International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Multi Column Convolutional Neural Network for Accurate Crowd Counting and Analysis in Highly Congested Urban Scenes

Manju G

Department of Computer Science, Govt. College, Ambalapuzha, Kerala, India
Corresponding Author Email: *manjualoshious[at]gmail.com*

Abstract: ***The estimation of highly congested, highly varied crowded scenes is a challenging vision task that has received a lot of interest in recent years. Crowd counting and analysis aims to count the number of people and make an analysis of the density of the crowded scene. Exponential growth in the world population and the resulting urbanization has led to an increase in the number of activities such as sporting events, political rallies, public demonstrations which would thereby result in a more frequent crowd gathering. In such situations, it is essential to analyze crowd behavior for better management, safety and security. In this paper, a Convolutional Neural Network (CNN) based approach is used for the application. Among the various approaches, Multi Column Convolutional Neural Network is used to train the network in order to estimate the number of people. Also, the crowd analysis task is associated with many challenges such as non-uniform density, intra-scene and inter-scene variations in scale, occlusions and perspective.***

Keywords: Convolutional Neural Network, Crowd counting, Digital image processing

1. Introduction

The exponential growth of the global population and the subsequent surge in urbanization have led to an increase in activities that gather large crowds, such as sporting events, political rallies, and public demonstrations. These crowded scenes pose significant challenges for public safety and security, necessitating effective crowd management and behavior analysis. Crowd counting and analysis, which involves estimating the number of individuals and understanding the density distribution within a crowd, is a crucial task in this context. Accurate crowd analysis helps in the efficient allocation of resources, prevention of overcrowding, and timely response to emergencies.

Traditional methods for crowd counting often rely on manual observation or simplistic algorithms that struggle to handle the complexities of real-world scenarios. These methods are frequently hampered by challenges such as non-uniform density, variations in scale within and between scenes, occlusions, and perspective distortions. As a result, there is a growing need for advanced techniques that can accurately and efficiently analyze crowded scenes under diverse conditions. Digital image processing is a growing technology that has experienced continuous and significant expansion in a period of years. The usefulness of this technology is apparent in many different applications covering medicine to remote sensing.

Crowd counting is an important application of image processing technology. Crowd counting otherwise called crowd estimating is a process used to count or estimate the number of people in a crowd. The most direct way is to actually count person by person in the crowd. Detection based approach is a supervised learning method. In this method, a classifier is trained by using a labelled set of training data which consists of full body shots of people. Detection based crowd counting using rescaling method aims to make a system which can count the people in a crowd irrespective of the size of the people. It can count very tiny faces to very large faces and can also correctly identify blurred people images. The number of people from a video input can be correctly obtained by this method. Frame by frame is extracted from the video to make the count of the crowd.

In recent years, Convolutional Neural Networks (CNNs) have emerged as a powerful tool for various computer vision tasks, including crowd counting. CNNs are capable of learning rich feature representations from raw pixel data, making them well-suited for handling the visual complexities of crowded scenes. Among the various CNN architectures, the Multi Column Convolutional Neural Network (MCNN) has shown particular promise for crowd counting applications. The MCNN architecture is designed to address the challenges of non-uniform density and scale variations by using multiple columns with different receptive fields to capture features at various scales.

In this paper, we propose a CNN-based approach for crowd counting and analysis using the MCNN architecture. Our method aims to accurately estimate the number of people in highly congested and varied crowded scenes. We train the MCNN to learn robust features that can handle the diverse challenges associated with crowd analysis, such as intra-scene and inter-scene variations, occlusions, and perspective distortions. By leveraging the strengths of MCNNs, our approach provides a reliable solution for real-time crowd management and safety applications.

2. Literature Survey

In Computer Vision Technology estimating crowds from images or videos accurately has become an increasingly important application for purposes of crowd control and public safety. Over the last few years, researchers have attempted to address the issue of crowd counting and density estimation using a variety of approaches such as detection-based counting, clustering-based counting and

regression-based counting. The initial works on the regression-based methods mainly uses handcrafted features and the more recent works uses Convolutional Neural Network (CNN) based approaches [1, 2].

a) Detection based crowd counting [3]

Detection based approach is a supervised learning method. Basic methodology involves scanning the image using sliding window detector Detection is usually performed either in the monolithic style [3] or parts-based detection[]. Min li *et al*. [4] combines a foreground segmentation algorithm based on MID (Mosaic Image Difference)[] and a head-shoulder detection algorithm based on HOG (Histograms of Oriented Gradients)[4].

b) Regression based crowd counting [5]

First Identifies the perspective map of the region of interest, then the extraction of low-level features such as the foreground pixels or edges from each segmented cell region and mapped to the learned regression model for generating a structured output that estimates the crowd count in each local region simultaneously. Different regression techniques such as linear regression [6], piecewise linear regression, ridge regression, Gaussian process regression and neural network [7] are used to learn a mapping from low-level feature to the crowd count.

c) CNN based crowd counting [8][9]:

CNN based techniques can be classified based on the property of the networks such as basic CNNs, Scale-aware models, Context-aware models, multi-task frameworks and based on inference methodology such as Patch-based inference, Whole image-based inference.

The CNN-based approaches have demonstrated significant improvements over previous hand-crafted feature-based methods, and thus, encouraging more researchers to explore CNN-based approaches further for related crowd analysis problems. While the CNN-based methods are very effective in large density crowds with a diverse scene conditions, the traditional approaches suffer from high error rates in such scenarios [10].

Table I: Existing Systems

Approach	Major Technique used	Advantages	Drawbacks
Detection based	Monolithic style or parts-based detection.	These systems work well for detection of faces.	It is not successful in the presence of extremely dense crowds and high background clutter. It is also time consuming.
Feature Regression based	Different regression techniques used are linear regression, piecewise linear regression, ridge regression, Gaussian process regression and neural network	Regression based methods help in extracting lower levels features.	The model trained is dependent on the perspective map. If the model were to be used in another scene of a different perspective map, it will have much inaccuracies in its result.
CNN based	Basic CNNs, Scale-aware models, Context-aware models, Multi-task frameworks.	Achieved drastically lower error rates and also the creation of new datasets has enabled learning of more generalized models.	Quality of the density maps obtained is poor although accurate count estimates are obtained.

Therefore, from the traditional approaches it can be found from the Table I that CNN-based methods outdo them across all datasets. From among the CNN-based methods, it is found out that most performance improvement is achieved by scale-aware and context-aware models [11].

3. Proposed System

Crowd counting system can estimate the count of people in the given input. The people are correctly detected using a method called Tiny Faces algorithm. The main advantage of this algorithm is that it can identify small faces and can also work in blurred faces. It is the best performing face detector from amongst other algorithms – given the difficulty of many of the faces, it gives a very good result. Estimations of the crowd density per pixel is always a challenging task due to the large variation of the crowd density values. The number of people in a frame can vary in large range. Some image frames may contain hundreds of people, while others might have only a few numbers. It is very difficult for a single CNN to handle the entire spectrum of crowd densities of varying scales. So, shared CNNs is used in this system.

(a) Single Image Crowd Detection and Counting

In the single image crowd detection method, system takes an image as the input feed. The Tiny Faces algorithm works as follows. The input image to the system is gone through several processes to obtain the count of the people and different analysis of the crowd scenario. The image frame is initially rescaled to multiple scales. This rescaling of images is known as interpolation. Each image is rescaled so that the faces in the images are in the range in which the classifier has the highest efficiency Image interpolation occurs when an image is resized or distorted from one-pixel grid to another. Image resizing is necessary to increase or decrease the total number of pixels, whereas remapping occurs when correcting for lens distortion or rotating an image takes place. Each image frame is passed as an array of pixels. The minimum scale is found by taking minimum of the log2 of cluster height / image height and cluster width / image width. Image Interpolation works by using known data pixels which is used to estimate values at unknown points. Image interpolation works in two different ways, and it tries to achieve a best approximation of a pixel's intensity, based on the values at the surrounding pixels.

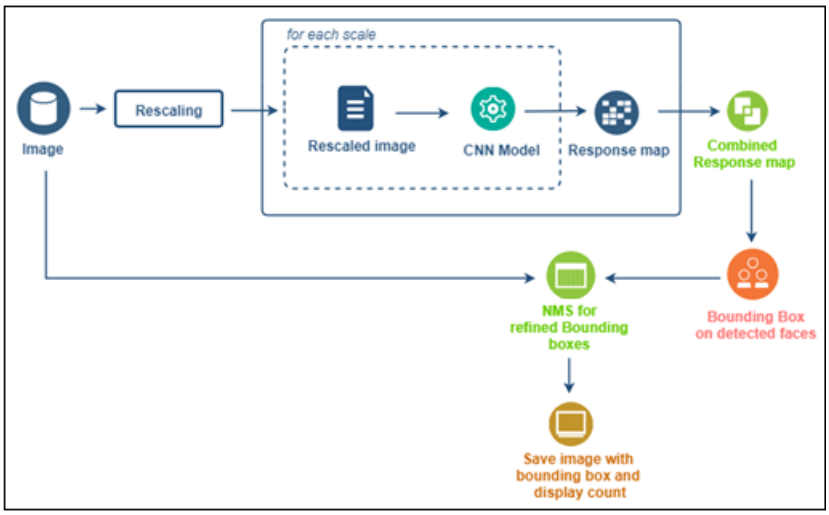

Figure 1: Block diagram of the system for single image

In the system, a small template is necessary to detect small faces and a large template is necessary to exploit detailed features facial parts to increase accuracy. Instead of using a single detector, separate detectors are trained to be tuned for different scales and aspect ratios. But, training a large collection of scale-specific detectors may have problems of lack of training data for individual scales and inefficiency from running a large number of detectors at test time. To address both these concerns, training and running of scale-specific detectors takes place in a multitask fashion so that it makes use of features defined over multiple layers of single feature hierarchy. For the objects of novel sizes, a simply strategy is employed that is to resize the images at test-time by interpolation and decimation. Interpolating the lowest layer of the finitely discretized image pyramid is particularly crucial for finding small objects.

There are seven rescaled inputs to the network. The interpolations are of size 0.0625X 0.125x, 0.5x, 1x, 1.4142x, 2.0x and 0.25x. The scaled input serves as entries to the Convolutional Neural Network (CNN) to predict the response maps at every resolution. ResNet101 is the Convolutional Neural Network used here. ResNet101 is trained on ImageNet dataset which contains over 1 million images. The model can classify nearly 1000 objects. The concept of Transfer learning is applied on this model. Each input image will pass through a series of convolution layers with filters, Pooling, fully connected layers (FC) and finally SoftMax function is applied to classify an object with the probabilistic values between 0 and 1.

The output is then mapped onto a feature map, which contains the coordinates of the classified face. A bounding box is drawn around the face region placing it best centered. By multiplying the coordinates of the feature mapped region by a spatial scale, it can be ensured that it fits the feature map just like the original bounding box would fit the original image. The size of the spatial scale is 1N where N is the sum of all the strides of convolutional layers that image was processed. The image is then rescaled to the original size. This is just the reverse process of giving rescaled input. All the rescaled images are merged onto a single image. In this case, there are chances of occurrence of overlapped bounding boxes around the same face. In order to bring to a single bounding box around the face, a technique called non maximal suppression is used. The idea behind non-maximum suppression is to reduce the number of detections in a frame to the actual number of objects present. The count of the number of people detected is returned as the output based on the number of bounding boxes in the image. It gives an estimation of the total crowd count.

(b) Crowd Counting and Analysis in Video

It takes the video file as input and extract each of the frames and process it. Then it produces analytics and insights based on the data. The video feed cannot be taken as such to give out the output. Therefore, from the video input, frames are extracted one by one. Each frame undergoes all the processes that an image is gone through.

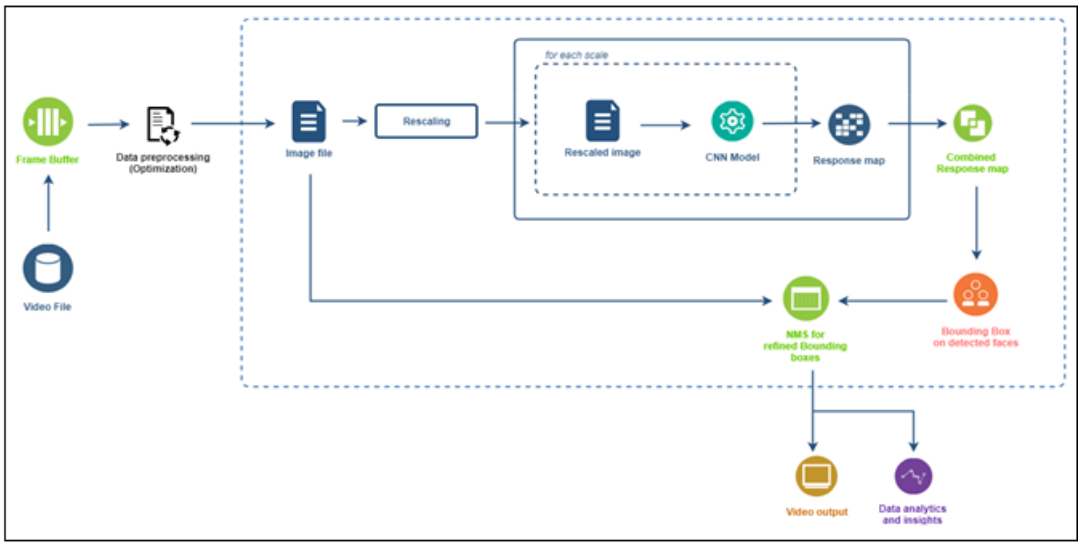

Figure 2: Block diagram of the system for video feed

Each and every frame are not taken for processing. For better optimization, some of the adjacent frames are skipped. This is because a video can have numerous frames and the adjacent frames are not much different. The count of the people will be almost same. The frame is then preprocessed and given to the system. Then the remaining working is same as that of the single image crowd counting method. The image frame is rescaled and given to the CNN network. The classified faces are obtained from the response map. Based on the response map, bounding boxes are drawn. Multiple bounding boxes are refined using Non-maximal suppression method. The count produced is displayed as output. We can get the count of all the frames selected.

4. Results and Discussions

The project was able to build an incremental count of the number of faces detected from single images as well as the video moves forward. The algorithm used was able to outperform most of the other face detection algorithm available. The experimental results are obtained by training lot of input images. WIDERFACE crowd dataset is used for training purpose. WIDER FACE dataset is a face detection benchmark dataset, of which images are selected from the publicly available WIDER dataset. It consists of 32,203 training set images and label 393,703 faces with a high degree of variability in scale, pose and occlusion as depicted in the sample images. WIDER FACE dataset is organized based on 61 event classes. The system was tested on the testing dataset of the WIDER FACE. The maximum size of rescaling applied was 2.

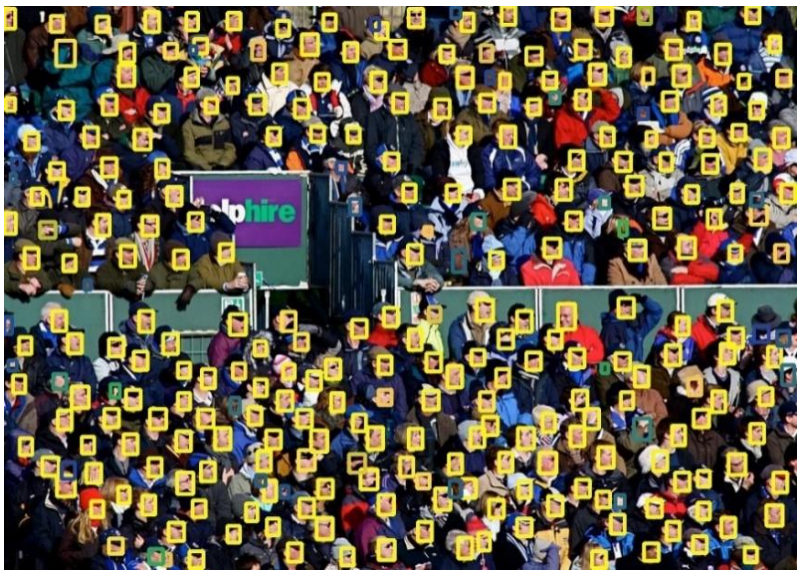

Figure 3: Crowd count image

From the results, the accuracy of the system could be calculated. The accuracy of the system is about 67% based on the given dataset. The MSE of this system is 332.15 on the WIDERFACE dataset, which is a good MSE value as compared to various other detection algorithms [16]. There are certain factors which is affecting the accuracy of the output. They are:

(i) Resolution of the Image

Higher the resolution of the images, better the results obtained. The reason behind this is when the algorithm resizes the input image it would preserve the quality of the image which makes the detection of faces easier. Higher resolution images will have less noise so that it gives the algorithm more data to work with.

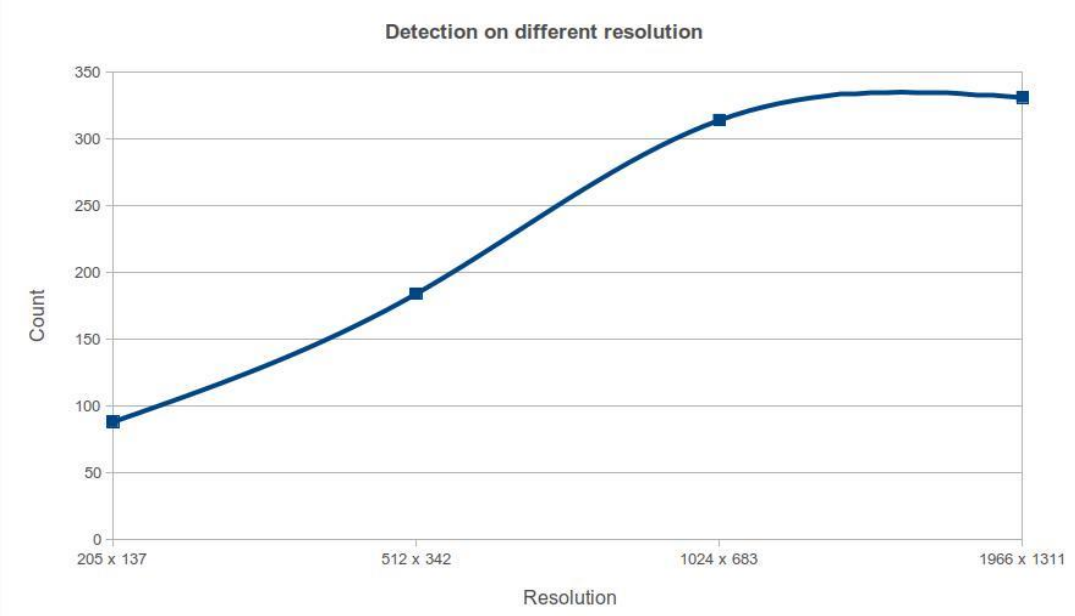

Figure 4: Count of people vs resolution of image

Table 2: No of faces detected in various resolution images

Resolution	Count	GT
205 x 137	88	381
512 x 342	184	381
1024 x 683	314	381
1966 x 1311	331	381

(ii) Size of faces in the image

The algorithm works best for a certain size range of faces. For finding the small (25x20) faces, building templates at 2x resolution improves overall accuracy by 6.3% whereas for finding large (250x200) faces, building templates at 0.5x resolution improves overall accuracy by 5.6%.

(iii) Maximum scale factor

Increasing the maximum scale factor can bring more faces in the optimum detection range which can thereby improve the detection rate. This only works for some range of images where the faces are very small. Increasing the maximum scale factor also increases the processing time for the image.

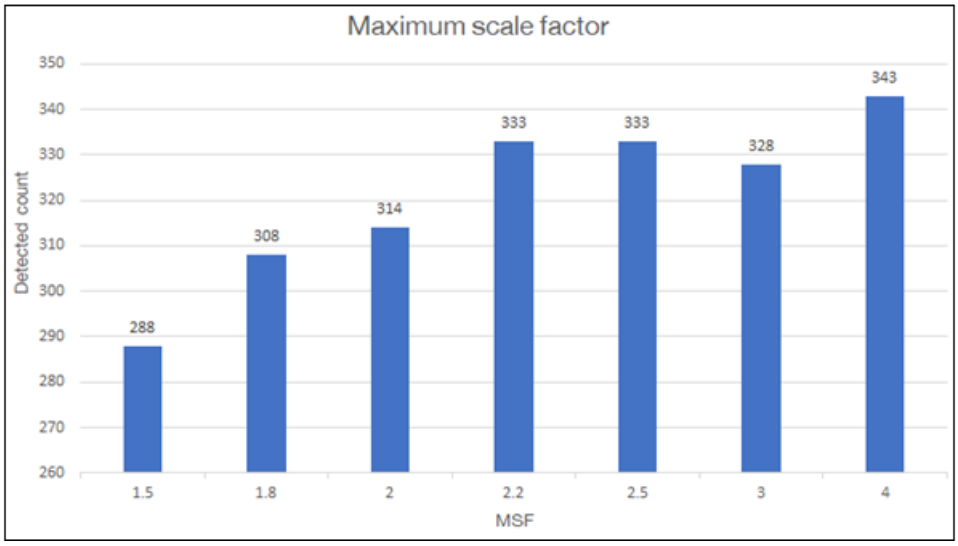

Figure 5: Maximum Scale factor vs Detected count

But as the Maximum Scale Factor increases the time taken for processing increases exponentially. Therefore, large scale factors are not preferable. The graph shows the relationship between Maximum Scale Factor and the time taken for processing. So, based on these results from the graph, a Maximum Scale Factor of 2 is selected. This would give a good result with comparatively low processing time.

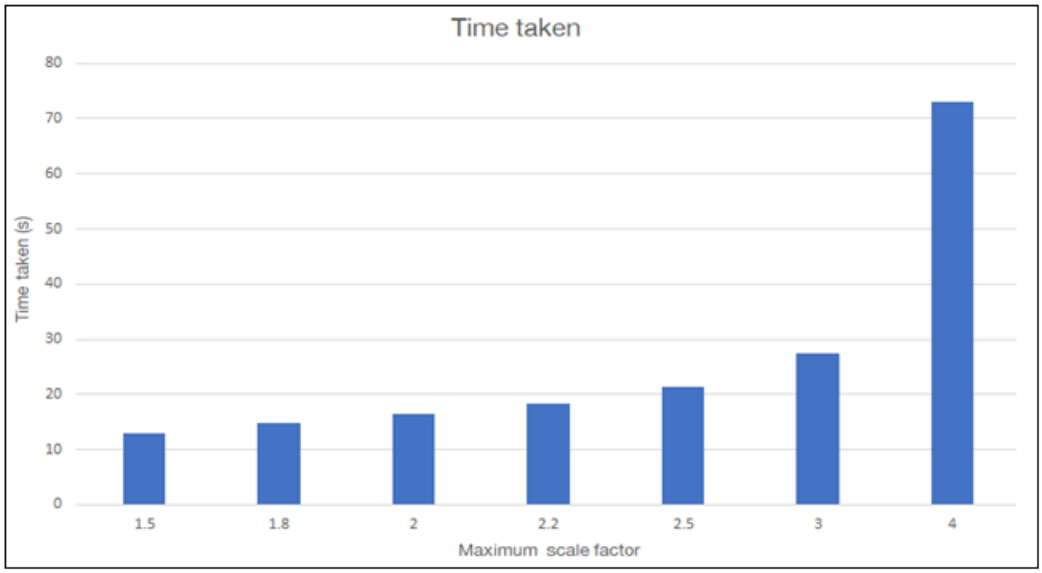

Figure 6: Maximum Scale factor vs Time taken for processing

(iv) Confidence threshold

This is the confidence of the network in classifying input as a face or not. Reducing this value will give more detections but it will also create more chances of false detections in image which reduces the accuracy.

(v) Face orientation

Face orientation plays a major role in the detection. The algorithm works better on faces that are facing straight. The algorithm sometimes fails to detect when the head is turned backward that is, facing backward.

When a video input is given, it outputs the total number of people present currently in the given frame. The algorithm splits the video input into single frames and process it for finding faces. The video source could be a recorded video file or a video input source such as a webcam or surveillance camera. Some outputs of the algorithm are given below

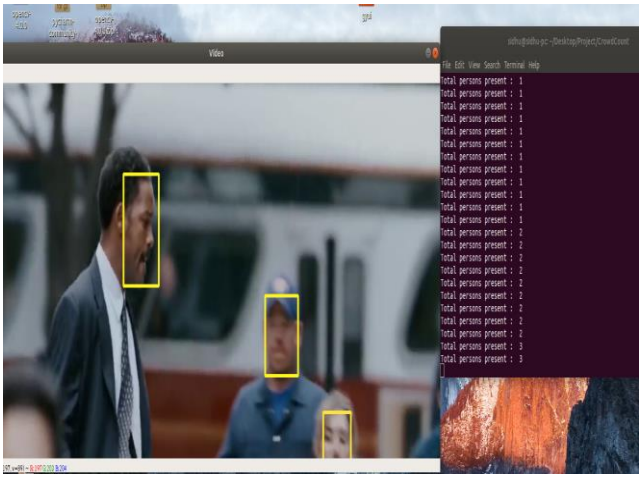

Figure 7: Counting of crowd from a video source

The analysis of the crowd count at a specific time interval is examined. A count versus time interval graph is generated as the output. This is useful data for obtaining peak time of the crowd. This graph can be used for analysing the crowd trends during some specified time. Also, it can be used for identifying any anomalous behaviour or for tracking a crowd.

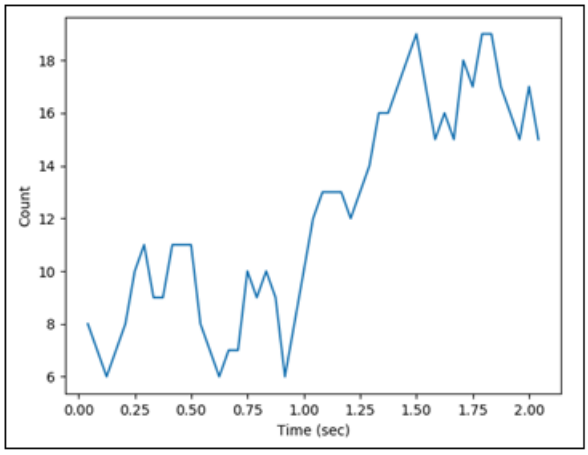

Figure 8: Crowd count versus time interval

5. Conclusion and Future Scope

This project is a simple yet effective framework for finding small objects, demonstrating that both large context and scale-variant representations are crucial. It is specifically shown that massively-large receptive fields can be effectively encoded as a foveal descriptor that captures both coarse context (necessary for detecting small objects) and high-resolution image features (helpful for localizing small objects).

One main problem to this algorithm is it can't detect duplicates, when someone re-enters into the frame the person is counted again as a new person. The algorithm could be extended to overcome this problem. For each face detected a feature vector for the face could be generated and saved temporarily. Using this the count could be made without any redundancy. The algorithm performs less accurate for images having faces that are turned other side. Using another CNN trained for counting head could be pipelined with the ResNet output.

The quality of estimated crowd density maps is another key issue ignored by earlier research. Most of the existing CNN-based approaches in their networks have a number of

max-pooling layers that force them to regress on down-sampled maps of density. Another important issue to address in the future would be to generate high-quality density maps along with low count estimation error. Given the challenge of training deep networks for new scenes, it would be important to explore how to leverage existing sources from models trained. Most of the existing methods retrain their models in a new scene and it is not practical to do so in real-world scenarios because it would be expensive to get annotations for each new scene. However, the idea of transfer learning or domain adaptation for crowd scenes is relatively unexplored and is a emerging area of research.

The given algorithm exceeds some of the recent algorithms for face detection. However, the most recent and accurate face detection algorithms such as FaceBoxes, Single Stage Headless Face Detector or RetinaNet[20] can be compared in future work. The general approach was also to detect small objects in images, but here the focus is on faces, but this approach can be applied in other small object such as calcifications in mammography pictures. In addition, a simple pipeline and approach is used here for tracking and counting people across frames. Combining with tracking algorithms like DeepSort can be done for future work.

From analyzing the results of the experiment, it is observed that a CNN works best for images that are similar to images that it was trained with. But a person who is turned back wouldn't be detected as a person, as the CNN is not trained in that manner. If another CNN model trained on back images or turned heads could be generated giving the output from the first CNN as the input then it would provide better detections thus improving the accuracy of the system.

The analysis from the output generated could be useful for numerous applications. The surveillance video could be processed to generate a density map, which points the areas in the video where the presence of crowd is higher. This can also be used to analyse the crowd interest of the region. Further it can be used to design streets, shops, cities etc. based on the data manipulated from the crowd. The flow of crowd in a region could be analyzed to develop better streets, pathways and roads so that can direct the crowd flow in a way that reduces congestion and make commuting faster for the crowd. Abnormal activities such as fights, accidents, outbreaks, hazards can be detected by analyzing the crowd. These activities could be controlled by triggering the required actions automatically by the system

References

[1] O. Shoewu and O.A. Idowu, "Development of Attendance Management System using Biometrics ", The Pacific Journal of Science and Technology, Vol. 13, Number1, pp.300-307, May 2012 (Spring).
[2] T. Lim, S. Sim, and M. Mansor," RFID based attendance system ", in Industrial Electronics and Applications, 2009. ISIEA 2009. IEEE Symposium on, vol. 2. IEEE, 2009, pp. 778782.
[3] S. Kadry and K. Smaili," A design and implementation of a wireless iris recognition attendance management system ", Information Technology and control, vol. 36, no. 3, pp. 323329, 2007
[4] Shilpi Singha ,S.V.A.V .Prasad"Techniques and Challenges of Face Recognition: A Critical Review" 8th International Conference on Advances in Computing and Communication (ICACC-2018),Elsevier
[5] Chen, K., Loy, C.C., Gong, S. and Xiang, T., 2012, September. Feature mining for localised crowd counting. In BMVC (Vol. 1, No. 2, p. 3).
[6] Dwi Sunaryono , Joko Siswantoro , Radityo Anggoro "An android based course attendance system using face recognition",ScienceDirect, Journal of King Saud University – Computer and Information Sciences, 18 January 2019.
[7] Nilesh D. Veer, B. F. Momin ,"An automated attendance system using video surveillance camera" IEEE International Conference On Recent Trends In Electronics Information Communication Technology, May 20-21, 2016.
[8] N. Kar, M. Kanti Debbarma, A. Saha, and D. Rudra Pa: Study of Implementing Automated Attendance System Using Face Recognition Technique, Volume 1, No. 2, international Journal of Computer and Communication Engineering (2012).
[9] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556..
[10] Using Face Recognition Technique," in *Multiple Access*, N. Abramson, Ed. Piscataway, NJ: IEEE Press, 1993, ch. 3, pp. 121-123.
[11] G. O. Young, "Synthetic structure of industrial plastics," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15-64.
[12] M. B. Kasmani, "A Socio-linguistic Study of Vowel Ha rmony in Persian (Different Age Groups Use of Vowel Harmony Perspective," *International Proceedings of E conomics Development and Research*, ed. Chen Dan, p p. 359-366, vol. 26, Singapore, 2011.
[13] W. D. Doyle, "Magnetization reversal in films with biaxial anisotropy," in *Proc. 1987 INTERMAG Conf.*, 1987, pp. 2.2-1-2.2-6.
[14] G. W. Juette and L. E. Zeffanella, "Radio noise currents n short sections on bundle conductors," presented at the IEEE Summer Power Meeting, Dallas, TX, June 22-27, 1990.
[15] J. Williams, "Narrow-band analyzer," Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.
[16] N. Kawasaki, "Parametric study of thermal and chemical nonequilibrium nozzle flow," M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.
[17] J. P. Wilkinson, "Nonlinear resonant circuit devices," U.S. Patent 3 624 12, July 16, 1990.
[18] *Letter Symbols for Quantities*, ANSI Standard Y10.5-1968.
[19] *Transmission Systems for Communications*, 3rd ed., Western Electric Co., Winston-Salem, NC, 1985, pp. 44-60.
[20] *Motorola Semiconductor Data Manual*, Motorola Semiconductor Products Inc., Phoenix, AZ, 1989.
[21] R. J. Vidmar. (August 1992). On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. 21(3). pp. 876-880. Available: http://www.halcyon.com/pub/journals/21ps03-vidmar