People Counting and Human Detection in a Challenging Situation

Ya-Li Hou, *Student Member, IEEE*, and Grantham K. H. Pang, *Senior Member, IEEE*

***Abstract*—Reliable people counting and human detection is an important problem in visual surveillance. In recent years, the field has seen many advances, but the solutions have restrictions: people must be moving, the background must be simple, and the image resolution must be high. This paper aims to develop an effective method for estimating the number of people and locate each individual in a low resolution image with complicated scenes. The contribution of this paper is threefold. First, postprocessing steps are performed on background subtraction results to estimate the number of people in a complicated scene, which includes people who are moving only slightly. Second, an Expectation Maximization (EM)-based method has been developed to locate individuals in a low resolution scene. In this method, a new cluster model is used to represent each person in the scene. The method does not require a very accurate foreground contour. Third, the number of people is used as *a priori* for locating individuals based on feature points. Hence, the methods for estimating the number of people and for locating individuals are connected. The developed methods have been validated based on a 4-hour video, with the number of people in the scene ranging from 36 to 222. The best result for estimating the number of people has an average error of 10% over 51 test cases. Based on the estimated number of people, some results of the EM-based method have also been shown.**

***Index Terms*—Expectation-maximum, human detection, neural network, people counting.**

I. Introduction

PEOPLE counting is a crucial and challenging problem in visual surveillance. An accurate and real-time estimation of people in a shopping mall can provide valuable information for managers. Automatic monitoring of the number of people in public areas is also important for safety control and urban planning.

In recent years, this field has seen many advances, but the solutions have restrictions: people must be moving, the background must be simple, or the image resolution must be high. However, real scenes always include both moving and stationary human beings, the background may be complicated, and most videos in a visual surveillance system have a relatively low resolution.

This paper aims to develop an effective method for estimating the number of people in a complicated outdoor scene, as shown in Fig. 1(a). A simple individual detection method based on the estimated result has also been introduced for subsequent video processing. The resolution of the target video is $640 * 480$ pixels. In this scenario, occlusions exist everywhere with people walking, sitting, and standing. The video was taken by a static camera overlooking a large area. This scene is very common, but few results have ever been demonstrated for such events.

(a) (b)

Fig. 1. (a) Typical scene to be processed. (b) Binary foreground image after background subtraction for (a).

The paper is organized as follows. Related work is reviewed in Section II. The methods for people counting and human detection are introduced in detail in Sections III and IV presents some evaluation results of our method on the target event.

II. Related Work

Usually, the methods for people counting can be classified into two categories: detection-based methods and map-based methods. Detection-based methods determine the number of people by identifying individuals in the scene. These methods determine the number of people and their locations simultaneously. Map-based methods exploit the relationship between the number of people and some features from the image. They can only count the number of people in a scene.

The detection-based methods can be further classified into two groups. Some methods try to segment the foreground blobs into individuals based on prior knowledge of human shapes and the characteristics of the foreground contour. The other methods detect individuals directly from the image.

The work by Zhao [1], [2] and Rittscher *et al.* [3] are two good examples of the first group. In [1], Zhao and Nevatia established an ellipsoid to describe the 3-D human shape. They detected heads by checking local vertical peaks on the foreground contour. The detected persons were then removed from the foreground blob, and peaks in the remaining foreground were checked. In their later work [2], a more accurate 3-D model composed of three ellipsoids was used. To deal with the occlusion problem, a joint probability for multiple humans

Manuscript received January 21, 2009; revised May 20, 2009 and September 29, 2009; accepted January 7, 2010. Date of publication September 20, 2010; date of current version November 10, 2010. This paper was recommended by Associate Editor M. Celenk.

The authors are with the Industrial Automation Research Laboratory, Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: ylhou@eee.hku.hk; gpang@eee.hku.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2010.2064299

has been considered. Finally, the human detection and tracking problem was formulated as a Maximum *A Posteriori* (MAP) problem simultaneously. A sophisticated sampling algorithm, Data Driven Markov Chain Monte Carlo, was employed to find the best configuration for the MAP problem. Some positive results for a crowd of a dozen people were obtained. To reduce the dependence on an accurate foreground contour, which may be easily corrupted by noise, Rittscher *et al.* [3] extracted some feature points from the contour. These features were annotated as top, left, right, and bottom based on local contour information. A variant of Expectation Maximum (EM) was developed to group these features into some human-sized rectangles. These methods work well even under low resolutions. However, these methods rely on an accurate foreground contour. When people are almost stationary, getting a good foreground is difficult.

Recently, some significant results have also been obtained for the methods detecting individuals directly from images [4]–[8]. Various features from a static image have been attempted: Haar wavelets [4], SIFT-like features [5], Histogram of Oriented Gradients [6], and contours [7], [8]. These methods may achieve more accurate counting and detection results when the crowd is small. However, most of them are time consuming and only show results for a small crowd with few occlusions. They also require high resolution images. In [4]–[6], detection results on the MIT or INRIA data set are shown. In this data set, human samples are images of $64 * 128$ pixels with few occlusions. Wu and Nevatia [7] presented their results on the CAVIAR data set, which includes more occlusions. However, the method requires the size of a human to be at least $24 * 58$ pixels. Worth mentioning is that back in 2001, Lin *et al.* [8] tried to estimate the number of people in a large crowd by only detecting human heads in the image. They achieved results for a crowd of 120 people in the model world and more than 1000 people in the real world. However, due to the difficulty of getting the ground truth for the real scene, no quantitative detection results were reported. The smallest detectable head size in the paper was $16 * 16$ pixels. Furthermore, motion features have proven effective for detecting individuals under various situations, particularly in [9], [10]. However, in most events, people just stand or sit, showing occasional articulated movements, which is difficult to use for individual detection.

In the map-based methods, edge or foreground pixels [11]–[15] and textures [16]–[21] inside the foreground are used to estimate the crowd density or the number of people.

Davies *et al.* [11] maintained that there is a linear relationship between foreground pixels and the number of people in situations with trivial perspective distortions and occlusions. Ma *et al.* [13] and Çelik *et al.* [14] investigated different strategies for perspective correction to establish the linear relationship under situations with serious perspective distortions. Using an accurate camera calibration, Kilambi *et al.* [15] transformed the foreground region to the projected area of the crowd on the ground plane in world coordinates. The average projected area for one person was learned at the initial training stage. Hence, the number of people was estimated. An accuracy of over 75% was reported for a group consisting of up to 10 people. The requirement for an accurate camera calibration limits the application of this method. In summary, methods based on foreground pixels can be easily implemented and developed for real-time applications. However, up to now, only moving people have been considered in this category.

Marana *et al.* [16] proposed a method for estimating crowd density with the Grey Level Dependency Matrix (GLDM). The GLDM is a texture measurement based on pixel distribution. In their later work [17], they obtained the Minkowski Fractal Dimension (MFD) by performing dilation operations on edge images using different sizes of structuring elements. Rahmalan *et al.* [18] used a new texture descriptor, Translation Invariant Orthonormal Chebyshev Moments (TIOCM), and compared its performance on a graduation scene with the GLDM and MFD. The results show that TIOCM outperforms the MFD and costs less time than the GLDM. Recently, Li *et al.* [19] transformed an image into multi-scale formats with wavelet transform and used the Support Vector Machine to classify them into different crowd density levels. They claimed that this method performs better than previous methods. However, the above methods [16]–[19] only provide a crowd density estimation. When textures from the background and human clothing are complicated, these methods may not work well.

Instead of using the total number of edges or foreground pixels, Kong *et al.* [20] used the edge orientation histogram and the foreground blob size histogram. They also considered feature normalization such that the method is invariant to different viewing angles. In a recent paper by Chan *et al.* [21], Gaussian Process Regression was adopted to ascertain the relationship between 28 different features and the number of people. To get more accurate results, the crowd was segmented into two components based on their moving directions before estimation.

In this paper, some methods are developed for a very complicated situation shown in Fig. 1(a). A method based on the neural network is used to estimate the number of people in Section III-A. To detect each individual person for subsequent video processing, a method based on the EM algorithm [22] is delineated in Section III-B.

III. Methodology

A neural network is used to estimate the number of people in real time. Then, with the estimated number of people, a method based on the EM algorithm is used to locate the individuals. Fig. 2 shows the two parts: neural-network-based people counting and EM-based individual detection.

A. People Counting

From Fig. 1(b), one can observe that even sitting people show some foreground pixels. This is because people always exhibit some movement whether they are standing or sitting. Motivated by this observation, we try to estimate the number of people by finding a relationship with the foreground pixels.

Due to illumination changes, camera movements, and objects removed from or introduced into the scene, getting a fixed background is usually difficult. A robust adaptive background estimation method based on the Gaussian Mixture Model [23],

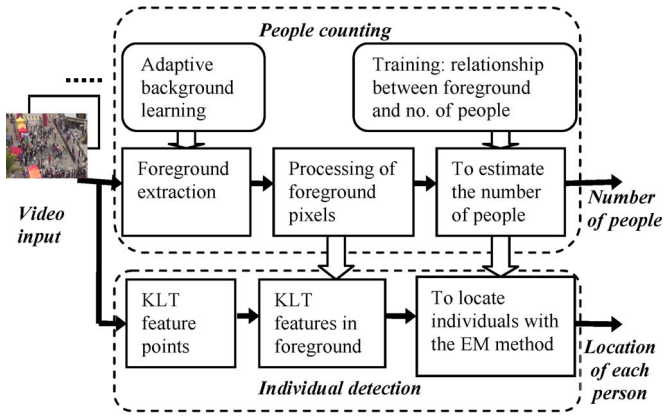


Fig. 2. Block diagram of the method, which includes two main parts: people counting and individual detection.

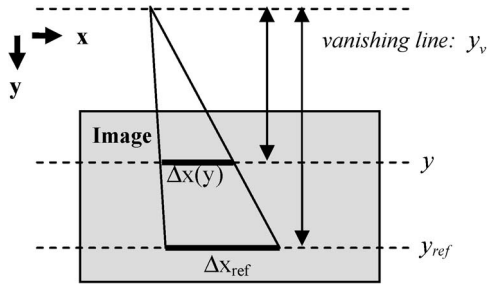


Fig. 3. Perspective correction method.

[24] is employed in this paper. To simplify the algorithm, only grayscale images are applied in our method. After getting the background image, a foreground image is obtained by subtracting the current image from the background image. The foreground image is then binarized based on a threshold to obtain the foreground pixels. The threshold should be set such that people moving slightly show some scattered pixels while keeping the noise low. The threshold in our evaluation is 40. When the intensity difference of a pixel between the current image and the background image is larger than 40, the pixel is viewed as a foreground pixel.

Perspective correction is an important step for foreground pixels-based estimation. Before estimating the number of people with foreground pixels, one must compensate for perspective distortion. We employ the same method as in [13]. We assume that the size of an object varies linearly as a function of the y-coordinate of the image. In this method, the objects at different locations are brought to the same scale. Equation (1) shows how to convert a scale at y to its scale at the reference location, y_{ref}. Fig. 3 is a simple illustration for (1). $\Delta x(y)$ is the horizontal (vertical) scale of an object at y, and Δx_{ref} is its horizontal (vertical) reference scale. $q(y)$ is the ratio for different locations

$$\Delta x_{ref} = \Delta x(y) * q(y) \text{ and } q(y) = \frac{(y_{ref} - y_v)}{(y - y_v)}. \quad (1)$$

The extension of parallel lines intersects at a vanishing point, which lies on y_v in the image. y_v can be easily estimated using the same object at two different coordinates in (1).

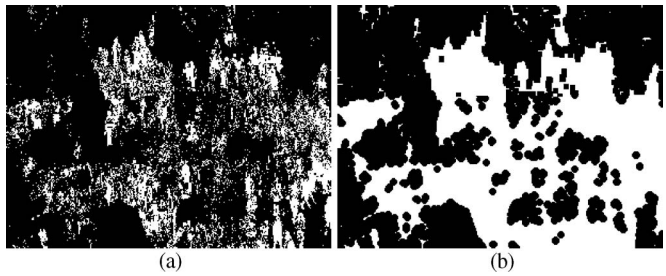


Fig. 4. (a) Foreground pixels extracted using an adaptive background. (b) The image after a closing operation is performed on (a). The maximum structuring element is a disk with a radius of four pixels.

After perspective correction, the number of foreground pixels is computed with (2), in which $imgY$ is the height of the processing image. $N(y)$ is the number of foreground pixels in the y_{th} row

$$N_{pixel} = \sum_{y=1:imgY} N(y) * q^2(y). \quad (2)$$

To determine the relationship between foreground pixels and the number of people, some manually annotated training images from a similar scene are needed. Several methods have been attempted for people counting in this section.

Method 1) Based on Foreground Pixels: First, the relationship between the number of foreground pixels after perspective correction and the number of people will be found directly. Suppose the number of foreground pixels after perspective correction is X, and the number of people is M. The relationship between M and X is shown in

$$M = f_1(X). \quad (3)$$

The training set will be used to build a neural network to ascertain the relationship f_1. Then, the trained neural network can be used to estimate the number of people.

Method 2) Based on Closed Foreground Pixels: From the extracted foreground [Fig. 1(b)], one can observe that some people show solid foreground blobs, while others only show some scattered pixels in the foreground image. Furthermore, the solid blobs are mainly from the moving people, and the scattered pixels come from the relatively stationary crowd.

To reduce the difference between moving people and stationary people, a closing operation is employed. After performing the closing operation, most areas occupied by people are covered with white pixels, while the other parts with black. An example of the foreground pixel image and its corresponding closing image is shown in Fig. 4. It should be noted that perspective effects also need to be considered during the closing operation. A smaller structuring element is used for image parts with lower y-coordinates. The maximum structuring element used in Fig. 4(b) is a disk with a radius of four pixels. Let C be the number of foreground pixels after the closing operation and M be the number of people. The relationship between C and M will be found and used for estimation

$$M = f_2(C). \quad (4)$$

Method 3) Based on Both Foreground Pixels and Closed Foreground Pixels: To keep more information about the original image, both foreground pixels and closed foreground pixels will be injected into the neural network. The relationship between the number of people and these two inputs is denoted as f_3

$$M = f_3(C, X). \tag{5}$$

B. Individual Detection

The methods based on foreground pixels can estimate the number of people easily, but they cannot provide any information on the location of each person. Individual detection is important for subsequent video processing. This section introduces a simple method for detecting individuals in such situation.

Since the image has a low resolution (a frontal human closest to the camera is only about 15 pixels wide in our test), human detection methods based on edges or gradients in the image will not work effectively. The methods based on segmenting foreground blobs do not require a high resolution. However, getting an accurate foreground contour for images with stationary people is almost impossible, as shown in Fig. 4(a).

As mentioned in the previous section, a closing operation can extract most areas occupied by human beings from a foreground image. The new image could be used as a rough foreground mask. To avoid the dependence on accurate foreground extraction, some corner-like feature points will be extracted from the whole image. After being filtered with the foreground mask, most feature points from the background will be filtered out while points from human beings remain. The key step for human detection is to cluster these feature points with some prior knowledge of human size. The details of this method will be introduced in the following sections.

Feature Detection: Kanade-Lucas-Tomasi (KLT) [25] is a popular corner detector and shows good performance for tracking. Briefly, the feature points are detected by examining the minimum eigenvalue of a $2 * 2$ gradient matrix, Z, at each location. Z is obtained with (6) and (7), in which I is the intensity value of the image. W is a $7 * 7$ window centered at the detected location in our test

$$Z = \sum_{(x,y)\in W} g(x,y)g(x,y)^T \tag{6}$$

$$g(x,y) = [\partial I/\partial x, \partial I/\partial y]^T. \tag{7}$$

In fact, any corner-like feature point can work in our evaluations. KLT features have been used due to KLT's good performance for tracking in subsequent video processing. Two parameters are important during feature detection. The first parameter is the number of features to be detected. In our methods, it is set to be large enough such that human beings show sufficient evidence of their existence. The second parameter is the minimum distance between two feature centers. For a human being, the features may come from the contour or clothing. KLT features from human clothing may not always appear, but points due to human shape appear on almost all. Head-shoulder parts offer crucial KLT features on human contours. Therefore, the minimum distance of two KLT features should be set such that the points from head-shoulder can be easily detected.

Foreground Mask: The foreground mask is obtained from the foreground pixel image after a closing operation. With an appropriate structuring element, the foreground image after a closing operation can cover almost all the areas occupied by human beings while cutting most of the cluttered background. The size of the structuring element is related to the density of scattered foreground pixels from the stationary people in the image.

After filtering with the foreground mask, almost all feature points from the background will be removed. The remaining feature points are mainly from human contours and different clothing. Hence, human detection is formulated as a problem of clustering these feature points.

Cluster Model: Before clustering these feature points to each individual person, a cluster model needs to be established. In our test, each cluster has a distribution as described in (8). To display it more clearly, the 2-D cluster model has been illustrated in 3-D space in Fig. 5(a). A profile along its short major axis is shown in Fig. 5(b)

$$h(s) = \begin{cases} m/(\pi * eh * ew) & inside{-}ellipse \\ \frac{\exp\left[-0.5(s-\mu)^T\Sigma^{-1}(s-\mu)\right]}{2\pi|\Sigma|^{1/2}} & outside{-}ellipse. \end{cases} \tag{8}$$

In this model, a vertical ellipse with semi-major axis, eh, and semi-minor axis, ew, is used to represent a prior human shape. $2 * eh$ and $2 * ew$ are the average height and width of a person. They can be estimated at one location in the image, and the values at other locations can be obtained with Equation (1). KLT features are assumed to be uniformly distributed over the entire human body. This is reasonable since there is a variety of clothing. m is a normalizing constant. It is set such that the integral of the proposed distribution is 1.

Outside the ellipse, a Gaussian distribution with mean, μ, and covariance matrix, Σ, is assumed. The center of the Gaussian distribution is the same as the center of the ellipse. The covariance of the Gaussian distribution is related to the size of the ellipse. When human beings to be detected are upright, the covariance matrix, Σ, is diagonal and it can be written as $\Sigma = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}$, where σ_x^2 and σ_y^2 are variances along two major axes.

To facilitate the computation, we assume the ellipse is coincident with the "30% ellipse" of the Gaussian distribution. The "30% ellipse" is composed of all points with a probability of 30% of the peak appearing in the Gaussian distribution, as shown in Fig. 5(b). "30%" is an empirical selection. Under these conditions, the total percentage of points falling inside the ellipse is 0.773. Hence, the normalizing constant m in the model is equal to 0.773. At the same time, the ellipse of size eh and ew and the covariance matrix of the Gaussian distribution Σ would satisfy the following relationship:

$$\exp\left[-0.5 * ew * \sigma_x^{-2} * ew\right] = 0.3$$

$$\exp\left[-0.5 * eh * \sigma_y^{-2} * eh\right] = 0.3.$$

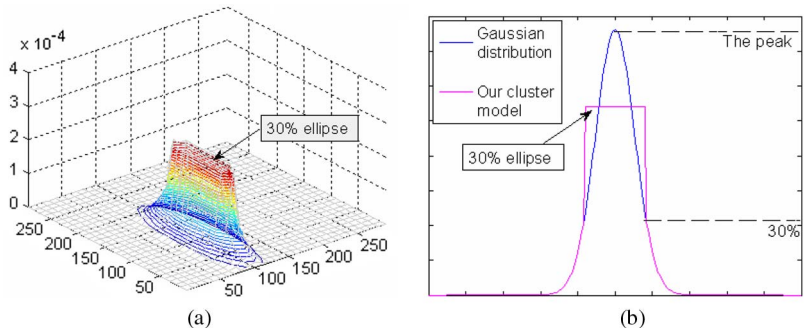


Fig. 5. (a) Three-dimensional illustration of our cluster model. (b) Profile along the short major axis of the ellipse. The blue curve is a standard Gaussian distribution, and the red curve is our cluster model.

TABLE I
EM Clustering Method

Initialization:

- Suppose the estimated number of people from the neural network-based method is L. Then, the number of clusters is initialized as $2*L$. Each cluster has an equal prior probability.

Iterate:

- **E-step:** The objective of the E-step is to obtain the assignment probabilities, which associate the feature points with each cluster. The probability that the feature point i generated by the j_{th} cluster is $\hat{p}(j\,|\,i)$ and can be obtained with (10). In (10), h is the cluster model, and $h(i\,|\,j)$ is derived from (8) using the parameters of the j_{th} cluster. $\hat{p}_j$ is the probability of each cluster, and k is the total number of clusters.

$$\hat{p}(j\,|\,i) = \hat{p}_j h(i\,|\,j) / \sum_{j=1}^{k} \hat{p}_j h(i\,|\,j) \qquad (10)$$

- **M-step:** The objective of the M-step is to maximize the likelihood with respect to the cluster model parameters. The parameters to be updated in our algorithm include the location $\hat{\mu}_j$ and the probability $\hat{p}_j$ of each cluster. n is the total number of feature points, and s_i is the location of feature point i in (11).

$$\hat{n}_j = \sum_{i=1}^{n} \hat{p}(j\,|\,i)$$

$$\hat{p}_j = \hat{n}_j / n, \qquad \hat{\mu}_j = \frac{1}{\hat{n}_j} \sum_{i=1}^{n} \hat{p}(j\,|\,i) s_i \qquad (11)$$

This means that

$$ew^2 = (-2\ln 0.3)\sigma_x^2 \quad eh^2 = (-2\ln 0.3)\sigma_y^2. \qquad (9)$$

EM Clustering: The EM algorithm was first proposed by Dempster *et al.* [22] and is widely used nowadays. In our method, the EM algorithm is used to cluster the feature points into each individual person. Table I shows the clustering algorithm. The EM algorithm mainly includes two steps, the E-step and the M-step, which are iteratively performed to obtain an optimal result.

A difficulty in the EM algorithm is the determination of the number of clusters. Usually, the number of clusters can be automatically determined with Bayesian Information Criterion [26]

$$F = -2\log f(S|k,\hat{\theta}) + v_k \log n. \qquad (12)$$

In (12), k is the number of clusters, $\hat{\theta}$ is a vector including all the model parameters in the k clusters, and f is the function of the mixture model composed of k clusters. Each cluster has a formulation as shown in (8) in our method. S represents all the feature points, n is the total number of the points, and v_k is the total number of free parameters in the mixture model with k clusters. The number of clusters should be the one that can achieve the maximum of F. However, from (12), one can observe that this criterion tends to use as few components as possible to explain all the foreground feature points. This is different from our application.

In our application, an ellipse that covers sufficient feature points is supposed to reasonably represent a human individual. We aim to cover almost all the feature points within those ellipses. The number of clusters in our methods will be provided before the EM algorithm starts. The estimation result obtained in the previous section can be used to indicate the number of clusters in the EM algorithm. However, more initial clusters are preferred for two reasons. First, some initial ellipses may be occupied by the unfiltered feature points from the background. Second, sufficient initial clusters can help to reduce the sensitivity of the EM algorithm to the initial cluster locations.

The clusters are initially placed on the KLT points with high densities. In our evaluations, feature density is defined as the number of feature points within a small neighborhood around each point. In Evaluation 3, the neighborhood is a circle with a radius of 8 pixels. Additionally, to avoid an overconcentration of initial clusters, the distance between two initial clusters must be larger than 8 pixels.

Postprocessing: After the EM clustering step, some post-processing operations need to be performed.

a) The EM clustering results may contain some redundant ellipses. The feature points falling in these redundant ellipses are also included in other ellipses. It is reasonable to remove the ellipses without sufficient evidence from the feature points. In our test, the candidate ellipses are checked one by one and the redundant ellipses removed.

b) A very simple occlusion analysis is performed in this step. For people close to each other, it is reasonable to assume that human beings with a low image y-coordinate would be occluded by those with a high image y-coordinate in most visual surveillance systems. In our evaluations, "occlusion" is simply defined as a 30% overlap of two ellipses. Humans not occluded by others should have more than three feature points, while two feature points are acceptable for those who are occluded.

TABLE II
RESULTS OF PEOPLE COUNTING

Inputs	**Mean error percentage** (for the 51 test cases)	**Accuracy** (% of cases with error percentage less than 10%)	**Accuracy** (% of cases with error percentage less than 15%)
X vs. M	16.36	45.10	60.78
C vs. M	10.68	60.78	80.39
X, C vs. M	10.03	68.04	80.39

IV. RESULTS AND ANALYSIS

Our focus is on a complicated scene, as described in the introduction. Comparing our results with others is difficult, as very little research has been done on complicated scenarios. As we know, [8] has obtained results using a large crowd in a scene similar to ours. Nonetheless, their methods require high resolution. They used a $16 * 16$ head template in their test. Human heads in our test set are only about five pixels wide. Besides, they did not provide the ground truth of their tested real images, so we are unable to compare the counting results. Some evaluations of our own methods are described in this section.

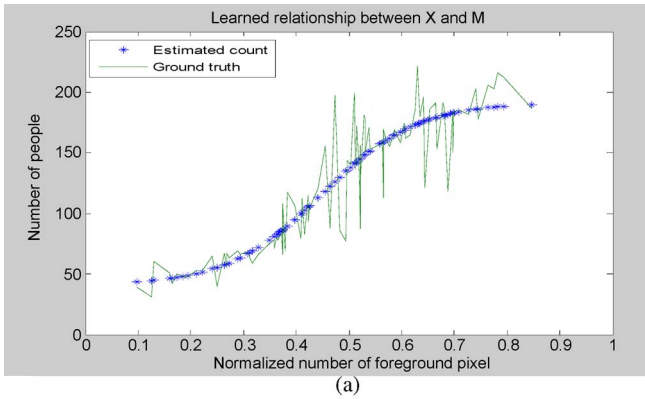
The evaluations were performed on a four-hour video taken at a public event. The video was taken at 10 fps, and the image resolution is $640 * 480$ pixels. One image scene every 100 seconds was used for the evaluation, and a total of 153 images were extracted from the original four-hour video. The ground truth, which is the actual number of people in the scene, for each image was obtained manually. In the set of images, the number of people in the scene ranges from 36 to 222. The training set consists of 102 images, which were formed by taking the first two images out of every three consecutive images. The test set is composed of the remaining 51 images. There is a wide range in the number of people for both the training set and the test set. To increase the speed of people counting, all the images were resized to $320 * 240$ pixels in Evaluation 1. The resolution of the image for human detection evaluations is $640 * 480$ pixels.

A. Evaluation 1

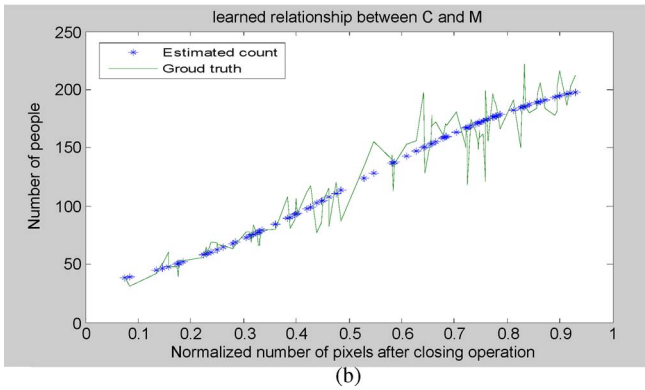
All three methods for people counting in Section III-A were tested. Their results are compared in Table II. The inputs were normalized to avoid bias due to their scale range.

From the table, one can observe that the estimation performance improved significantly after the closing operation. In fact, the relationship between the foreground pixels and the number of people becomes quite simple after the closing operation, as shown in Fig. 6. Different sizes of disks have been attempted in our evaluations. The results show that the linear relationship is not very sensitive to the size of the disk. However, when the disk size is too large, many areas without people will also be identified as foreground pixels, which can result in false estimations.

The best estimation results come from method 3, in which both foreground pixels and closed foreground pixels are considered. The mean error percentage is around 10%. The percentage of test cases with error less than 10% also increases when compared with method 2.



(a)



(b)

Fig. 6. (a) Relationship between the number of people and the number of foreground pixels. (b) Learned relationship between the number of people and the number of pixels after the closing operation.

(a)

(b)

Fig. 7. Sudden decrease in the number of people results in false foreground pixels. (a) The 30th sample in the test set and the extracted foreground pixels from this sample. (b) The scene 100 seconds before (a) and its foreground image.

One cause of errors in the test set is the sudden decrease of people. Take the 30th sample [see Fig. 7(a)] in the test set as an example; the scene 100 seconds before is shown in Fig. 7(b). A large number of people were sitting in chairs for a long time and were thus considered as the background. However, they suddenly moved away. So in the 30th sample, the background had not been adapted completely yet. The wrong foreground

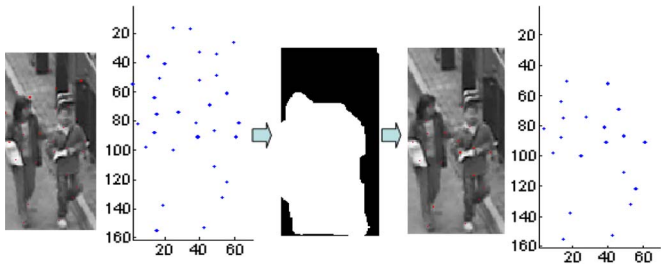


Fig. 8. After using a foreground mask, almost all the feature points from the background are removed.

pixels at the back of the sitting area caused large errors in the test.

The movement of non-human objects also results in wrong foreground pixels. At the end of the four-hour video, the ground truth of the number of people is relatively low. Hence, the movement of some boxes caused large error percentages.

B. Evaluation 2

In this evaluation, a simple test of the human detection method is shown on an image with two persons, as shown in Fig. 8.

During the test, KLT feature detection was performed with the implementation in [27]. Fig. 8 illustrates the process for using a foreground mask. Initially, many KLT feature points come from the railings behind. After using the foreground mask, the feature points mainly come from the two human beings.

After removing those background feature points, the EM algorithm was used to cluster the remaining points. Clustering with different numbers of initial clusters was tested, and the results are shown in Fig. 9. Each cluster is indicated with an ellipse. In this evaluation, the initial location of each cluster is uniformly random in the image. Fig. 9(a), (d), and (g) show the locations of 3, 5, and 8 initial clusters, respectively. The corresponding clustering results after 15 iterations with the EM algorithm are shown in Fig. 9(b), (e), and (h). Some clusters overlap in these figures. Take Fig. 9(h) as an example. Five initial clusters have converged to the left ellipse, and three other clusters overlap on the right ellipse. They are so close that differentiating them from the figure is difficult. After the postprocessing operations introduced in Section III-B, the results are shown in Fig. 9(c), (f), and (i). Although a large number of initial clusters were provided, the clustering results always converged to an accurate number of persons in the test image after postprocessing.

C. Evaluation 3

Some human detection results on a typical outdoor scene will be shown in this evaluation. Since our focus is upright human detection, a scene with few sitting people, as shown in Fig. 10, will be used as an example.

First, the most significant 1000 KLT feature points were extracted from the image. The minimum distance between two feature points is eight pixels. To include almost all the feature

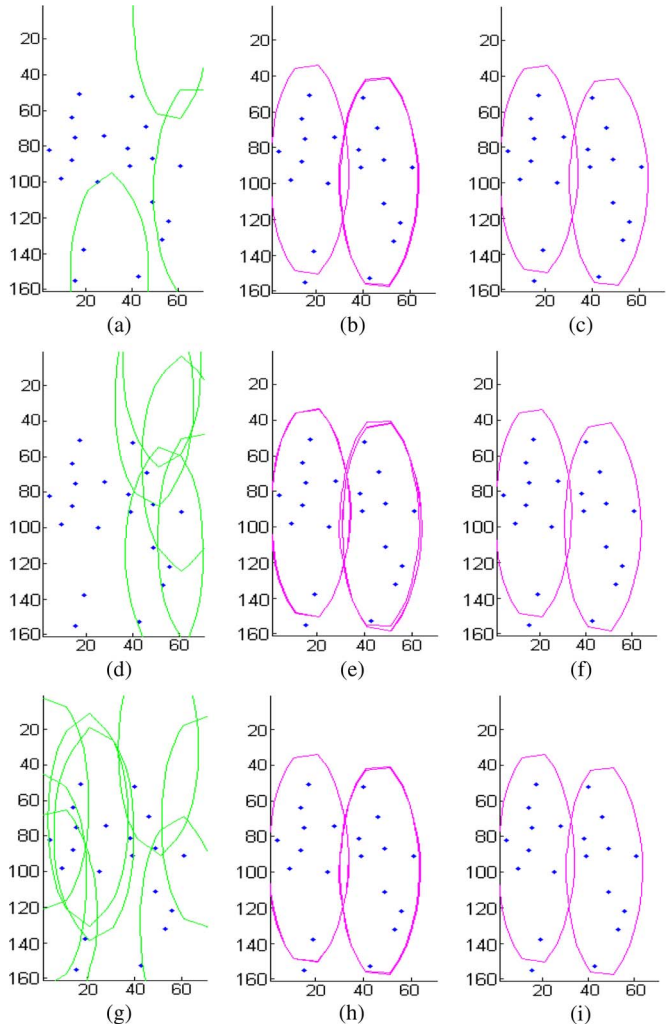
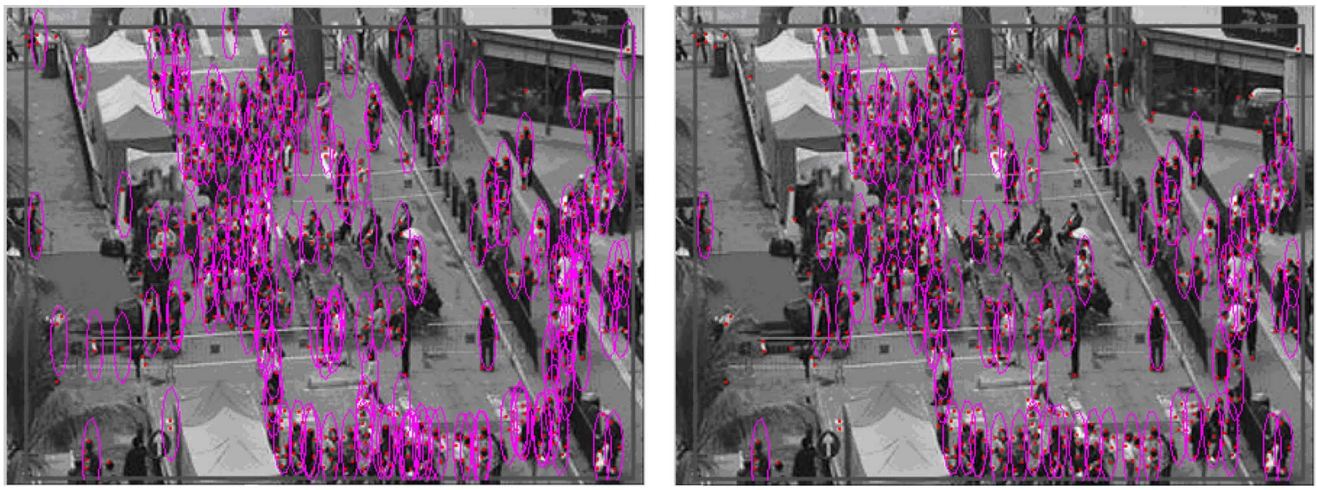


Fig. 9. Clustering results for different initial numbers of clusters. Initial clusters 3, 5, and 8 are shown in (a), (d), and (g); the corresponding clustering results after 15 iterations with the EM algorithm are shown in (b), (e), and (h); the final cluster results after removing unreasonable ellipses are shown in (c), (f), and (i).

points from human beings within the foreground mask, a disk with a radius of five pixels was used for the closing operation. After using the foreground mask, the number of remaining foreground feature points is 653.

The estimation result from the neural network method for this image is 126. The number of clusters for the EM algorithm was initialized as twice the estimate, which is 252. The initial clusters were located according to the density of the feature points, and the probability of each cluster was initially set to be equal.

Fig. 10(a) shows the clustering results when the EM algorithm converges. The ellipses indicate the location and the size of each person. Some ellipses have a large overlap with others, and some contain very few feature points. Hence, two postprocessing measures introduced in Section III-B were taken to remove those unreasonable ellipses. Fig. 10(b) is the result after the postprocessing steps were performed. Overlapped ellipses were combined. Ellipses caused by scattered background feature points were removed. Although the resolution of the test



(a) (b)

Fig. 10. Human detection results. (a) Results after EM clustering. (b) Results after postprocessing. The red dots show the feature points after using the foreground mask, and the ellipses show each detected human being.

(a) (b)

Fig. 11. (a) Result based on the Gaussian model. (b) Result based on our proposed model.

(a) (b)

Fig. 12. (a) Result based on the Gaussian model. (b) Result based on our proposed model.

image is very low (about 15 pixels wide for a frontal person closest to the camera), from Fig. 10(b), one can observe that the method produces effective results in most areas. Human beings lining up were detected correctly with our method, even though they only show some scattered foreground pixels in the extracted foreground.

(a) (b)

Fig. 13. (a) Result based on the Gaussian model. (b) Result based on our proposed model.

(a) (b)

Fig. 14. (a) Result based on the Gaussian model. (b) Result based on our proposed model.

However, the detection results are still far from the ground truth. The number of the remaining ellipses in Fig. 10(b) is only 110. The difference between the estimated number of people, 126, and the located people, 110, is mainly due to the following: First, stationary people may be excluded from the foreground. Second, some people show very few feature points. Third, similar to the other methods based on foreground segmentation, there may be some misdetections in very dense areas. Additionally, since the foreground mask is not very accurate, some background areas near the crowd may cause false feature points and result in some false detections. To improve the method, more texture features inside the crowd need to be studied. However, this will require higher resolution images.

Finally, for better comparison of our model to the Gaussian model, more results based on our test sequence are shown in Figs. 11 and 12. Additional results based on INRIA's CAVIAR scenarios are shown in Figs. 13 and 14. At the locations indicated by the green ellipses, one can observe that with our proposed method, the detection is more accurate, with less overcounting.

V. Conclusion

The scenario analyzed in this paper is quite common in public areas. Yet, little research has been carried out in such scenes.

In this paper, foreground pixels from both moving people and near stationary people have been considered to estimate their number. After a closing operation over foreground pixels, one can observe a linear relationship between the number of people and foreground pixels. The best estimation results, with a 10% average error, were achieved when both foreground pixels and closed foreground pixels are learned in a neural network.

With the estimated number of people, a human detection method based on the EM algorithm has been attempted for subsequent video processing. By clustering the KLT feature points in a foreground mask, the requirement for an accurate foreground contour has been reduced. The application of methods based on segmenting the foreground has been extended to detection of people who are moving only slightly. This new cluster model has been shown to be more accurate in both counting and detection than the Gaussian model.

In the future, texture inside the foreground region can be used as another input for the neural network. This addition will be an extension of the present paper. As for the detection algorithm, foreground pixels will be combined with the feature point clustering method to avoid the insufficiency of feature points on some human beings. More detailed analysis of the distribution of KLT features in each ellipse would help handle the non-human moving objects. However, to distinguish human and non-human objects more accurately, a high-resolution video is needed to provide sufficient data. A high-resolution video is also necessary to handle a denser crowd.

REFERENCES

[1] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1208–1221, Sep. 2004.

[2] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1198–1211, Jul. 2008.

[3] J. Rittscher, P. H. Tu, and N. Krahnstoever, "Simultaneous estimation of segmentation and shape," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 486–493.

[4] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 4, pp. 349–361, Apr. 2001.

[5] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 69–82.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 886–893.

[7] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, Nov. 2007.

[8] S. F. Lin, J. Y. Chen, and H. X. Chao, "Estimation of number of people in crowded scenes using perspective transformation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 31, no. 6, pp. 645–654, Nov. 2001.

[9] G. J. Brostow and R. Cipolla, "Unsupervised Bayesian detection of independent motion in crowds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 594–601.

[10] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2006, pp. 705–711.

[11] A. C. Davies, J. H. Yin, and S. A. Velastin, "Crowd monitoring using image processing," *Electron. Commun. Eng. J.*, vol. 7, no. 1, pp. 37–47, Feb. 1995.

[12] S.-Y. Cho, T. W. S. Chow, and C.-T. Leung, "A neural-based crowd estimation by hybrid global learning algorithm," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 29, no. 4, pp. 535–541, Aug. 1999.

[13] R. Ma, L. Li, W. Huang, and Q. Tian, "On pixel count based crowd density estimation for visual surveillance," in *Proc. IEEE Conf. Cybern. Intell. Syst.*, 2004, pp. 170–173.

[14] H. Celik, A. Hanjalic, and E. A. Hendriks, "Towards a robust solution to people counting," in *Proc. IEEE Int. Conf. Image Process.*, 2006, pp. 2401–2404.

[15] P. Kilambi, O. Masoud, and N. Papanikolopoulos, "Crowd analysis at mass transit site," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2006, pp. 753–758.

[16] A. N. Marana, S. A. Velastin, L. F. Costa, and R. A. Lotufo, "Estimation of crowd density using image processing," in *Proc. IEE Colloq. Image Process. Security Appl.*, 1997, pp. 11/1–11/8.

[17] A. N. Marana, L. Da Fontoura Costa, R. A. Lotufo, and S. A. Velastin, "Estimating crowd density with Minkowski fractal dimension," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1999, pp. 3521–3524.

[18] H. Rahmalan, M. S. Nixon, and J. N. Carter, "On crowd density estimation for surveillance," in *Proc. Inst. Eng. Technol. Conf. Crime Security*, 2006, pp. 540–545.

[19] X. Li, L. Shen, and H. Li, "Estimation of crowd density based on wavelet and support vector machine," *Trans. Inst. Meas. Control*, vol. 28, no. 3, pp. 299–308, Aug. 2006.

[20] D. Kong, D. Gray, and T. Hai, "A viewpoint invariant approach for crowd counting," in *Proc. Int. Conf. Pattern Recog.*, 2006, pp. 1187–1190.

[21] A. B. Chan, Z. S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–7.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

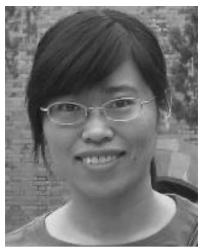
[23] W. E. L. Grimson, C. Stauffer, R. Romano, and L. Lee, "Using adaptive tracking to classify and monitor activities in a site," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1998, pp. 22–29.

[24] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[25] C. Tomasi and T. Kanade, "Detection and tracking of point features," Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-91-132, 1991.

[26] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[27] S. Birchfield, Source Code of the KLT Feature Tracker, 2006. [Online]. Available: http://www.ces.clemson.edu/~stb/klt/



Ya-Li Hou (S'09) received the B.Eng. degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 2004, and the Master degree from the Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai, China, in 2007. She is currently working toward the Ph.D. degree in the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong.

Her current research interests include visual surveillance, pattern recognition, and computer vision.

Grantham K. H. Pang (S'84–M'86–SM'01) received the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 1986.

He was with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, from 1986 to 1996. Currently, he is an Associate Professor in the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong. He has published more than 160 technical papers and has authored or coauthored six books. He has also obtained five U.S. patents. His research interests include machine vision for surface defect detection, optical communications, expert systems for control system design, intelligent control, and intelligent transportation systems. He acts as consultant to a number of local and international companies, and has served as expert witness for the Courts of the Hong Kong Special Administrative Region.