Low Level Crowd Analysis Using Frame-Wise Normalized Feature For People Counting

Hajer Fradi, Jean-Luc Dugelay
EURECOM
Sophia Antipolis, France
Hajer.Fradi@eurecom.fr
Jean-Luc.Dugelay@eurecom.fr

***Abstract*—People counting is a crucial component in visual surveillance mainly for crowd monitoring and management. Recently, significant progress has been made in this field by using features regression. In this context, perspective distortions have been frequently studied, however, crowded scenes remain particularly challenging and could deeply affect the count because of the partial occlusions that occur between individuals. To address these challenges, we propose a people counting approach that harness the advantage of incorporating an uniform motion model into Gaussian Mixture Model (GMM) background subtraction to obtain high accurate foreground segmentation. The counting is based on foreground measurements, where a perspective normalization and a crowd measure-informed corner density are introduced with foreground pixel counts into a single feature. Afterwards, the correspondence between this frame-wise feature and the number of persons is learned by Gaussian Process regression. Experimental results demonstrate the benefits of integrating GMM with motion cue, and normalizing the proposed feature as well. Also, by means of comparisons to other feature-based methods, our approach has been experimentally validated showing more accurate results.**

I. Introduction

There is currently significant interest in visual surveillance systems for crowd analysis with the steady population growth. In this context, people counting is one of the crucial parameters of the crowd and its automatic monitoring is receiving much attention in surveillance and security communities. In fact, accurate estimation of number of persons in public areas is extremely important information for safety control to prevent crowd disasters mainly when the number of persons flooding some areas exceeds a certain level of crowd. Many stadium tragedies could illustrate this problem, also the Love Parade stampede in Germany and the Water Festival stampede in Colombia. To prevent the succession of such mortal accidents, accurate estimation of number of persons is required and appropriate decisions for security reasons have to be taken in case of large scale crowd. Also, the estimation of number of passengers is relevant to economic applications such as optimizing the schedule of public transportation systems. Hence, many recent works in the domain of automatic video surveillance have been proposed to address the problem of people counting.

While significant progress has been achieved in the field of people counting, crowded scenes still remain challenging because of the spatial overlaps that make delineating people difficult. Therefore, recent works typically bypass the task of detecting people and instead focus on learning a mapping between the number of persons and a set of low level image features. In this perspective, intensive study has been conducted by varying and increasing the number of features; some other works address this problem by applying different regression functions to select the one fitting better the features (e.g. linear in [1] and [2], ϵ-SVR regressor and ANFIS in [3], Bayesian Poisson in [4] and Gaussian Process regression in [5]). Ideally, the number of persons is simply proportional to the features, but some factors are affecting this relationship which leads to a deviation from the proportionality. Therefore, varying the features or the trainable functions are just applied as an implicit way to cope with this deviation and to infer more information about the frame contents.

Unlike these proposals, we expose that there is no need to use several features, however, we are more interested in revealing the factors that affect the relationship between the features and the number of persons. In particular, we intend to explore distance and crowd density cues. The first cue is employed to handle the problem of perspective distortions, whereas, the second cue is used as a crowd feature to detect and to measure the overlap between individuals. To achieve this goal, we adopt an integration of GMM background subtraction with an uniform motion model into a single overall system that has the potential to better segment foreground entities. Then, we propose to apply a perspective map normalization and to weight the feature by a crowd measure in order to compensate the variations in distance and density.

The remainder of the paper is organized as follows: In Section II, a taxonomy of relevant works to people counting is presented. Then, we introduce our approach for people counting in Section III. The proposed approach is evaluated using PETS dataset and the experimental results are summarized in Section IV. Finally, we conclude briefly in Section V.

II. Related works

The first paradigm of people counting is detection-based methods, where the number of persons and their locations are provided simultaneously. By applying these methods, the count is not affected as long as people are correctly segmented.

However, the difficulty is that detecting people is by itself a complex task, mainly in the presence of crowds and occlusions. This problem has been addressed by adopting part-based detectors [6], or by detecting only heads [7] or the Ω-shape formed by heads and shoulders [8]. These attempts to mitigate occlusions could be effective in low crowd scenes, however, they are not applicable in very crowd cases which are of primary interest for people counting. Therefore, feature-based methods have become complementary solution when it is nearly impossible to isolate each person in crowd areas. A brief description of these methods is provided next, along with some representative approaches.

The second paradigm of people counting consists of estimating the number of persons from various low level features. These methods are more efficient since it is easier to detect features than to detect persons. For this purpose, many features of foreground pixels (e.g. total area, textures and edge count [5], [9], [10], [11]) and also features based on interest points measurements (e.g. corner points [1] and SURF features [12]) are introduced into counting methods. To perform the counting, a regression function has to be applied. It is required to learn the relationship between features and number of persons.

More in details, Hou and Pang [13] addressed this problem by using a neural network to map the foreground pixels to the number of persons. In this work, the foreground pixels are extracted by subtracting each frame from a learned statistical background model. In [1], Albiol *et al.* proposed to use Harris corner points as features. Then, the count is performed by assuming a direct proportional relation between the number of moving corner points and the number of persons. This method has shown good performance using PETS dataset, whereas, its application is limited because it does not consider the difference between the perceived size of persons at different distances from the camera and with different densities as well. These limitations were not revealed in the PETS contest since only videos characterized by short depth range and trivial occlusions were required for the tests.

Differently from the two aforementioned works, some other methods take into account the effects of perspective distortions. To handle that, different techniques have been investigated. For instance, in [2], this problem is addressed by weighting foreground pixels according to geometric information. In [14], Ma *et al.* proposes a geometric correction to bring all the objects at different distances to the same scale.

While different techniques have been proposed to address the problem of perspective distortions, only few attempts have been made to handle the problem of occlusion that prevalently exists in the crowd and could deeply affect the count. In [5], Chan *et al.* applied dynamic texture motion models for foreground segmentation. Then, 28 features from each crowd segment are extracted and weighted according to an estimated perspective map. These features varies between geometric, edges and texture. The reason behind using all these features is to better interpret the image contents, in particular, to have a deep idea about the level of the crowd. Also, in [12], both of perspective and crowd problems have been addressed by applying a clustering algorithm to partition different groups of persons. Then, the distance from the camera is computed using an Inverse Perspective Mapping (IPM) and the density of each cluster is obtained as the ratio between the number of the detected points and the area of the bounding box. Recently, an explicit estimation of the crowd density levels is involved in [15], and the number of persons is estimated through a scaling factor which is learned for different levels of the crowd.

III. Proposed approach

In this section, our proposed approach for people counting is presented, we follow the recent methods based on features regression. One major advantage of applying these methods is to not depend on intermediate steps of individual detection or tracking. To infer the contents of each frame under analysis, only foreground pixels have to be extracted. Given the importance of foreground segmentation and its impact on the next steps, an efficient solution based on integrating GMM background subtraction with motion cue is employed. Afterwards, only two holistic features are used: foreground pixel counts and corner density. The first feature is weighted according to an estimate perspective map in order to compensate the effects of perspective distortions. We additionally explore density cue to handle partial occlusions due to the crowd. Under the assumption that images of low density crowd tend to present less dense corners compared to images of high density crowd, we propose to associate dense or sparse corners to the crowd size. For this purpose, local features are extracted and synthesized for global corner density, which is employed in a further step to normalize the first feature stated earlier. Finally, the two normalization are introduced into a single frame-wise feature. An overview of the feature extraction modules and their interaction is shown in Fig. 1.

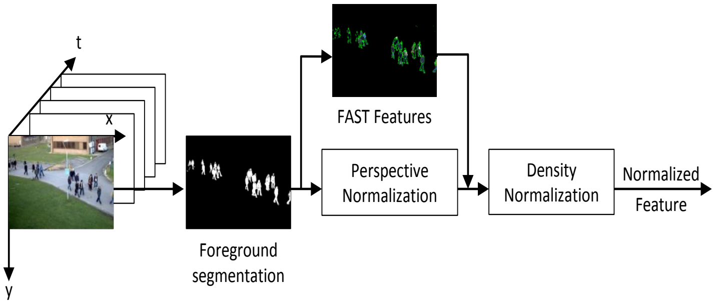

Fig. 1. *Schematic for Frame-Wise Normalized Feature Extraction*

A. Foreground segmentation

The first step of our proposed approach is to segment foreground entities. In this context, Gaussian Mixture Model (GMM) background subtraction [16] has been widely employed. It is based on a probabilistic approach that achieves satisfactory performance to handle complex scenes thanks to its ability to model various background distributions. Therefore, GMM based background subtraction is considered as a standard method and it has become the basis for a large number of extensions. Despite this, GMM includes some weakness. First, there is no consideration of spatial information. Second, the background model estimation step is problematic; the main difficulty is to decide which distributions of the mixture belong to the background. GMM assumes that the often occurring pixels are deemed to model the background which is not always true. Also, to adapt variations in the background (to maintain good precision), the detection rate is decreased. To overcome these limitations, we apply an integration of GMM background subtraction with an uniform motion model [17]. For this purpose, the improved adaptive GMM [18] is used, it has the advantage of constantly updating not only the parameters of the Gaussians but also the number of the mixture components using the Dirichlet prior. The second cue of this method is motion information, it is obtained by computing the optical flow [19] between each two adjacent frames. The optical flow field is defined by its magnitude and its direction. The magnitude of motion is convoluted with the difference between each current frame and the mean of the background to get precise boundaries. After that, a measure of uniformity of motion is applied to distinguish different connected components with the same velocity and orientation of the optical flow. Finally, the labeling process is updated by favoring pixels moving together to be classified as foreground entities. The goal of this integration is to improve the detection rate of GMM and to avoid outliers caused by the optical flow as well. It could also add spatial and temporal coherence since the labeling process using GMM is done only at pixel level.

After performing the foreground segmentation, we note that using only the total number of foreground pixels to estimate the number of persons is not enough. Therefore, further enhancements for this feature are necessary to improve its invariance to distance and crowd density.

B. Perspective normalization

At this stage, the objective is to compensate for changes in number of foreground pixels due to perspective distortions. The effects of perspective can be simply explained by the fact that objects far from the camera appear smaller than the closest ones. This makes any extracted feature from farther away persons account for a smaller portion compared to closer persons. This problem could be addressed by weighting each foreground pixel according to a perspective map with assigning larger weights for farther points in the scene.

Similar to [5], we estimate the perspective map by linearly interpolating between the two extremes of the scene. First, the ground plane is marked. Then, the distance d_1 and d_2 of the two extreme lines are measured. After that, the difference between the perceived height of persons in these two lines can be derived by manually calibrating two frames, where the center of a reference person belongs to the first line in the first frame while belonging to the second extreme line in the second frame. A weight of 1 is assigned to pixels on the first line, and the pixels on the second line are weighted by $\frac{h_1 * d_1}{h_2 * d_2}$, where h_1 and h_2 denote the two heights of the reference person in the two frames. A linear interpolation is applied to compute the remaining weights between the two extreme lines. Finally, the weights W_p are assigned according to the y-coordinate of each foreground pixel.

After perspective normalization, the total number of foreground pixels in each frame i under analysis is updated as follows:

$$FeatN_i = \sum_{y=1}^{Y} W_p(y) * N_T(y) \qquad (1)$$

Where $N_T(y)$ is the total number of foreground pixels in the y^{th} row.

C. Corner density estimation for crowd measurement

In addition to perspective distortions, the foreground pixel counts are also extremely sensitive to the "crowdedness" (level of the crowd density). When people are closer to each other, less foreground pixels are extracted due to the partial occlusions that occur. Therefore, we intend to estimate the density of people by measuring how close local features are.

1) Features from accelerated segment test: For local features, we extract *features from accelerated segment test* (FAST) [20]. FAST is developed for corner detection in a fast and a reliable way. It depends on wedge model style corner detection. Also, it uses machine learning techniques to find automatically optimal segment test heuristics. The segment test criterion considers 16 surrounding pixels of each corner candidate P. Then, P is labeled as corner if there exist n contiguous pixels that are all brighter or darker than the candidate pixel intensity.

The reason behind applying FAST as local features for crowd measurement is its ability to find small regions which are outstandingly different from their surrounding pixels. The selection of this feature is also motivated by the work in [21], where FAST is used to detect dense crowds from aerial images. The derived results in [21] demonstrate a reliable detection of crowded regions using FAST.

2) Crowd measurement: The objective of extracting FAST is to handle the problem of variations in crowd density, but without involving explicit estimation of the crowd level. This inspires us to search for a way that can directly weight the feature defined in (1). Therefore, we propose to synthesize FAST local features for a global corner density by computing the ratio between the number of corners and the number of foreground pixels. Then, we aim at formulating a weighting function by using the corner density as a crowd measure. Precisely, our goal is to weight the proposed feature defined in (1) by inflating its value in high crowd situations, while reducing it in low crowd situations. Thereby, we use the estimated corner density values d_i, $i = 1...M$, where M is the total number of frames for the video sequences. And we define the weight function as:

$$W_d(i) = \frac{d_i - \mu}{\sigma_{max}} + 1 \qquad (2)$$

Where $\mu = \frac{1}{M}\sum_{i=1}^{M} d_i$ and σ_{max} is the maximum of standard deviation values σ_i.

This weight function ensures crowd normalization. It is achieved by setting $W_d = 1$ if the crowd is medium $(d_i = \mu)$, $1 < W_d \leq 2$ if the crowd is high, and $0 \leq W_d < 1$ otherwise. Consequently, to take into account the effects of the crowd on the extracted foreground pixels, our proposed feature defined in (1) is again updated as follows:

$$FeatN_i = W_d(i) * \sum_{y=1}^{Y} W_p(y) * N_T(y) \qquad (3)$$

D. Gaussian Process regression

Our proposed frame-wise feature defined in (3) has been formulated to be invariant to perspective and to crowd density. This could ensure the linearity of the trainable function mapping the feature to the number of persons. In order to have more flexibility, we suggest to consider any eventual errors that could occur in the crowd segmentation or in any other step of our counting system. Therefore, we propose to use Gaussian Process (GP) regression which is well adopted for linear features with local non-linearities (more details about GP can be found in [22]).
The entire system architecture is illustrated in Fig. 2.

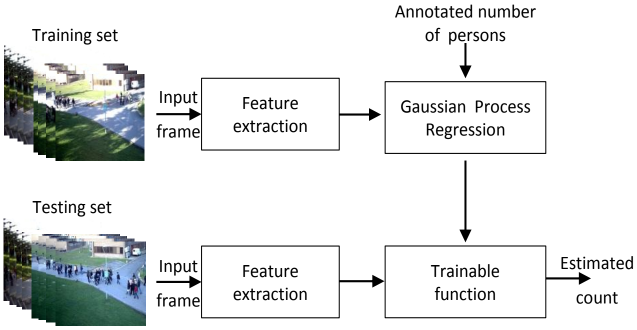

Fig. 2. *Flowchart of the proposed counting system*

In this chart, there are two flows: the training and the testing flows. In the training flow, the trainable function is learned from a set of labeled examples by using GP regression. Once the trainable function is estimated, the number of persons could be predicted from the value of the feature for each frame under analysis in the testing flow.

IV. Experimental Results

In this section, we present the experimental results on the PETS 2009 public dataset [1] to evaluate our proposed approach for people counting. From this dataset, the section S_1 is used to assess *Person count and Density estimation* algorithms. Only 4 videos from the first view were tested in people counting contest held in PETS 2009. Since more tests under situations with serious perspective distortions and occlusions are required to evaluate our proposed approach, we also employ other videos from the second view in our experiments. The main characteristics of these videos are summarized in Table I.

			Number of people	
Video Sequence	View	Length	Min	Max
S1.L1.13-57	1	221	5	34
S1.L1.13-59	1	241	3	26
S1.L2.14-06	1	201	0	43
S1.L3.14-17	1	91	6	41
S1.L1.13-57	2	221	8	46
S1.L2.14-06	2	201	3	46
S1.L2.14-31	2	131	10	43
S3.MF.12-43	2	108	1	7

TABLE I
Characteristics of 8 sequences from the PETS 2009 dataset used for the counting experiments.

Additionally, the counting regression function is learned from a training set built by other videos from section S_1. The training frames are carefully selected to guarantee different situations in terms of number of persons, distance and crowd density. The ground-truth of the count is obtained by annotating the number of persons by hand in every 5^{th} frame. Linear interpolation is applied to count the number of persons in the remaining frames.

To assess our people counting approach, we compare the estimated number of persons to the ground truth using the Mean Absolute Error (MAE) and the Mean Relative Error (MRE) metrics which are defined as:

$$MAE = \frac{1}{M} . \sum_{i=1}^{M} |E(i) - G(i)| \qquad (4)$$

$$MRE = \frac{1}{M} . \sum_{i=1}^{M} \frac{|E(i) - G(i)|}{G(i)} \qquad (5)$$

Where M is the total number of frames in a video sequence. $E(i)$ and $G(i)$ denote, respectively, the estimated and the ground-truth number of persons in the $i-$th frame. The MAE metric was used to compare algorithms submitted to the PETS contest. But, the same error could be negligible if the number of persons is high. Therefore, in [12], the authors propose to also use the MRE metric, which relates the error to the number of persons.

For comparisons, unfortunately, we are not able to compare our proposed method to Chan's method [5]. In fact, for their work [23] submitted to PETS 2009, only tests with videos from the first view were provided. Since we are interested to test more challenging videos; our results are compared to the results of Albiol and Conte methods [1], [12] which are reported in [12]. A summary of poeple counting results, with respect to our hand-annotated ground-truth, are given in Table II. In this table, it is shown that the results of [1] are not accurate mainly for videos from the second view. That could justify the incapability of this method to deal with challenging situations. Whereas, the method in [12] proposes to handle perspective distortions and crowd density which are the two

[1] http://www.cvg.rdg.ac.uk/PETS2009/

Video Sequence		Albiol et al.[1]		Conte et al.[12]		Our approach		
		MAE	MRE	MAE	MRE	MAE	MRE	WMRE
View1	S1.L1.13-57	2.80	12.6%	1.92	8.7%	1.78	8.62%	7.81%
	S1.L1.13-59	3.86	24.9%	2.24	17.3%	3.16	19.19%	19.66%
	S1.L2.14-06	5.14	26.1%	4.66	20.5%	2.89	37.18%	10.97%
	S1.L3.14-17	2.64	14.0%	1.75	9.2%	1.60	8.53%	6.29%
View2	S1.L1.13-57	29.45	106.0%	11.76	30.0%	3.26	11.61%	8.90%
	S1.L2.14-06	32.24	122.5%	18.03	43.0%	6.83	19.96%	18.07%
	S1.L2.14-31	34.09	99.7%	5.64	18.8%	3.35	14.30%	9.98%
	S3.MF.12-43	12.34	311.9%	0.63	18.8%	2.75	96.83%	54.83%

TABLE II
QUANTITATIVE EVALUATION OF OUR PROPOSED APPROACH COMPARED TO OTHER METHODS

major problems that usually affect the results of feature-based methods. That could also justify the better results of Conte's method as compared to [1].

A comparison of our results with the results of [12] reveals the effectiveness of our proposed approach. As stated earlier, Conte's method [12] is the only work that dealt with the two aforementioned factors, but this approach is still problematic as it is shown in the results. One of the drawbacks of [12] is that it assigns one distance value to each group of persons which is less accurate than processing the perspective normalization at pixel level. It also includes other weakness such as the clustering algorithm which is not well adapted for separating different groups of persons, and the bounding box used to define the boundaries of interest points which fails to accurately delineate that by leaving large gaps. All these problems could amply deteriorate the estimated density. It is also important to note that Conte's method requires three parameters (number of detected points, distance, and density) for each cluster separately, which causes burdensome annotation task.

All these reasons could justify that our proposed approach outperforms the two others methods with respect to MAE and MRE metrics. In particular, the tests with S1.L1.13-57(2) and S1.L2.14-06(2) show the effects of the proposed crowd measure to compensate the underestimation of number of persons due to the dense crowd that occurs at several frames. From the same table (Table II), we notice that some MRE values are not relevant to the error made in the estimation of the number of persons, mainly, this problem arises with S3.MF.12-43(2). This video is characterized by small number of persons (as it is indicated in Table I, the maximum number of persons there is 7), so it is expected that a small error in the estimation could bring to a high value of MRE, but this is not the only reason that makes MRE in this video reaches around 97%. The problem occurs at the first frames of the video with a series of small denominators (the ground truth), where even a singularity error could cause very large changes in MRE value. That is why, we propose an other metric called Weighted MRE (WMRE), where the ground truth is replaced by the average of all ground truth values. Hence, the distortions in MRE values are smoothed out using WMRE metric, as it is depicted in the last column of Table II.

The best results that we obtained compared to [1], [12] are thanks to many factors. First, we want to demonstrate the effectiveness of the foreground segmentation method. One of the problems that we faced using PETS dataset is the moving grass that occurs in several frames. GMM succeeds to handle this problem, but at the same time, adapting more variations in the background yields to decline in the detection rate. Since we applied an integration of GMM background subtraction with motion information into a single framework, better segmentation of the scene into foreground and background entities is achieved and it is expected to bring a good performance to people counting.

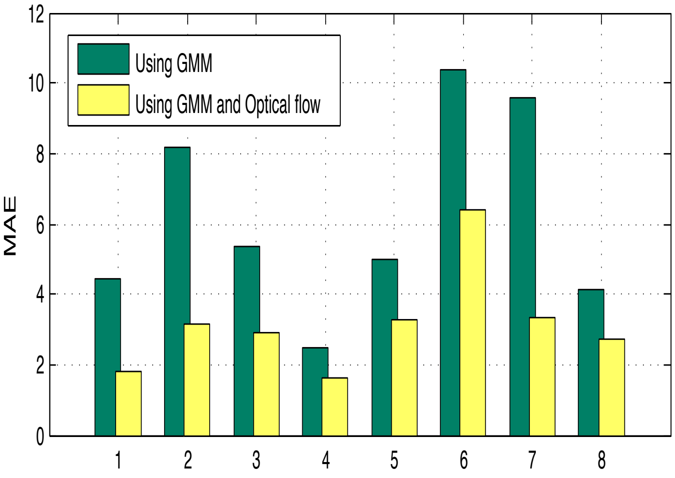

Fig. 3. *Improvement made by using motion cue with GMM background subtraction*

In Fig.3, we demonstrate the impact of the foreground segmentation step on the accuracy of people counting results by comparing MAE metric for the 8 videos (ordered by the same way as in Table II) between applying the improved GMM [18] and the integration of the improved GMM with motion cue [17]. This comparison highlights an overall performance using the approach presented in [17].

Likewise, we prove the effectiveness of our proposed approach by showing that the two normalizations (perspective normalization and crowd density normalization) could significantly improve the accuracy of the counting results, see Fig. 4.

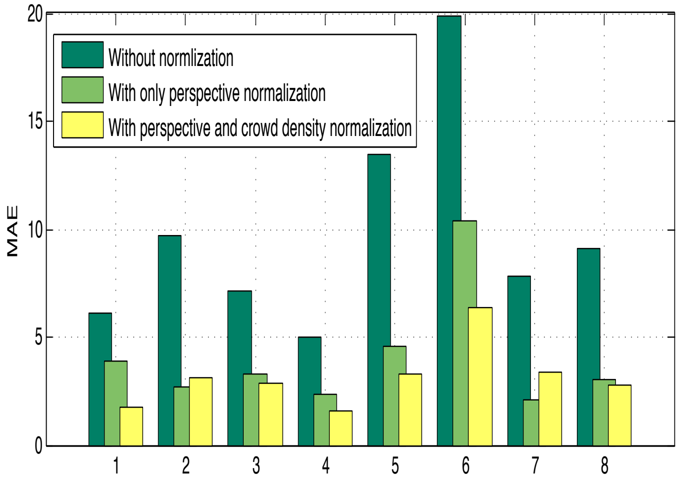

Fig. 4. *Improvement made by normalizing the foreground pixels counts against perspective distortions and crowd density variations*

V. Conclusion

In this paper, we propose a people counting approach, which is important for crowd monitoring in intelligent visual surveillance systems. Additionally to the problem of perspective distortions which is widely addressed in the literature, we handle the problem of crowd density variations in a slightly different way by formulating a new weight function based on corner density estimation for crowd normalization. Our proposed approach consists of regressing a single frame-wise feature independent from variations of perspective and crowd density. Experiments on PETS dataset demonstrate that our approach achieves good results under situations of heavy occlusions and important perspective distortions. By means of comparisons with other existing feature-based methods, our results demonstrate the ability of our approach to improve significantly the counting accuracy. Also, we show other experiments that highlight the role of the two normalizations and the integration of motion cue with GMM background subtraction as well.

References

[1] A. Albiol, M. J. Silla, A. Albiol, and J. M. Mossi, "Video analysis using corner motion statistics," in *IEEE international Workshop on PETS*, 2009, pp. 31–37.

[2] N. Paragios and V. Ramesh, "A mrf-based approach for real-time subway monitoring," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 1034–1040.

[3] G. Acampora, V. Loia, G. Percannella, and M. Vento, "Trainable estimators for indirect people counting: A comparative study," in *FUZZ-IEEE*, 2011, pp. 139–145.

[4] A. B. Chan and N. Vasconcelos, "Bayesian poisson regression for crowd counting," in *IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 545–551.

[5] 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.

[6] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *International Conference on Computer Vision (ICCV)*, vol. 1, 2005, pp. 90–97.

[7] S. Lin, J. Chen, and H. Chao, "Estimation of number of people in crowded scenes using perspective transformation," in *IEEE Trans. System, Man, and Cybernetics*, vol. 31, no. 6, 2001, pp. 645–654.

[8] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," in *International Conference on Pattern Recognition*, 2008, pp. 1–4.

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

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

[11] 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. 1–8.

[12] D. Conte, P. Foggia, G. Percannella, F. Tufano, and M. Vento, "A method for counting people in crowded scenes," in *International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2010.

[13] Y. Hou and G. Pang, "Automated people counting at a mass site," in *IEEE International Conference on Automation and Logistics*, 2008, pp. 464–469.

[14] R. Ma, L. Li, W. Huang, and Q. Tian, "On pixel count based crowd density estimation for visual surveillance," in *IEEE Conference on Cybernetics and Intelligent Systems*, 2004, pp. 170–173.

[15] S. Srivastava, K. K. Ng, and E. J. Delp, "Crowd flow estimation using multiple visual features for scenes with changing crowd densities," in *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2011, pp. 60–65.

[16] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1999, pp. 246–252.

[17] H. Fradi and J. L. Dugelay, "Robust foreground segmentation using improved gaussian mixture model and optical flow," in *International Conference on Informatics, Electronics Vision*, 2012.

[18] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *International Conference on Pattern Recognition*, 2004, pp. 28–31.

[19] G. Farneback, "Two-frame motion estimation based on polynomial expansion," in *Proc. of 13th Scandinavian Conference on Image Analysis*, 2003, pp. 363–370.

[20] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, pp. 105–119, 2010.

[21] M. Butenuth, F. Burkert, F. Schmidt, S. Hinz, D. Hartmann, A. Kneidl, A. Borrmann, and B. Sirmacek, "Integrating pedestrian simulation, tracking and event detection for crowd analysis," in *ICCV Workshops*, 2011, pp. 150–157.

[22] C. E. Rasmussen and C. K. I. Williams, "Gaussian processes for machine learning," December 2006.

[23] A. B. Chan, M. Morrow, and N. Vasconcelos, "Analysis of crowded scenes using holistic properties," in *IEEE International Workshop on PETS*, 2009.