People Counting System with C-Deep Feature in Dense Crowd Views

Htet Htet Lin
University of Computer Studies, Mandalay
Myanmar
htethtet.linnnnn@gmail.com

Kay Thi Win
University of Computer Studies, Mandalay
Myanmar
kthiwin11@gmail.com

***Abstract*—People counting in a crowded scene is an urgent and vital task of monitoring the surveillance systems. Accrual guesses of a dense crowd views are effected from a different illuminations and inter-class variations, so comes to be a complicated issue and still remains as an active research area. To tackle this fact, this paper establishes an effective people counting framework in dense crowd views that automatically appraisals the accurate number of people. In this paper, a new intuition Color Deep system which utilizes based on the color-based feature and convolutional neural network (CNN)-based feature is proposed for detecting and estimating the people numbers. Unlike the other, this paper proposes C-Deep feature by contributing the color transformation matrix and segmentation. Firstly, the color transformation matrix is introduced and then C-Deep features is calculated by using the Deep CNN with color feature matrix to handle the occlusion, inter-class variations and density levels. Calculation experiments on the challenging public crowd counting dataset achieve the lowest miss rate than state-of-the-art results. This shows the effectiveness of the proposed framework.**

Keywords—C-Deep Feature, Deep CNN, Crowd Counting

I. INTRODUCTION

The population of the world is emerging day per day. Therefore, monitoring images or videos becomes very important when monitoring the security systems. The number of people in the scenario with a heavy cluster is a vital task in various applications, such as visual surveillance, operational, traffic and safety monitoring processes. Counting many people, especially in a large, dense crowd area is very essential and an active area of research. Because the person movement in dense crowd clusters foundation can cause an unwanted or accident dangerous happenings and situations. Thus crowd analysis task is still active field of research. It should create a system for monitoring and maintaining the safety of personnel in many sectors (personnel detection and monitoring, that is, fall detection, action recognition, tracking, etc., as well as bandwidth capacity, that is, count people space for disaster, audience counting, student attendance counting system, personages counting system, meeting room management, etc.) .

According to the literature, they were divided into three groups: counting by detection based model, counting by regression based (map based) model and counting by density estimation based model. Counting by detection approaches occupy training the various object detectors to find person individually and count each person in the scenes. Counting by regression techniques on the other hand attempt to study the automatically related mapping of low-level features with the overall number of person in the scenes or within a region of that scene. Counting by feature regression techniques attempt to study the centroid distances among objects. Individual people are not explicitly detected or tracked in the rest two approaches, meaning visual occlusions and illuminations have less impact on counting accuracy. Due to the lack of varied training data, regression-based techniques have suffered greatly from over-fitting in the past although the computational efficient is usually higher than the computational efficient of the detection methods.

In addition, existing methods based on patch-based or the whole image. They also extract local or global features. These features included feature fusion, frequency domain analysis, interest points or edge orientations were used in people counting. But the performance of detectors was inability in complex and noisy scenes due to appearances inter-class intra-class variations, heavy occlusion, illuminations and various perspective distortions. Illuminations and variations are the most important issues. This fact motivates to learn a framework that can avoid illuminations and can handle the inter-class variations to take the task of counting in crowd scenes especially under heavy occlusion condition. Given an input data, the two main objectives of the proposed system are to find out more significant and distinctive features and to obtain the accurate person numbers.

The proposed framework first transforms the color space to extract an accurate foreground color detection. Then, the color feature matrix is extracted to avoid illuminations. This highlights the first point contribution of the proposed system. For an accurate counting, the proposed system focuses on Deep CNN for attaining a new insight, C-Deep features. This also highlights the second major core contribution point. Deep CNN can also be used as the classifier for attaining the decreasing estimation errors by taking a color transformation of all the images as the input and then automatically output the counting results.

This paper also shows the distinct progress that has been achieved by proposed methods. This paper organizes as follows. Section 2 presents the earlier methods and technologies of the crowd counting system. Section 3 describes the two main phases of the proposed system and also explains the detail of step by step process. Section 4 investigates and

978-1-5386-5892-5/18/$31.00 ©2018 IEEE
ICIS 2018, June 6-8, 2018, Singapore

discusses the experimental results compared with state-of-art results. The rest section, Section 5 has been attempted the conclusion and concerns.

Fig. 1. Example images of PET 2009 dataset. Challenges include severe occlusion, clutter and similar appearance of people

II. Related Work

A. Counting by Detection Based Model

Most of the state of art emphasized on detection framework that performed features (such as Haar wavelets, histogram oriented gradients, edgelet and shapelet) extracted from the entire body or part of the body [16, 5, 18, 15]. Then these features are trained into various classifiers (such as adaboost boosting, random forest and Support Vector Machines) [17]. These methods have only four or five people on a low density scenes that receive high performance on that scenes. Although they used part based or shape based detector, they do not mitigate the problem of occlusion, illuminations and not suitable for crowded scenes.

B. Counting by Regression (Map-Based) Based Model

To overcome the problem of occlusion, most of the researchers emphasized on regression or mapping. They suggest calculating the regression-based method of learning the mapping comparison between the actual counts with features extracted from local image patches [14, 4]. They can independent on learning detectors that a relatively complex task. In these frameworks, various classifiers are used (such as linear regression, piecewise linear regression, ridge regression, Gaussian process regression, and neural network) [8] to study low-level feature with crowd counts.

Their framework focused on the concept of discriminatory attributes used to solve sparse training data. Their methods can effectively handle unbalanced data. But, they have less estimate of people region in real time video.

C. Counting by Density Estimation Based Model

Most of the state-of-arts observed that most of the regression based methods ignored the global spatial information. Lempitsky et al [9] presented a linear framework by learning a mapping between object density maps with corresponding local patch features. Their issue is to optimize the convex plane of cutting. Pham et al. [12] presented a non-linear framework by using random forest to map local patch features with density maps. The generating of high quality density and error maps with an estimation error will be another important issue that must be addressed in the future. This model is the most sophisticated and time consuming among three models.

In order to tackle the problem of inter-class variations and develop low count estimation error, this paper proposes to establish the most accurate people counting framework using color feature segmentation of the foreground area and learning C-Deep feature.

III. Proposed Approach

This paper describes the framework of the people counting, where the input is a video or image and the result is an image, and each image is labeled by the number of people count. The quality measure of the output is based on count error rate techniques. In the experiments, the proposed system introduces a lower mean absolute error rate and mean relative error rate to capture our concerns.

A. Overview

The system overview consists of two main phases: Color Transformation and Deep CNN Framework as in Fig. 2. The step by step process is explained more detail in Section 3.2 and 3.3. The original image is firstly set and then transformed the color space on each input frame into Hue-Saturation-Intensity (HSI) to avoid illuminations. These color matrix features are extracted the gradients to the corresponding color information. Unlike [6], this paper focused most cells in the overlapping blocks in the detecting window needs to remove the redundant operations. After the color matrix is clearly calculated, these matrixes have putted into the input of Deep CNN approach. Then C-Deep feature is extracted by combining the color matrix and Deep CNN feature. Finally, Deep CNN is also used as the classifier for attaining the decreasing estimation errors by taking a color transformation of all the images as the input and then automatically output the counting result. As the proposed system is a color-based Deep CNN method, it intends to avoid the illumination, occlusion and noise issues. Unfortunately, the shortcoming exist in cases of non-human objects appearing in the scene, the system may overestimate the crowded size.

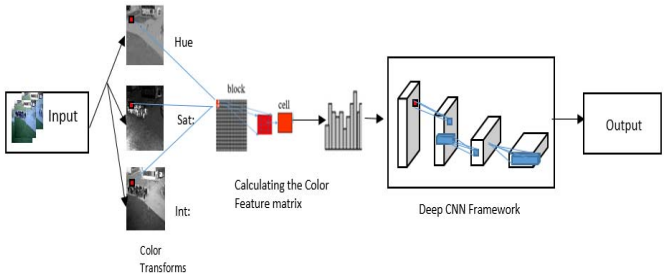

Fig. 2. Overview of the proposed system

B. Color Transformation

Given the video, the input image is originally RGB color space. Although these three color channels represent the value of red, green and blue descriptors. These tri-tuples are not suitable for the retrieving objects because it is not well for the

representation of feature extraction. Problem of illumination and noise are observed in the RGB color space, which is a serious problem for the researchers. To tackle this fact, the proposed system first converts the original RGB input image into HSI color space. Unlike [19], the proposed system takes the entire image frame into an HSI space and computes color matrix features of all three channels.

Fig. 3 shows a graphical representation of the HSI color space. The hue is the directional as well as orientation and the saturation is the size of the gradient as well as magnitude values, respectively. This gives the saturation histograms over hue bins field, which can pronounce the distribution of colors in the images. The matrix of color feature is calculated by the following equation:

$$In = \frac{1}{3}(r+g+b) \tag{1}$$

$$Sa = 1 - \frac{3}{I}\min(r+g+b) \tag{2}$$

$$\theta = \cos^{-1}\left[\frac{\frac{1}{2}[(r-g)+(r-b)]}{\sqrt{(r+g)^2+(r-b)(g-b)}}\right] \tag{3}$$

$$Hu = \begin{cases} \theta & g \le b \\ 2\pi - \theta & g \ge b \end{cases} \tag{4}$$

where In is the pixel value of the intensity of the image object, Sa is the saturated pixel value, r is the red pixel value, g is the green pixel value, b is the blue pixel value, and θ is the angular value to calculate the hue value Hu.

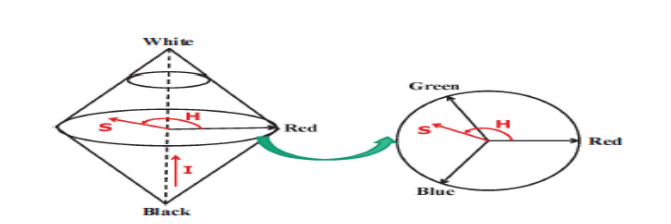

Fig. 3. Graphical representation diagram of Hue-Saturation-Intensity color space

After converting the color space and obtaining the HSI pixel values, the proposed system calculates the matrix of color feature. This is due to differences in the appearance and form of the local person, which is due to the local distribution of the intensity gradient or distribution along the edge, and they do not have a special facts that relates the positions of the gradient or the edges. The color gradient feature matrix is extracted the gradients from the whole frames of images. This converted HSI input is firstly applied with one dimensional derivative mask. These mask values are -1, 0, and 1. The values of magnitude and angle of each pixel are calculated by the following equation:

$$Mag = \sqrt[2]{(gra\_x)_2 + (gra\_y)_2} \tag{5}$$

$$A = \arctan\frac{gra\_x}{gra\_y} \tag{6}$$

where Mag is the magnitude value, A is the angle value, gra_x and gra_y are the pixel values in the horizontal and the vertical gradient, and arctan is the tangent inverse of the gradient values. The proposed system accumulates gradient in the corresponding cell, marks the histogram value on each cell, and normalizes the histogram along the four directions. Unlike the histogram of other gradient approaches, this paper normalized the cell feature along the four sum together, instead of reducing the dimensional feature vector to one-forth. These detected people gradients are grouped to produce the matrix of color feature located throughout the human body. Then, these are fed into Deep CNN framework. This highlights the first contribution.

C. *Deep CNN Framework*

CNN is one of the most popular types of deep neural networks as well as deep learning. They eliminate the need for manual feature extraction. The previous works of CNN are directly extracted features from input images. According to the literature related to CNN, the main points of two broad categories are Network based model and Training Process. The first category is focused on the networks model such as hydra, VGG-16, GoogLeNet, AlexNet, and ResNet. The basic model is used the initial CNN layer. Scale-aware model used many columns or many resolution architectures. Context-aware model fused local and global information. Therefore, their evaluation error is lower. Multi-task learned various vision tasks and different approach estimated the crowded counting and density. But the weakness of the first category is a strict dependence on hardware and network models. Thus, they need high-end graphics processors and a lot of large datasets.

The second category is based on the methodology of patch-based or the whole image. Patch-based cropped of the input image by using different approaches and whole-based used the whole image. They avoid computationally expensive sliding window but the issue is highest average absolute error rate due to the learning of the whole image (both people region and non-people region). This fact inspired the proposed system.

Motivated by the success of CNN for various computer vision tasks, various approaches have been developed but this paper has developed to join crowd counting and estimation with color transformation with whole-image based inference. This is a distinct factor from other existing works.

C-Deep Feature: In this paper, the color feature matrix is put into the Deep CNN framework to extract the C-Deep features for effectiveness and efficiency of features.

Unlike the other, the framework pre extracted the color feature matrix as matrix proposal. Unlike also the other neural network based work, this paper firstly extracted the color transformation to reduce not only time but also inefficiency

due to illumination issue. This is the highlighted contribution of this paper contract with another. To get pedestrians, the size of each patch in diverse positions is selected according to the perspective value of its center pixel. Then these are distorted into 32 pixels by 32 pixels.

Classifier: The structure of a deep CNN model with switchable targets is shown in Fig. 4. This structure is also the second type of end-to-end method.

In the proposed system, two layers are used as the convolution layers and three layers are used as fully-connected to extract the C-Deep feature. The goal of the extracting this feature is to find out more significant and distinctive features. This paper used four hidden layers. The convolution layer 1 contains 20 filters with size of 5×5×3 and layer2 convolution has 50 counterparts with size of 5×5×26. The corresponding feature map is 6×6×20 and 12×12×30, respectively. The purpose of using Deep CNN as the classifier is to obtain accurate estimates of the count population. For a two-dimensional image H and a convolution kernel of 5×5, is used like as the following equation:

$$G = F \otimes H \tag{7}$$

$$G[i,j] = \sum_{u=-k}^{k} \sum_{v=-k}^{k} H[u,v]F[I+u, j+u] \tag{8}$$

where G is the convolution operator, F is the kernel size and H is the data of input layer.

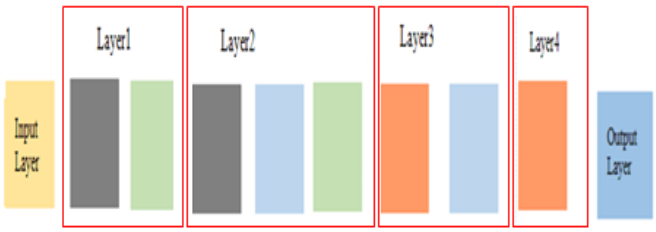

Fig. 4. Deep CNN framework

After localization the receptive fields in the framework, the important weight and biases is calculated. The weight of the four layers perceptron can be arranged into the matrix as the following equation:

$$wei = (\sum_{i=1}^{n} wei_{ij})_{j=1}^{n} \tag{9}$$

where wei is the weight of the proposed system and n is the neurons number of the hidden layer. The corrected rectify linear unit is also calculated as follow:

$$F = \max(0, x) \tag{10}$$

where x is the pixel value. Then 2×2 max pooling is used, as shown in Fig. 5. This is made the representations smaller and more manageable one. It is operated over each activation map independently. Fully connected layers took the high-level filtered images and translate them into votes. The system compute the error rate of loss function as follows:

$$L = -\sum_{j} y^j \log \theta(o)^j \tag{11}$$

where L is loss value, y is true label, o is the last output layer of the network, j is the dimension of vector and ɵ is the probability estimate value.

To avoid over-fitting issue, the system surveys the different values of seek, epochs and batch sizes. As one of the contributions, this paper used the best potion and seed number by analyzing various number range. An epoch of options is important for calculating the number of batches. For time reducing, the system reduces the number of pixel patch. The epoch's number (passes over the training set) is very important. Unlike the traditional neural networks, a tiny region proposal neuron of input layer is only associated to the hidden layer. Therefore, the proposed system can provide efficiency and tolerance.

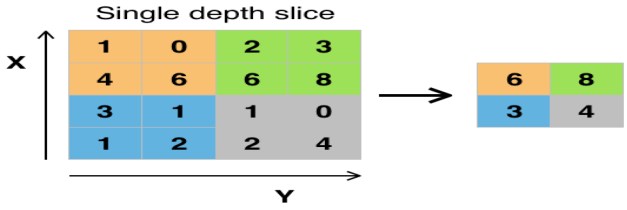

Fig. 5. Max pooling process

IV. Experimental Results

To calculate the framework, this paper evaluates the experiments on the challenging PET 2009 crowd counting dataset [13] that includes serious occlusion, clustering and appearance problem in Fig. 1.

A. Dataset

This paper evaluates the accuracy on the challenging PET 2009 dataset. This consisted of several sections to test the different surveillance scenes. In this experiment, the proposed system applies the "S1" section. This video sequences have good crowded properties and challenging variations such as people walking and running, lighting changes and diverse crowd densities and sizes. This sequence also contains the ground truth annotations [11].

Since there is no formal set of trainings and a set of tests, the proposed system arranges the necessary information. The frames of each test video is used as an annotated subset to train the proposed system, and the rest frame of the video sequences are used as the test set. To compute the effectiveness of the proposed system by comparing the previous method [6], the training frames is almost chosen in the same way as the order of the frames chosen in [6]. In contrast to [6], for the training set specified in Table I, the last frame from every six consecutive frames is selected.

B. Experiment

The proposed system uses MAE (mean absolute error) and MRE (mean relative error or average relative error) to evaluate the system performance. The MAE can clearly give an insight

that the predicted value coincides with the actual situation (ground truth). The MRE, like as relative square error is the ratio of the predictor value to the actual value. This is calculated as using the following equations:

$$MAE = \frac{1}{TN}.\sum_{t=1}^{TN} | h_a - h_b | \quad (12)$$

$$MRE = \frac{100}{TN}.\sum_{t=1}^{TN} | \frac{h_a - h_b}{h_a} | \quad (13)$$

where TN is the total testing frames, h_a is the ground truth and h_b , the result of output layer related to the accrual people number, is the estimated people count in frame t. A highlighted fact is that h_b is more important to exactly quantify the counting error in an entire video.

TABLE I. TEST VIDEOS AND TRAINING SETS

Size and Frame Number	S1.L1.13-57	S1.L1.13-59
Length	221	241
Training frames no.	36	40
Training frames	1:6:221	1:6:241

C. Results

The performances of the proposed system compares the previous different methods on PETS2009 videos are reported in Table II and III. An important fact is that the whole process is necessary for calculation of getting the predicted value. The final performance of the error rate is mainly depend on the predicted value. If the value is close to the situation, the error rate will decrease. The reported result [6, 13, 10, 1, 3, 19, and 2] were noted directly from the corresponding papers.

TABLE II. COMPARISON OF VARIOUS PREVIOUS APPROACHES ON "PET 2009" DATASET (S1.L1.13-57)

Methods	MAE	MRE%
The proposed system	**0.53**	**2.07**
Hashemzadeh [6]	0.89	2.34
Liang et al. [13]	1.01	4.97
Hashemzadeh et al. [10]	1.03	9.31
Albiol et al. [1]	1.72	N/A
Rao et al. [3]	2.17	37.61
Li et al. [19]	1.91	N/A
Chan et al. [2]	2.30	N/A

From the results in Table II and III and IV, it can be clearly seen that in all test videos, the proposed method is achieved by the highest performance. In the contract, the MAE rate was 0.36 which was 0.34% and the MRE was 0.27% that lower than [6]. In addition, 0.48 MAE rate lower than [13], 0.50 than [10], 1.19 than [1], 1.64 than [3], 1.38 than [19] and 1.77 than [2]. The MRE rate is 0.27 less than [6]. Comparing MRE with other, 2.9 below [13], 0.50 than [10], and 35.54 than [19]. The MAE and MRE result of the proposed system are also lower than the previous work in Table III and IV. Table V is the average results of the three Table II, III and IV.

TABLE III. COMPARISON OF VARIOUS PREVIOUS APPROACHES ON "PET 2009" DATASET (S1.L1.13-59)

Methods	MAE	MRE%
The proposed system	**0.50**	**3.89**
Hashemzadeh [6]	0.84	4.47
Liang et al. [13]	1.17	9.30
Hashemzadeh et al. [10]	1.14	12.84
Albiol et al. [1]	1.89	N/A
Rao et al. [3]	1.62	35.41
Li et al. [19]	2.02	N/A
Chan et al. [2]	1.64	N/A

TABLE IV. COMPARISON OF VARIOUS PREVIOUS APPROACHES ON "PET 2009" DATASET (S1.L2.14-06)

Methods	MAE	MRE%
The proposed system	2.00	**7.69**
Hashemzadeh [6]	2.32	8.12
Liang et al. [13]	4.33	18.76
Hashemzadeh et al. [10]	**1.68**	14.23
Albiol et al. [1]	1.95	N/A
Rao et al. [3]	2.47	41.11
Li et al. [19]	3.28	N/A
Chan et al. [2]	4.32	N/A

Based on C-deep feature operation on color transformation and CNN, foreground pixel area is extracted. Deep CNN is also trained to learn a mapping between the extracted features

and the related pixel. Finally, total crowd estimation comes out and it is equal to the summation of the group sizes. In this way, the system can receive the accurate person numbers.

Fig. 6 shows some results, including frame numbers, counting results and the ground truth on all the testing videos. The estimated count number produced by the proposed counting system is almost the same as the ground truth number found in the benchmark PET2009 dataset [11]. However, some frames need correction. This can be seen in this Fig. 6 (a), the correct calculation of estimate count and (b) the calculation of miss estimate count.

TABLE V. COMPARISON OF VARIOUS PREVIOUS APPROACHES ON "PET 2009" DATASET (AVERAGE OF THREE SETS)

Methods	MAE	MRE%
The proposed system	**1.01**	**4.55**
Hashemzadeh [6]	1.35	4.97
Liang et al. [13]	2.17	6.2
Hashemzadeh et al. [10]	1.28	12.12
Albiol et al. [1]	5.56	N/A
Rao et al. [3]	2.08	38.04
Li et al. [19]	2.40	N/A
Chan et al. [2]	2.75	N/A

Fig. 6. Example of the proposed counting results on (*a*) S1.L1.13-57 and (b) S1.L1.13-59

V. CONCLUSION

A robust system for crowded person counting from C-Deep feature is intended to solve the problems of occlusion, illumination and various inter-class variations. The experimental results tested on the challenging crowd dataset are significantly outperformed than the state of art methods. Firstly, input image is converted to avoid noise and illumination. The hue-saturation-intensity color transformation is taken out to get discriminate color features. The color gradient matrix containing people region is extracted by calculating the pixel value in the horizontal and vertical gradients. Finally, these matrix is set to Deep CNN framework to obtain the accurate person number estimations. The proposed system also has a good trade-off between the detection rate and the processing time. As a future work, the proposed system will evaluate on more dense complex crowded datasets.

REFERENCES

[1] A. Albiol, M.J. Silla, A. Albiol, J.e. M. Mossi, "Video analysis using corner motion statistics", in: Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2009, pp. 31-38.

[2] A. B. Chan, M. Morrow, N. Vasconcelos, "Analysis of crowded scenes using holistic properties", in: Proceedings of the Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2009, pp. 101-108.

[3] A. Rao, J. Gubbi, S. Marusic, M. Palaniswami, "Estimation of crowd density by clustering motion cues", in: Vision Computer, 2015, pp. 1533-1552.

[4] Chen, K., Loy, C.C, Gong, S., Xiang, T., "Feature mining for localised crowd counting", in: European Conference on Computer Vision, 2012.

[5] Dalal, N., Triggs, B., "Histograms of oriented gradients for human detection", in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 886-893.

[6] Hashemzadeh, M., & Farajzadeh, N., "Combining keypoint-based and segment-based features for counting people in crowded scenes", in: Information Sciences, 2016, pp. 199-216.

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

[8] Kowcika A. And Sridhar S., "A Literature Study on Crowd (People) Counting With the Help of Surveillance Videos", International Journal of Innovative Technology and Research, 2016, pp. 2353-2361.

[9] Lempitsky, V., Zisserman, A., "Learning to count objects in images", in: Advances in Neural Information Processing Systems, 2010, pp. 1324-1332.

[10] M. Hashemzadeh, G. Pan, M. Yao, "Counting moving people in crowds using motion statistics of feature-points", in: Multimed Tools Application, 2014, pp. 453-475.

[11] Milan, A., in: Milan (Ed.), Data, 2012, http://research.milanton.de/data.html.

[12] P ham, V.Q., Kozakaya, T., Yamaguchi, O., Okada, R., "Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation", in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3253-3261.

[13] Y. Zhu, H. Wang, "Counting crowd flow based on feature points", in: Neurocomputing, 2014, pp. 377-384.

[14] Ryan, D., Denman, S., Fookes, C., Sridharan, S., "Crowd counting using multiple local features", in: Digital Image Processing: Techniques and Applications, 2009, pp. 453-475.

[15] Sime, J.D., "Crowd psychology and engineering", in: Safety Science, 1955, pp. 1-14.

[16] Viola, P., Jones, M.J., "Robust real-time face detection", in: International Journal of Computer Vision, 2004, pp. 137-154.

[17] Viola, P., Jones, M.J., Snow, D., "Detecting pedestrians using patterns of motion and appearance", in: International Journal of Computer Vision, 2005, pp. 153-161.

[18] Wu, B., Nevatia, R., "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors", in: Tenth IEEE International Conference on Computer Vision, 2005, pp. 90-97.

[19] Y. Li, E. Zhu, X. Zhu, J. Yin, J. Zhao, "Counting pedestrian with mixed features and extreme learning machine", in: Cognition Computer, 2014, pp. 462-476.