Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

A Density-Based Approach for Effective Pedestrian Counting at Bus Stops

Guillermo García-Bunster, Miguel Torres-Torriti
Dept. of Electrical Engineering
Pontificia Universidad Catolica de Chile
Santiago, Chile
grgarcia@uc.cl, mtorrest@ing.puc.cl

***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 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. A linear model, which is employed to correct the densities due to perspective scaling for people far from the camera position, yields the final people count estimates. The approach is compared to the well-know 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. The proposed method is general and can be employed to count people in other public spaces, such as buildings.**

***Index Terms*—pedestrian detection, pedestrian counting, background subtraction, Haar-features, density-based demand estimation.**

I. Introduction

Modern public transportation systems requires accurate real-time information of route conditions and demand for optimal fleet scheduling and control [3], [9]. Existing technologies involving vehicle localization and telecommunications allow to monitor traffic flow in real-time, nonetheless measuring the demand in real-time is a challenging problem for which there is still no adequate solution. Traditionally demand information has been obtained using statistical models built off-line from manually collected data, therefore, developing automated methods to reliably count passengers at bus stops and within the buses is necessary. This paper proposes an approach based on computer vision techniques to count passengers waiting at regular bus stops without passageway turnstiles. Usually pedestrians stand at different places around the bus stop and do not wait in line (see fig. 1), consequently most of the conventional people counters based on pressure mats, infrared detectors or ultrasonic sensors [2] cannot be applied. On the other hand, the ever more popular surveillance video cameras provide an economical and minimally invasive solution to acquire images that enclose a large area around the bus stop and that can be programmed to automatically detect and count people. One of the main challenges to the solutions based on image processing is the development of algorithms capable of adequately handling the varying conditions proper of bus stops, such as illumination changes, shadows, occlusions and crowdedness, among other situations. The first attempts to detect pedestrians in non-static images considered the human body as a whole and employed shape models or edge templates [10], [5]and [4]. These approaches have proved quite useful for detecting partially invariant objects, such as faces and cars. However, detection of the whole body using templates often yields poor results in practice due to the many different poses pedestrians can assume. In addition to the lack of invariance, occlusions and clothing colors also difficult the extraction of appropriate edges. This fact is illustrated in figs. 2 and 3, which show the edges corresponding to fig. 1 extracted using a LoG filter. The edges in fig. 2 are obtained without any pre-processing of the original image, while edges in fig. 3 result after applying the LoG filter to a color-reduced version of fig. 1. Figure 3 shows that the main contours can be detected more effectively by reducing the number of colors and building larger uniformly colored areas. However, due to the evident occlusions and lack of complete contours, it should be clear that applying any of the existing template matching methods to figs. 2 or 3 would yield a low number of correct detections.

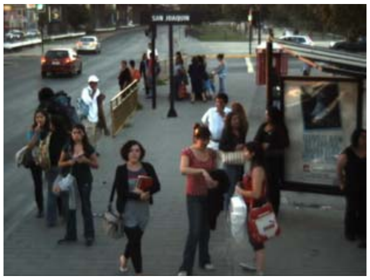
Fig. 1. People standing at a bus stop.

Viola and Jones [13] proposed a different framework for object detection based on the selection of weak Haar features to build a cascade detector. Their approach does not require solving a template matching problem and has been widely used in several applications involving still objects with rates of detection of about 95% and false positives rates below 1% when applied to face detection [13], [14]. However, their original approach also 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 fea-

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

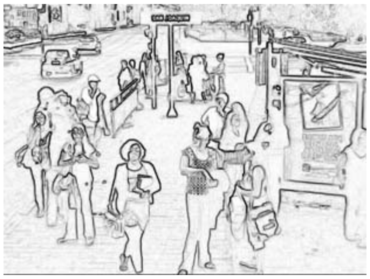

Fig. 2. Edges extracted from figure 1 using a LoG filter.

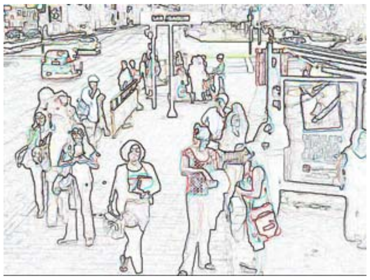

Fig. 3. Edges extracted from a color-reduced version of figure 1.

tures from motion patterns [14]. Other researchers have instead 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 [8], [7], [1]. Each of these part detectors are trained using the Viola-Jones framework or a Support Vector Machine. These authors conclude that detecting humans by their components allows finding partially occluded pedestrians and increase detection rates. For a recent and exhaustive review of pedestrian detection techniques see [6].

This paper proposes an approach to people counting from 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. The only two basic assumptions required are i) that the background can be extracted 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 II. 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 and the reference method based on the Viola-Jones detector are implemented as two-stage processes (see fig. 4). 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.

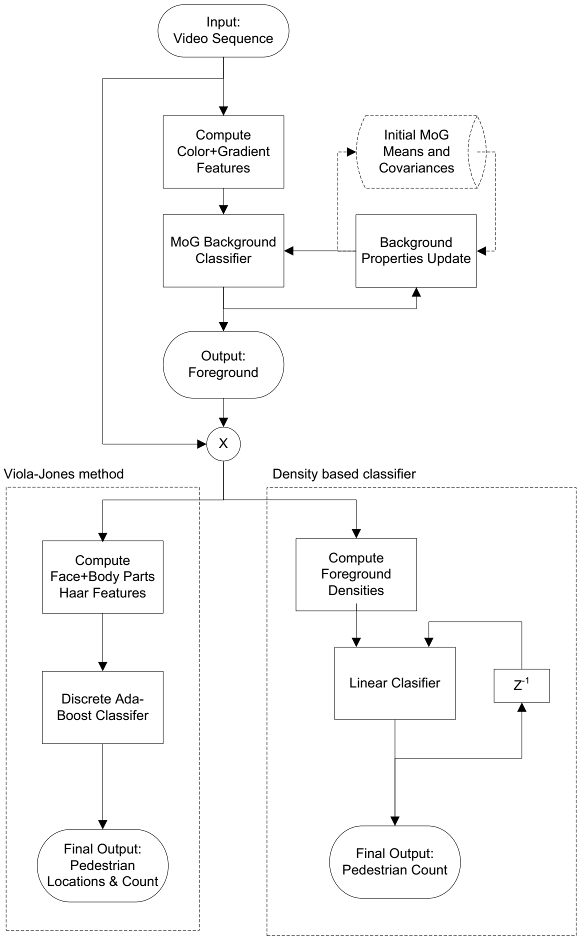

Fig. 4. Detection scheme.

The contribution of this paper can be summarized in a simple method for pedestrian counting at bus stops or any open environment based on density computations. The paper shows this approach is more accurate than the method based on a combined action of a background subtraction process and the Viola-Jones approach. Both approaches are tested using image sequences from a real bus stop.

The paper is organized as follows. The background identification approach is explained in the next section. Section three describes the pedestrian count using density estimates. Section four presents the testing methodology and results. The conclusion and ongoing research is discussed in section five.

II. 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. The reduced search does not only saves valuable processing time, but also allows to increase detection certainty because it is 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 and gradient feature space. The feature vector is defined as a five component vector $f = [r, g, b, g_o, g_m]^T$ containing the red, green and blue color values, gradient orientation and gradient magnitude at a given pixel. The gradient magnitude and orientation at pixel with coordinates (x, y) is calculated as

$$\begin{aligned} g_m &= \sqrt{g_x^2 + g_y^2} \\ g_o &= \arctan\left(g_y, g_x\right) \end{aligned} \qquad (1)$$

where

$$\begin{aligned} g_x &= I_{x+1,y} - I_{x-1,y} \\ g_y &= I_{x,y+1} - I_{x,y-1} \end{aligned} \qquad (2)$$

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

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 [11]. 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:

$$B_k = (1 - \alpha)B_{k-1} + \alpha I_k \qquad (3)$$

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 background colors can have multimodal distributions over time due to variant lighting conditions.

An approach that copes with multimodal distributions is the method by Stauffer and Grimson [12], which models each pixel's feature vector f as a mixture of L Gaussians with mean and covariance parameters $\mu_k^l \in \mathbb{R}^n$, $\Sigma_k^l \in \mathbb{R}^{n\times n}$, $l = 1, 2, \ldots, L$ at time k. More specifically, the probability of observing a pixel feature f_k is given by:

$$P(f_k) = \sum_{l=1}^{L} \omega^l P(f_k|\mu^l, \Sigma^l) \qquad (4)$$

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

$$P(f_k|\mu^l, \sigma^l) = \frac{1}{(2\pi)^{n/2}|\Sigma^l|^{\frac{1}{2}}} e^{-\frac{1}{2}(f_k-\mu^l)^T {\Sigma^l}^{-1}(f_k-\mu^l)} \qquad (5)$$

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 $L = 5$. In [12] B given L is calculated as:

$$B = \arg\min_k \left(\sum_{l=1}^{k} \omega^l > \delta_b \right) \qquad (6)$$

where each Gaussian is sorted in decreasing order according to $\omega_k^l / \left\|\Sigma_k^l\right\|$, 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 $\sqrt{(f_k - \mu_k^l)^T {\Sigma_k^l}^{-1}(f_k - \mu_k^l)} < \lambda$ for some l, where λ represents the number of standard deviations from the mean that defines the matching threshold ($\lambda = 2.5$ in [12]). 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 MMoG probability density function parameters are updated by means of the following incremental form of the Expectation Maximization algorithm:

$$\mu_k^l = \frac{S_{k-1}}{S_k}\mu_{k-1}^l + \frac{1}{S_k}\tilde{\omega}^l \qquad (7)$$

$$\Sigma_k^l = \frac{S_{k-1}}{S_k}\Sigma_{k-1}^l + \frac{1}{S_k}\tilde{\omega}^l \left(I_k - \mu_k^l\right)\left(I_k - \mu_k^l\right)^T \qquad (8)$$

$$\omega_k^l = \frac{1}{k}S_{k-1} + \frac{1}{k}\tilde{\omega}^l \qquad (9)$$

where

$$S_k = S_{k-1} + \tilde{\omega}^l \tag{10}$$

$$\tilde{\omega}^l = \frac{P\left(I_k \left|\mu^l, \sigma^l\right.\right)\omega^l}{\sum_{c=1}^{L} P\left(I_k \left|\mu^c, \sigma^c\right.\right)\omega^c} \tag{11}$$

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.

III. Pedestrian Count Using Density Estimates

In order to estimate the total amount of pedestrians, the problem of counting people is solved by counting foreground pixels and applying a linear regression model to the pixel count. The linear regression model is employed to account for scaling changes due to perspective, which make pedestrians far from the camera to appear smaller than closer ones. More specifically, the regression over the foreground pixels is implemented by defining N evenly distributed horizontal sections and assigning to each section a people density index $a_i,\ i = 1, 2, \ldots, N$. Moreover, since the total amount of pedestrians shows low variance over short periods of time, the model includes terms that depend on pedestrian count estimates $\hat{y}(k-2)$ and $\hat{y}(k-1)$ corresponding to sampling periods $k-2$ and $k-1$, which allow to filter out sudden variations due to noisy measurements from the estimate $\hat{y}(k)$ at current instant k. If the density indeces are treated as the number of people per area (measured in pixels), then a simple linear model for the total number of people can be stated as:

$$\hat{y}(k) = \sum_{i=1}^{N} a_i x_i + a_0 + b_1 \hat{y}(k-1) + b_2 \hat{y}(k-2) \tag{12}$$

where x_i are the total number of pixels labeled as people in region i, for $i = 1, 2, \ldots, N$, and $\hat{y}(k)$ is the estimate of people in the image at instant $k \in \mathbb{N}$. The coefficients a_i, b_1 and b_2 are obtained by solving a least squares optimization problem to minimize the overall mean square error between the estimated and the actual number of people. Practical experiments have showed that a good trade-off between computation simplicity and estimation accuracy is achieved when employing five regions, i.e. $N = 5$.

IV. Testing Methodology and Experimental Results

The algorithms developed were tested using data from 28 video sequences acquired at 15 fps between 17 and 18 pm on March 26^{th}, 2008. At this time of the year the sun's location is close to the horizon producing challenging conditions of lights and shadows, as well as increased reflections on surrounding structures and passing cars. The camera employed is a standard FireWire® IEEE1394 camera with a 1024x768 pixels 1/2" CCD and a Tamron varifocal lens with focal distances in the range 6-12 mm, corresponding to an angle of view in the range $30.4° \times 23.1°$ (telephoto) – $58.7° \times 44.4°$ (wide). The camera was located at $3\,m$ from the bus stop at a height of $3.2\,m$ above the ground, as shown in fig. 5. This configuration allows to cover an area in which pedestrians would normally stand (see fig. 1).

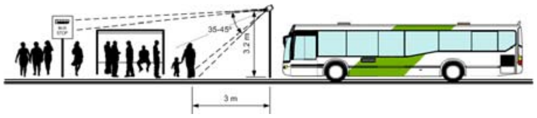

Fig. 5. Bus stop and camera configuration.

Figure 6 shows the results corresponding to frames 100, 700, 1500, 1900, 2600 and 3400. In these frames white pixels correspond to background. The background identification scheme responds rapidly to lighting variations and updates the corresponding class parameters according to the incoming background features. Successful detections of people using the VJ approach are indicated by boxes surrounding the face or the body whenever the upper/lower extremities are detected. As may be seen in the fifth frame of fig. 6, several false detections occur when there is a crowded scene. Most of the time the VJ approach failed 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 classifiers, 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.

The performance statistics of the PDM and VJ methods are summarized in Table I, which presents the root mean-squared errors, percentage of correctly counted people (CCP) percentages, false positives averages ($\overline{FP}$) and false negative averages ($\overline{FN}$). In terms of the root-mean-squared error $RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \|y(k) - \hat{y}(k)\|^2}$, where $y(k)$ is the real number of pedestrians in frame k, it is clear from Table I that the PDM performs better than the VJ-based approach with RMSE values of 2.91 against 4.11 pedestrians, respectively. This better performance of PDM may also be appreciated from fig. 7, which illustrates the evolution over time of the pedestrian count estimates and the real value.

The estimated pedestrian count versus the real amount of pedestrians is shown in figs 8 and 9. These results show that the PDM is more accurate and precise, since the average standard deviation of the VJ method is 3.02 pedestrians, while this figure is only 1.85 pedestrians per frame on average for the proposed solution. On the other hand, the VJ approach shows less false positives over time than the PDM, but the latter yields a smaller false negatives average.

It is to be noted that 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

TABLE I

Results using Viola-Jones Detector and PDM (CCP is the percentage of frames in which people was correctly counted and $\bar{FP}$,$\bar{FN}$ are the false positive, false negative averages).

	PDM	VJ
RMSE	2.91	4.11
CCP	13.87%	11.51%
$\overline{FP}$	1.14	0.45
$\overline{FN}$	1.13	1.91

often consider sample images of pedestrians in scenes that are not very crowded or taken under controlled lighting conditions.

V. Conclusion and Future Work

An algorithm for pedestrian counting was presented. 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 and could be supplied to an automated public transportation management system to dispatch buses according to demand measured on-line.

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 focus-of-attention mechanisms and ways to remove the background.

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. In terms of under and over estimation, the VJ approach tends to underestimate while the proposed approach overestimates the total amount of pedestrians. This difference could be adjusted by penalizing the false positive averages ($\overline{FP}$) and false negative averages ($\overline{FN}$) in the objective function when fitting the parameters of the model, or by adding constraints in terms of $\overline{FP}$ and $\overline{FN}$.

Ongoing research aimed at improving the robustness and accuracy of the proposed approach considers the use of color-based segmentation for skin detection, as well as using body-parts detectors based on gradient distribution as suggested in [1]. The use of texture is also being investigated in order to improve the background identification process.

Acknowledgment

This project has been supported by the National Commission for Science and Technology Research of Chile (Conicyt) under PBCT Grant ACT-32 and Doctoral Scholarship #21090189.

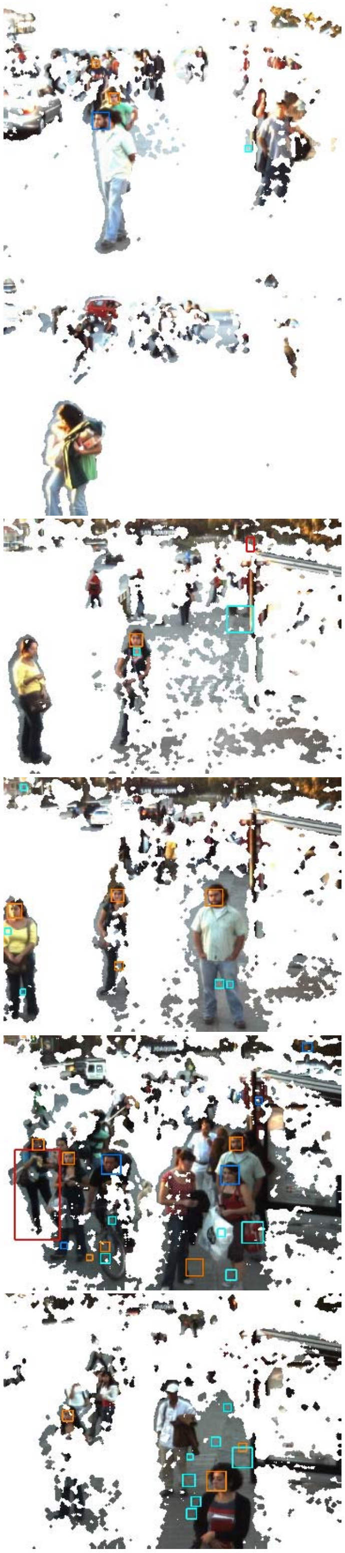

Fig. 6. Pedestrian detection results for frames 100, 700, 1500, 1900, 2600 and 3400 (from top to bottom).

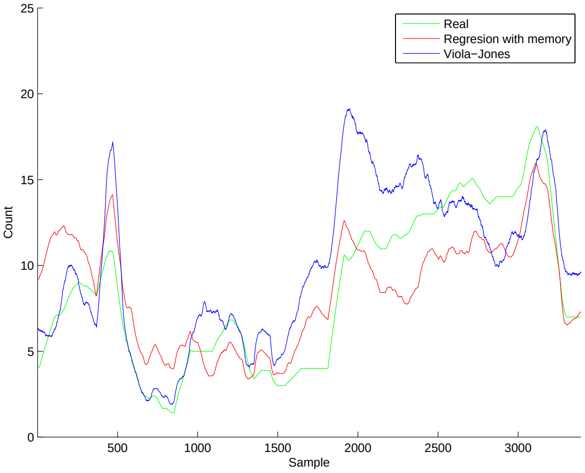

Fig. 7. Pedestrian estimation over time

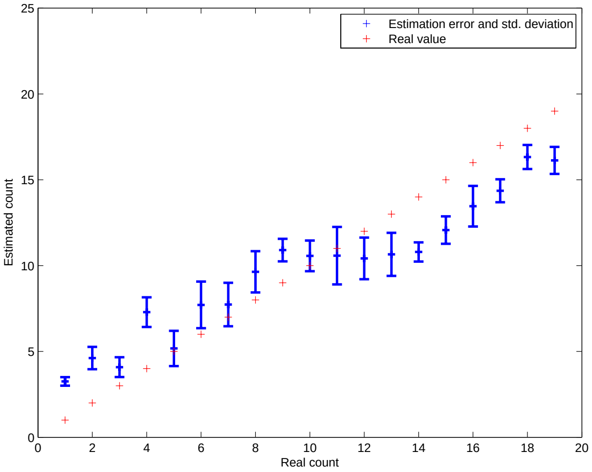

Fig. 8. Pedestrians counted by PDM versus the real number of pedestrians.

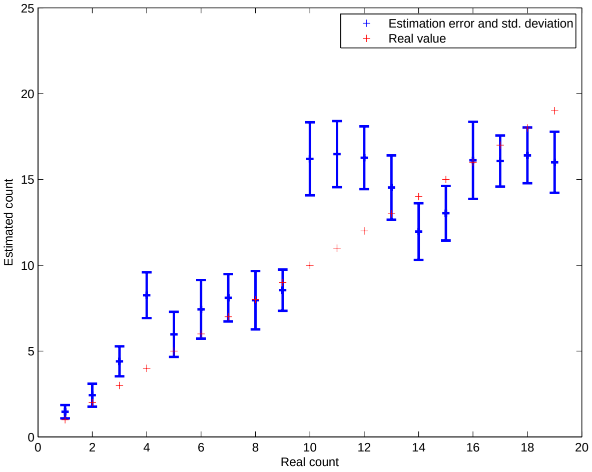

Fig. 9. Pedestrians counted by Viola-Jones approach versus the real number of pedestrians.

References

[1] R. N. B. Wu. 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, 2007.

[2] F. Bu and C.-Y. Chan. Pedestrian detection in transit bus application: sensing technologies and safety solutions. In *Intelligent Vehicles Symposium, 2005. Proceedings. IEEE*, pages 100–105, June 2005.

[3] C. Cortés, D. Sáez, E. Sáez, A. Núñez, and A. Tirachini. 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*, June 2007.

[4] P. F. Felzenszwalb. Learning models for object recognition. In *In CVPR*, pages 56–62, 2001.

[5] D. Gavrila. Pedestrian detection from a moving vehicle. In *ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part II*, pages 37–49, London, UK, 2000. Springer-Verlag.

[6] Z. Li, K. Wang, L. Li, and F.-Y. Wang. A review on vision-based pedestrian detection for intelligent vehicles. In *Vehicular Electronics and Safety, 2006. ICVES 2006. IEEE International Conference on*, pages 57–62, Dec. 2006.

[7] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV*, volume I, pages 69–81, 2004.

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

[9] A. Núñez, C. Cortés, D. Sáez, and M. Riquelme. 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*, May 2008.

[10] C. Papageorgiou and T. Poggio. Trainable pedestrian detection. In *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, volume 4, pages 35–39 vol.4, 1999.

[11] M. Piccardi. Background subtraction techniques: a review. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 4, pages 3099–3104 vol.4, Oct. 2004.

[12] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2, page 252 Vol. 2, 1999.

[13] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511–I–518 vol.1, 2001.

[14] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, July 2005.