IOSR Journal of Engineering (IOSRJEN)
e-ISSN: 2250-3021, **p**-ISSN: 2278-8719
Vol. 3, Issue 1 (Jan. 2013), ||V1|| PP 05-10

Literature survey for people counting and human detection

G.Thomas Prathiba[1], Y.R.Packia Dhas[2]
[1]*PG Scholar, PET Engineering College, India*
[2]*Associate Professor, PET Engineering College, India*

Abstract:- People counting is a crucial and challenging problem in visual surveillance. Automatic monitoring of the number of people in public areas is important for safety control and urban planning. For this purpose many techniques and methods have been proposed. These techniques are not producing high performance and better accuracy for complicated scenes counting. Recently Foreground Extraction and Expectation Maximization (EM) based methods are proposed, which provides a better accurate solution for people counting and finding individual location. This literature survey discusses some of the existing methods and their performance.

Keywords - ACC, DDMCMC, HOG, Neural Network, Privacy preserving system.

I. INTRODUCTION

Reliable people counting and human detection is an important problem in visual surveillance. An accurate and real time estimation of people in a shopping mall can provide valuable information for managers. 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. In [10], a key factor in the solutions described in the use of global or semiglobal pixel intensity values to infer crowd behaviour avoiding recognition and tracking of individual pedestrians. Human beings perceive images through their properties like colour, shape, size, and texture described in [15]. Main goal is a vision system that monitors activity in a site over extended periods of time is described in [21], [22]. In [16], describes a technique for crowd density estimation based on Minkowski fractal dimension. A neural based crowd estimation system for surveillance in complex scenes at underground station platform is presented in [11]. In [12] surveillance systems for public security are going beyond the conventional CCTV. In [1] first objective is to segment multiple human objects and track their global motion in complex situations. Second is to estimate the locomotion modes and the coarse 3D body postures. In [3], main work is the integration of feature grouping and model based segmentation into one consistent framework. A supervised data driven bayesian clustering algorithm is described in [8] which has detection of individual entities. Trajectory set clustering method is used in [9], to identify the number of moving objects in a scene. The possibilities of developing a robust statistical method for people counting are investigated in [13]. In [14], groups are tracked in the same manner as individuals, using Kalman filtering techniques. In [17], image features were extracted using Grey Level Dependency Matrix, Minkowski Fractal Dimension and a new method called Translation Invariant Orthonormal Chebyshev Moments. An estimation method of the crowd density based on multi-scale analysis and support vector machine is described in [18]. In [19], describes a viewpoint invariant learning–based method for counting people in crowds from a single camera. In [7], body part detectors are learned by boosting edgelet feature based weak classifiers. Different methods are handled, but it produces less accuracy and performance than proposed [5] method.

II. LITERATURE REVIEW

2.1. Component based people detection

Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio [4] proposed a general example-based framework for detecting objects in static images by components. The technique is demonstrated by developing a system that locates people in cluttered scenes. In particular, the system detects the components of a person's body in an image, i.e., the head, the left and right arms, and the legs, instead of the full body by using four distinct example based detectors. The system then checks to ensure that the detected components are in the proper geometric configuration (i.e. shown inTable 1).

Table 1 Geometric constraints placed on each component

Component	centroid		Scale		Other Criteria
	Row	Column	Min	Max	
Head and Shoulders	23 ±3	32 ±2	28×28	42×42	
Lower Body		32 ±3	42×28	69×46	Bottom Edge: Row:124±4
Right Arm Extended	54 ±5	46 ±3	31×25	47×31	
Right Arm Bent		46 ±3	31×25	47×31	Top Edge: Row:31±3
Left Arm Extended	54 ±5	17 ±3	31×25	47×31	
Left Arm Bent		17 ±3	31×25	47×31	Top Edge: Row:31±3

We calculated the geometric constraints for each component from a sample of the training images, tabulated in Table 1, by taking means of the centroid and top and bottom boundary edges of each component over positive detections in the training set. There are two sets of constraints for the arms, one intended for extended arms and the other for bent arms. Haar wavelet functions are used to represent the components in the images and Support Vector Machines (SVM) to classify the patterns. Four component-based detectors are combined at the next level by another SVM. The results of the component detectors are used to classify a pattern as either a "person" or a "nonperson". For this purpose uses one classifier, named as Adaptive Combination of Classifiers (ACC) that improves accuracy of people detection. This system performs significantly better than a similar full-body person detector. This suggests that the improvement in performance is due to the component-based approach and the ACC data classification architecture. While this paper establishes that, this system can detect people who are slightly rotated in depth, it does not determine, quantitatively, the extent of this capability. This is the main drawback of the method and also more time consuming task.

2.2 Histograms of Oriented Gradients Approach for Human detection

Navneet Dalal and Bill Triggs [6] proposed that the grids of Histograms of Oriented Gradient (HOG) descriptors significantly outperform existing feature sets for human detection. We study the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks are all important for good results. The new approach gives near-perfect separation on the original MIT pedestrian database, so we introduce a more challenging dataset containing over 1800 annotated human images with a large range of pose variations and backgrounds.

An Overview of our feature extraction and object detection chain is as shown in fig 1. The method is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid. The basic idea is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. In practice this is implemented by dividing the image window into small spatial regions, for each cell accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. The combined histogram entries form the representation. For better invariance to illumination, shadowing, etc., it is also useful to contrast-normalize the local responses before using them. This can be done by accumulating a measure of local histogram "energy "over somewhat larger spatial regions and using the results to normalize all of the cells in the block. We will refer to the normalized descriptor blocks as Histogram of Oriented Gradient (HOG) descriptors. Tiling the detection window with a dense grid of HOG descriptors and using the combined feature vector in a conventional SVM based window classifier gives our human detection chain (in Fig 1).

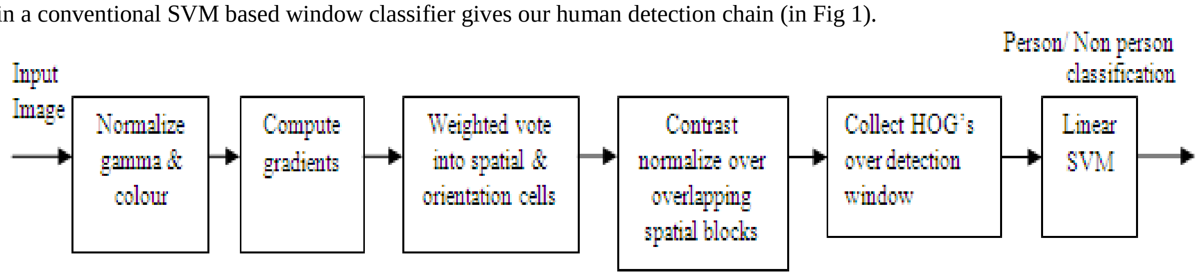


Figure 1. An overview of our feature extraction and object detection chain.

We have shown that using locally normalized histogram of gradient orientations features similar to SIFT descriptors in a dense overlapping grid gives very good results for person detection, reducing false positive rates by more than an order of magnitude relative to the best Haar wavelet based detector. Histograms of oriented Gradients may achieve more accurate counting and detection results when the crowd is small. Disadvantage of this process is that time consuming task and shows only results for a small crowd with few occlusions, it also required high resolution images.

2.3 DDMCMC approach

Tao Zhao and Bo Wu [2] proposed a model based approach to interpret the image observations by multiple partially occluded human hypotheses in a Bayesian framework. This approach to segmenting and tracking multiple humans emphasizes the use of shape models. An overview diagram is given in figure.2.

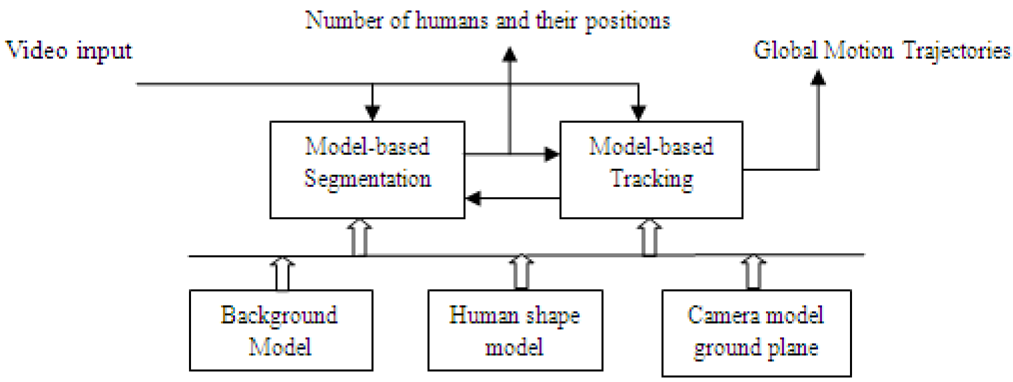


Figure 2. An overview diagram of Segmentation and tracking

In fig 2.based on a background model, the foreground blobs are extracted as the basic observation. By using the camera model and the assumption that objects move on a known ground plane, multiple 3D human hypotheses are projected onto the image plane and matched with the foreground blobs. Since the hypotheses are in 3D, occlusion reasoning is straightforward. In one frame, we segment the foreground blobs into multiple humans and associate the segmented humans with the existing trajectories. Then, the tracks are used to propose human hypotheses in the next frame. The segmentation and tracking are integrated in a unified framework and interoperate along time. We formulate the problem of segmentation and tracking as one of Bayesian inference to find the best interpretation given the image observations, the prior models, and the estimates from the previous frame analysis that is, the maximum a posteriori (MAP) estimation. The optimal solution is obtained by using an efficient sampling method, data-driven Markov chain Monte Carlo (DDMCMC), which uses image observations for proposal probabilities. Knowledge of various aspects, including human shape, camera model, and image cues, are integrated in one theoretically sound framework. To improve the computational efficiency, we use direct image features from a bottom-up image analysis as importance proposal probabilities to guide the moves of the markov chain. This method is able to successfully detect and track humans in the scenes of complexity with high detection and low false alarm rates. Here a more accurate 3-D model composed of three ellipsoids was used. To deal with the occlusion problem, a joint probability for multiple humans 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, is used 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, this may be easily corrupted by noise. This is a time consuming task.

2.4. Counting people without people models

Antoni B.Chan, Zhang-Sheng John Liang and Nuno Vasconcelos [20] proposed a privacy-preserving system for estimating the size of inhomogeneous crowds, composed of pedestrians that travel in different directions, without using explicit object segmentation or tracking or models. First, the crowd is segmented into components of homogeneous motion, using the mixture of dynamic textures motion model. Second, a set of simple holistic features is extracted from each segmented region, and the correspondence between features and the number of people per segment is learned with Gaussian Process regression. We adapt the mixture of dynamic textures to segment the crowds moving in different directions. The mixture model is learned with the Expectation- Maximization (EM) algorithm. Video locations are then scanned sequentially; a patch is extracted at each location, and assigned to the mixture component of largest posterior probability. The location is declared to belong to the segmentation region associated with that component. Before extracting features from the video segments, it is important to consider the effects of perspective. Because objects closer to the camera appear larger, any feature extracted from a foreground object will account for a smaller portion of the object than one extracted from an object farther away. This makes it important to normalize the features for perspective. Features such as segmentation area or number of edges should vary linearly with the number of people in the scene. These features capture segment shape and size. In this approach textures inside the foreground are used to estimate the crowd density or the number of people. 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. The class of functions that the GP can model is dependent on the kernel function used. For the task of pedestrian counting, we note that the dominant trend of many of the features is linear (e.g. segment area), with local non-linearities. To capture both trends, we combine the linear and the squared-exponential (RBF) kernels, i.e:

$$k(x_p, x_q) = \alpha_1(x_p^T x_q + 1) + \alpha_2 e^{\frac{-\|x_p - x_p\|^2}{\alpha_3}} + \alpha_4 \delta(p,q) \tag{1}$$

With hyperparameters $\alpha = \{\alpha 1, \alpha 2, \alpha 3, \alpha 4\}$. The first and second terms of the kernel are the linear and RBF components, while the third term models observation noise. The system tracks the changes in pedestrian traffic fairly well. Most errors tend to occur when there are very few people like less than two in the scene. This is the drawback of this approach.

2.5 Neural Network and EM based people counting and individual detection

Although there are many human detection methods, most of them are not producing high accuracy and the image resolution must be high. Ya-Li Hou and Grantham K.H. Pang [5] is developed neural network based people counting and EM based individual detection in a low resolution image with complicated scenes that is shown in Figure 3.

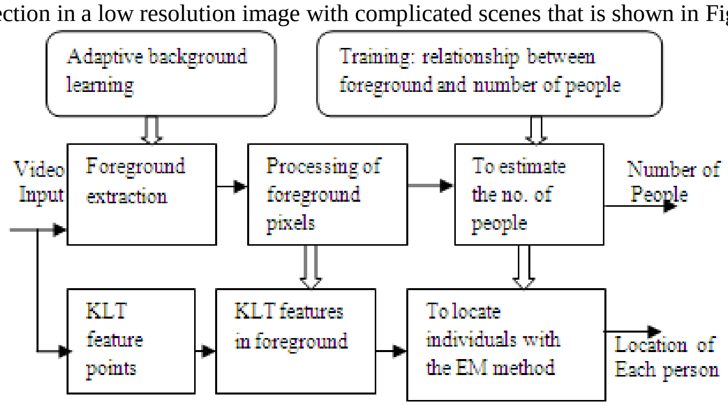


Figure 3. Block diagram of people counting and human detection

2.5.1 People Counting

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. 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.

$$\Delta x_{ref} = \Delta x(y) * q(y) \quad \text{and} \quad q(y) = (y_{ref} - y_v)/(y - y_v) \tag{2}$$

Equation (2) shows how to convert a scale at y to its scale at the reference location, y_{ref} . $\Delta x(y)$ 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 .After perspective correction, the number of foreground pixels is computed with (3), in which imgY- is the height of the processing image. N(y) is the number of foreground pixels in the y_{th} row

$$N_{pixel} = \sum_{y1:imgY} N(y) * q^2(y) \tag{3}$$

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 equation (4):

$$M = f_1(X) \tag{4}$$

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. It should be noted that perspective effects also need to be considered during the closing operation. The relationship between C and M will be found and used for estimation

$$M = f_2(C) \tag{5}$$

Let C be the number of foreground pixels after the closing operation and M be the number of people. 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 f3

$$M = f_3(C, X) \tag{6}$$

Table 2. 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

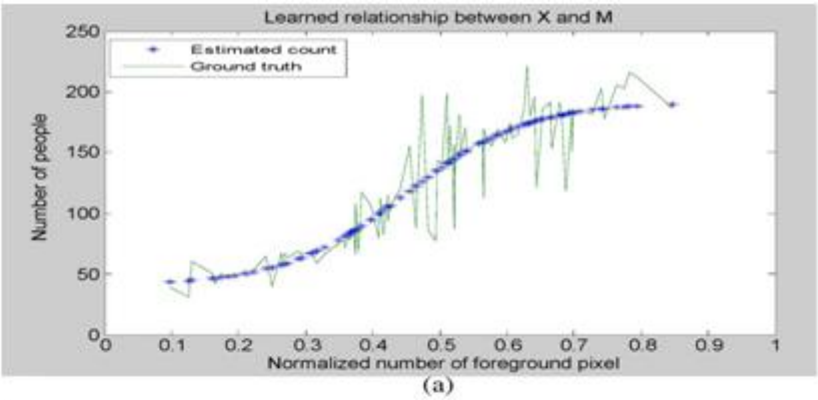
Table 2. Describes the results of people counting for the 51 test cases

2.5.2 Individual Detection

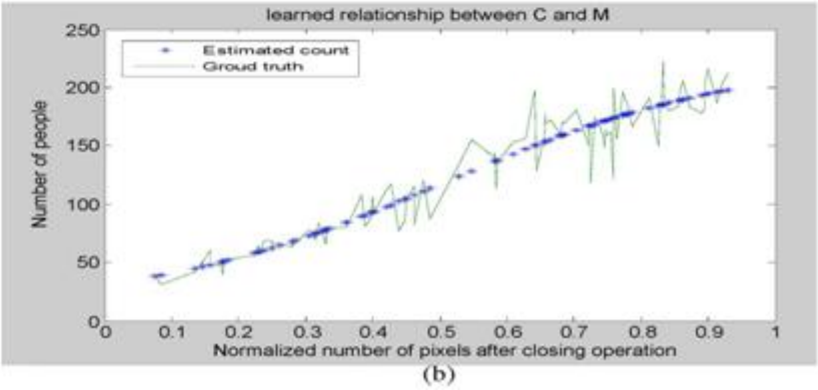
Individual detection is important for subsequent video processing. Kanade-Lucas-Tomasi is a popular corner detector and shows good performance for tracking. The foreground mask is obtained from the foreground pixel image after a closing operation. Before clustering feature points to each individual person, a cluster model needs to be established. In this process each cluster has a distribution as described in equation (7). To display it more clearly, the 2-D cluster model has been illustrated in 3-D space.

$$h(s) = \begin{cases} \frac{m}{(\pi * eh * ew)} & inside - ellipse \\ \frac{\exp\left[-0.5(s-\mu)^T \Sigma^{-1}(s-\mu)\right]}{\left[2\pi|\Sigma|^{\frac{1}{2}}\right]} & outside\ ellipse \end{cases} \tag{7}$$

In this model, 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. EM method is used to cluster the feature points into each individual person. 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. This 3-D cluster model is more accurate in both counting and detection than the Gaussian model. An advantage of this process is that, it produces high accuracy.



(a)



(b)

Figure 4. (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.

The relationship between the foreground pixels and the number of people becomes quite simple after the closing operation that is shown in Fig 4. Fig 4. (b) gives better performance than Fig 4.(a).This process produces high accuracy than existing methods.

III. CONCLUSION

In this paper, a brief literature survey for people counting and human detection methods are discussed elaborately and neural network based people counting and EM based individual detection methods are studied The best results for estimating the number of people has an average error of 10% over 51 test cases are shown in Table 2. These methods provide better performance and high accuracy than existing methods. Fig 4. (b) gives better performance than Fig 4.(a). This people counting and human detection process is very useful for safety control in public areas by using static camera.

ACKNOWLEDGEMENT

I pledge my thanks to my parents and friends who encouraged me to complete this paper successfully. I thank all my staff members especially my supervisor for giving me valuable suggestions throughout the completion of this work.

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.,vol30,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 Cnf.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]. Ya-Li Hou, and Grantham K.H.Pang, People counting and human detection in a challenging situation, *IEEE Trans.sys.man and cybernetics.,vol.41,No.1*, pp.24-33, Jan.2011.

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

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

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

[10]. 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, Feb1995.

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

[12]. 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*.pp.170-173,2004.

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

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

[15]. A.N.Marana, S.A.Velastin, L.F.costa, and R.A.Lotufo, Estimation of crowd density using image processing, *in Proc.IEEE colloq.Image Process.Security Appl.*,pp.11/1-11/8,1997

[16]. 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*.,pp.3521-3524,1999.

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

[18]. 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.

[19]. D.Kong, D.Gray, and T.Hat, A viewpoint invariant approach for crowd counting, *in Proc. Int. Conf. Pattern Recog*, pp. 1187-1190, 2006.

[20]. 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*., pp.1-7, 2008.

[21]. 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*, pp. 22-29, 1998.

[22]. 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.