Crowd Counting Using CNN

Divyesh Tharakan[1], Nilima Kulkarni[2], Harshil Shah[3], Bilwa Vyas[4]
[1,2,3,4]*Department of Computer Science and Engineering, MIT School of Engineering, MIT Arts Design and Technology University, Pune,412201, India*

***Abstract* - Increase in population all over the world has opened a pathway to large gatherings for many causes. It might be grand openings of hotels and institutions, award ceremony, local football matches, riots and much more. Analysing the crowd formation before is quite a difficult task and therefore would not come under control. Due to this, many accidents have occurred in the past taking away innocent lives. The technique, which is Crowd Counting, is an approach to analyse and detect the crowd, that is, to estimate the number of people gathered in a certain area. In this paper, a CNN model is presented to estimate the count of a particular crowd. We have implemented the VGG-16 model for a deep layered image classification and henceforth a better understanding of each image. Shanghai Crowd Counting Dataset has been used which contains more than 400 images for training and testing.**

***Index Terms* - CNN, Crowd Counting, Estimation, Prediction.**

1.INTRODUCTION

As the world is evolving day by day with new technology, there is also evolution/increase in population in world and seeing the new normal conditions created due to pandemic, avoiding huge crowd gatherings at public places (maintaining social distance) has become top priorities.

There are many algorithms proposed in the literature for crowd counting [2,5,8]. Some algorithms are not quite efficient to identify each and every person or an object in an image [6,8]. They tend to leave out the ones which overlap with some parts thereby not predicting the exact or a roundabout count. Some methods intended to use the technique for video sequences specifically [5]. Proper alarms or a scenario to reach out to authorities were not being able to setup and accidents like the stampede on the banks of Godavari river in 2015 occurred which created a lot of chaos.

Counting the crowd task is a challenging one because the model has to overcome those differentiating factors like pose, expression, orientation, height of a particular person to properly count in an image. Consider the below image and let us predict the count of the people present in the gathering.

Fig.1: Crowd Counting Example Image (Source: Shanghai Dataset)

It is clear from figure 1 that crowd is in unpredictable number and if one decides and start counting physically, maybe it may take a day or two. A CNN model can do the same task within a few minutes [1,2,11].

There are various methods for Crowd counting in the literature:

1. Detection-Based Methods: Here, we use a moving window-like detector to identify people in an image and count how many there are. Although these methods work well for detecting faces, they do not perform well on crowded images as most of the target objects are not clearly visible.
2. Regression-Based Methods: We first crop patches from the image and then, for each patch, extract the low-level features.
3. Density Estimation-Based Methods: We first create a density map for the objects. Then, the algorithm learns a linear mapping between the extracted features and their object density maps.
4. CNN-Based Methods: Instead of looking at the patches of an image, we build an end-to-end regression method using CNNs. This takes the entire image as input and directly generates the crowd count. CNNs work really well with

regression or classification tasks, and they have also proved their worth in generating density maps.

CNN is being used to deploy a deeper model to train every image and generate density.

Fig.2: Crowd Counting Image (Source: Google)

This scenario is of a public live performance. Manually it would take days to estimate the exact count. These types of functions actually need the previous data to analyse and understand the density of the gathering and accordingly decide their next grand show. An accurate and efficient count would do well in this case and hence avoid chaos.

Table I: Literature Survey

Sr. No	Author Name	Method	Observation
1	Sabrina H. et.al., 2019	Erosion method	Erosion technique could not accurately identify all head regions of the individuals.
	Akbar K. et.al., 2020	Deep CNN	Scale Driven Convolutional Neural Networks are considered as best models with highest accuracy.
	Muhammad W. et.al., 2017	Mobile phones and Internet data Network.	Less Accuracy as compared to others.
	Junyu G. et.al., 2019	C^3 Framework in CNN	This framework helps in reducing the human cost in training process.
	Jugal K. et.al., 2020	Raspberry Pi with WiFi	In this model, it easily counts the humans in dense areas by using RFID tags whereas camera module easily identify the humans body.
	Mayur D. et.al.,2018	CNN	Performs whole image-based count but requires a large dataset training.

II. PROPOSED SYSTEM

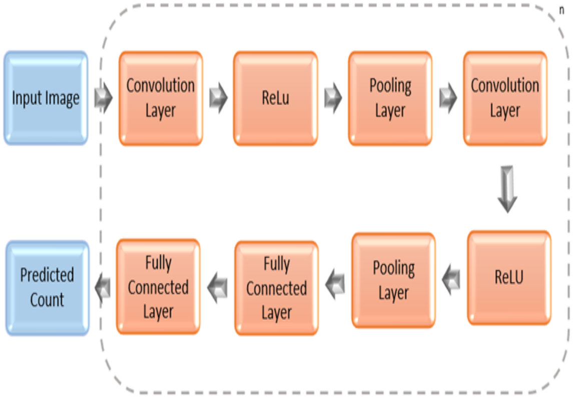


Fig.3: Flow Chart Representation of CNN

CSRNet technique, a deeper architecture, implements VGG-16, which captures high-level features and comes up with 1/8th of the initial size of the image.

At the back end, dilated convolution layers are used whose main aim is to increase the size of the kernel without changing the parameters or in other words keeping it constant. Dilation rate will be increasing but size remains the same.

Convolution Layer: An input image creates patches which are then passed on to the convolution layer for filtering. This layer itself has many inner layers which helps to filter according to the size of the image, the filter, also known as kernel, crops image into 9 patches. One single patch takes around only 1/4th of the original size. It takes input as parameters with a 3x3 kernel, for example, and using those bunch of filters it creates some activation layers.

Another factor which comes into existence is stride. If we provide a stride of, let us say, value 1, and then the output dimension will be same as the original one. Incrementation of the stride value checks for every other patch. It can also be called as down-sampling. The main aim here is to extract activation layers for the uniformity of the process.

Nonlinear Activation layer: Here it activates ReLU (Rectified linear unit layer) which reduces the size or dimension of the image created previously in the convolution layer. Non-linearity is often considered while performing CNN on patches. These activations are most commonly used while applying the algorithm. The volume remains unchanged, thus increasing the size of the kernel.

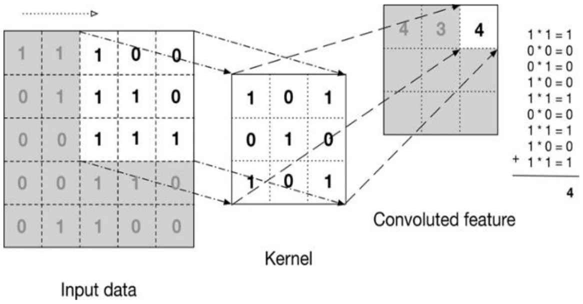


Fig.4: Working of Convolution operation (Image Source: Medium)

Pooling layer: Pooling layer further reduces the size of the image into half. If we have a window size of 2x2, it will check for every 2x2 kernel size and create pixels. Going through this, it will be able to focus more on every feature of the image and work on it.

Max pooling is the most common approach while pooling is implemented. CNN is known for its computary and memory heavy issues. It down samples and hence reduces the size and computational complexity and memory complexity.

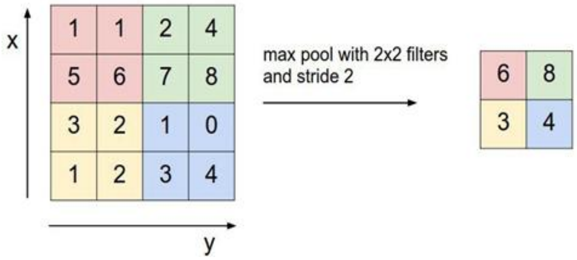


Fig.5: Max Pooling On 4x4 Layer (Image Source: GitHub)

Fully connected layer: The objective over here is to identify or detect different and final output categories and hence it lies in the output section. Every node at the input is connected to the co-efficiency due to which heavy data is loaded here. As the size reduces on and on, this FC layer picks the top three from them and performs probability distribution algorithms like the SoftMax layer.

Above all can be understood using this example:

The major disadvantage for this layer is that it is low driven as huge sum of data is handled and processed which leads to non-time consumption. Rather than being data heavy, people move to compute heavy to balance the architecture.

VGG-16 model:

VGG-16 is a model which comes under very deep convolution networks and is most commonly used to extract deep features of an image or input and to classify them while reducing the size of the input. As we had mentioned earlier, it takes only 1/8th of the original size because of its strong learning ability.

For example, if we take 224 x 224 x RBG input, then after applying convolution layers with 64 filters, and max pooling, the filters double up from 64 to 128 and then again max pooling occurs which doubles up again from 128 to 256 and so and so forth. Then Fully Connected layers comes in phase with around 4096 filters.

The HP laptop has been used to implement the work. Below are its specifications:

OS: Windows 10

RAM: 8GB

Storage: 1TB

III. EXPERIMENTAL RESULTS

A. DATASET

The model trained in proposed work uses the famous crowd counting dataset, Shanghai Tech Dataset[6] which has been clearly grouped and divided into two parts, each containing around 400 images for training along with ground truth values and 125 images for testing. Generating density on each and every training image takes around 3 hours so that a threshold value or the ground truth can be considered for further analysis.

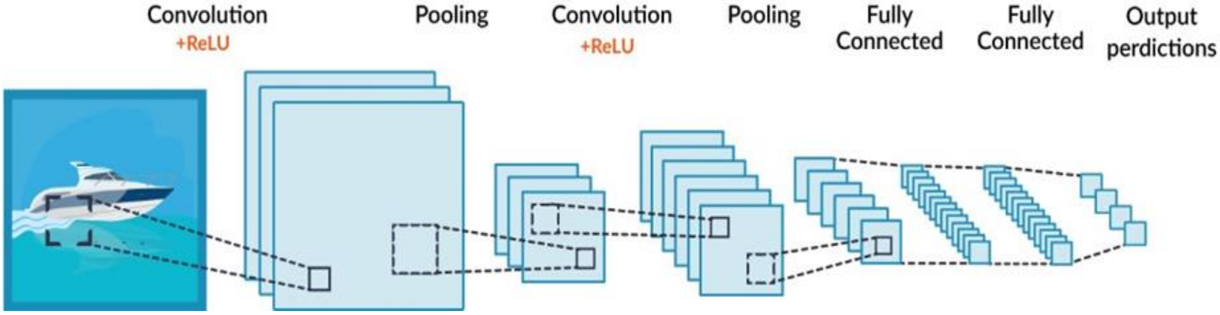
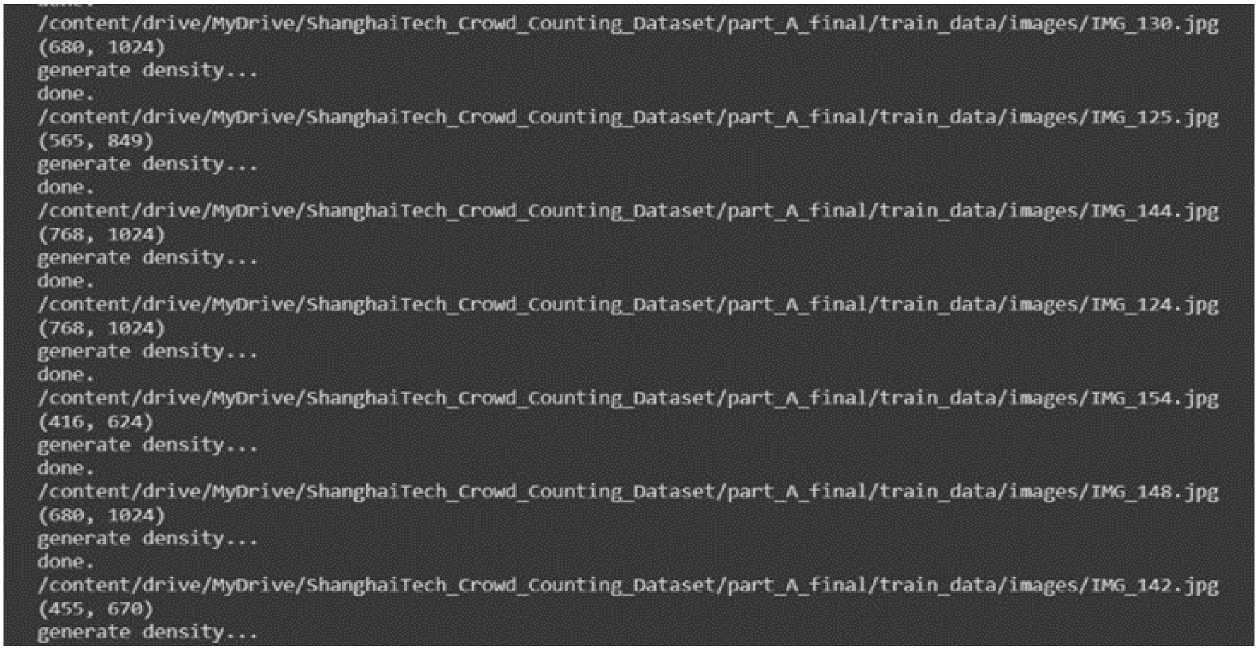


Fig.6 : Convolution Neural Network Breakdown

After training and collecting MAE value, we can test the model on various different images to check the predicted count.



```
/content/drive/MyDrive/ShanghaiTech_Crowd_Counting_Dataset/part_A_final/train_data/images/IMG_130.jpg
(680, 1024)
generate density...
done.
/content/drive/MyDrive/ShanghaiTech_Crowd_Counting_Dataset/part_A_final/train_data/images/IMG_125.jpg
(565, 849)
generate density...
done.
/content/drive/MyDrive/ShanghaiTech_Crowd_Counting_Dataset/part_A_final/train_data/images/IMG_144.jpg
(768, 1024)
generate density...
done.
/content/drive/MyDrive/ShanghaiTech_Crowd_Counting_Dataset/part_A_final/train_data/images/IMG_124.jpg
(768, 1024)
generate density...
done.
/content/drive/MyDrive/ShanghaiTech_Crowd_Counting_Dataset/part_A_final/train_data/images/IMG_154.jpg
(416, 624)
generate density...
done.
/content/drive/MyDrive/ShanghaiTech_Crowd_Counting_Dataset/part_A_final/train_data/images/IMG_148.jpg
(680, 1024)
generate density...
done.
/content/drive/MyDrive/ShanghaiTech_Crowd_Counting_Dataset/part_A_final/train_data/images/IMG_142.jpg
(455, 670)
generate density...
```

Fig.7: Training the model with Shanghai-Tech Dataset

B. PERFORMANCE

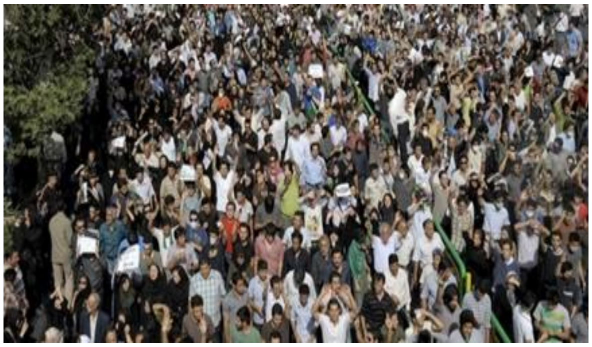


Fig.8 (a): Original Image

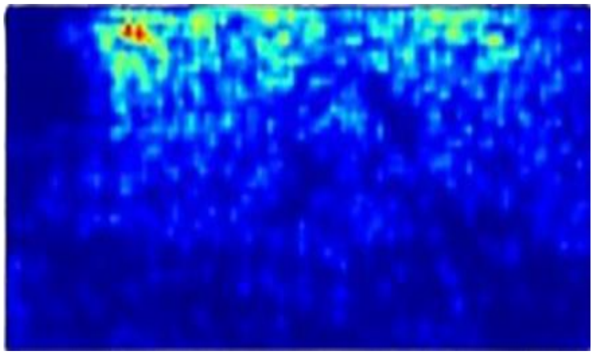
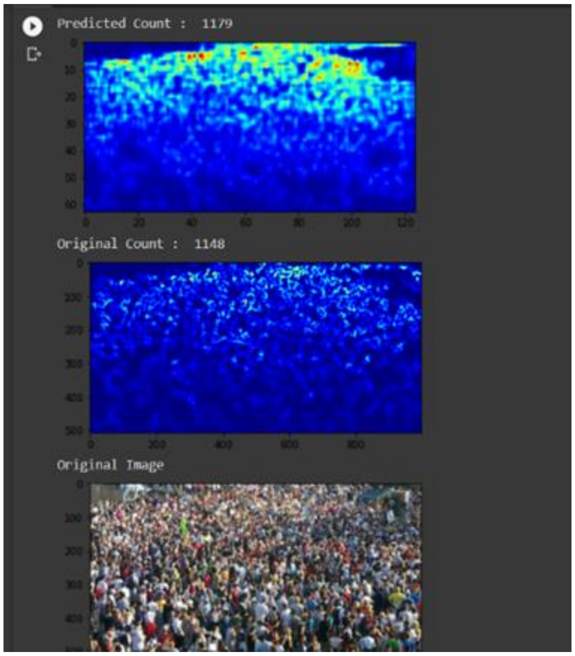


Fig.8(b) After Application of Gaussian Filters

After Gaussian filters are applied, the image is converted like this in Fig. 8 (b)

C. OUTPUT



Predicted Count : 1179

Original Count : 1148

Original Image

Fig. 9: Output Screenshot

The image above clearly depicts the predicted output along with the original count and the corresponding images.

The first image is the image trained by our CNN model along with VGG-16 model and has predicted the count of people as 1179. The second one is just a representation of Gaussian filters along with the original count which is 1148. The final one is the original image which was chosen from the dataset.

The predicted count is more than the original count which conveys that the model is able to succeed the original one and hence being more accurate and efficient.

Table II: Comparison of original count and predicted count from various images

Original Count	Predicted Count	Original Count	Predicted Count
734	766	169	207
890	934	273	346
431	429	1043	1263
1200	1347	539	566
270	318	401	473
377	429	175	200
264	308	353	405
205	216	1164	1220

Table III: Comparison of proposed work with literature work

Paper Author	Original Count	Literature Count	Predicted Count
Miaojing S. et.al., 2017	101	134	125
Miaojing S. et.al., 2017	71	92	96
Miaojing S. et.al., 2017	471	426	424
Liping Z. et.al., 2020	1226	1232	1232
Liping Z. et.al., 2020	127	141	148

IV. CONCLUSION

The accuracy of the model is around 90% in accordance with the above table. The techniques such as CSRNet and VGG-16 model has helped us gain a in-depth filtered analysis over quite number of images. An end-to-end training algorithm which helps us gain better performance and higher accuracy.

Further we aim to explore different techniques and usage of classification of data to reduce memory and increase efficiency at the same time.

REFERENCES

[1] Vishwanath A. Sindagi, Vishal M. Patel, "A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation", Pattern Recognition Letters, elsevier 2017.

[2] Lingke Zeng, Xiangmin Xu, Bolun Cai, Suo, Qiu, Tong Zhang, "Multi-Scale Convolutional Neural Networks for Crowd Counting", IEEE 2017.

[3] Archana Gotkar, "A Study on Crowd Detection and Density Analysis for Safety Control", IJCSE 2018.

[4] Sabrina Haque, Muhammad Sheikh Sadi, Md. Erfanul Haque Rafi,Md.Milon Islam and Md. Kamrul Hasan, "Real Time Crowd Detection to Prevent Stampede".

[5] Jugal Kishor Gupta, Sanjay Kumar Gupta, "IoT Based Statistical Approach for Human Crowd Density Estimation-Design and Analysis", 2020.

[6] UjwalaBhangale, Suchitra Patil, Vaibhav Vishwanath, ParthThakker, AmeyBansode, Devesh Navandhar, "Near Real-time Crowd Counting using Deep Learning Approach", 2019.

[7] Muhammad Waqar Aziz, Farhan Naeem, Muhammad Hamad Alizai, Khan Bahadar Khan, "Automated Solutions for Crowd Size Estimation", 2017.

[8] Min-hwanOh, Peder Olsen, Karthikeyan Natesan Ramamurthy, "Crowd Counting with Decomposed Uncertainty" ,2020.

[9] JunyuGao, WeiLin, BinZhao, DongWang, ChenyuGao, JunWen, "C3 Framework: An Open-source PyTorch Code for Crowd Counting", 2019.

[10] Guangshuai Gao, JunyuGao, Qingjie Liu, Qi Wang, Yunhong Wang, "CNN-based Density Estimation and Crowd Counting", IEEE, 2020.

[11] Mayur D. Chaudhari, Archana S. Ghotkar, "A Study on Crowd Detection and Density Analysis for Safety Control", 2018.

[12] Akbar Khan, Jawad Ali Shah, Kushsairy Kadir, Waleed Albattah and Faizullah Khan, "Crowd Monitoring and Localization Using Deep Convolutional Neural Network", 2020.

[13] MiaojingShi, Lu Zhang, Qiaobo Chen, "Crowd counting via scale-adaptive convolutional neural network", 2017.

[14] Liping Zhu, Hong Zhang, Sikandar Ali, Baoli Yang, and ChengyangLi, "Crowd counting via Multi-Scale Adversarial Convolutional Neural Networks", 2020.