Crowd Counting Estimation in Video Surveillance Based on Linear Regression Function

Maying Shen[1, 2]
1 School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
Shanghai, China

Tanfeng Sun[1, 2], Xinghao Jiang[1, 2], Ke Xu[1, 2]
2 National Engineering Lab on Information Content Analysis Techniques, GT036001
Shanghai, China

Abstract—**Nowadays the estimation of crowd density and people counting is a great focus under public security affairs in surveillance. A scheme for crowd counting estimation based on linear regression function is proposed in this paper. In the proposed scheme, the Light Effect Suppression Model (LESM), which effectively reduces the sensitivity to illumination change, is applied to extract the foreground. Besides the number of foreground pixels, a novel feature called Oriented Inner Edges (OIEs) is proposed in this paper to deal with the problem of occlusion. Moreover, a new method of perspective normalization is applied during the feature extraction procedure to reduce the deviation caused by perspective distortion to the greatest extent. The final experimental results show that the proposed scheme is effective and feasible.**

Keywords- foreground extraction; Oriented Inner Edges (OIEs); perspective normalization; regression estimation; density estimation

I. INTRODUCTION

As the urban population grows, population monitoring and management are facing enormous challenges and pressures. Stampede, crowd unrest and other security incidents are common these days, so there is a great demand for monitoring the crowd density or quantity timely in intelligent video surveillance system.

There are already many research results for crowd density estimation in the field of intelligent video surveillance analysis. From the view of final result, the existing algorithms usually can be divided into quantitative statistics methods and qualitative estimation methods. Quantitative statistics method aims to get the exact number of people. It includes methods that directly count number by identifying parts of bodies, and methods that estimate density by trajectories analysis with feature tracking[1]. In addition, some researchers approximate the counting by finding the linear relationship between the local low-level features and the ground truth[2, 3]. In general, it is still a great challenge to reduce the impact of occlusion for quantitative counting. Qualitative estimation method usually divides the crowd density into several levels, such as “low”, “median” and “high”. In these methods, a classifier such as Support Vector Machine (SVM) is always applied to make a distinction between different levels of density[4]. Considering the implementation technology, the algorithms of density estimation can be roughly divided into two categories. One is pixel-based methods, based on the fundamental idea that the number of foreground pixels is in direct proportion to the population[5,6]. The other is textual-based methods, based on the theory that the texture of crowd image is fine mode when the crowd density is dense while it is coarse mode when the density is sparse on the contrary[7]. Among the above existing methods, human detection and tracking framework doesn’t work well when the crowd is dense. Compared to the textual-based method, the pixel-based method can obtain a more accurate counting number, but it faces the difficulty of foreground extraction and the problem of perspective distortion, which makes it hard to get a high accuracy. There are also some researchers combining multiple methods mentioned above together to achieve a better performance[8]. With the development and wide use of deep learning, researchers have applied convolutional neural network to model the crowd and approved that it has better capability for describing crowd scenes than other hand-craft features[9]. However, these methods usually have high computation cost.

In this paper, a scheme of pixel-based regression analysis is proposed to estimate the crowd counting. The superiority of this scheme is that the Light Effect Suppression Model (LESM) is used to extract the foreground and it applies a novel feature called Oriented Inner Edges (OIEs) to indicate the occlusion area. Besides, a new method of perspective normalization is applied to furthest reduce the impact of perspective distortion. The scheme can achieve real time with the low computation cost. The proposed scheme mainly aims to estimate the crowd counting in the pedestrian area where people walk all the time or stand for just a while. The experimental results have shown the effectiveness of the proposed scheme. The rest of the paper is organized as follows. In section II we give an introduction to the proposed scheme. Section III shows the experimental results and the conclusions are drawn in Section IV.

II. PROPOSED SCHEME

A. Overview

The proposed scheme is introduced in this section. It is separated into three main modules. Fig. 1 shows an overview of the proposed scheme. The first step is to establish the Light Effect Suppression Model (LESM) and use the model to detect the moving foreground. Then calculate the number of foreground pixels and Oriented Inner Edges (OIEs) as the

978-1-5090-3710-0/16/$31.00 ©2016 IEEE

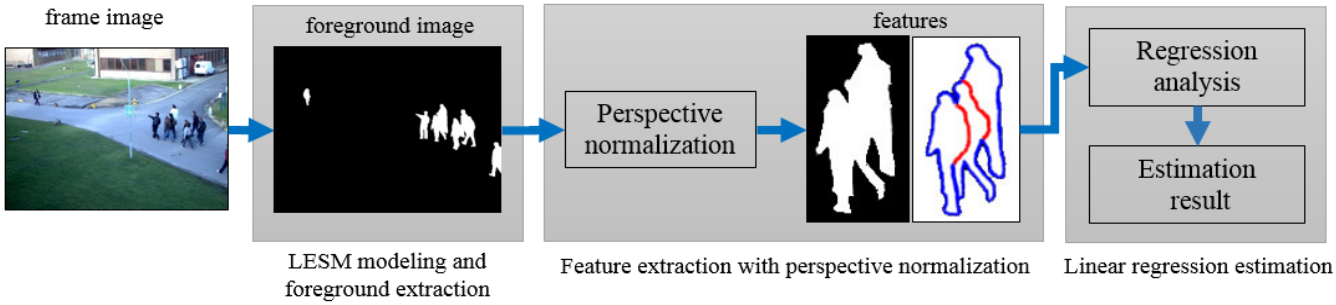


Fig.1. Overview of the proposed scheme

crowd features under the perspective normalization. During the training phase, the input is the video sequences selected for regression modeling, and the regression analysis is carried out between the ground truth and the extracted features to model the crowd density. During the test phase, the crowd counting then can be estimated by calculation based on the linear equation prior trained.

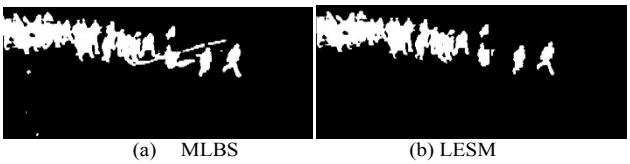
B. LESM Modeling and Foreground Extraction

In this paper, the LESM is a model that combines Multi-Layer Background Subtraction (MLBS) in [10] and Multi-Frame Difference Method (MFDM) in [11], following the shadow processing. The method of MLBS is robust and performs well in foreground extraction, but we found that sometimes it misjudges the background pixels as foreground because of the complex texture. Moreover, the extracted foreground always contains the areas of shadow when under the situation with strong sunlight. To address these problems, LESM is proposed in this paper and it is effective in dealing with the complex background texture and in reducing the impact of illumination change.

For a coming frame image, it is firstly processed by MLBS and MFDM separately. Two foreground images are then obtained after the process and each pixel owns two detection results. The pixel would be regarded as a foreground one if and only if it belongs to foreground both in the two detection results. By applying this strategy, a much cleaner foreground is extracted especially in the area with complex texture. As for the shadows under sunlight, it can be found that they share a common characteristic that they are always long and locate in an oblique direction. Thus, the connected components are detected on the extracted foreground. The shape and direction of each component are analyzed according to its outer contour. More specifically, detect the contour of the connected component and calculate the inclination angle with the positions of points on the contour. We note the position of the left-most point of a connected component as (x_l, y_l) and position of the right-most point as (x_r, y_r). The inclination angle and aspect ratio of this component are then calculated by (1) and (2) separately.

$$Angle = \arctan\frac{y_r - y_l}{x_r - x_l} \tag{1}$$

$$Ratio = \frac{|y_r - y_l|}{|x_r - x_l|} \tag{2}$$



(a) MLBS (b) LESM

Fig.2. Comparison of foreground extraction result.

It is believed that a pedestrian is in an upright position and owns an aspect ratio of 2:1. Thus, when the detected inclination angle is away from vertical and meanwhile the aspect ratio is an unexpected number, the component is considered as a shadow and is eliminated from the foreground image. Fig. 2 shows the comparison of foreground image with MLBS and the proposed LESM. It is shown that LESM indeed helps to eliminate the shadow area from the foreground image.

C. Features Extraction with Perspective Normalization

In our scheme, the number of foreground pixels is selected as the basic crowd feature. Ideally, the feature is proportional to the crowd counting. To reduce the impact of occlusion, the concept of OIEs is proposed and the number of OIEs is chosen as another crowd feature in this paper.

1) Oriented Inner Edges

Inspired by the inner edge in [6], the OIEs are used to indicate occluded areas between pedestrians. The inner edge in [6] does help to improve the algorithm performance, but it always contains the edges caused by the colorful dress. Since the inner edges we concern about are generated when some parts of a person's body are blocked by someone else, the edges are always lines that close to the vertical. Thus we improve the inner edge in this paper by calculating the orientation of each edge and checking whether it is a vertical line.

The gradients of each pixel in x-direction and y-direction are calculated separately on the origin image, and the foreground is applied as a mask to filter out the background pixels. Each foreground pixel owns two components of gradient after gradient calculation. Then the amplitude $A(x, y)$ and orientation $O(x, y)$ are obtained by (3).

$$A(x,y) = \sqrt{dx^2 + dy^2}, O(x,y) = dy / dx \tag{3}$$

where dx and dy are gradients in x-direction and y-direction separately. Set a threshold on the amplitude to determine whether point (x, y) belongs to an edge since the gradient

values of the point on an edge are usually higher than values of other points. As introduced in [6], both sides of the inner edge belong to foreground region while one side of the non-inner edge belongs to background. Then the OIEs are defined as a set of the points on the inner edges and whose value of *O(x, y)* is close to zero. As shown in Fig. 3, where the area filled with number "0" is the background area and that filled with "1" belongs to foreground on the contrary, and the red arrows indicate the orientation of the pixels. According to the definition, the OIEs in Fig. 3 are the edges that are marked as red.

II) Perspective Normalization

There is always a phenomenon that the pedestrian looks much larger when it is close to the camera and has more pixels in the foreground due to the various angle of camera and the different distances between the targets and camera. In this paper, to compensate the difference caused by perspective effect, every foreground pixel is endowed with a weight based on the principle that the further the pixel is, the larger weight it has. Mark a pedestrian firstly both in the closest positon and the furthest position in the image sequences using rectangles. Measure the length and width of the borders, then note them as $\overline{AC}$ and $\overline{AB}$ in the far end, $\overline{A'C'}$ and $\overline{A'B'}$ in the near end. There is a linear projection relationship between them, which is shown as a perspective map like Fig. 4(a) where O_1 and O_2 represent the center of the two pixel groups respectively. Choose an arbitrary point and note it as O_3. Extract the profiles which across through O_3 and label the vertices as shown in Fig. 4(b) and Fig.4(c).

To reduce the error of manual measurement, the value of $\overline{AB}/\overline{A'B'}$ and $\overline{AC}/\overline{A'C'}$ is always determined by the average value of several measurements of different individual targets.

Then the width and height of the pixel group, whose center is O_3, can be calculated by geometric formula of trapezoid as (4) and (5).

$$\overline{EF} = \frac{1}{m_1 + n_1}\left(n_1\overline{RS} + m_1\overline{R'S'}\right) \tag{4}$$

$$\overline{GI} = \frac{1}{m_2 + n_2}\left(n_2\overline{PQ} + m_2\overline{P'Q'}\right) \tag{5}$$

where m_1, m_2 are the distances between point O_3 and the furthest point in horizontal and vertical respectively, n_1 and n_2 are the perpendicular distances to the nearest point.

In order to correct the distortion caused by perspective effect, the weights in vertical and horizontal direction are respectively defined as the reciprocal of the height and width. Thus, the final weight of the pixel at point O_3 is calculated by (6), and the weight of every pixel can be obtained in this way.

$$w = \frac{1}{\overline{EF}} \cdot \frac{1}{\overline{GI}} \tag{6}$$

After obtaining the foreground image with LESM and the image of OIEs, the perspective normalization method is applied to compensate the difference of pixels caused by perspective. Since every pixel has its weights in both vertical and horizontal direction, for a target image whose size is M*N, the features are obtained by the following equations:

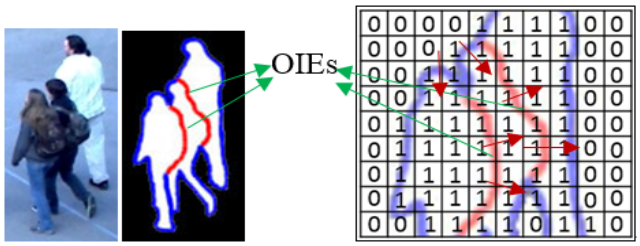
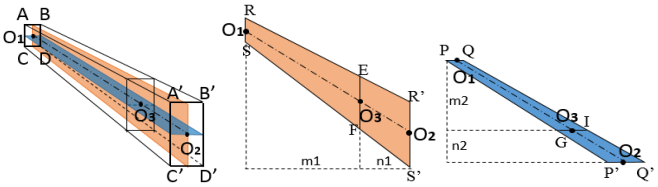


Fig.3. Sketch of the OIEs



(a) (b) (c)

Fig. 4. The sketch for calculation of width and height at point O_3 (a) the perspective map of pixels, (b) height in horizontal direction, (c) width in vertical direction.

$$L(i,j) = \begin{cases} 0, & (x,y)\ belongs\ to\ FG\ or\ OIEs \\ 1, & otherwise \end{cases} \tag{7}$$

$$n = \sum_{i=0}^{M-1}\sum_{j=0}^{N-1} w(i,j)\cdot L(i,j) \tag{8}$$

where "FG" means foreground, *w(i, j)* is the final weight of pixel *(i, j)*. Then the number of foreground pixels and OIEs can be obtained, denoted as n_{fg} and n_{eg} separately.

D. Linear Regression Estimation

During the phase of training, the regression model is trained with the ground truth and the extracted features of the training set, by using the least square method. The obtained linear regression equation is shown as (9).

$$C_H = f\left(n_{fg}, n_{eg}\right) = b_1 + b_2 \cdot n_{fg} + b_3 \cdot n_{eg} \tag{9}$$

where C_H is the approximate value of crowd counting, and b_1, b_2, b_3 are the trained regression coefficients.

During the testing phase, the crowd features can be extracted from the target image in the same way. And the estimation result can be obtained by taking the extracted features as the input of the trained linear equation.

III. Experimental Analysis and Discussion

A. Experimental Dataset

Several datasets for crowd counting estimation are mentioned in [12]. Among these datasets, the Mall Dataset and Grand Central Dataset are not suitable for our scheme since the proposed scheme mainly aims to estimate the crowd counting

in the pedestrian area where people walk all the time or stand for just a while. By contrast, the PETS2009 dataset[16] is the most appropriate for our scheme. As for the recent PETS benchmarks, PETS2014[17] is mainly used for abnormal detection and behavior understanding while PETS2015 is not available on the website, so they are not used in our experiment. PETS 2009 dataset contains four subsets, each subset contains several sequences and each sequence contains different views. Among these subsets, "S0" is the subset of training data and "S1" is for person count and density estimation, where the density of the crowd in the walkways ranges from sparse to dense under different illuminations, and the distortion of perspective is quite severe. We then choose totally six sequences of view_001 as the test data.

B. Performance Analysis of Our Scheme

Table I shows the estimation results of our scheme. We use mean relative error (MRE), mean absolute error (MAE) and mean square error (MSE) to evaluate the performance. The estimation results are reported in Table I. It is shown that the proposed scheme performs well generally. Compared to the method that only considers the distortion in vertical direction, the values of these three metrics of the proposed scheme are much lower since putting the distortion in horizontal direction into consideration makes the distortion correction more reasonable. Thus, the result shows that our scheme is feasible and it proves that:

1) The proposed crowd counting estimation scheme is also applicable to the situation where crowd is dense, such as crowd in sequences 14-06, 14-17 and 14-31. The obtained MREs are all lower than 10% even the occlusion is quite severe in these situations. It mainly benefits from the use of OIEs, which puts the occlusion situation into consideration.

2) The new perspective normalization method indeed helps to reduce the estimation deviation. It promotes the performance both in accuracy and stability with much lower error values compared to the method that only corrects the perspective distortion in vertical direction.

C. Comparison Analysis

The authors of [13, 14, 15] also introduced algorithms of crowd density estimation and tested them over the sequences of PETS2009 dataset. These algorithms are all pixel-based method. The relationship between the number of SURF points and the crowd counting was considered in the methods of [13] and [14]. These two methods are similar to our method. The authors of [13] took no measures to reduce the impact of perspective, while in [14], the distortion was corrected with the distance between the camera and people. The authors of [15] estimated the counting by SURF point clustering and normalized the perspective only in vertical direction.

The performance comparison of our scheme and the methods of [13, 14, 15] is shown in table II. The results over sequences that the algorithms in the references didn't test over are not listed in the table or marked as "NA". In general, it is shown that our scheme performs well with lower MRE and MAE when compared to the similar methods of [13, 14]. It is because that we choose OIEs to indicate the occlusion and a better perspective normalization method is applied in the proposed method. It can be seen that the MRE of our method is a little higher compared to the value shown in [15] over sequence 14-17. It is mainly because that in our method, the shadows of pedestrians join together to form a large connected area in the foreground image and it is not eliminated, which causes the number of foreground pixels is a little bit higher. However, our method performs much better when tested over sequences 13-59 and 14-06. Specifically, the relative improvement obtained using the proposed method is around 50% on these two sequences. Besides, our results obtained on sequences 14-03 and 14-31, where the crowd is quite dense, also prove the effectiveness of the proposed scheme. Compared with other methods, our scheme performs well in accuracy and stability in general. Fig. 5 represents the estimated results and the ground truth of the tested sequences with respect to time. In each graph, the green line is the estimation result of the proposed scheme, and the red one is the ground truth.

TABLE I. PERFORMANCE OF OUR SCHEME

("p-v" means "correct perspective in vertical only", and "p-vh" means "correct perspective in vertical and horizontal", namely our method)

	p-v	p-vh	p-v	p-vh	p-v	p-vh
	13-57		13-59		14-03	
MRE%	11.67	**5.01**	6.54	**5.90**	44.61	**4.22**
MAE	2.20	1.01	1.01	0.83	5.28	0.50
MSE	1.96	1.14	1.04	0.70	4.96	0.33
	14-06		14-17		14-31	
MRE%	20.14	**9.56**	30.81	**9.58**	23.53	**9.41**
MAE	3.62	1.87	6.34	1.88	7.47	2.64
MSE	9.31	1.56	2.99	1.59	8.17	3.01

TABLE II. COMPARISON WITH DIFFERENT METHODES

	13-57	13-59	14-06	14-17
	MRE% / MAE			
proposed	5.01/1.01	5.90/0.83	9.56/1.87	9.58/1.88
[13]	11.50/2.28	12.70/1.81	NA/NA	NA/NA
[14]	5.81/1.20	11.00/1.39	21.80/5.21	9.60/1.92
[15]	4.97/1.01	9.30/1.17	18.76/4.33	7.28/1.39

IV. CONCLUSION

In this paper, a crowd counting estimation scheme in video surveillance is presented. During the process of the scheme, we introduced LESM, which is applicative under changing illuminations, to extract the foreground. To reduce the impact of occlusion, a novel feature called OIEs is proposed in this paper. Besides, a new method for perspective correction is applied and it helps to improve the estimation accuracy. The experimental results have proved the validity of the normalization theory and demonstrated that the proposed scheme is effective and feasible.

ACKNOWLEDGMENT

Thanks to Prof. Dawen Xu for giving his selfless contribution to promote our work. This work was supported by the National Natural Science Foundation of China (No. 61272249, 61272439, 61572320, 61572321). Corresponding

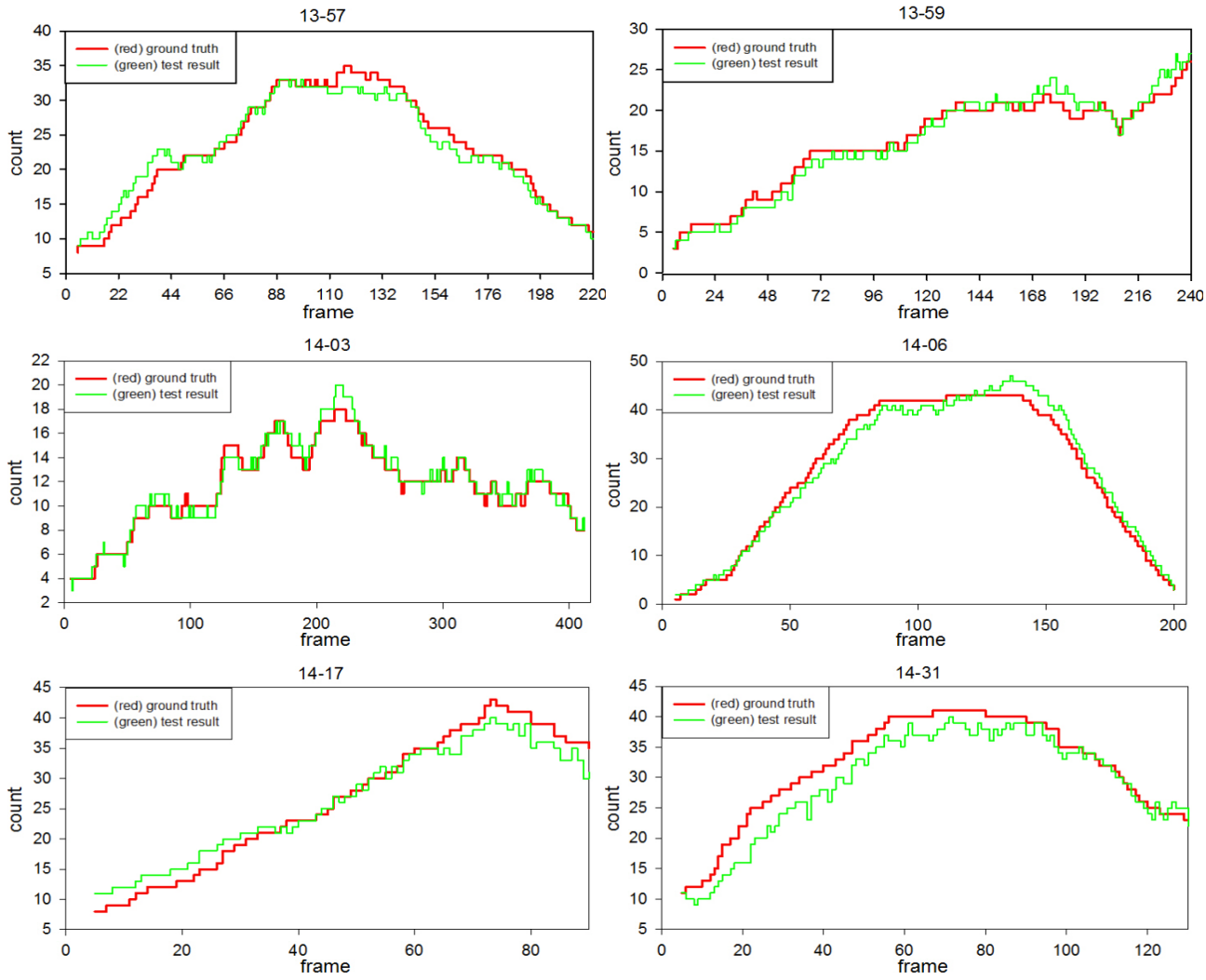


Fig. 5. Estimation result over image sequences of PETS2009

author is Tanfeng Sun, any comments should be addressed to tfsun@sjtu.edu.cn.

REFERENCES

[1] H. Fradi, V. Eiselein, J.L. Dugelay, I. Keller, and T. Sikora. "Spatio-temporal crowd density model in a human detection and tracking framework". Signal Processing: Image Communication, 2015, Vol. 31: pp.100-111

[2] M. Jiang, J.C. Huang, X.Q. Wang, J.F. Tang, and C.M. Wu. "An Approach for Crowd Density and Crowd Size Estimation". Journal of Software, 2014, Vol. 9, Iss.3: pp.757-762.

[3] A.B. Chan, M. Morrow, and N. Vasconcelos. "Analysis of crowded scenes using holistic properties". Performance Evaluation of Tracking and Surveillance workshop at CVPR. 2009: pp.101-108.

[4] J.J. Yang, J. Li, and Y. He. "Crowd Density and Counting Estimation Based on Image Textural Feature". Journal of Multimedia, 2014, Vol. 9, Iss.10: pp.1152-1159.

[5] A.C. Davies, H.Y. Jia, and S.A. Velastin. "Crowd monitoring using image processing". IEE Electronic & Communication Engineering Journal. 1995, vol. 7: pp.37-47.

[6] L. Zhang, T. Deng, Y. Song, and Y. Fan. "Density Estimation of Unidirectional Crowds". 2014 International Conference on Computer, Communications and Information Technology, CCIT-14. Atlantis Press, 2014.

[7] A.N. Marana, S.A. Velastin, L.F. Costa, and R.A. Lotufo. "Estimation of crowd density using image processing". IEEE Colloquium Image Processing for Security Applications, 1997, vol. 11, no. 4: pp.111-118.

[8] H. Idrees, I. Saleemi, C. Seibert, and M. Shah. "Multi-source Multi-scale Counting in Extremely Dense Crowd Images." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 2013: pp.2547-2554.

[9] C. Zhang, H. Li, X. Wang, and X. Yang. "Cross-scene crowd counting via deep convolutional neural networks." IEEE Conference on Computer Vision & Pattern Recognition IEEE, 2015: pp.833-841.

[10] J. Yao, and J.M. Odobez. "Multi-Layer Background Subtraction Based on Color and Texture", IEEE Conference on Computer Vision & Pattern Recognition. 2007: pp.1-8.

[11] JE. Ha, and WH. Lee. "Foreground objects detection using multipe difference images". Optical Engineering, Apr, 2010, Vol. 49(4): pp.047201.

[12] S.A.M. Saleh, S.A. Suandi, and H. Ibrahim. “Recent survey on crowd density estimation and counting for visual surveillance”. Engineering Applications of Artificial Intellifence. May, 2015, Vol. 41: pp.103-114.

[13] R. Shirine, K. Walid, and G. Hanna. “An Improved Real-time Method for Counting People in Crowded Scenes Based on a Statistical Approach”. Informatics in Control, Automation and Robotics (ICINCO), 2014 11th International Conference on. IEEE, Sept. 2014, Vol. 2: pp.203-212.

[14] D. Conte, P. Foggia, G. Percannella, F. Tufano, and M. Vento, “Counting moving people in videos by salient points detection”[C], 20th International Conferenceon Pattern Recognition (ICPR), Aug, 2010, pp.1743–1746.

[15] R.H. Liang, Y.G. Zhu, and H.X. Wang. “Counting crowd flow based on feature points”. Neurocomputing, 2014, Vol. 133: pp.377-384.

[16] PETS2009: http://www.cvg.reading.ac.uk/PETS2009/

[17] http://www.pets2014.net/