ISSN No. 0976-5697

Volume 3, No. 2, March-April 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Applicaton of ViBe Algorithm for People Counting in a crowded Environment

Minu.S*
CSE Department
Dr. M.G.R. Educational and Research Institute
University, Chennai, India
minu999@gmail.com

Dr.V.Cyril Raj
CSE Department
Dr. M.G.R. Educational and Research Institute
University Chennai, India
cyrilraj@hotmail.com

Abstract: People counting is a very important problem in visual surveillance. An accurate and real time estimation of people in a crowded place can provide valuable information. Here video is given as input and outputs the average number of people passing over the video. The video input is separated to number of frames and some processing steps are performed on background subtraction results to estimate the number of people in a complicated scene. Foreground pixel extraction is done with ViBe (VIsual Background Extractor) algorithm. The extracted foreground image's pixels count is calculated and gives as input to the neural network. In learning phase, the people count is calculated manually with test dataset and while testing phase remaining test cases are tested by adjusting weight parameters to obtain relative to the target result.

Keywords: background subtraction, ViBe (VIsual Background Extractor) algorithm, neural network, and people counting.

I. INTRODUCTION

Automatic monitoring of the number of people in public areas is also important for safety control. The paper aims to develop an effective method for estimating the number of people in a complicated outdoor scene. In this scenario, the video has been taken by a static camera in a public place. The video is segmented to number of frames based on a frame rate and background subtraction algorithms are used for extracting the human occupied areas. Simple motion detection algorithms compares static background frame with the current frame in order to detect zones where a significant difference occurs. The purpose of this algorithm is therefore to detect moving objects(called foreground) from static or slow moving parts of the scene (called background). Another definition for the background is that it corresponds to a reference frame with values visible most of the time, which is with highest appearance probability.

Here a recent background subtraction algorithm used which is ViBe Algorithm. ViBe stands for VIsual Background Extractor. In this technique, a set of values taken in the past for each pixel stores at the same location or in the neighborhood. It then compares this set to the current pixel value for determining whether that pixel belongs to the background and adapts the model by choosing randomly which values to substitute from the background model. The value of the pixel is propagated in to the background model of neighboring pixel when it is found to be the part of the background.

The Resulted foreground extracted is binarized and found out the foreground pixels count which is given as input to the neural network. The network is already trained manually with sample test cases and found out the average people count in that video. In this Experiment, 3 different methods are adopted for people counting. Last method showed better results compared to other two. Section II describes the related works regarding counting and background subtraction methods.

Section III explains about the ViBe algorithm and counting methodologies. After that Results and analysis are shown (Section IV).

A video of 400 frames are processed in the evaluation time and 50 frames used as the trained data which is manually counted to train network. Rest 350 frames are used as a test data which will count the average people based on the neural network training.

II. RELATED WORK

The work done by Ya-Li Hou, Student Member, IEEE, and Grantham K. H. Pang , Senior Member, IEEE [1] was developed a method based on the neural network to estimate the number of people. They used a robust adaptive background estimation method based on the Gaussian Mixture Model (GMM) for extracting background mean. This foreground image binarized based on a threshold and produce foreground pixels. The foreground pixel count used as input to neural network for people estimation. Sudden increase or decrease of people resulted large variations in the people count.

Olivier Barnich and Marc Van Droogenbroeck, Member, IEEE [2] introduced a new background subtraction algorithm ViBe on IEEE Transactions on Image Processing, 20(6) :1709-1724, June 2011. The paper entitled "ViBe: A universal background subtraction algorithm for video sequences". It gives a detailed description of the ViBe algorithm and comparison with other background subtraction algorithms.

Another paper brought out by [3] Lijing Zhang and Yingli Liang was "Motion human detection based on back ground subtraction". They established a reliable background updating model based on statistical and use a dynamic optimization threshold method to obtain a more complete moving object. Now morphological filtering is introduced to eliminate the noise and solve the background disturbance problem. They

focused only on the human location detection based on background subtraction.

T. Zhao, R. Nevatia, and B. Wu [4] researched on "Segmentation and tracking of multiple humans in crowded environments" .A model-based approach used here to interpret the image observations by multiple partially occluded human hypotheses in a Bayesian framework.

The work of V.Rabaud and S. Belongie [5] was motivated in this paper which is "Counting crowded moving objects," presented in *Proc.IEEE Conf. Comput. Vis. Pattern Recog.*, in 2006. It was based on a highly parallelized version of the KLT tracker in order to process the video into a set of feature trajectories.

S.Y. Cho, T. W. S. Chow, and C.T. Leung, [6] developed "Neural-based crowd estimation by hybrid global learning algorithm". It was a neural-based crowd estimation system for surveillance in complex scenes at underground station platform. Estimation is carried out by extracting a set of significant features from sequences of images. Those feature indexes are modeled by a neural -network to estimate the crowd density. The learning phase is based on the least-squares and global search algorithms which are capable of providing the global search characteristic and fast convergence speed.

P. Kilambi, O. Masoud, and N. Papanikolopoulos, [7] researched on "Crowd analysis at mass transit site." They proposed a novel method for detecting and estimating the count of people in groups, dense or otherwise, and tracking them. Using prior knowledge obtained from the scene and accurate camera calibration, the system learned the parameters required for estimation. This information can then be used to estimate the count of people in the scene, in realtime. There were no constraints on camera placement and Groups are tracked in the same manner as individuals, using Kalman filtering techniques. Results are provided for groups of various sizes moving in an unconstrained fashion in crowded scenes. The techniques used here, gave an idea about counting people in crowded environment.

The work of R. Ma, L. Li,W. Huang, and Q. Tian [8] was "On pixel count based crowd density estimation for visual surveillance." They derived the relation for geometric correction for the ground plane and proved formally that it can be directly applied to all the foreground pixels. They also proposed a very efficient implementation because it was important for a real-time application. Finally a time-adaptive criterion for unusual crowdedness detection is described.

D. Kong, D. Gray, and T. Hai [9] did paper named "A viewpoint invariant approach for crowd counting," in *Proc. Int. Conf. Pattern Recog.*, 2006, pp. 1187–1190. It describes a viewpoint invariant learning-based Method for counting people in crowds from a single camera. It takes into account feature normalization to deal with perspective projection and different camera orientation. The training features include edge orientation and blob size histograms resulted from edge detection and background subtraction. A density map that measures the relative size of individuals and a global scale measuring camera orientation are estimated and It is used for feature normalization. The relationship between the feature histograms and the number of pedestrians in the crowds is learned from labeled training data.

A. B. Chan presented a privacy-preserving system [10], that was "Privacy preserving crowd monitoring: Counting people without people models or tracking," It was used for estimating the size of inhomogeneous crowds, composed of pedestrians that travel in different directions, without using explicit object segmentation or tracking. First, the crowd is segmented into components of homogeneous motion, using the mixture of dynamic textures motion model. Second, a set of simple holistic features is extracted from each segmented region, and the correspondence between features and the number of people per segment is learned with Gaussian process regression.

III. METHODOLOGIES

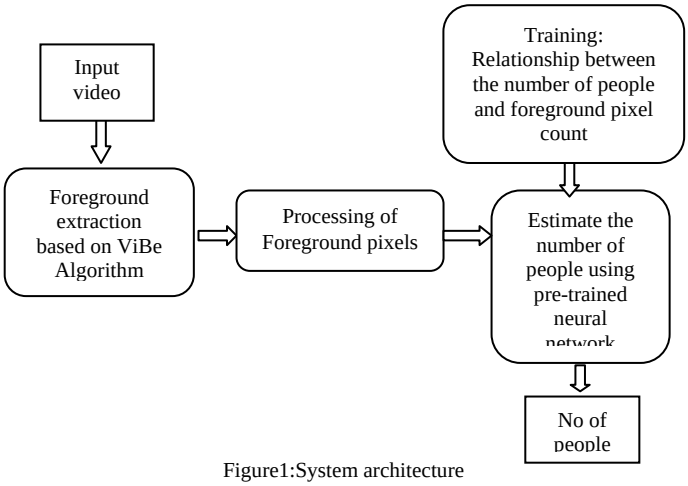

Figure1:System architecture

People always exhibit some movement whether they are standing or sitting. Motivated by this observation, it's possible to estimate the number of people by finding a relationship with the foreground pixels.

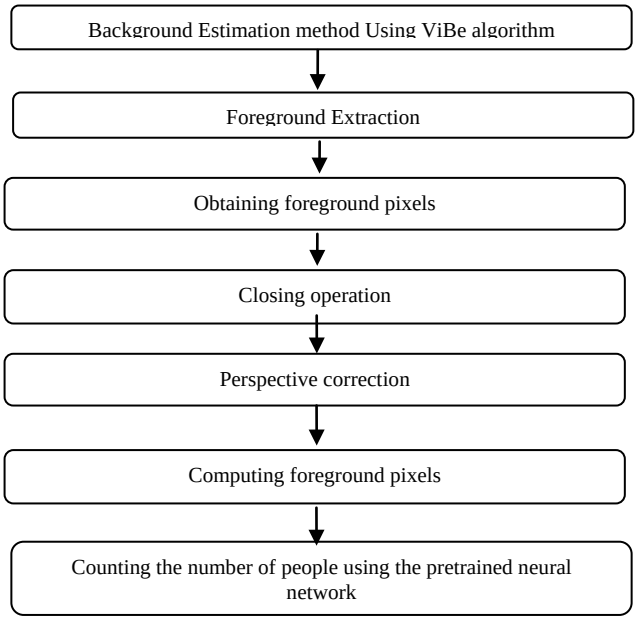

Figure2 :flow chart which shows the main procedures for counting.

A foreground image is obtained by subtracting the current frame (image) from the background frame (image). The background subtraction can be performed in different ways. In Ya-Li-Hou's paper they used GMM method (Gaussian Mixture Model) which produced a Gaussian distribution model of background image in HSV color space and generated the background mean. This background mean is subtracted with the next frame's mean which is resulted from its Gaussian distribution model. But In this paper , ViBe (VIsual Background Extractor) Algorithm is used for motion human detection.

A. Background Estimation Method:

Vibe Algorithm involves:

a. Pixel model and classification process:

Each background pixel is associated with a set of samples instead of an explicit pixel model. The current value of the pixel is compared to its closest samples within the collection of samples. This is the difference with other algorithms.

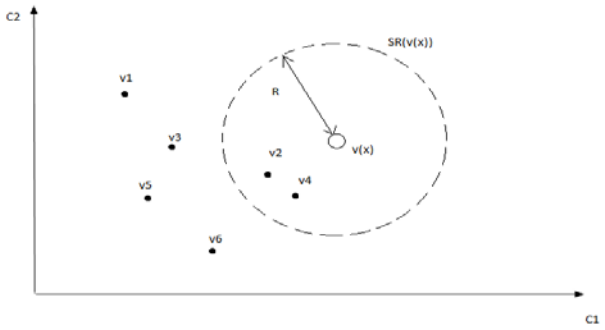

Figure3: shows the comparison of pixel value with a set of samples in a two dimensional Euclidean color space (C1,C2).

If v(x) denotes the value in a given Euclidean color space taken by the pixel located at x in the image. Vi is the background sample value with an index i. each background pixel x is modeled by a collection of N samples.

$$M(x) = \{v1, v2, .. vN\} \qquad (1)$$

To classify a pixel value v(x) with respect to model m(x), compare it to the closest values within the set of samples by defining a sphere SR(v(x)) of radius R centered on v(x). The pixel v(x) is classified as background if the cardinality (#) of the set intersection of the sphere and the set of model samples M(x) is greater than or equal to a given threshold #min.

$$\#\{SR(v(x)) \cap \{v1, v2, .. vN\}\} \quad (2)$$

The classification of v(x) involves computation of N distances between v(x) and model samples and N comparisons with threshold Euclidean distance R. this segmentation process of a pixel can be stopped once #min matches have been found.

b. Background model initialization (from single frame):

Neighboring pixels share a similar temporal distribution. Based on this fact, populate the pixel models with values found in the spatial neighborhood of each pixel. That is filling them with values randomly taken in their neighborhood in the first frame. If t=0 indexes the first frame and NG(x) is a spatial neighborhood of a pixel location x.

$$M^0(x) = \{ v^0 (y/y \in Ng(x))\} \quad (3)$$

c. Updating the background model over time.:

It means continuously updating the background model with each new frame. Update method incorporates three important components.

a) Memory less update policy:- ensures a smooth decaying life span for the samples stored in the background pixel model.
b) Random time sub sampling method:- To extend the time windows covered by the background pixel models.
c) A mechanism that propagates background pixel samples spatially to ensure spatial consistency and allowing the adaptation of background pixel models which is masked by the foreground.

Algorithm describes below.

a. Initialize the number of samples per pixel (N=10).
b. Initialize number of closed samples for being part of the background.(nmin=2).
c. Initialize amount of random sub sampling (phi=4).
d. Create an image array img[width][height].
e. Create a background model array samples[width][height][N]
f. Create an image segmentation map array segmap[width][height].
g. Assign values for background and foreground identifiers.(bg=0 and fg=1);
h. Initialize all sample values.
i. Initialize x=0, y=0, count=0, index=0,dist=0.
j. Segmentation step:
k. For each pixel x=0 to width and y= 0 to height.

i. Compare pixel to background model:-
While (count<#min) and
(Index<N) then
Compute Euclidean distance.
Dist= EuclidDist(img[x][y], samples[x][y][index]);
If dist < Euclidean sphere radius then increment value of count and index.

ii. Classify pixel and update model:-
Compare a new pixel to background samples to find two matches (nmin=2). Once 2 matches have been found, step over to the next pixel. Then ignore the remaining background samples.
If count>= nmin
Store that img[x][y] as background.
Segmap[x][y]=background.
Get a random number (rand) between 0 and phi-1.
Replace randomly chosen sample samples[x][y][rand]=img[x][y].
Else count < nmin
Store that img[x][y] as foreground.
Segmap[x][y]=foreground.

Thus the foreground image obtained is binarized based on a threshold to return the foreground pixels. Compute the all Foreground pixels from the foreground image and give that count as input to the neural network. It will find out the average number of people based on the pre-trained data set.

Figure 4: a)Typical scene to be processed. (b) Binary foreground image after foreground extraction for (a).

B. Obtaining Foreground Pixels:

The foreground image obtained after extraction is Binarizing based on a threshold. A Foreground pixel is a pixel whose intensity difference between current image & background image greater than the threshold.

C. Perspective Correction:

The Size of an object varies linearly as functions of the y-coordinate of the image. The objects at different locations are brought to the same scale in this method.

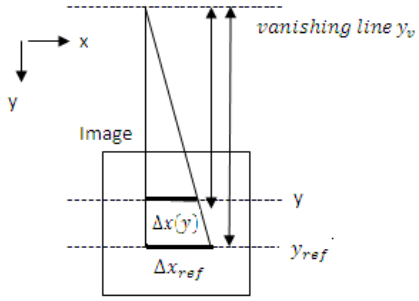

Figure 5: perspective correction

$\Delta xref = \Delta x(y) * q(y)$ and

$q(y) = (yref - yv)/(y - yv)$. (4)

Equation (4) shows how to convert a scale at y to its scale at the reference location, *yref* . Fig. 4 is a simple illustration for (1). $\Delta x(y)$ is the horizontal (vertical) scale of an object at y, and $\Delta xref$ is its horizontal (vertical) reference scale. $q(y)$ is the ratio for different locations. The extension of parallel lines intersects at a vanishing point, which lies on *yv* in the image. *yv* can be easily estimated using the same object at two different coordinates in (4).

D. Computing Foreground Pixels:

Perspective corrected foreground pixels are computed together using the following equation.

$$Npixel = \sum_{y=1:imgY} N(y) * q(y) \qquad (5)$$

imgY = *height of processing image*

Npixel = total no of foreground pixels

N(y) = *no of foreground pixels in yth row*

q(y) = *ratio for different locations.*

E. Closing Operation:

Extracted foreground results solid blobs and some scattered pixels. Those solid blobs are from the moving people and scattered pixels are from the stationary crowd.

To reduce the difference between moving people and stationary people, a closing operation is employed.

Closing operation means, most areas occupied by people are covered with white pixels, while the other parts with black. Perspective effects also need to be considered during the closing operation.

F. Counting the number of people using pre-trained neural network:

Some manually annotated training images from a similar scene are needed to find the relationship between foreground pixel count & no of people. 3 methods are discussed below.

Method 1)-based on foreground pixels

M=f1(X)

Where M is the no of people and X is the no of foreground pixels after perspective correction and f1 is the manually annotated training set is used to ascertain the relationship f1.

Method 2) Based on Closed Foreground Pixels:

M = f2(C)

Let C be the number of foreground pixels after the closing operation and M be the number of people. The relationship between *C* and *M* will be found and used for estimation.

Method 3) Based on Both Foreground Pixels and Closed Foreground Pixels:

To keep more information about the original image, both foreground pixels and closed foreground pixels will be injected into the neural network. The relationship between the number of people and these two inputs is denoted as *f*3.

$M = f3(X,C)$.

IV. RESULTS AND ANALYSIS

Here Test data used was a video of 400 frames extracted from this. Training set consists of 50 images of that video, which is counted manually. The test set is composed of the remaining 350 images. Image resolution taken was 768 * 576. To increase the speed of people counting, all the images were resized to 256 * 256 pixels.

Figure:6 : training of neural network

The above figure shows the training of neural network which produces a training set of 50 image frames.

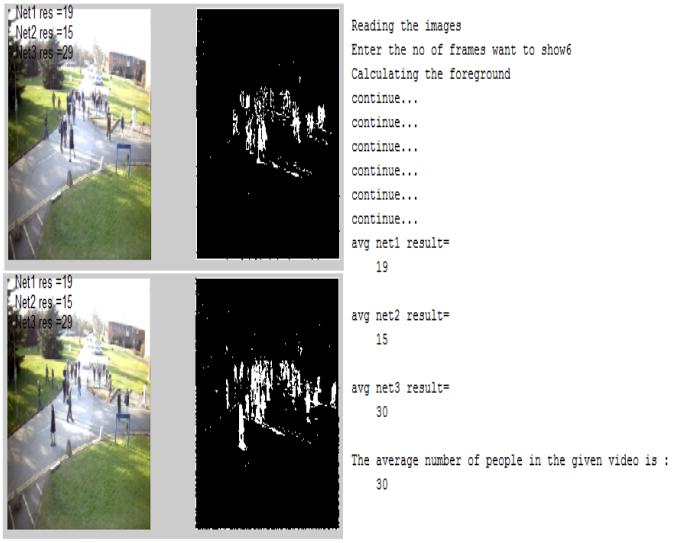

Figure 7: Testing video frames and shows the people count.

The above results show the people counting of some random image frames and the average count of people based on three methods.

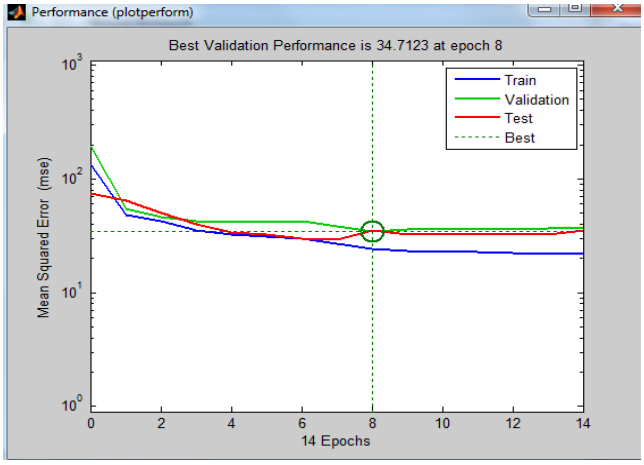

Figure 8: Performance graph while training, validation and testing.

The above graph shows the performance graph while training, validation and testing. The dotted lines show the best performance and it coincides at one point which shows the best validation performance. x axis represents the number of epochs and y axis represents the mean squared error(mse) , ie, the mean difference between the target output and network output.

V. CONCLUSION

In this paper, foreground pixel extraction is carried out by the recent algorithm ViBe which is faster than other background subtraction algorithms. This can be embedded in digital cameras. Closing operation over foreground pixels used for increasing the accuracy of people count. Some optimization methods are also performed over output and the best estimation results, with a 14% average error, were achieved when foreground pixels and closed foreground pixels, are learned in a neural network.

VI. ACKNOWLEDGMENT

We are thankful to beloved God first and each and every one who supports and motivates for our work and also IJARCS Journal for the support to develop this document.

VII. REFERENCES

[1] Ya-Li Hou, *Student Member, IEEE*, and Grantham K. H. Pang, *Senior Member, IEEE"* People Counting and Human Detection in a Challenging Situation.

[2] O. barnich and M. Van Droogenbroeck. "ViBe : A universal background subtraction algorithm for video sequences." IEEE Transactions on Image Processing, 20(6) :1709-1724, June 2011. doi :10.1109/TIP.2010.2101613 1

[3] Lijing Zhang and Yingli Liang "Motion human detection based on back ground subtraction".

[4] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments"

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

[6] S.-Y.Cho, T. W. S. Chow, and C.-T. Leung," neural-based crowd estimation by hybrid global learning algorithm,"

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

[8] R. Ma, L. Li,W. Huang, and Q. Tian, "On pixel count based crowd density estimation for visual surveillance," in *Proc. IEEE Conf. Cybern. Intell. Syst.*, 2004, pp. 170–173

[9] D. Kong, D. Gray, and T. Hai, "A viewpoint invariant approach for crowd counting," in *Proc. Int. Conf. Pattern Recog.*, 2006, pp. 1187–1190.

[10] 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