Real-Time Crowd Detection to Prevent Stampede

Chapter · January 2020

DOI: 10.1007/978-981-13-7564-4_56

5 authors, including:

Sabrina Haque

3 PUBLICATIONS **9** CITATIONS

SEE PROFILE

Muhammad Sheikh Sadi

Khulna University of Engineering and Technology

72 PUBLICATIONS **295** CITATIONS

SEE PROFILE

Md Erfanul Haque Rafi

Texas State University

2 PUBLICATIONS **7** CITATIONS

SEE PROFILE

Md. Milon Islam

Khulna University of Engineering and Technology

51 PUBLICATIONS **1,013** CITATIONS

SEE PROFILE

Chapter 56
Real-Time Crowd Detection to Prevent Stampede

Sabrina Haque, Muhammad Sheikh Sadi, Md. Erfanul Haque Rafi, Md. Milon Islam and Md. Kamrul Hasan

1 Introduction

By crowd, we mainly refer to the average number of individuals present in a particular place. A place becomes crowded if the total population of the certain area becomes much more than the capacity. As a result of such crowd, various accidents may take place. Extreme crowd results individuals in losing control and turning the place into disaster. Often miscreants use such crowd to do various inhuman activities like harassing women. Presently, crowd counting is of severe importance in order to maintain human safety in crowded situations. [1, 2]. One such occurrence happened on 14 April 2015. That day, several women have been sexually harassed at Dhaka University premises while celebrating the first day of the Bengali year, Pohela Boishakh. Again, old people may become uncomfortable in overly crowded places. Another crowd catastrophe took place in the year 2015 during Hajj pilgrimage. That time, 2236 hajjis lost their lives being rushed during the Hajj in Mecca on September 24. There are many such incidents massive tragedy due to overcrowded situations. The crowd may frequently occur in modern society. To avoid hazardous situations

S. Haque (✉) · M. S. Sadi · Md. E. H. Rafi · Md. M. Islam · Md. K. Hasan
Department of Computer Science and Engineering, Khulna University of Engineering and Technology, Khulna 9203, Bangladesh
e-mail: sabrina.hq@gmail.com

M. S. Sadi
e-mail: sadi@cse.kuet.ac.bd

Md. E. H. Rafi
e-mail: ehrafi@gmail.com

Md. M. Islam
e-mail: milonislam@cse.kuet.ac.bd

Md. K. Hasan
e-mail: mhgolap11@gmail.com

M. S. Uddin and J. C. Bansal (eds.), *Proceedings of International Joint Conference on Computational Intelligence*, Algorithms for Intelligent Systems, https://doi.org/10.1007/978-981-13-7564-4_56

due to the crowd, the need for crowd analysis system for handling dense crowds is at a raise. For any crowd analysis system, crowd counting is a must. This includes evaluating the aggregate number of people in the crowd, along with the density of the crowd in the different parts of the area. Certain potential mishaps may be easily avoided by giving advance warning whenever the crowd density of a certain area went over the safe limit. This may help to maintain the overall management and infrastructure of the area as well.

Crowd counting techniques have been tried many times in the past too. Ma et al. [3] and Atnic et al. [4] proposed methods that work for medium-sized places only. Shao et al. [5] proposed methods intend to count crowd from video sequences. Real-time parallel processing was not performed efficiently in any such methods.

In this paper, we propose realistic approaches to detect the crowd density of a place. This model works with real-time images captured from the area under observation using Raspberry pi camera module. The number of people in an area is counted, and if the crowd goes beyond the limit, certain authorities are warned using a WAN technology so that proper steps can be taken to handle the growing crowd immediately. Our approach intends to count the crowd not only for large areas, but it also helps practical implementation by using real-time images. The use of WAN technology to spread the news of upcoming threats allows the related authorities to take necessary actions. Two techniques are used here to estimate the crowd density. The technique uses to count the head regions. The second technique uses Convolutional Neural Network (CNN) in order to detect crowd. The resultant crowd detection will measure the likelihood of a stampede and generate alert signal to take necessary steps.

The remaining part of the paper is organized as follows: The related work that covers the recent developments in this area is described in Sect. 2. The proposed methodology including image processing and convolution neural network techniques are illustrated briefly in Sect. 3. Section 4 outlines the experimental outcomes of the proposed system. Section 5 concludes the paper.

2 Related Works

Previously various approaches were taken to measure the crowd conditions. In this section, we get acknowledged with those. Shao et al. [5] used a deep learning approach for understanding crowded scene from video sequences. The method introduced by Ankan et al. [4] can detect crowd count for high-density images, but it is inefficient for images containing mutual occlusion. On the contrary, the crowd counting technique proposed by Fu et al. [6] works for low-density images only but cannot work in high-density one.

A promising crowd counting method was proposed by Huiyuan Fu [7] where probable head regions can be found using depth camera. But the system was infeasible due to cost overhead because of the depth camera. Besides, it did not work for large regions too. Lin et al. [8] utilized counting by detection technique that has low-level attributes to identify human heads or dynamic objects. Since the training samples

are normally of a high determination without impediments, the finders' execution falls apart basically once the main target on people is incompletely blocked or in blur pictures. Besides, the procedure overhead of the detection stage is simply too high to sustain period responses.

The count of moving objects was estimated in the methods proposed in [9, 10]. These methods used the pattern of moving objects obtained from video streams requiring a good frame rate which is pretty tough to achieve and also do not work in case of still images. Zhang et al. [11] introduced a method utilizing a deep network trained using perspective maps of images. The training methodology for our model is much simpler but can acquire a more accurate result.

3 The Proposed Methodology to Detect Crowd

This section presents an overview of the proposed methodology of this paper. Crowd mainly refers to the population count of a certain place. So, by crowd detection, we mean counting the number of people present in a given area. Our proposal here is to capture still pictures of a certain area at first and then detect the crowd density of the area from the captured image. We propose two techniques for this purpose. The first technique uses erosion. Erosion is computationally lightweight which makes the process faster. The second technique uses a Convolutional Neural Network to detect the crowd density of a place. The result of our approach is then compared the predefined threshold value which represents the number of people assumed to be safe for the area. The estimated safe number of people for a place is predetermined by the authority. Finally, if the compared result indicates overflow, then this message is conveyed to proper authorities.

3.1 Crowd Detection using Image Processing

This technique involves some basic image processing operations. To count the number of people in an image, we have to ensure that each head is identified individually. Figure 1 shows the flow diagram of crowd reckoning using image processing technique. For further analysis, an image is retrieved which is then converted into a grayscale image as depicted in Fig. 2a. Several steps of the proposed technique are then carried out, and these are described briefly as follows.

3.1.1 Thresholding

Thresholding [12] is generally used for image segmentation. This method is a kind of image segmentation that separates objects by altering grayscale images into binary

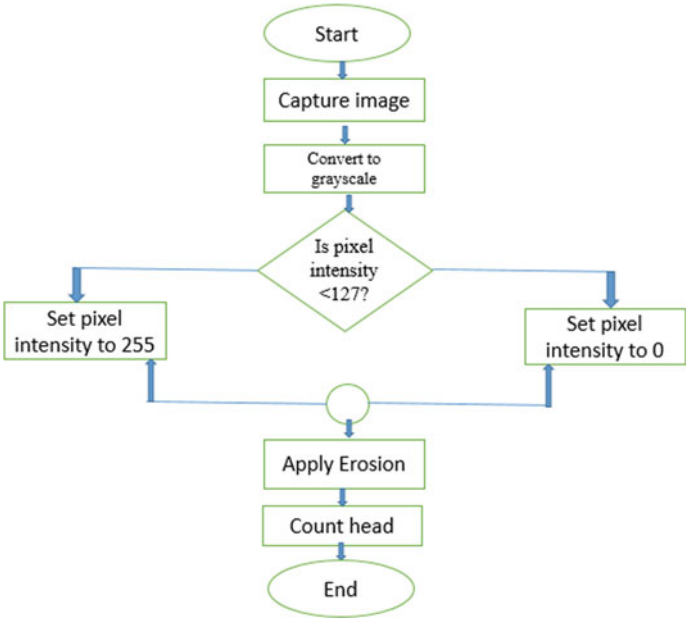

Fig. 1 Flow diagram of crowd detection using image processing technique

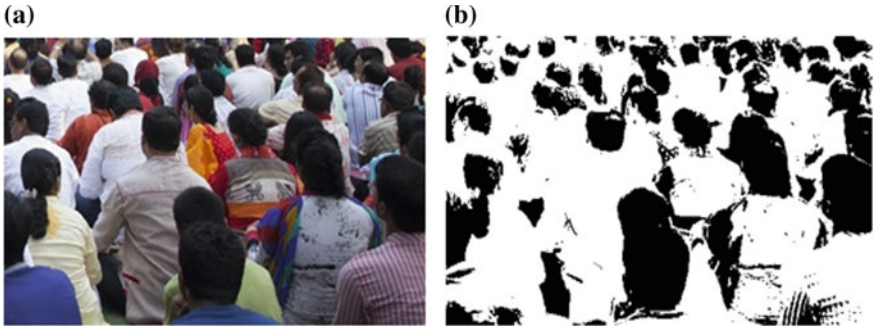

Fig. 2 Thresholding a sample image

images. Image thresholding technique is the most appropriate in images with high stages of contrast. The thresholding procedure can be stated as

$$T = T[a, b, p(a, b), f(a, b)] \tag{1}$$

where T represents the threshold value, the coordinates points of threshold value are (a, b), and the grayscale image pixels are p (a, b), f (a, b). The resulting binary thresholding image is displayed in Fig. 2b.

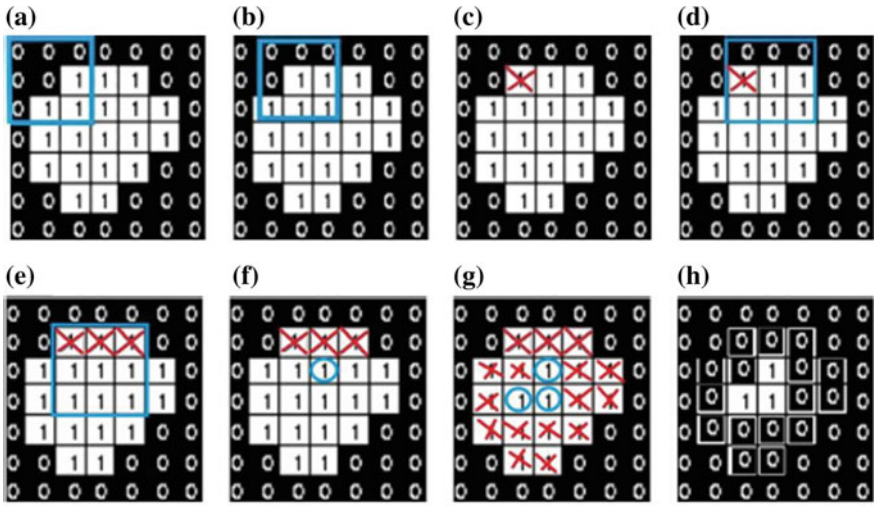

Fig. 3 Erosion technique

3.1.2 Erosion

The basic effect of the erosion on a binary image is to erode the boundaries of regions of foreground pixels. Erosion is applied here to segment the head regions of individuals.

Initially, the kernel is fitted over the image. The center pixel is ignored when the value is zero as illustrated in Fig. 3a. Then, the kernel is shifted to the right column and check the pixel condition as illustrated in Fig. 3b. In this case, the pixels are of low-density as the pixel intensity is set to zero. The same procedure is applied in Fig. 3c, d. The kernel pixels are equal to one after some iterations as illustrated in Fig. 3e. The similar technique is applied until the iterations are completed. After applying erosion, we count the total number of individual heads that provide the crowd density.

3.2 Crowd Detection Using Convolutional Neural Network

In the case of crowd images, individuals closer to the camera are often captured with clearer features while people away from the camera are represented only as head blobs. This happens due to wide variety of perspectives and viewpoints. In order to detect crowd efficiently, we need to consider both these cases. In our proposed methodology using Convolutional Neural Network, we are combining both deep and shallow neural network to achieve this efficiency.

3.2.1 Required Dataset

Generally, a huge training dataset is required for deep learning-based approaches. But most of the crowd datasets have a very limited number of samples (<100). In our research, we have used UCF_CC_50 dataset which contains images of extremely dense crowds. The images are collected mainly from the FLICKR. In order to get a large number of images from this dataset, we have performed multiscale data augmentation.

Primarily, two types of augmentations are performed here. The first crop patches from the multiscale pyramidal representation of each training image. The image pyramid is constructed by considering a scale of 0.5 to 1.2 incrementing 0.1 at each step. 225_225 patches with 50% overlap are cropped from this pyramidal representation. This helps in tackling the problem of scale variations in crowd images. In order to improve CNN's performance for highly dense crowd regions, we augmented the training data by sampling high-density patches more often.

3.2.2 Convolutional Neural Network Architecture

Convolutional Neural Networks (ConvNets or CNN) [13] are a category of Neural Networks that have been demonstrated very real in the fields such as image recognition and classification. Because of its classifying visual patterns such as pixel images with a very few preprocessing, it has added a new dimension to image classification tasks [14]. A CNN entails of some layers. The layers of CNN are Convolution layer, Pooling layer, Activation function, and fully connected layer. Our proposed network architecture for crowd counting using CNN is illustrated in Fig. 4. It uses both deep and shallow fully convolutional neural networks and combines the result of both to give accurate crowd count. The network architecture is discussed in detail below.

3.2.3 The Deep Network

An architectural design similar to VGG-16[8] network is used here. The deep network is used to capture high-level semantics required for crowd counting. The VGG-16 network was originally used for image classification, but here, we modify it slightly to make an efficient approach for crowd detection. Originally, VGG-16 network had 5 max-pool layers each having a stride of 2. Thus, the output image was only 1/32 times the input image. We modified this architecture by removing the 5th pooling layer and setting the stride of the 4th max-pool layer to 1. As a result, the output image has a spatial resolution of 1/8 times the input image.

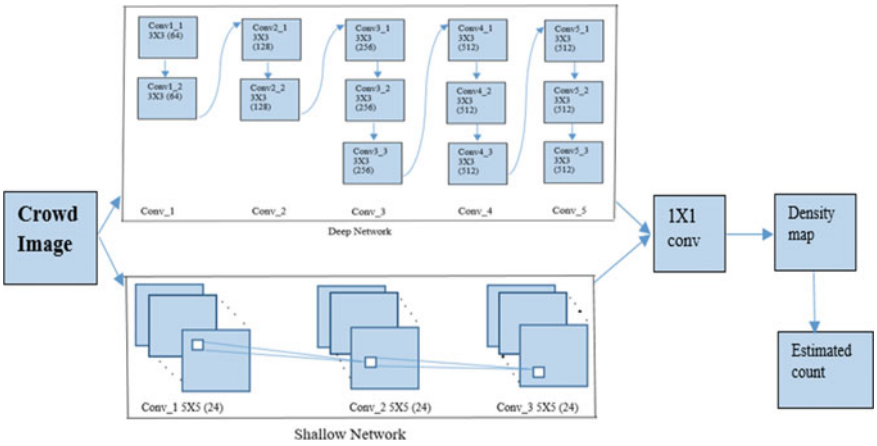

Fig. 4 Proposed architecture of crowd detection using Convolutional Neural Network

3.2.4 The Shallow Network

The people far away from the camera are captured mostly as low-level head blobs, which does not require high-level semantics. So they can be recognized using a shallow convolutional network. Here, we have designed the shallow network of the depth of only three convolutional layers. Each layer is pooled after each convolution in order to match the output resolution with that of the deep network.

3.2.5 Ground Truth

Ground truth is of foremost importance in order to train a fully convolutional network. We have used Gaussian filtering technique to blur each head annotations and generate our ground truth. As a result of such blurring, the summation of the resultant density map becomes equal to the total number of individuals in the image. The density mapping makes the task of training CNN easier since it does not need to locate the exact point of head annotation. It likewise gives the data about the contribution of the crowd in different regions and the CNN to predict both crowd density and crowd count precisely.

3.3 Detecting Overcrowded Situation

The proposed technique process still images captured from the area under observation and then estimating the number of people in the scene. The number of people may vary from place to place, i.e., each place may have a different safe threshold value

Fig. 5 A typical image which is captured actually by the camera

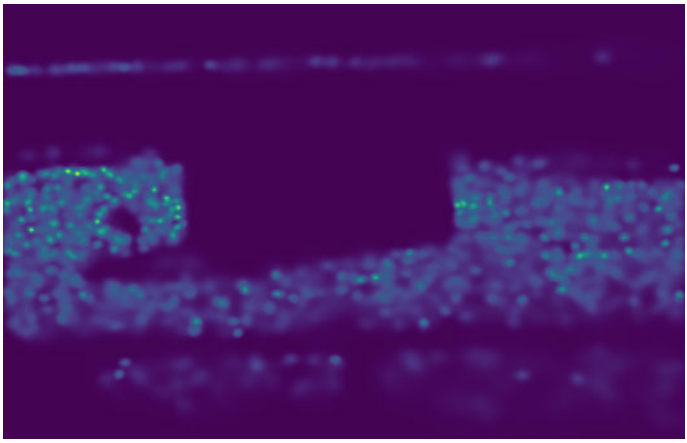

Fig. 6 Generated density map of the captured image

for crowd density. To detect whether the situation is out of control or not, the number of people in a scene is needed to be checked with the threshold value. The captured image and generated density map are depicted in Figs. 5 and 6, respectively. In that case, the control centers are notified, and they will take the necessary steps.

The notification is sent to proper authorities by using a wide area network technology. A typical networked model is shown in Fig. 7. The overall process of capturing an image, processing it to know the estimated count and then sending the information about crowd density is done using Raspberry pi. A network of Raspberry pi is maintained, and WAN technology is used to send the necessary information from one Raspberry pi to another. Several Raspberry pi is connected in the above manner using WAN technology for the proper circulation of information.

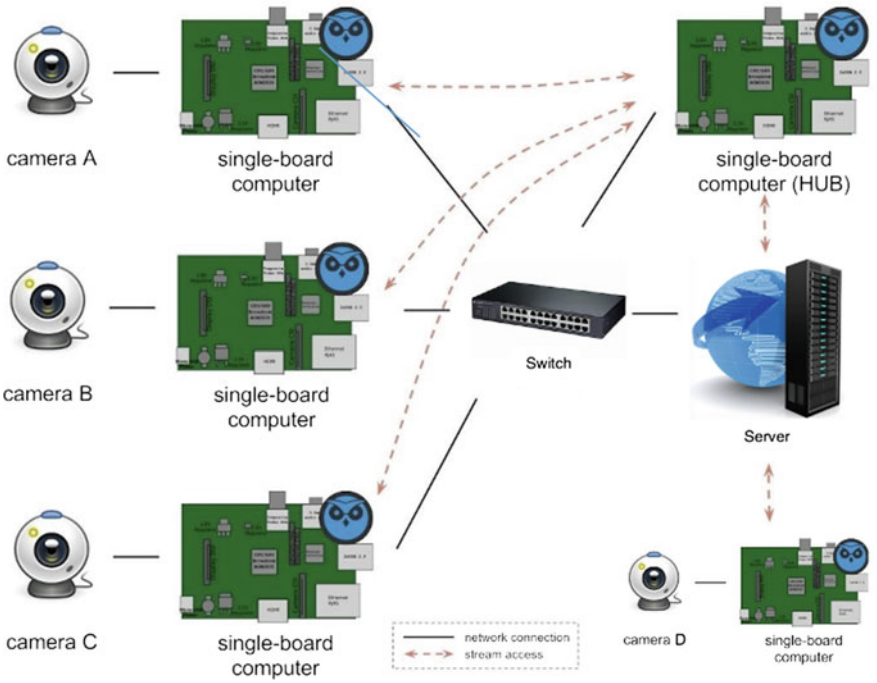

Fig. 7 A typical networked model of the proposed system

4 Experimental Analysis

In this section, we discuss our experimental setup and overall analysis of our proposed technique. The outcome of the implementations of both techniques is discussed here briefly.

4.1 Experimental Tools

The tools that were used to evaluate the performance of the proposed method are as follows:

(i) A Raspberry Pi 2 running raspbian operating system
(ii) A category 6 LAN cable
(iii) Picam 2 camera module for Raspberry Pi
(iv) A computer connected in the same network
(v) Nvidia GeForce GTX 950M GPU.

Raspberry Pi is a small single board computer. It is a quad-core ARMv7 CPU and 1 GB of RAM. Picam is a module for Raspberry Pi which allows the computer

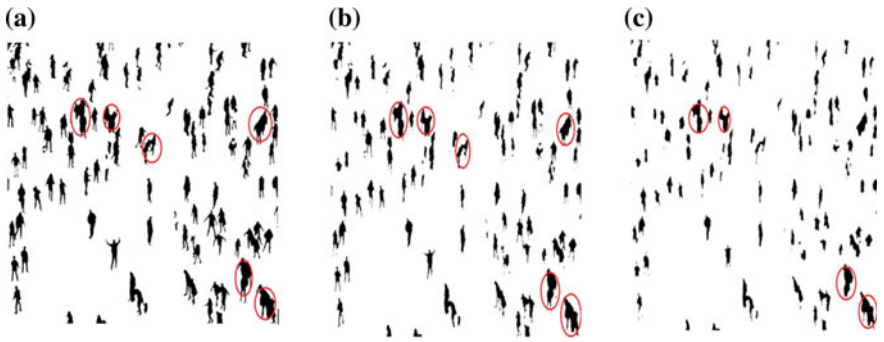

Fig. 8 Noise element. **a** 1st iteration **b** 2nd iteration **c** 3rd iteration

to take photos and record videos. Raspberry Pi and a computer must be in the same network which allows connecting to raspberry pi through SSH service.

4.2 Analysis of Image Processing Technique

From Fig. 8a, we can observe that our first iteration has maximum noise as indicated by the red circles. The second iteration also contains noise but less than the first image which is depicted in Fig. 8b. The last and third iteration shows much better and promising result which is illustrated in Fig. 8c. So, we take an average of each step counts.

4.3 Analysis of Convolutional Neural Network Technique

UCF CC 50 [15] dataset is a challenging dataset of dense crowd scenes from a wide range of scenarios of various gatherings. As perceptible from the name, this dataset contains 50 images converted to grayscale along with head annotations for each image. With varying number of people per image, the average individual per image was found out as 1280. Fivefold cross-validation was used by dividing the dataset randomly into five groups with 10 images in each group.

In each fold of the cross-validation, four groups with a total of 40 images were used to train the network, and the group of 10 images was used for validating its performance. Following previous data augmentation method, 225_225 patches were formed from each 40 training images creating an average of 50,292 patches per fold. Caffe deep learning framework was used to train our deep convolutional network.

If a true head is perfectly identified then it is counted as True Positive (*TP*). But if a non-head region is identified as a head region, then it is called False Positive (*FP*).

Table 1 The Precision–Recall observed for different sample images using Erosion and CNN technique

Image No.	Ground Truth	Erosion Technique	CNN Technique	CNN		Erosion	
				Precision	Recall	Precision	Recall
1	10	11	4	1.00	0.63	0.90	1.00
2	54	59	48	0.95	1.00	0.89	0.98
3	59	49	62	0.87	0.94	0.92	0.86
4	38	21	25	0.88	0.60	0.71	0.60
5	153	122	143	0.80	0.74	0.78	0.89
6	117	129	92	0.91	0.70	0.66	0.72
7	72	57	73	0.88	0.98	0.83	0.76
8	22	39	31	0.64	0.84	0.54	0.69
9	133	106	121	0.78	0.73	0.82	0.58
10	204	149	227	0.69	0.86	0.85	0.60

Also, there is False Negative (FN) associated with the output which occurs if our algorithm misses a head region.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The precision–recall value observed for different sample images using erosion technique and CNN technique is demonstrated in Table 1. Both precision and recall value is important in people counting results. Erosion technique is not able to accurately identify all the true head regions. But CNN technique is more successful in detecting true positives and also the precision is higher than erosion technique. It is also clear that CNN technique has a higher recall value than Erosion technique.

The comparison of the estimated count for each image in the dataset with its actual count is shown in Fig. 9. It also shows the comparison of proposed approaches with the Cross-Scene Counting [10]. Though for most cases the estimated count lies close to the actual count, the model may underestimate the count due to an insufficient number of training images for the very large crowd. Figure 9 indicates that CNN technique is much closer to the ground truth than the other technique.

The Mean Absolute Error (MAE) computes the mean of the absolute difference between the actual count and the estimated count of individuals for every image in the dataset. We calculated the MAE to determine the accuracy of our result. The results shown here do not include any postprocessing techniques. The result of our approach is compared with Cross-Scene Counting [10] which is shown in Table 2.

Thus, the accurate comparison of both techniques is shown in Fig. 10. The accuracy is measured by taking the deviation of the proposed methodology from ground

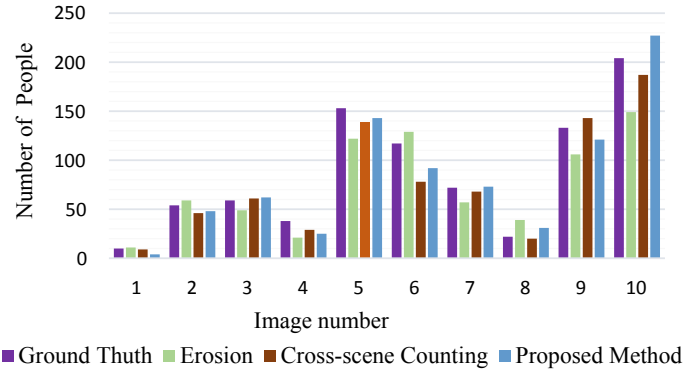

Fig. 9 Comparison of the proposed techniques and Cross-Scene Counting [10]

Table 2 Quantitative results of our approach along with Cross-Scene Counting [10] approach on the UCF_CC_50 dataset

Technique	Mean Absolute Error(MAE)
Cross-Scene Counting [10]	467.0
Proposed technique	436.0

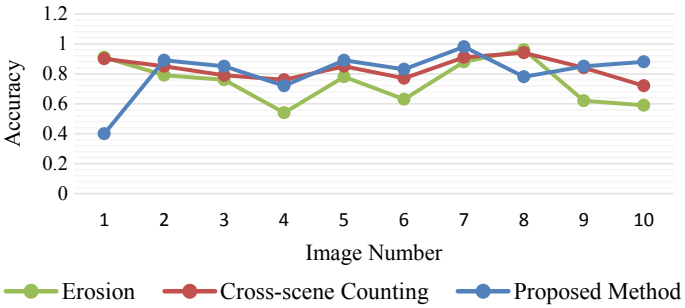

Fig. 10 Accuracy comparison of proposed techniques and Cross-Scene Counting [10]

truth. Figure 10 shows the accuracy curve for test images using the proposed techniques.

Both techniques show promising results. However, CNN with deep and shallow network technique shows more accuracy. The average accuracy of CNN technique is 88%. On the other hand, the average accuracy of Cross-Scene Counting technique is 84%.

5 Conclusion

This paper utilized two different methodologies to estimate the crowd density of an area from its static image. Each of these methodologies has some confinement of its own. Despite the fact that the approach using erosion is computationally fast, it does not give a precise outcome for every scenario. In overcrowded area, the erosion-based technique does not provide much accuracy. In fact, the second technique using Convolutional Neural Network provides more precise outcomes in those situations. Though CNN-based technique may take more time due to several levels of processing, the accuracy of dense crowd detection using CNN is much higher than erosion. As observed from the paper, the CNN methodology obviously outperforms the existing approaches of crowd detection for intensely dense crowd scenes.

References

1. Tang NC, Lin Y-Y, Weng M-F, Liao H-YM (2015) Cross-camera knowledge transfer for multiview people counting. IEEE Trans Image Process 24(1):80–93
2. Liu B, Vasconcelos N (2015) Bayesian model adaptation for crowd counts. In: Proceedings of IEEE international conference on computer vision (ICCV), Santiago, pp 4175–4183
3. Ma HD, Zeng CB, Ling CX (2012) A reliable people counting system via multiple cameras. ACM Trans Intell Syst Technol 3(2):1–22
4. Antic B, Letic D, Culibrk D, Crnojevic V (2009) K- means based segmentation for real time zenithal people counting. In: Proceedings of IEEE international conference on image processing, pp 2565–2568
5. Shao J (2017) Crowded scene understanding by deeply learned volumetric slices. IEEE Trans Circuits Syst Video Technol 27(3):613–623
6. Bansal A, Venkatesh KS (2009) People counting in high density crowds from still images. In: Proceedings of. IEEE international conference on computer vision and pattern recognition, pp 1093–1100
7. Fu H, Ma H, Xiao H (2014a) Crowd counting via head detection and motion flow estimation. In: Proceedings of 22nd ACM international conference on Multimedia, Florida, pp 877–880
8. Fu H, Ma H, Xiao H (2014b) Real-time accurate crowd counting based on RGBD information. In: Proceedings of IEEE international conference on image processing, pp 2585–2568
9. Chauhan RV, Kumar S, Singh SK (2016) Human count estimation in high density crowd images and videos. In: Proceedings of fourth international conference on parallel, distributed and grid computing (PDGC), Waknaghat, pp 343–347
10. Brostow GJ, Cipolla R (2006) Unsupervised bayesian detection of independent motion in crowds. In: Proceedings IEEE computer society conference on computer vision and pattern recognition (CVPR'06), pp 594–601
11. Zhang C, Li H, Wang X, Yang X (2015) Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, pp 833–841
12. Sezgin M, Sankur B (2004) Survey over image thresholding techniques and quantative performance evaluation. J Electron Imaging 13(1):146–165
13. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105

14. Sun T, Wang Y, Yang J, Hu X (2017) Convolution neural networks with two pathways for image style recognition. IEEE Trans Image Process 26(9):4102–4113
15. Reddy K, Shah M (2012) Recognizing 50 human action categories of web videos. Mach Vis Appl 1–11