those in l, then l in equation (2.18) can be replaced with the objects' mean height l_0. Since the objects in the image are mainly people it is valid to assume the existence of l_0. Solving for x in (2.18), and replacing in (2.17) yields:

$$A_p \;=\; \frac{4\rho_y}{\rho_x(l_0 - 2h)^2}(u - u_0)^2 A \quad [pixels^2] \tag{2.19}$$

The area of a segmented object in the image is proportional to the amount of objects. Therefore, if a scene contains n pedestrians, the total area of a foreground region will be

given by:

$$A_p \quad = \quad kA_0(u-u_0)^2 n \quad [pixels^2] \tag{2.20}$$

where $k = \frac{4\rho_y}{\rho_x(l_0-2h_{cam})^2}$, A_0 is the average visible area of an individual and n is the amount of pedestrians in the analyzed foreground region.

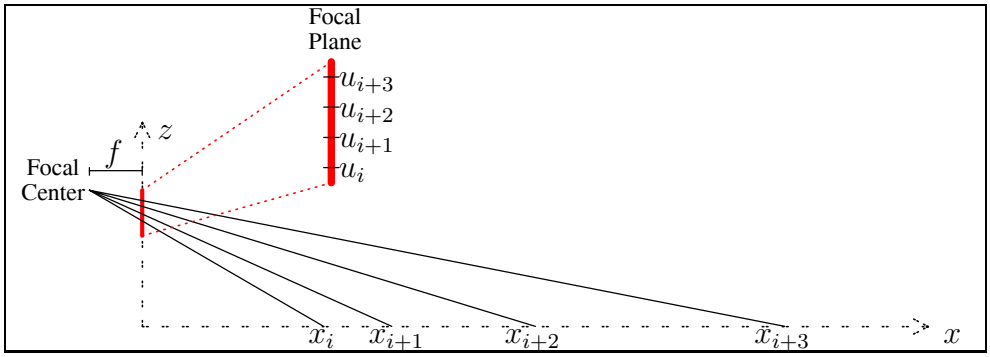


FIGURE 2.4. Horizontal sections in the focal plane representing different distance intervals.

It is to be noted that equation (2.20) is valid for pedestrians standing at a given distance x from the camera. However, since in a real bus stop pedestrians may stand at different locations, their projected areas will be dependent on x. It is therefore convenient to divide the image into N horizontal sections. Figure 2.4 shows that for each distance interval $[x_i, x_{i+1}]$ there is a horizontal section with coordinates $[u_i, u_{i+1}]$ in which (2.20) can be assumed to hold. In each of these sections the count of persons can be computed as:

$$n_i \quad = \quad \eta_i A_i \quad [individuals] \tag{2.21}$$

where $\eta_i = \frac{1}{k(u_i-u_0)^2}$ is the pedestrian density coefficient for each vertical section located at u_i.

In order to estimate the total amount of pedestrians, the problem of counting people is solved by measuring foreground pixels areas per section and finding the model parameters i.e. the density coefficients η_i, that minimize the estimation error for the total amount of

pedestrians is given by:

$$n \;=\; f(n_1, \ldots, n_N, \theta) \quad [individuals] \tag{2.22}$$

where n is the total amount of pedestrians and θ is the vector of parameters that describe the model. A reasonable expression of (2.22) is to consider n as the summation of individuals in each region, which by (2.21) is given by:

$$n = \sum_{i=1}^{N} n_i = \sum_{i=1}^{N} \eta_i A_i \;\; [individuals] \tag{2.23}$$

where the parameters to be found are $\theta = [\eta_1, \ldots, \eta_N]$. Note that since the pedestrian count depends on A_i it is not necessary to obtain explicitly the camera parameters because they are taken into account by the coefficients η_i, which can be found by linear regression from available training data.

Considering that the total amount of pedestrians per frame is subject to continuity constraints, the model takes into account past measurements to limit the amount of instantaneous change in the estimated number of pedestrians. The output of the fitted model at frame k is thus of the form:

$$\hat{n}(k) = \sum_{i=1}^{N} \hat{\eta}_i A_i(k) + \sum_{p=1}^{P} \hat{\beta}_p \hat{n}(k-p) \;\; [individuals] \tag{2.24}$$

The previous equation is commonly known as a linear regression model (LRM). Each of the elements of θ are found as those that minimize the overall root mean square error (RMSE):

$$\hat{\theta} \;=\; \underset{\theta}{\operatorname{argmin}} \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left\| \hat{n}\left(k, \theta\right) - n\left(k\right) \right\|} \tag{2.25}$$

where K is the total amount of training samples, $n(k)$ is the real amount of pedestrians at sample k and $\hat{\theta} = [\hat{\eta}_1, \ldots, \hat{\eta}_N, \hat{\beta}_1, \ldots, \hat{\beta}_P]$.

2.4. Other Classification Schemes

In addition to LRM, three other approaches are considered concerning the structure of the function f in (2.22) relating A_i to n. These approaches are Linear Discriminant Analysis (LDA), Probabilistic Neural Network (PNN) and k-Nearest Neighbors (KNN). Each model is trained and tested with the same training and testing samples and cross-validated for tolerance analysis in the same fashion.

The LDA approach consists in projecting the feature vector to a subspace of lower dimension. In contrast to PCA, the objective of LDA is to find a subspace in which classes are more distinguishable, rather than finding a subspace in which samples have maximum variance. Following the methodology described extensively in (Duda, Hart, & Stork, 2001), this is done by finding a linear transformation matrix that describes the features subspace in which the Fisher Discriminant is maximum. If $\mathbf{A} = [A_1, \ldots, A_N]$ is the original sample, then its subspace representation $\mathbf{A}^{\subset}$ is:

$$\mathbf{A}^{\subset} \quad = \quad W^T \mathbf{A} \tag{2.26}$$

where W, $\left\{W : \mathbf{A} \subset \mathbb{R}^N \rightarrow \mathbf{A}^{\subset} \subset \mathbb{R}^M, M < N\right\}$, is the subspace transformation matrix ($N \times M$), found while maximizing the interclass separability function (or Fisher discriminant):

$$W \quad = \quad \underset{w}{\operatorname{argmax}} \frac{\left|w^T S_B w\right|}{\left|w^T S_W w\right|} \tag{2.27}$$

where S_B and S_W are the between-class and within-class covariance matrices. In this application, each class is defined as the number of pedestrians, where C is the maximum number of pedestrians plus one. If N_j is the number samples in class j, $\bar{\mathbf{A}}$ is the mean value of the feature vectors on the training set and $\bar{\mathbf{A}}_j$ is the mean value of the feature vectors belonging to class j in the training set, then S_B is given by:

$$S_B = \sum_{j=1}^{C} N_j (\bar{\mathbf{A}}_j - \bar{\mathbf{A}})(\bar{\mathbf{A}}_j - \bar{\mathbf{A}})^T \tag{2.28}$$

The within-class covariance matrix S_W is the sum of the covariance matrices of each class S_j:

$$S_W = \sum_{j=1}^{C} S_j \tag{2.29}$$

where S_j is given by:

$$S_j = \sum_{\mathbf{A}(k) \in \text{class} j} (\mathbf{A}(k) - \bar{\mathbf{A}}_j)(\mathbf{A}(k) - \bar{\mathbf{A}}_j)^T \tag{2.30}$$

The optimization problem enunciated in 2.27 is solved by finding the generalized eigenvectors and eigenvalues (w_i and λ_i, $i = 1, \ldots, N$), of S_W and S_B, i.e. solving:

$$S_B w = \lambda S_W w \tag{2.31}$$

W is the collection of the M eigenvectors with larger eigenvalues as described in (Duda et al., 2001). For purposes of pedestrian counting, the subspace dimension is chosen to be $M = 1$, so that the final classification is done over a single variable.

If each sample in the new subspace is $\mathbf{A}^{\subset}(k) = W^T \mathbf{A}(k)$, the class means $\mathbf{A}_{\mathbf{j}}^{\subset} = W^T \bar{\mathbf{A}}_j$ and the class scatter is $S_j^{\subset} = W S_j W^T$, then the amount of pedestrians on the scene at sample k given sample $\mathbf{A}(k)$ is approximated by:

$$\hat{n}_{LDA}(k) = \underset{j}{\operatorname{argmin}} \left[\mathbf{A}^{\subset}(k) - \mathbf{A}_{\mathbf{j}}^{\subset}\right] S_j^{\subset -1} \left[\mathbf{A}^{\subset}(k) - \mathbf{A}_{\mathbf{j}}^{\subset}\right]^T \tag{2.32}$$

Once the classification process is completed, the filtering process is carried out using the following time series:

$$\hat{n}(k) \;\; = \;\; b_0 \hat{n}_{LDA}(k) + \sum_{p=1}^{P} b_p \hat{n}(k-p) \tag{2.33}$$

where $\hat{n}(k)$ is the final output and the parameters $b_0, ..., b_P$ are found by a simple linear regression.

The PNN approach (Duda et al., 2001) consists in a non-parametric model that can be represented as a two layer network, as shown in figure 2.5. The first layer after the input layer is composed of one radial basis neuron (Rb_j) for each training sample. If $\mathbf{A}$ is the input and $\mathbf{A}_t(j)$ is the training sample corresponding to neuron Rb_j, then the output of this neuron is:

$$y_j = e^{-\frac{\|\mathbf{A}-\mathbf{A}_t(j)\|^2}{\sigma_j^2}} \tag{2.34}$$

Each of the neurons that correspond to a sample of a given class k are connected to one neuron (S_k) on the second layer, where there is one neuron for each class. Each of these neurons simply average their inputs, i.e.

$$g_k = \frac{1}{N_k} \sum_{j \in C_k} y_j \tag{2.35}$$

The final output is the neuron number in the second layer with highest output:

$$\hat{n} = \underset{k}{\operatorname{argmax}}\, g_k \tag{2.36}$$

This kind of network resembles a voting scheme, where each training sample votes according to its similarity with respect to the input feature vector. For further references on this particular type of network see Wasserman (1993) and Duda et al. (2001).

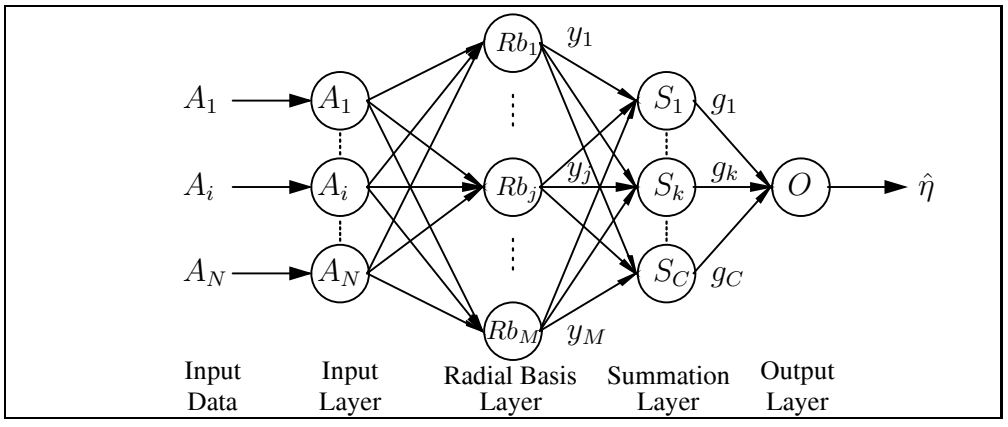


FIGURE 2.5. Probabilistic Neural Network for PDM

The last type of model evaluated, the KNN approach (Duda et al., 2001), is also a non-parametric technique that is widely used in many pattern recognition applications. It consists on finding the K most similar training samples to a given sample using an Euclidean distance. The final output can be either the most frequent output from those K samples or a weighted sum of the outputs.

3. TESTING METHODOLOGY

3.1. Data Acquisition

Video sequences using conventional, omnidirectional and thermal cameras were captured late in the afternoon when the sun was close to the horizon and a few ours after sunset, thus involving challenging lighting conditions.

The camera employed, for both normal and omnidirectional sensing is a standard Firewire® camera with a 1024x768 pixels 1/2" CCD and a Tamron varifocal lens with focal distances in the range 6-12 mm, corresponding to a field of view in the range $30.4° \times 23.1°$ (telephoto) – $58.7° \times 44.4°$ (wide). The regular camera camera was located at $3[m]$ from the bus stop at a height of $3.2[m]$ above the ground, as shown in fig. 3.1. This configuration allows to cover an area in which pedestrians would normally stand (see fig. 1.2).

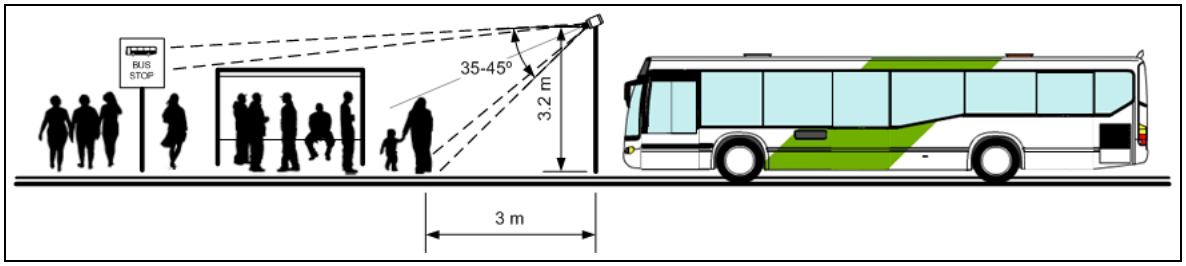
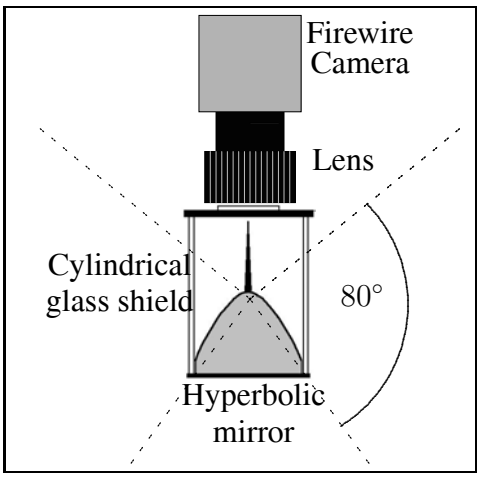


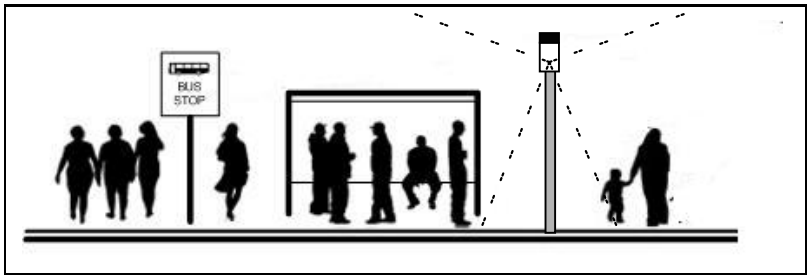
FIGURE 3.1. Bus stop and camera configuration.

An ACCOWLE hyperbolic mirror with a $360° \times 80°$ field of view was employed for omnidirectional sensing. The camera is facing towards the mirror as shown in figure 3.2a while the mirror is facing upward with it principal axis vertically oriented and coinciding with the camera's focal axis. One advantage of this sensor is that, because of its large field of view, it can be mounted in the bus stop area rather than outside, as illustrated on figure 3.2b.

The thermal camera employed was an Electrophysics PV320L with a 2-14$[\mu m]$ spectral sensitivity band, a resolution of 320×240 pixels and a 50mm F1 lens. The field of view of this camera is $18° \times 13°$ and has thermal sensitivity of $0.08°C$. The thermal camera



(A) Detailed Omnidirectional Sensor and its field of view



(B) Bus Stop and camera configuration

FIGURE 3.2. Omnidirectional camera configuration

is oriented as shown in figure 3.1, but $10m$ away from the bus stop rather than $3[m]$ as the regular camera because of its narrower field of view.

3.2. Results Analysis and Validation

Each of the classification schemes was trained and tested with randomly generated partitions of a data set. Table 3.1 shows the amount of samples available for each type of camera. In each experiment 70% of the samples were used for training and the remaining 30% for testing.

The validation experiments were repeated $N_e = 1000$ times. The results so obtained were averaged in order to obtain a mean statistic and its respective confidence interval in

TABLE 3.1. Number of available samples for training and testing.

Type of camera	Number of available samples
Conventional	3490
Omnidirectional	4493
Infrared	5400

term of its standard deviation. Given a confidence level c, the interval of confidence (IoC) of a set of data with standard deviation σ, is calculated as:

$$IoC = \Phi^{-1}\left(\frac{1+c}{2}\right)\frac{\sigma}{\sqrt{N_e}} \tag{3.1}$$

where Φ^{-1} is the inverse normal cumulative distribution. In this application the selected confidence level is of 95%, i.e. $c = 0.95$, so the interval of confidence is simplified to $1.96\frac{\sigma}{\sqrt{N_e}}$.

Mean error, false positives and false negatives are calculated and tabulated for each classifier and type of camera with their respective interval of confidence. Estimated count versus real amount of people are illustrated using error bars with its corresponding standard deviation.

The proposed method, using conventional perspective cameras is compared to the object detection scheme proposed by Viola and Jones (2001) applied to face and human body recognition. The VJ approach employed relies on the OpenCV implementation of the algorithm with previously trained classifiers. The VJ scheme is aided with the background subtraction process in order to reduce the amount false positives in each frame. The performance of this approach is measured using the same criteria as for the proposed approach.

4. EXPERIMENTAL RESULTS

Figure 4.1 shows the background removal results for frames 100, 700, 1500, 1900, 2600 and 3400, in which white pixels correspond to background. Similar results were achieved using the omnidirectional and thermal cameras. People correctly or wrongly detected with the VJ approach are indicated by boxes surrounding the face or around the body (whenever upper/lower extremities were detected). As may be seen in fig. 4.1e, several false detections occur when there is a crowd of people. The VJ approach failed most of the time to detect some people facing to the side or wearing caps. The full-body detector would only find people standing far away in less crowded areas of the scene. The face detector employed both frontal and lateral face detectors, but performed poorly when the intensity differences of features within the face (eyes, nose, mouth) was small or whenever there were dominant light-shadow effects.

Using a simple classifier, the PDM method exhibits a monotonically decreasing error rate (see fig. 4.2) as the number of horizontal sections N increases. An appropriate selection of the number of regions is $N = 5$, since there is only a 1% or less error reduction when employing 6 or more regions.

TABLE 4.1. Summary of detection results using conventional perspective camera.

	VJ	PDM-LRM	PDM-LDA	PDM-PNN	PDM-KNN
RMSE	$4.11 \pm 0.12^{(1)}$	2.94 ± 0.01	2.55 ± 0.05	3.29 ± 0.15	2.99 ± 0.36
$\overline{FP}$	3.97 ± 0.11	2.52 ± 0.05	2.25 ± 0.10	1.56 ± 0.10	2.78 ± 0.83
$\overline{FN}$	2.64 ± 0.12	2.32 ± 0.06	1.41 ± 0.05	2.95 ± 0.13	1.90 ± 0.57

(1) $c = 95\%$

VJ: Viola-Jones
PDM: People Density Method
LRM: Linear Regression Model
LDA: Linear Discriminant Analysis
PNN: Probabilistic Neural Network
KNN: K Nearest Neighbors

The performance statistics of the PDM and VJ methods are summarized in table 4.1, which presents the RMSE, false positives ($\overline{FP}$) and false negative ($\overline{FN}$). In terms of the RMSE, it is clear from table 4.1 that PDM-LDA performs better than the approach based on VJ detector with RMSE values of 2.55 against 4.11 pedestrians, respectively. The better performance of the PDM may also be appreciated from fig. 4.3, which illustrates the evolution over time of the pedestrian count estimates and the real value.

(A) Frame 100 (B) Frame 700

(C) Frame 1500 (D) Frame 1900

(E) Frame 2600 (F) Frame 3400

FIGURE 4.1. Pedestrian detection results for different frames using the VJ approach aided with background subtraction.

The estimated pedestrian count versus the real amount of pedestrians is shown in figs. 4.4a and 4.4b. Results using different classifiers are illustrated in A.1. These results show that the PDM is more accurate and precise, since the average standard deviation

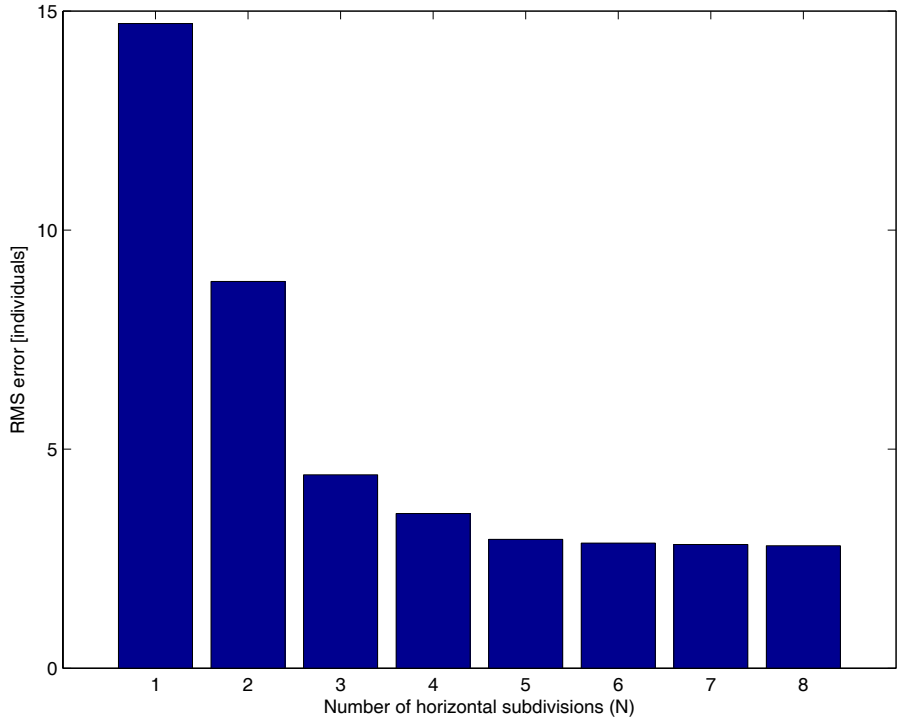


FIGURE 4.2. Error rate for different number of foreground subdivisions.

of the VJ method is 3.02 pedestrians, while the standard deviation for PDM is only 1.15 pedestrians per frame on average. Furthermore, the mean error per class for the PDM is on average 2% lower than the real value, while the VJ approach overestimates the real value by 24% on average. Considering the results presented in table 4.1, the precision of the PDM and its low averaged percentual error, it is possible to conclude that PDM provides a very effective solution. On the other hand, the VJ approach shows more false positives over time than the PDM. It is to be noted that the PDM also yields a smaller number of false negatives on average.

In many frames detection rates achieved with the Viola-Jones approach are very low compared to those found in the literature. This is because the testing conditions are more challenging than those usually reported in the literature, which often consider sample images of pedestrians in scenes that are not very crowded or taken under controlled lighting conditions.

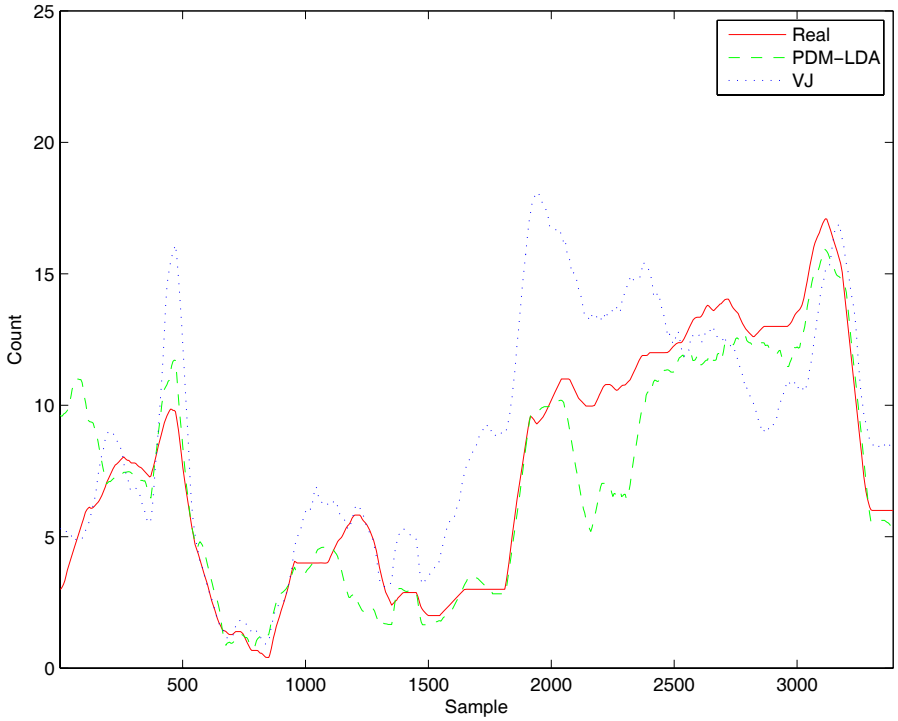
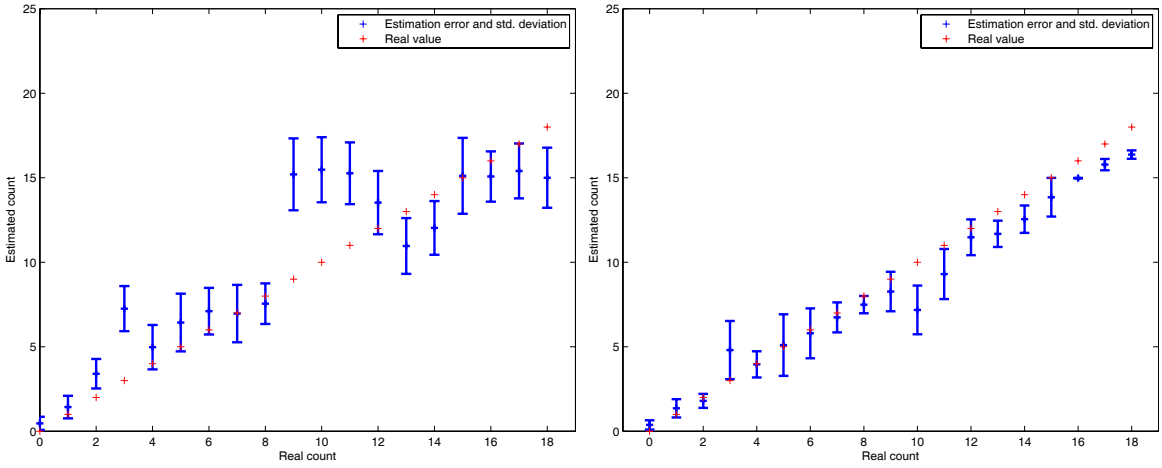


FIGURE 4.3. Pedestrian estimation over time

The loss of actual resolution due to the larger field of view of the omnidirectional mirror significantly limits the applicability of the VJ approach, which performed poorly. On the other hand, the PDM approach applied to the omnidirectional images maintains similar levels of performance to those obtained using conventional cameras, as shown in table 4.2. In this case the number of circular regions was kept the same, i.e. $N = 5$. In contrast to the results obtained using a conventional camera, in which the PDM-LDA yielded the best performance, the PDM-LRM turned out to be the best approach when using omnidirectional images. The estimated pedestrian count versus the real amount of pedestrians is shown in fig 4.6, where a persistent error of 2 individuals can be noticed. Results using different classifiers are illustrated in A.3.

Finally the application of the PDM to infrared images revealed that slightly more accurate pedestrian counts can be achieved because of the robustness of the background segmentation process, which due to the fact that the thermal camera is insensitive to varying lighting conditions and because it is easier to distinguish warm objects from the colder



(A) (B)

FIGURE 4.4. Pedestrians counted versus the real number of pedestrians using a normal camera: (a) VJ approach and (b) PDM-LDA.

FIGURE 4.5. Detail of omnidirectional image of people standing at a bus stop, .

TABLE 4.2. Results of the PDM on omnidirectional images using different classifiers.

	PDM-LRM	PDM-LDA	PDM-PNN	PDM-KNN
RMSE	2.01 ± 0.01	2.52 ± 0.06	3.02 ± 0.08	2.41 ± 0.59
$\overline{FP}$	1.72 ± 0.02	1.26 ± 0.03	1.99 ± 0.09	1.72 ± 0.67
$\overline{FN}$	1.71 ± 0.03	2.69 ± 0.12	2.49 ± 0.10	2.04 ± 0.66

background. Table 4.3 shows that the PDM-LRM once again yields the smallest RMS error. However, even if this error is on average of only 2 pedestrians, fig. 4.7 shows that using the thermal camera may yield very inaccurate results when only one person is waiting at the bus stop. Results using different classifiers are illustrated in A.2.

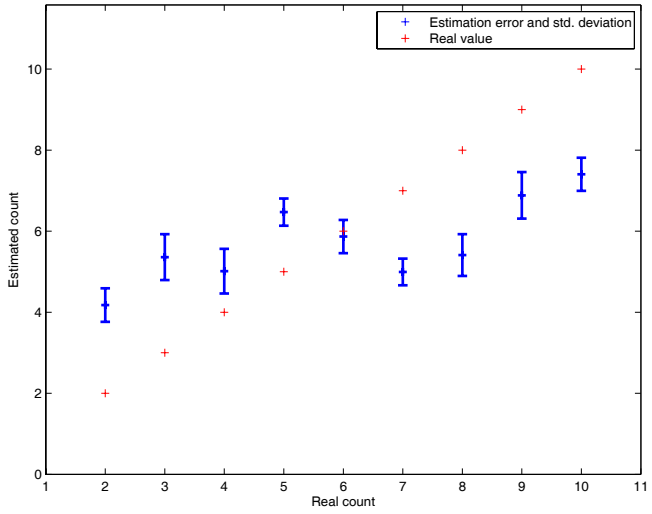


FIGURE 4.6. Pedestrians counted by PDM versus the real value using employing omnidirectional images.

TABLE 4.3. Results of the PDM on thermal images using different classifiers.

	PDM-LRM	PDM-LDA	PDM-PNN	PDM-KNN
RMSE	1.89 ± 0.01	2.01 ± 0.04	3.05 ± 0.07	2.37 ± 0.46
$\overline{FP}$	1.60 ± 0.02	1.43 ± 0.03	1.82 ± 1.71	2.23 ± 1.39
$\overline{FN}$	1.16 ± 0.01	1.03 ± 0.03	2.50 ± 0.33	1.54 ± 0.41

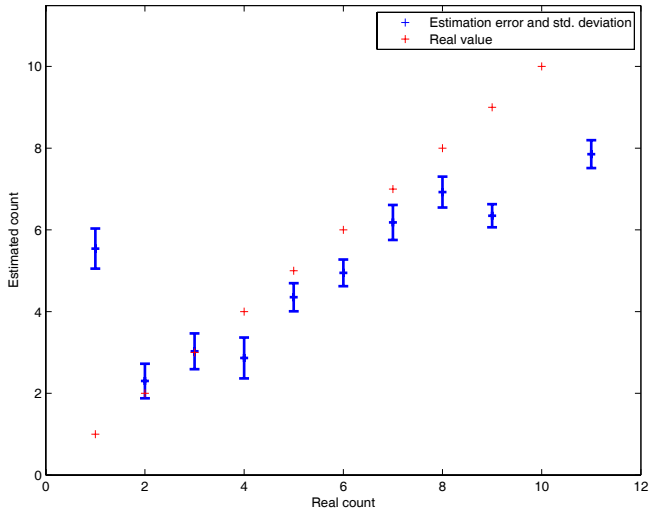


FIGURE 4.7. Pedestrians counted by PDM versus the real value using employing thermal images.

5. CONCLUSION AND FUTURE RESEARCH

An algorithm for pedestrian counting was presented based on the analysis of foreground areas and the estimation of density coefficients. The approach yields good count estimates despite challenging illumination and crowdedness conditions. The count estimates are more accurate than those obtained with the VJ approach using a conventional camera. Minor improvements are possible if an infrared camera is employed instead. Results show that also an omnidirectional camera can be employed with similar levels of accuracy. Although in principle this camera can capture panoramic 360° view of the area surrounding the bus stop, the practical advantage is unclear due to the loss of resolution which must be sacrificed in exchange for a larger field of view.

Our experiments demonstrated that full-body and body-parts identification using the VJ method is a much more difficult task because people may be wearing clothes having colors similar to the background or stand in positions that differ from the set of training poses. This lack of invariance, especially in real outdoor settings, limits significantly the performance of the VJ classifier and motivates the development of approaches that incorporate ways to remove the background and focus-of-attention mechanisms.

The PDM is more robust to occlusions and pose changes because it does not attempt to find body parts. The results presented in the previous section are quite encouraging considering that the final goal of the algorithm is to obtain the total number of pedestrians waiting at bus stops rather than identifying each person individually.

Ongoing research aimed at improving the robustness and accuracy of the current detection system considers the use of body-parts detection and tracking based on histograms of oriented gradients as proposed in (Wu & Nevatia, 2007). In order to improve the background removal process the use of texture and stereo disparity analysis is also being investigated.

REFERENCES

Beyer, W. (1987). *CRC standard mathematical tables* (28 ed.). Boca Raton, FL: CRC Press.

Bi, L., Tsimhoni, O., & Liu, Y. (2009, March). Using image-based metrics to model pedestrian detection performance with night-vision systems. *IEEE Transactions on Intelligent Transportation Systems*, *10*(1), 155-164.

Bu, F., & Chan, C.-Y. (2005, June). Pedestrian detection in transit bus application: sensing technologies and safety solutions. In *Proceedings of the IEEE Intelligent Vehicles Symposium* (p. 100-105).

Cheng, J., Yang, J., Zhou, Y., & Cui, Y. (2006). Flexible background mixture models for foreground segmentation. *Image and Vision Computing*, *24*(5), 473 - 482.

Cortés, C., Sáez, D., Sáez, E., Núñez, A., & Tirachini, A. (2007, JUN). Hybrid predictive control strategy for a public transport system with uncertain demand. In *Proceedings of Sixth Triennial Symposium on Transportation Analysis (TRISTAN VI).* Phuket Island, Thailand.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. In (2 ed., p. 44-47). New York, NY: John Wiley & Sons.

Elgammal, A. M., Harwood, D., & Davis, L. S. (2000). Non-parametric model for background subtraction. In *Proceedings of the 6th European Conference on Computer Vision-Part II* (pp. 751–767). London, UK: Springer-Verlag.

Felzenszwalb, P. (2001). Learning models for object recognition. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 1, p. I-1056-I-1062 vol.1).

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference Machine Learning* (pp. 148–156).

Friedman, N., & Russell, S. (1997). Image segmentation in video sequences: A probabilistic approach. In *Proceedings Thirteenth Conference on Uncertainty in Artificial Intelligence* (pp. 175–181).

Fuerstenberg, K., Dietmayer, K., & Willhoeft, V. (2002, June). Pedestrian recognition in urban traffic using a vehicle based multilayer laserscanner. In *Proceedings of the IEEE Intelligent Vehicle Symposium* (Vol. 1, p. 31-35 vol.1).

Gandhi, T., & Trivedi, M. (2006, Sept.). Pedestrian collision avoidance systems: a survey of computer vision based recent studies. In *Proceedings of the IEEE Intelligent Transportation Systems Conference* (p. 976-981).

Gavrila, D. (2000). Pedestrian detection from a moving vehicle. In *Proceedings of the 6th European Conference on Computer Vision-Part II* (pp. 37–49). London, UK: Springer-Verlag.

Gavrila, D. (2007, Aug.). A bayesian, exemplar-based approach to hierarchical shape matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(8), 1408-1421.

Gavrila, D., & Giebel, J. (2002, June). Shape-based pedestrian detection and tracking. In *Proceedings of the IEEE Intelligent Vehicle Symposium* (Vol. 1, p. 8-14 vol.1).

Haritaoglu, I., Harwood, D., & Davis, L. (2000, Aug). W4: real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 809-830.

Hartley, R., & Zisserman, A. (2004). *Multiple view geometry in computer vision*. Cambridge University Press.

Heikkilä, J., & Silvén, O. (2004, July). A real-time system for monitoring of cyclists and pedestrians. *Image and Vision Computing*, *22*(7), 563–570.

Mikolajczyk, K., Schmid, C., & Zisserman, A. (2004). Human detection based on a probabilistic assembly of robust part detectors. In *Proceedings of the 8th European Conference on Computer Vision, 2004* (pp. 69–82).

Mittal, A., & Paragios, N. (2004, June-2 July). Motion-based background subtraction using adaptive kernel density estimation. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2, p. II-302-II-309 Vol.2).

Mohan, A., Papageorgiou, C., & Poggio, T. (2001, Apr). Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(4), 349-361.

Noriega, P., & Bernier, O. (2006). Real time illumination invariant background subtraction using local kernel histograms. In *British Machine Vision Conference* (p. III:979).

Núñez, A., Cortés, C., Sáez, D., & Riquelme, M. (2008, May). Hybrid predictive control for real-time optimization of public transport systems operations based on evolutionary multiobjective optimization. In *10th International Conference on Applications of Advanced Technologies in Transportation.* Athens, Greece.

Oliver, N., Rosario, B., & Pentland, A. (2000, Aug). A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 831-843.

Panahi, S., Sheikhi, S., Hadadan, S., & Gheissari, N. (2008, Dec.). Evaluation of background subtraction methods. In *Digital Image Computing: Techniques and Applications DICTA '08* (p. 357-364).

Papageorgiou, C., & Poggio, T. (1999). Trainable pedestrian detection. In *Proceedings of the International Conference on Image Processing ICIP 99* (Vol. 4, p. 35-39 vol.4).

Piccardi, M. (2004, Oct.). Background subtraction techniques: A review. In *IEEE International Conference on Systems, Man and Cybernetics* (Vol. 4, p. 3099-3104 vol.4).

Polivka, M., Svanda, M., Hudec, P., & Zvanovec, S. (2009, May). UHF RF identification of people in indoor and open areas. *IEEE Transactions on Microwave Theory and Techniques*, *57*(5), 1341-1347.

Rosin, P. L. (1998). Thresholding for change detection. In *Proceedings of the Sixth International Conference on Computer Vision ICCV 98* (p. 274). Washington, DC, USA: IEEE Computer Society.

Rymel, J., Renno, J., Greenhill, D., Orwell, J., & Jones, G. (2004, Oct.). Adaptive eigen-backgrounds for object detection. In *International Conference on Image Processing ICIP 04* (Vol. 3, p. 1847-1850 Vol. 3).

Stauffer, C., & Grimson, W. (1999). Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2, p. 252 Vol. 2).

Sun, J., Zhang, W., Tang, X., & Shum, H.-Y. (2006). Background cut. *European Conference on Computer Vision, ECCV 06*, 628–641.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001* (Vol. 1, p. I-511-I-518 vol.1).

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

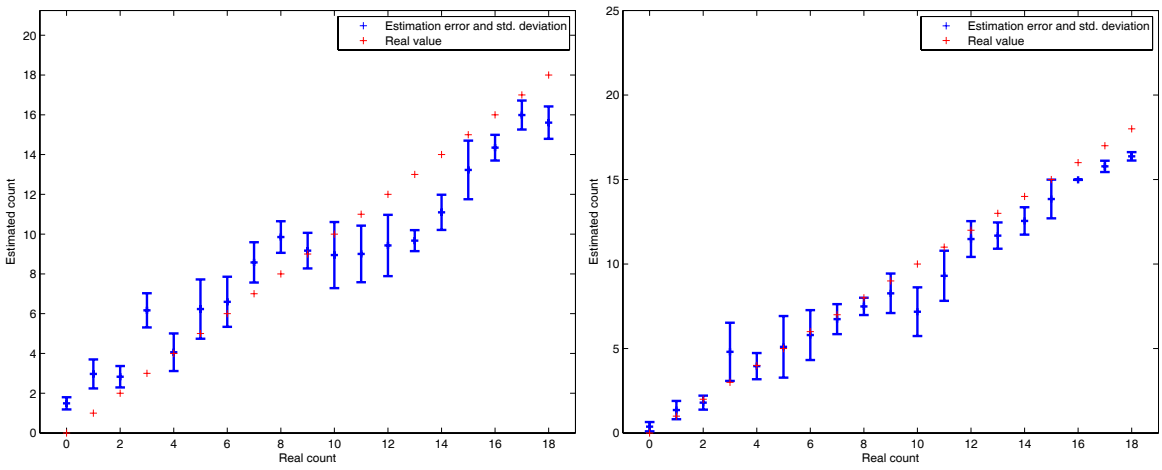
Wasserman, P. D. (1993). Advanced methods in neural computing. In (p. 35-55). New York, NY, USA: John Wiley & Sons, Inc.

Wren, C., Azarbayejani, A., Darrell, T., & Pentland, A. (1996, Oct). Pfinder: real-time tracking of the human body. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition* (p. 51-56).

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), 247-266.

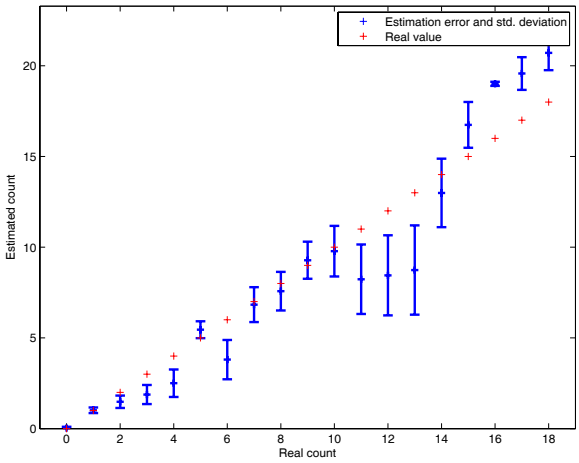
Zhang, J., & Zhuang, Y. (2007, July). Adaptive weight selection for incremental eigen-background modeling. In *IEEE International Conference on Multimedia and Expo*. (p. 851-854).

APPENDIX A. ADDITIONAL RESULTS

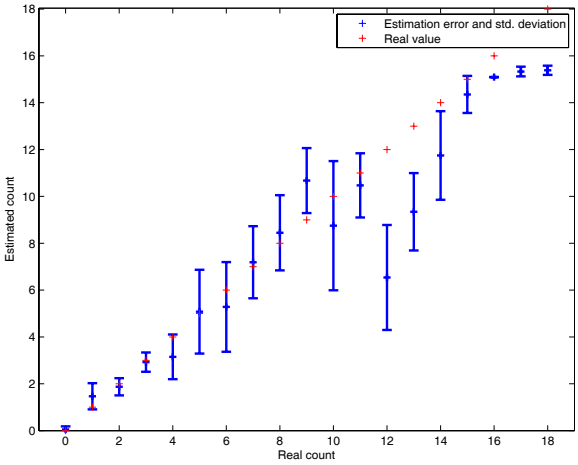


(A) LRM

(B) LDA

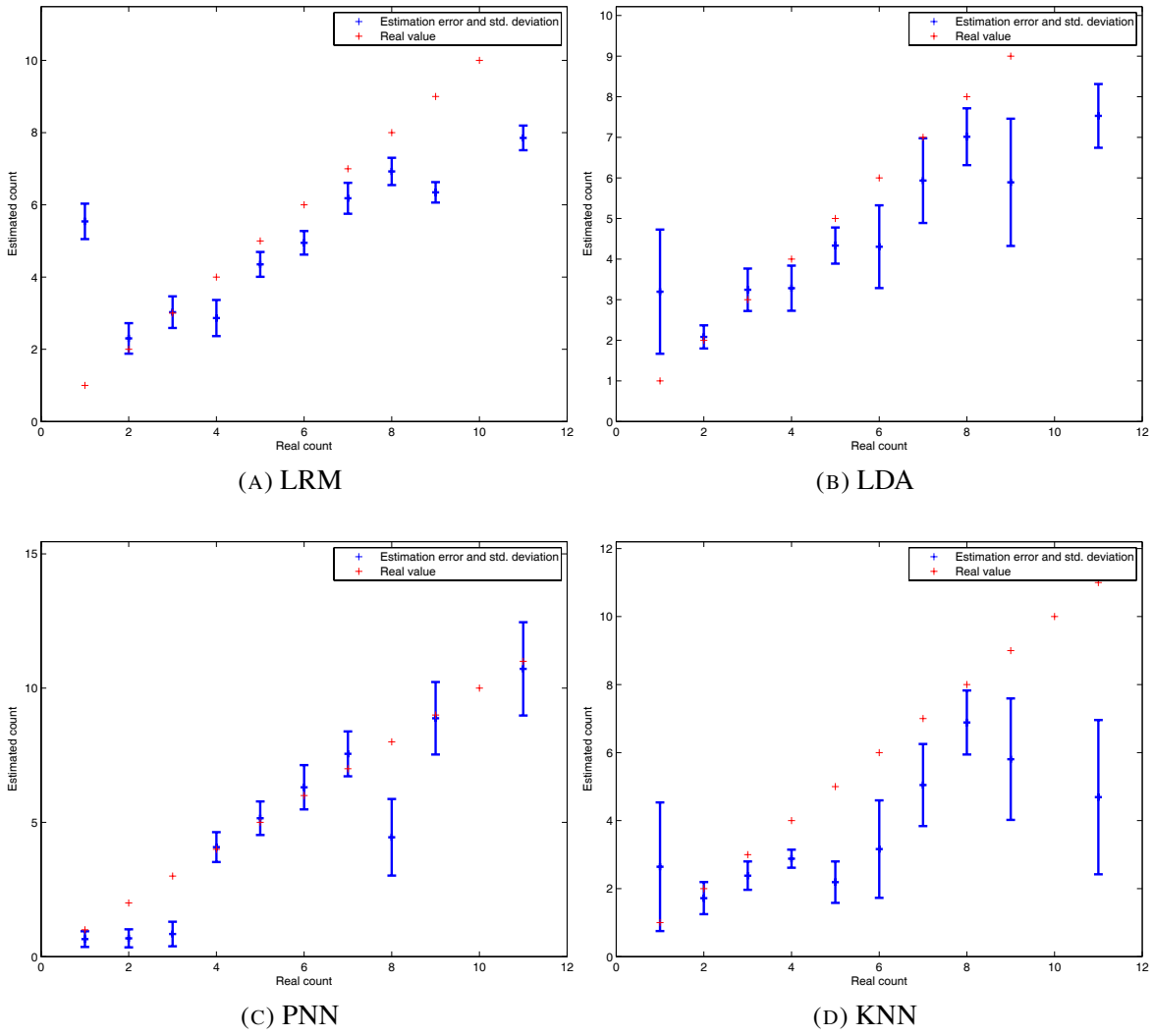


(C) PNN



(D) KNN

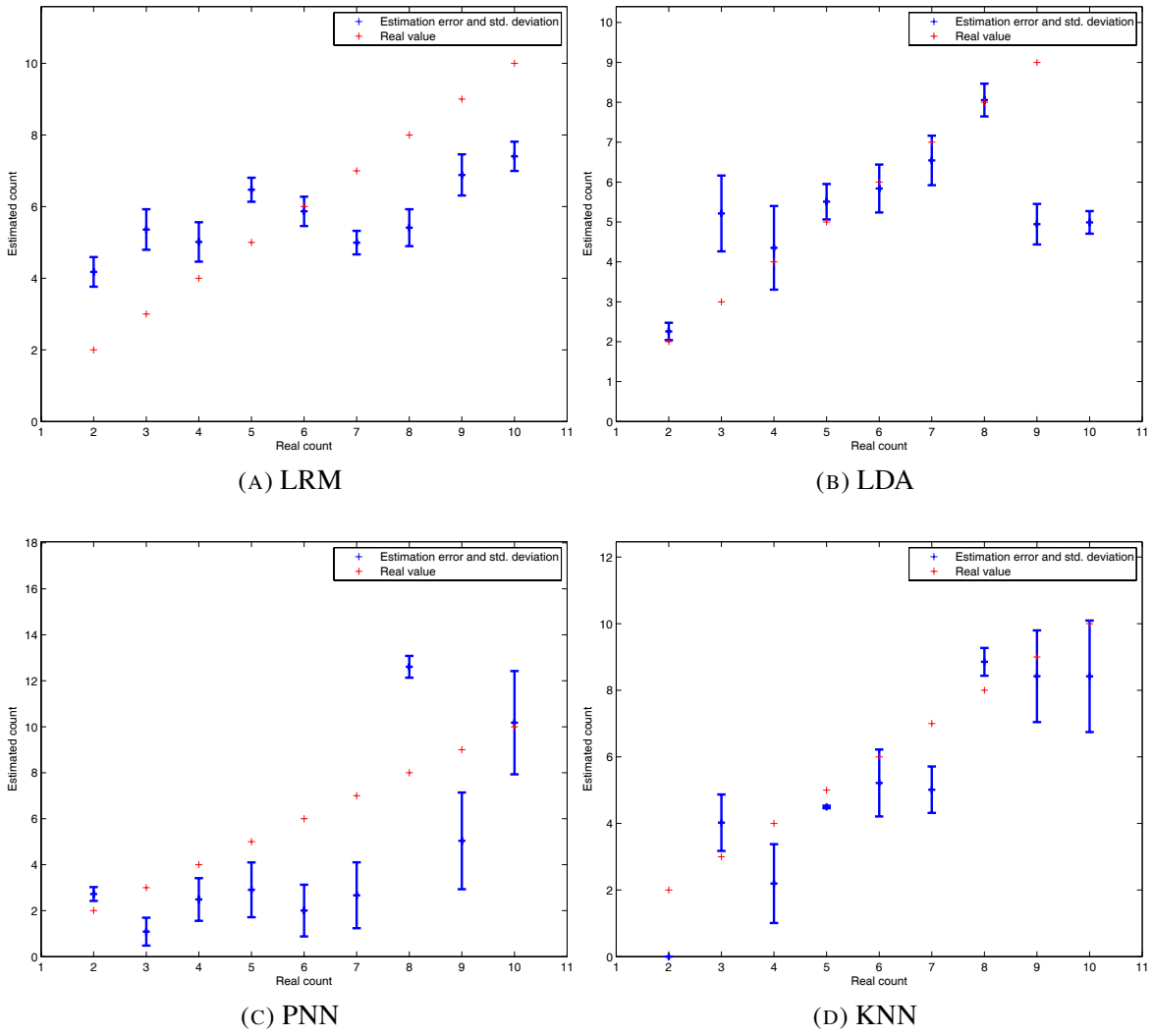
FIGURE A.1. Pedestrians counted by PDM with different classification approaches versus the real value using conventional perspective camera.



(A) LRM (B) LDA

(C) PNN (D) KNN

FIGURE A.2. Pedestrians counted by PDM with different classification approaches versus the real value using thermal camerea.



(A) LRM (B) LDA

(C) PNN (D) KNN

FIGURE A.3. Pedestrians counted by PDM with different classification approaches versus the real value using omnidirectional camerea.