People Count from the Crowd using Unsupervised Learning Technique from Low Resolution Surveillance Videos

Kowcika A
Assistant professor
RV College of Engineering
Bangalore
akowcika@rvce.edu.in

Abstract
Higher density crowded areas becomes difficult for officials to identify the number of people present within certain area. For the purpose of counting the number of people in an area, image processing concepts are highly supported. Several methods were involved for this counting, but it was capable only to be performed with still images. It is not that much complex to identify the number of people from an image in which it looks visually static. Further numbers of people are counted from the running videos. In this paper we introduce with the identification of people count in a crowd from an ongoing video, in which all the living objects in the video keeps moving. Our proposed work involves with algorithms for background extraction, detection of people, tracking of people and finally counting of the people. We perform background extraction by using ViBE algorithm, then unsupervised learning neural network is executed for detecting people and we use kalman filter for tracking the people. All these advanced algorithms involved here, results with better performances than the state-of-the-art works. Our proposed experimental result obtained by executing our work shows better accuracy results.

Keywords – People Detection, Background Extraction, Counting people, Trajectory analysis, surveillance videos.

1. INTRODUCTION

People counting is significant, that is used for the purpose of traffic management, number of tourists estimation, evacuating people in case if hazards, crowd monitoring,etc,.By performing certain image processing techniques and algorithms, we could identify the presence of people and also count the number of people present in a crowd. This system of counting the number of people was started with the images [1], [2]. In [1] and [2], still images are considered as input, from which the number of people in an image is detected. The image is partitioned into cells and then people are detected within the cell which includes with the use of SIFT features, Fourier analysis, GLCM features and wavelet decompositions[1].Then texture features like dissimilarity, homogeneity, energy and entropy are considered. For determining the numbers of people, many authors have involved with head detection process, since on detecting the number of heads it becomes simpler to perform people count. And also head detection procedures involves with addition of robustness for the entire algorithm.

Deep learning framework was also involved in detection of people in the crowd [2], [3]. Deep Convolution Neural Network is used for the estimation of the density maps. In this procedure three convolutions is applied and then a loss layer is present which reduces the density map and the global count loss. Further the numbers of people are counted from videos [4], [5], [7]. In [4] the authors have used background subtraction and simple-fast tracking algorithm. Simple-fast tracking algorithm is performed based on the construction of virtual lines as entry line and exit line. With these lines, based on the centroid, distances are estimated. Further Merge-Split counting strategy is applied and then the people present are counted.

Statistically Effective Multi-scale Block Local Binary Pattern (SEMB-LBP) features were involved for the determination of the number of people [5]. As per SEMB-LBP, the classifier gives with positive and

978-1-5386-1887-5/17/$31.00 ©2017 IEEE

negative samples. The positive samples are retrieved from targets that are focused and the negative samples were considered to be the extracted background. Further a matching method is used for tracking. People are counted by using head detection methods [7], [8]. Head detection is performed with the formation of skeleton graph. Background subtraction is performed using Adaptive Gaussian Mixture Model followed by the skeleton computing, then head detection and then head pose detection.Head detection is also performed by intersect point based gradient orientation feature [7]. Head of the people is identified by using adaptive boost classifier with soft cascade.The major problem on using this algorithm is that it becomes very complex on using integral images.

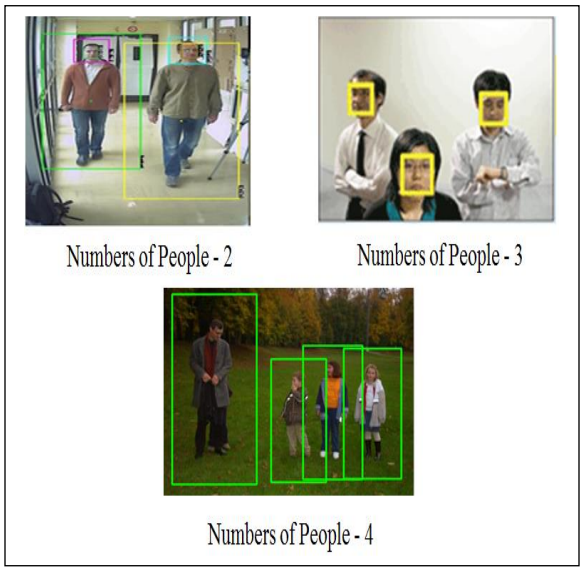

Figure1. People Count

Figure 1 illustrates counting of people from the videos by using several image processing techniques as Head detection, feature extraction, foreground extraction, morphological operation, segmentation, people tracking, classification, etc,. each process is involved with some algorithms and techniques as HOG, Meanshift algorithm, Cascade Ada Boost algorithm, Gaussian Mixture Model, etc,.The numbers of people are estimated by utilizing a Histogram of Oriented Gradients (HOG) [6], [8]. Authors Ibrahim SayginTopkaya, HakanErdogan, FatihPorikli, have involved with people counting using clustering scheme and HOG [6]. The clustering algorithm used in this paper was based on Dirichlet Process Mixture Models (DPMMs).DPMM involves with the formation of HOG which is extracted for both positive and negative images. DPMM clustering is applied based on the features as color, spatial and temporal. Similarly in [8] HOG was involved for feature extraction and also support vector regressionis used to determine the numbers of people. Mixed Gaussian background model is used for foreground extraction. The problem involved in this procedure is the use of HOG which may fail to differentiate targets in case if crowd increases. Due to this, the result of counting the numbers of people was not effective.

Contributions of our proposed work are listed as follows,

- Accurately obtain the numbers of people present in the on-going video
- Background extraction by using Visual Background Extractor (ViBe)
- Perfect detection of people by using Unsupervised learning Neural Network
- Kalman filter's efficient performance is utilized for tracking people

Then rest of our paper is organized with the following sections, section 2 gives brief explanation of previous algorithms and techniques, section 3 deals with the complete processing of our proposed work, then in section 4 we discuss about the experimental results and finally section 5 concludes with our proposed work.

2. RELATED WORKS

Authors *Pirah Noor Soomro, UfraMemon, SheerazMemon*[9], introduced with the processing of classifiers as Viola and jones algorithm. In this algorithm features are computed using integral image values (i.e.) location (x,y). Detection of numbers of people majorly involves with haar feature. Set of positive and negative images are involved in classification. Cascade Ada Boost is used for classifier, in which it makes the classifier stronger with three cascaded classifiers. In this procedure a loop is applied over the specified number of frames that are present in the video. Ada boost classifier is used for several classifications in image processing [9], [12]. In [12] authors have used Meanshift algorithm for tracking the local head in the video. Initially the video files are given, which is involved with the following process as image acquisition, foreground extraction, morphological processing, classification, tracking, apply crossing-line judgment and then perform counting. The use of crossing-line

judgment was to identify the positions of the head and then decides whether to take that head in count or not. In some cases the head detection may fail, which means that it have the possibilities to wrongly detect other object as human head.

In [11] pylon grid algorithm is followed with certain steps that are performed sequentially. The crowded video is converted into frames and the frames are preprocessed by using median filter which is followed by segmentation and background subtraction. The frames are segmented and borders are set, then non head pylons are discarded. Further the adjacent head pylons are grouped and then head centers are calculated. This process was lengthier with all the mathematical computations. Then authors in paper [10] performed people counting method by generating ground-truth data and greedy algorithm for matching and for identifying the best tracked people bipartite graph is used. Ground truth data is generated based on the events that are detected from the video. The greedy matching algorithm is performed with the constructed bipartite graph, the edges are sorted in the descending order, in case if it matches then mark it as matched. In this method, the expected numbers of people in the video is obtained initially. Bipartite graph becomes complex to construct of the number of inputs are increased at higher numbers.

Bingyin Zhou, Ming Lu, Yonggang Wang et al [13] was involved with the process of counting people from videos by using holistic properties of the videos. In this process the crowded regions with different directions are segmentedfrom the videos and then the features are extracted. Features involves with size, shape, edges and texture. The size and shape features includes area, perimeter, perimeter-area ratio and perimeter edge orientation, then the internal edge features includes edge length, edge orientation and Minkowski dimension and then the texture features includes energy, homogeneity and entropy. After this, classification is performed by using Gradient Boosted trees. This entire procedure involves with several mathematical computations and also the accuracy was less when compared with previous works.

People detection and counting is executed with a set of three main processes as background subtraction, morphological operation, blob detection [14].For background subtraction Gaussian Mixture Model is used which supports with robustness. Then the morphological operations involves with erosion, dilation, closing and opening. Morphological operation is performed for obtaining the best count result. The objects that are detected on the video were at different sizes as small or big, which are filtered out. After filtering out the objects at specified size, counting is performed. But this process was not checked with identification of numbers of people present in the crowd, it just enabled to detected few numbers of person passing by the way. Authors A.JaysriThangam, PadminiThupalliSiva, B.Yogameena, in [14] was introduced with robust people counting approach. This approach involves with color based segmentation and generic head detection. Further the whole video is partitioned into four quadrants. Initially skin tone segmentation is performed, then enhancement using morphological operations further detection of heads and then quadrant partitioning and finally people counting. Skin tone is detected by different color spaces for distinguishing skin regions and non-skin regions. Hue (H), Saturation (S), Value (V), (i.e.) HSV is also a color space which is used here. The major problem in this process was the identification of skin tone, in which at some cases the other object will also match with skin tone color, hence the person identification with only skin tone will not produce accurate result in people counting.

All the previous algorithms and techniques were discussed in this section and certain problematic issues in those algorithms were also insisted. Such problems are overcome in our work and finally our result shows better accuracy than the other algorithms and techniques used.

3. PROPOSED SYSTEM

3.1 System Overview

People count with the effective procedural steps followed in the videos. Our proposed work involves with the following major steps,

- Background Extraction
- People Detection
- People Tracking
- Trajectory Analysis

Each process involves with an algorithm, firstly the input video is converted into frames then we perform background extraction over the frames using ViBE algorithm and then unsupervised neural network is involved for people detection and further kalman filter is used for people tracking and finally we count

the numbers of people. All these processes are performed sequentially one after the other.

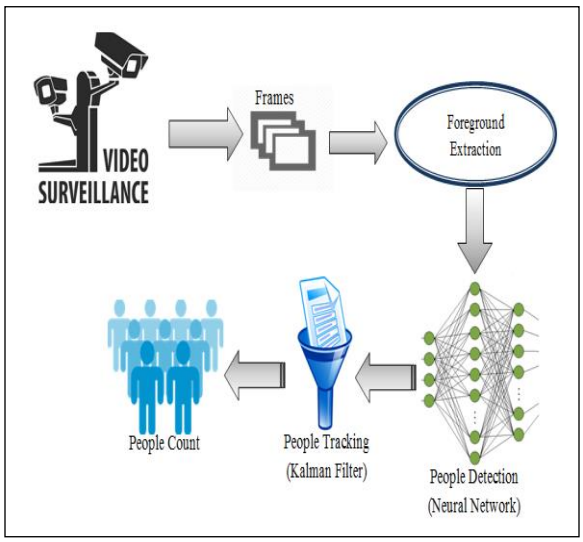

Figure2. System Architecture

Figure 2 implies the step by step process that is involved in our proposed work for identification of the numbers of people moving in the area. In the next sections we have detailed description of each algorithm that has been used here.

3.2 Background Extraction

Background of the frames is extracted by using Visual Background Extractor (ViBE) algorithm. Frames from the video involve with pixels values which are significant for comparing frames with one another. This background model is designed based on the probability distribution function.

Let '$v(x,y)$' represent the pixel values that are obtained in color space at the positioning point '(x,y)' in the frame, then 'v_i' denotes the 'i^{th}' sample from the background model. Now we define the model with the corresponding pixel '(x,y)' with the set of samples 'N' as,

$$M(x,y) - \{v_1, v_2 \dots \dots, v_N\} \qquad (1)$$

For the purpose of classifying the pixels '$v(x,y)$' a sphere represented as '$S_R(v(x,y))$' which is with the radius 'R' (i.e.) centered at the point '$v(x,y)$'. The pixel that is analyzed is taken in account as background with the minimum samples from model '$M(x,y)$' which are considered to be located within the sphere. Further the distance between pixels is estimated using Euclidean distance formula. The pixel values are compared with the historical pixels. Based on the threshold value the frame is classified into foreground and background.

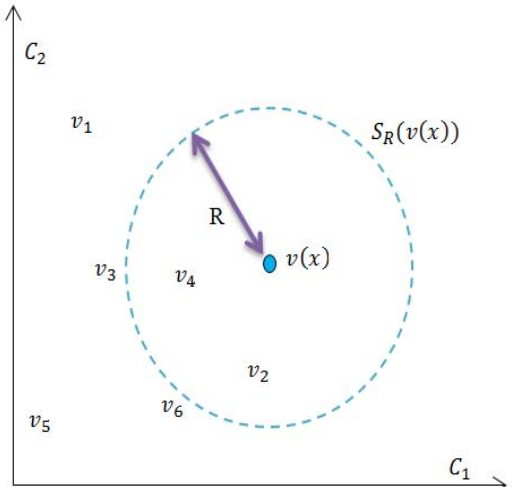

Figure3. ViBE algorithm

Figure 3 illustrates the comparison of pixel values for a set of samples at Euclidean color space 'C_1, C_2', Hereby we count the sample '$M(x)$' that are intersecting the sphere at radius 'R' (i.e.) center of '$v(x)$'.For classifying the pixel '$v(x)$' with the model '$M(x)$', we compare all the closest values that are present within the samples present at the sphere '$S_R(v(x))$'. The background is differentiated from foreground and then the background is extracted.

3.3People Detection

People detection is performed by using neural network. Neural Network involves with the processing of three layers. The result obtained from the ViBE algorithm is input to neural network. The multilayer perceptron is classified into three layers as input layer, hidden layer and output layer. The result of the input layer is given to the hidden layers and then the result of the hidden layer is given into the output layer.

Usually the process is proceeded by summation of inputs by utilizing the activation function for generating output. Activation function is of either linear or non-linear. Each neuron is connected to each other by the weights 'w_i'. The raw data information

is provided to the input layer, then the hidden layer predicts the activities of all the input layers, further the activity of the output layer depends upon the hidden layer and the weights. Based on the weighted values, the layers are executed. So the background extracted frames are classified with the presence of people. Then we follow up with the procedure of people tracking after completion of people detection in the frames.

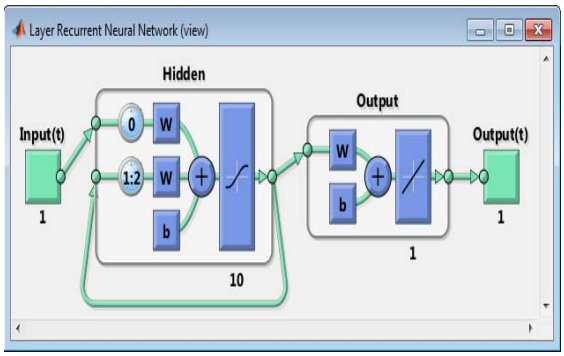

Figure4. Neural Network

3.4 People Tracking

People in the frames are detected and then they are tracked by using kalman filter.Each people are modeled into a rectangular patch, those are moving with same acceleration. The size changes are in lines, since there is not much difference in size of the people between the consecutive frames. Now let us derive people model by assuming a sampling interval of 't'. The discrete – time dynamic system is formulated as,

$$X_t = \emptyset\, X_{t-1} + \eta_{t-1} \qquad (2)$$

$$M_t = H_t X_t + V_t \qquad (3)$$

Where,

$M_t = [X_t Y_t \Delta X_t \Delta Y_t]^T$ – denotes measurement vector

$H_t = [E, 0]$ – Measurement function

$V_t -$ denotes measurement error vector

$$\eta_t = \left[\eta_{\Delta X_t}\eta_{\Delta Y_t}\eta_{C_t}\eta_{K_t}\eta_{\Delta X\prime_t}\eta_{\Delta Y\prime_t}\eta_{C\prime_t}\eta_{K\prime_t}\right]^T \qquad (4)$$

So we obtain results for the video frames '$t-1$' and the target area in frame 't' is given as,

$$X_C = X_{t-1} + \Delta X_{t-1} \qquad (5)$$

$$Y_C = Y_{t-1} + \Delta Y_{t-1} \qquad (6)$$

$$w_x = 1.5\, C_{t-1} \qquad (7)$$

$$w_y = 1.5\, K_{t-1} \qquad (8)$$

After completion of searching of the people in 't^{th}' frame, we compute '$V(i,j)$' mathematically by,

$$v(i,j) = \alpha D(i,j) + \beta H(i,j) + \gamma A(i,j) \qquad (9)$$

The components used in the above equation (9) are given as,

$$D(i,j) = \frac{\left|C_t^i C_{t+1}^j\right|}{Max_n\left|C_t^i C_{t+1}^n\right|} \qquad (10)$$

$$H(i,j) = \frac{\left|G_t^i - G_{t+1}^j\right|}{Max_n\left|G_t^i - G_{t+1}^n\right|} \qquad (11)$$

$$A(i,j) = \frac{\left|S_t^i - S_{t+1}^j\right|}{Max_n\left|S_t^i - S_{t+1}^n\right|} \qquad (12)$$

The terms 'X_t^i', 'Y_t^i' are defined as the mass of 'i^{th}' object at 't^{th}' frame, then 'G_t^i', 'S_t^i' represents the presence of average gray levels and values of 'α', 'β' and 'γ' denotes the weights which should be,

$$\alpha + \beta + \gamma = 1 \qquad (13)$$

The values of these weights are assigned initially, which is applicable to be modified later. The term 'n' in equation (12) represents the number of detected objects which is to produce,

$$1 \le n \le target(t+1) \qquad (14)$$

From the above equation '$t+1$' denotes the total number of detected objects that are presented in the frame '$t+1$'. We perform tracking of people based on the object-chain formation. The people present in the current frame may not exist in previous frame.Based on this chain the numbers of people are tracked in the video.

With the confirmation of maximum similarities distance is computed based on Euclidean formula. This Euclidean distance is estimated for the purpose of updating the object-chain. Hence the locations are given back to kalman filter as feedback for obtaining new state. This kalman filter can be effectively used for object tracking at different kinds of linear dynamic systems. Here also we have considered videos in which, each frame differs from the other.

Kalman filter gives better results in tracking people and so this filter is used at many image processing applications. People moving in each frame is tracked by using this kalman filter which gives us effective results and so further the trajectory analysis process can be followed and finally the number of people present could be identified.

3.5 Trajectory Analysis

In this section we perform with the counting of people present in the video. We define the trajectory with '(x, y)' which the center of the object, 'w, h' represents the width, height and 'i', '$i+1$' denotes the frame number of object that have been appeared and disappeared. So the trajectory is given as,

$$T = \{(x_i, y_i, W_i, h_i), (x_{j+1}, y_{j+1}, W_{j+1}, h_{j+1}), \ldots, (x_{j+n}, y_{j+n}, W_{j+n}, h_{j+n})\} \quad (15)$$

Then we apply Euclidean distance formula for determining the distance between initial point and the end point. With this distance we could finalize whether the trajectory is present or not. Euclidean distance is given as,

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (16)$$

From the above distance formula, '(x_1, y_1)' and '(x_2, y_2)' represents the coordinates.

Finally we could obtain the number of people present in the video on following the sequential steps that are involved in our proposed work.

4. EXPERIMENTAL RESULTS

4.1 Dataset and Simulation Setup

In this section we discuss about all the requirements with which we execute all our proposed algorithms and techniques. Here we use MATLAB-R2013a with the windows 7 operating system. MATLAB is one of the perfect software to work with videos which supports all vectors and matrices based analysis. We have used PETS dataset which consists of vides with which we perform our entire process. Initially we covert the query video into frames and then these frames are considered for our sequential processing of foreground extraction, people detection, people tracking and count the number of people.

4.2 Performance Evaluation

Our proposed work involves with major steps as frame conversion, foreground extraction, detection of people, people trajectory and people count. Each step involves with the use of an algorithm that accurately matches the performance and produces higher accuracy results.

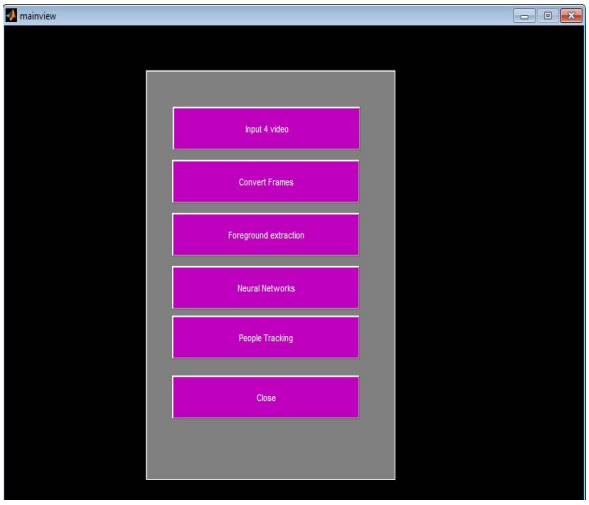

Figure5. Main View

Figure 5 shows the way in which our designed main view of simulation looks. From PETS data set we have taken up four video for our process. The second button is used to convert the given videos into frames and so we could start our process with the algorithms. Each button represents the step which is involved to obtain the final result about the numbers of people moving in the video.

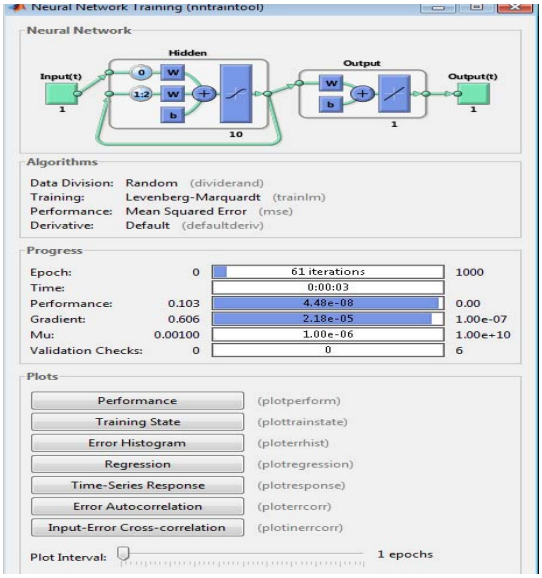

Figure6. People detection using Neural Network

Figure 6implies theneural network which is used for the process of people detection. This neural network includes three layers as input layer, hidden layer and output layer. Based on the processing of neural network we detect the people present in the frames or not. The results from the neural network are given into kalman filter for determining the trajectory of the people.

Each video file is first converted into frames and the number of frames depends upon the size of the video that is selected for processing. In case if the size of the video is small then the number of frames will also be small and hence the execution time will also be smaller and the results would be accurate. It is simpler to perform comparison on people trajectory in case if the number of frames is lesser. The entire process gives accurate results, which does not mean that the size of the file is a constraint.

(b)

Figure7. (a) And (b) People Count

Figure 7 (a) and (b) shows the final result obtained from our proposed work's simulation. In figure 7(a) the current video shown is comprised with three people, which shows the people count as 3. Similarly in figure 7(b) the video is with seven people and our result shown is also people count of 7. This implies that our entire proposed work is accurate in the identification of numbers of people moving in the playing video. Based on the video playing time, the people count varies.

Table1. Accuracy result

Accuracy	Error Rate	Actual Numbers of People Present in frame	People count by our proposed work
100	0	8	7
95	5	8	11
80	20	9	9
100	10	10	8

Table 1 implies the accuracy of our proposed work with its error rate. Accuracy depends on the perfect identification of the people moving on the screen. With this result we prove that our proposed work results with better accuracy in detection of people. So this procedure will support several applications for counting the number of people in dense crowds.

CONCLUSION

In this paper we have proposed with counting numbers of people from a dense crowd. The numbers of people in a crowd in counted for several reason either for safety or security purposes. Here we have used significant algorithms for foreground extraction, detection of people and tracking people. For foreground extraction we have used ViBE algorithm, then for People detection we have used Unsupervised learning technique (i.e.) Neural Network and for people tracking kalman filter is used and finally

perform trajectory analysis and count the number of people present. The kalman filter used here is capable to tolerate small occlusion. With this our procedure is completed with higher level of accuracy in the detection of numbers of people.

REFERENCES

[1] AnkanBansal, K S Venkatesh, "People Counting in High Density Crowds from Still Images", Cornell University Library, 2015.

[2] LokeshBoominathan, Srinivas S SKruthiventi, R VenkateshBabu, "CroedNet: A Deep Convolution Network for Dense Crowd Counting", Cornell University Library, 2016.

[3] Cong Ahanf, Hongsheng Li, XiaogangWang, Xiaokang Yang, "Cross-scene Crowd Counting via Deep Convolution Neural Networks", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[4] Jianzhao Cao, Lianglinag Sun,, Manfred Gilbert Odoom, Fangjun Luan, Xiaoyu Song, "Counting People by Using a Single Camera without Calibration", IEEE, Control and Decision Conference (CCDC), 2016.

[5] ZebinCai, Zhu Liang Yu, Hao Liu, Ke Zhang, "Counting People in Crowded Scenes by Video Analyzing", IEEE 9th Conference onIndustrial Electronics and Applications (ICIEA), 2014.

[6] Ibrahim SayginTopkaya, HakanErogan, Faith Porikli, "Counting People by Clustering Person Detector Outputs", 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2014.

[7] Djamel MERAD, Kheir-Eddine AZIZ, Nicolas THOME, "Fast People Counting Using Head Detection from Skeleton Graph", Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2010.

[8] VenkateshBalaSubburaman, "Counting People in the crowd using generic head detector", IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS), 2012.

[9] Pirah Noor Soomre, UfraMemon, SheerazMemon, "Human Detection and Counting in Crowded Scenes", Asian Journal of Engineering, Science and technology, volume 4, issue 1, pp 18 – 22, 2014.

[10] V H C Melo, S SAlmedia, J C Mendes, D Menotti, "A Methodology for Evaluation of People Counting Methods based on Video Analysis", 2012.

[11] P Karpagavalli, A Vinoth Nelson, A V Ramprasad, "Human Tracking and Counting In Range Images by Using Mean Shift and Pylon Gris Algorithm" International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Issue 3, pp 1231 – 1237, 2014.

[12] Bin Li, Jian Zhang, Zheng Zhang, Yong Xu, "A People Counting Method Based on Head Detection and Tracking", International Conference on Smart Computing (SMARTCOMP), 2014.

[13] Bingyin Zhou, Ming Lu, Yonggang Wang, "Counting People using Gradient Boosted Trees",IEEE Information Technology, Networking, Electronic and Automation Control Conference, 2016.

[14]PongsakonBamrungthai, SuteeraPuengsawad, "Robust People Counting Using a Region Based Approach for a Monocular Vision System", International Conference on Science and Technology, IEEE, pp 308 – 312, 2015.