International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 703-714 www.ijaem.net ISSN: 2395-5252

A System to Estimate Crowd and Detect Violence

Dr. Sharath Kumar Y H[1], Lakshmeesh S[2], Nisargashree S P[3], Sudhanva D J[4], Yashaswi P[5]

[1] *Professor and The Head of the department,* [2345] *Engineering Students,*
Department of Information Science and Engineering
Maharaja Institute of Technology Mysore, India

Date of Submission: 08-07-2020 Date of Acceptance: 23-07-2020

ABSTRACT

One of the major concerns throughout the world in all the places of large gatherings during an event is crowd control. The event can be of any form ranging from gatherings of few hundreds to millions. When large number of people assemble or move in an area which might not be suitable to handle the number it can create congestion which would lead to unexpected events such as stampede, riots or emergencies in an unfavourable situation. It would be extremely time consuming and hectic for manual monitoring. Henceforth an attempt is made to build an automated system that would count the number of people assembled and compare it with the thresholds provided by the security management and also would try to identify physical violence in the region of interest. This system would also help in maintaining law and order during protests, curfew and imposed social distancing. These would alert the security personnel to take necessary actions. The system is built to estimate crowd and detect violence based on the street camera's feed which are usually positioned at the strategic places of crowd gatherings. The model is built using Convolutional Neural Network (CNN). All the results which are provided by this system is based on image processing. A real time alert system in practicality has it's own challenges and limitations. Therefore, this paper focuses on the model which could help the system perform better.

INDEX TERMS - crowd control, crowd estimation, image processing, violence detection, street camera, social distancing, security management.

I. INTRODUCTION

With the increasing number of population mainly in the cities it has become quintessential to monitor the crowd for numerous reasons such as thefts, fights, stampede, harassment etc,. One of the major reason for all these unfortunate events to occur is large crowd gatherings in a small or unsuitable area. It would be very difficult for security department to monitor every surveillance camera in real time; also it would require a lot of manual work and efficient monitoring but if the department is under staffed it would cause problems. Stampede has become a major reason for deaths at religious events,protests and rallies; whereas in political campaigning and other such events riots are caused which should be detected at the earliest and actions must be taken. This paper is an attempt to build a system to cater than need. This would be of great help to analyse the crowd shape and size and allocate the required staff for the task.

The existing systems are not conducive to be used in every places as few of the systems are expensive and needs high end monitoring systems to adapt to their program. These surveillance systems must handle the computation which can be complex and heavy on the system. Many real time system are built to learn continuously with the new data but these are on the highest end of technology and infrastructure. The proposed system is a simple model which can adapt to most of the locally available systems and provide reasonable accuracy which is sufficient to achieve the desired goal. The trade - off between the computation power, time and accuracy has been one of the major focus in our paper.

II. RELATED WORK

The related works of others with regards to crowd estimation were of different varieties each varied based upon various scenarios, environment and parameters. The paper [1] keeps track of both the crowd density and their direction. The UCF crowd dataset was used in this paper. The direction was determined With the help of feature extraction methods like simple blob detector, SURF, MSER, SIFT. The density was set to a threshold of a factor of 0.6 above which it classified as densely populated area but does not give the exact count. The paper [2] partitions the frame to grids and then

applies gabor kernel to determine the density and maps into 3 categories such as empty region, low and high density. The papers [3] and [4] is based on detection and estimation of crowd density by the count of the heads in crowd with the input source as CCTV surveillance with respect to their region of interest, though the approach works well it has many real world limitations and challenges. Few papers proposes a different novel method built with the help of LSTM (Long short-term memory).The paper [5] uses of LSTM along with RNN (Recurring neural network). The paper [6] uses of LSTM along with CNN-RNN Crowd Counting Neural Network (CRCCNN). This method showed improvements than the previous other methods and not only produced density but also produced the count of people with better accuracy. This paper used the ShanghaiTech dataset which was promising for our study. The paper [7] used Multi - column convolution neural network (MCNN) to produce density maps. The extension of this method is in the paper [8] where Dilated Convolutional Neural Network was used. The paper [9] concentrated on the indoor crowd estimation whereas most of the papers focus on outdoor environment. Overall this paper attempts to obtain accurate estimation of crowd indoor and outdoor and in the methods of the papers [7][8] and focus is not only on classification based on density but also to provide the accurate count of the crowd. This paper further explores these methods in this paper.

The related works of others with regards to violence detection were of different varieties each varied based upon various scenarios, environment and parameters. The papers [10][11] focuses on the macro level anomaly detection in the crowd movement, it is achieved by considering the speed and direction of their movement and the region or space designated for their movement. This is a wide-scale approach whereas our focus is on static micro camera feed mostly for a small region of interest. The paper [12] proposes a holistic approach for real time violence detection by partitioning captured frames into grids and further mapping them as a scene of friction and congestion among people in the frame and by training the model to detect violence in such cases. The paper [13] proposes a novel method using LSTM and CNN to detect violence from the hockey fights dataset[15] and movie fight dataset[16]. This method performed better than the previous comparable methods in it's approach. This paper attempts a different method to build a model for a more generalized crowd violence dataset[17].

III. METHODOLOGY

1) Crowd estimation

There are many datasets that are open sourced online for the purpose of developing deep learning models for crowd count estimation[14]; But all of these data sets have the annotated bodies of people and the drawbacks of these are in detecting the relative distance between any two people from a live visual. Therefore, a better option is to find out the density. One of the dataset that has this option is the ShanghaiTech dataset and it also provides the true count in it's dataset which would help in detemining accuracy. The ShanghaiTech dataset is made up of two parts. Part A consists of 500 images collected from the internet and made up of all ranges of resolution. Part B is made up of 700 images collected directly from the streets of Shanghai and is of a standard resolution of 1280x720 pixels. Both these parts have unique images. An image from both these parts and their heatmap can be seen in Fig. 1. Each part has 3 folders, images, ground truth and groundtruth-h5. Images folder has the jpeg files, the ground truth folder has matlab files contain annotated head (coordinate x, y) for that image. And the groundtruth-h5 folder is having the density map of that image. The density map of the image is calculated by using both matlab file and the jpeg file by using gaussian filter on the places where heads are annotated, pixels where the head is present add up to one and the other pixels are all equated to zero.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 703-714 www.ijaem.net ISSN: 2395-5252

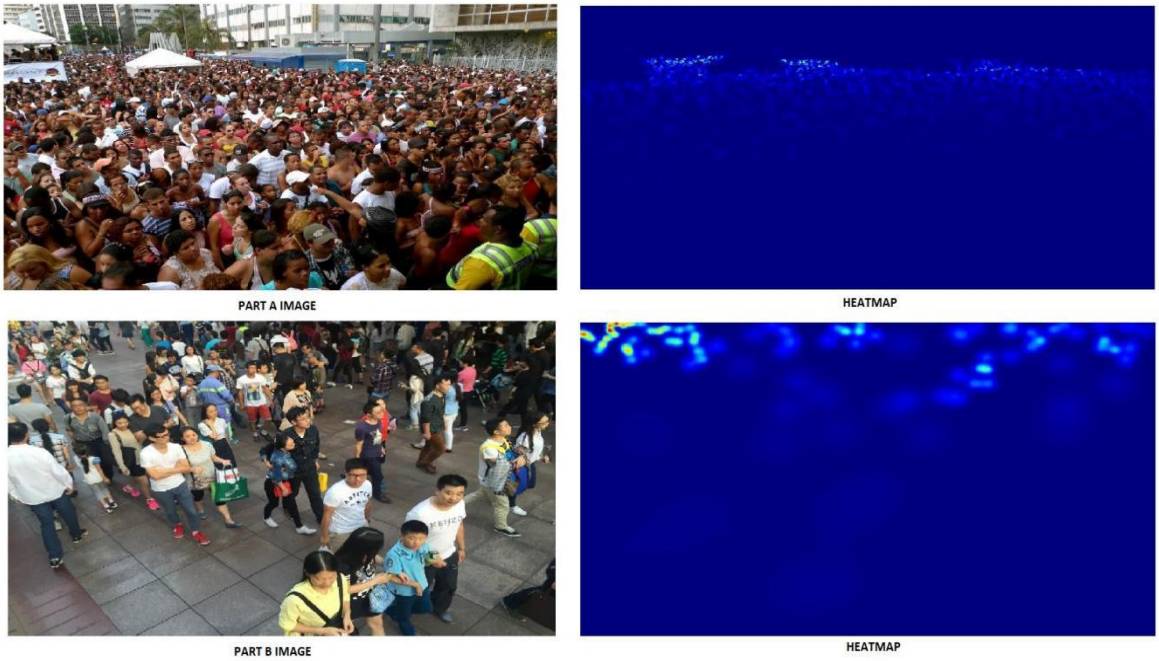

Fig. 1 : A datset sample of shanghaiTech datset.

The crowd estimation model follows[8] with few changes in it's model. The original model used VGG - 16 for first 10 layers of it's 16 layered model. This paper attempts in building 3 different related models and experiment. The first model for crowd estimation is built up of six convolution layers and two max pooling layers. All convolution layers have a kernel size of 3x3 and for maxpooling layer 2x2 filter size is used. The first two layers have 64 kernels followed by another two convolution layers having 128 kernels, after every 2 of such layers a maxpooling layer is present. The model consists of two dilated convolution layers which have 128 and 64 kernels respectively with a dilated rate of two. The output image is reduced to 1/4th the size of the input image using two maxpooling layers. dilated convolution layers is used to preserve the further loss of resolution and spatial information of the image. The Rectified Linear Unit (ReLU) activation function used is used. The model can be visualised as in Fig. 2. After the 6 layered model, 8 is attempted by adding a 256 kernel layer and a dilated convolution layer and 12 layered model by adding 256,256,512 kernel layers and 2 dilated convolution layers.

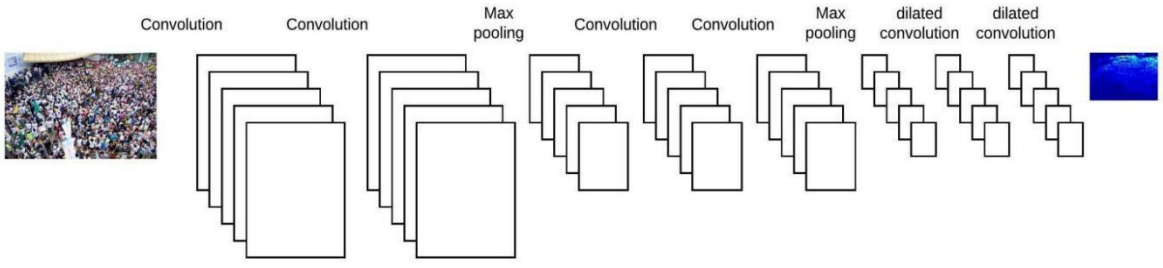

Fig. 2 : 6 layered Crowd estimation Model.

If the input is an image no conversion is needed. It is directly given as input to the model. If the input is a video it is converted into frames with the help of OpenCV.The VideoCapture() function is responsible for converting the video into image sequences. Every second of the video is made up of 30 image frames. It may vary with the type and quality of camera used to capture the video. The input image or frame is first taken as an RGB image and converted into a tensor which is a form of

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 703-714 www.ijaem.net ISSN: 2395-5252

n-dimensional array. The tensor is normalized and then fed into the CNN model. The output is also in the form of a tensor which is converted into a numpy array. Density map is obtained by converting the numpy array into an image file. This is achieved by using the matplotlib library. Crowd count is obtained by the summation of all the values in the numpy array. A density map is generated for all the images in the dataset by using gaussian filter on the places where heads are annotated. The pixels in which the head is present add up to one and the other pixels are all equated to zero. The input images are normalized before being fed into the model with a mean and standard deviation values of [0.485, 0.456,0.406] and [0.229, 0.224, 0.225] respectively. The model is trained using Stochastic Gradient Descent(SGD) with momentum. A single data point is taken from the dataset for which the weights of the model are updated because SGD is used and the batch size is 1. Mean square loss is taken as the loss function with the learning rate fixed at 1e-5 and momentum at 0.95 which is a standard value. The Mean Absolute Error (MAE) and the computation time taken for each model is discussed in the results and analysis section of this paper.

2) Violence detection

A model is built to detect violence in a video by using image processing technique with the help of Keras from Tensorflow library. The code for the model is of python language. The model is trained from a video dataset called "real-life-violence-situations-dataset"[17] obtained from the kaggle datasets library which consists of 2000 videos in total; where it is classified into 2 classes namely violence and non violence with 1000 videos in each class. This datset consists of all the features of both the hockey fights dataset[15] and movie fight dataset[16] and much more varieties. The videos are ranging from the short duration of few seconds to few minutes and majority of the videos are between 25-30 frames per second. The model is trained using images or frames obtained from the video at the rate of 10 frames per second which had better data when compared to our trial using 2,3,5,7,12,15 frames per second. The train-test split was in the 80-20 ratio which is one of the the standard general split. Totally for training 23000+ images are used and for testing 5700+ images are used with the goal of binary classification. Few sample images can be seen in the Fig. 3.

Fig. 3 : Few samples of Violence detection dataset.

Each frame undergoes data augmentation and resizing before subjected to Convolutional neural network(CNN) model. Frames are rescaled at 1./255 with shear range of 0.1 and zoom range of 0.1 along with horizontal flip. The Hyper parameters of image width x image height are taken as 200x200 and the batch size as 32. The batch sizes of 8, 16, 32, 64, 128 and 256 was tried and batch size of 32 was efficient.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 703-714 www.ijaem.net ISSN: 2395-5252

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 198, 198, 32)      896
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 196, 196, 32)      9248
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 98, 98, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 96, 96, 64)        18496
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 94, 94, 64)        36928
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 47, 47, 64)        0
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 45, 45, 128)       73856
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 43, 43, 128)       147584
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 21, 21, 128)       0
_________________________________________________________________
conv2d_6 (Conv2D)            (None, 19, 19, 256)       295168
_________________________________________________________________
conv2d_7 (Conv2D)            (None, 17, 17, 256)       590080
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 8, 8, 256)         0
_________________________________________________________________
flatten (Flatten)            (None, 16384)             0
_________________________________________________________________
dense (Dense)                (None, 256)               4194560
_________________________________________________________________
dropout (Dropout)            (None, 256)               0
_________________________________________________________________
dense_1 (Dense)              (None, 256)               65792
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 257
_________________________________________________________________
activation (Activation)      (None, 1)                 0
=================================================================
Total params: 5,432,865
Trainable params: 5,432,865
Non-trainable params: 0
```

Fig. 4 : The Violence detection model summary.

The model consists of multiple CNN layers, it can be seen in the Fig. 4. All the convolutional layers has it's kernel size as 3x3 and 2x2 filter size for MaxPooling layer. First 2 convolutional layers consists of 32 kernels, followed by a MaxPooling layer. Input shape is fed to The first convolutional layer. Next 2 convolutional layers consists of 64 kernels, followed by a MaxPooling layer. Next 2 convolutional layers consists of 128 kernels, followed by a MaxPooling layer. Next 2 convolutional layers consists of 256 kernels, followed by a MaxPooling layer. All these layers have an Rectified Linear Unit (ReLU) function which is an activation function. Input is flattened and fed into 2 fully connected dense layers. Each fully connected dense layer consists of 256 neurons followed by a dropout layer with value of 0.5. The output layer has a single neuron with sigmoid activation function where if the output is greater 0.5 it's classified to the non violence class if not then violence class. The model is trained using Stochastic Gradient Descent (SGD) optimizer with momentum. RMSProp and Adam optimizer with different learning rates (lr) was tried but SGD optimizer with the learning rate of 0.001 and momentum of 0.9 performed better.

The best model is obtained using early stopping criterion. Early stopping function is used to stop the epochs if there is no further minimization in validation loss with a patience of 10 epochs. The best model is saved with the help of model checkpoint function where it saves the model with the highest validation accuracy as a h5 file. The number of epochs was set to 100 with a patience of 10 epochs. The validation steps and steps per epoch is dependent on the batch size.

IV. IMPLEMENTATION

This system which estimates crowd and detects violence has a Graphical User Interface (GUI) built using web technologies such as HTML and Bootstrap library for the front end and Flask framework which is a micro framework of web written in python for the back end, The data is

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 703-714 www.ijaem.net ISSN: 2395-5252

stored using file structures. The GUI can be seen in the Fig. 5, the web page is divided into 2 major blocks one for crowd estimation and another for violence detection. Both of these accept images and videos of given formats and produces the output. If in case of any error it shall display it to the user. This is a prototype but can be extended to implement for the real time camera feed.

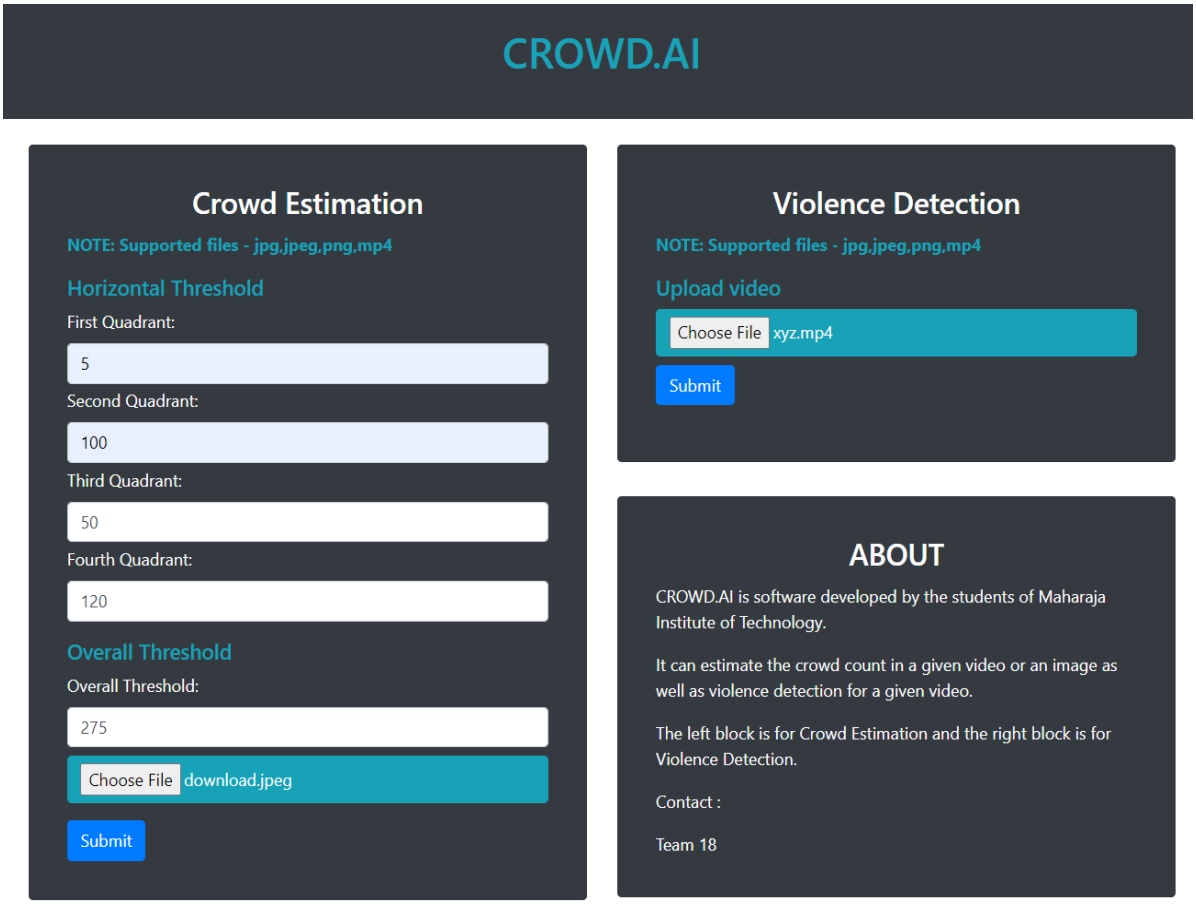

Fig. 5 : The Graphical User Interface (GUI)

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 703-714 www.ijaem.net ISSN: 2395-5252

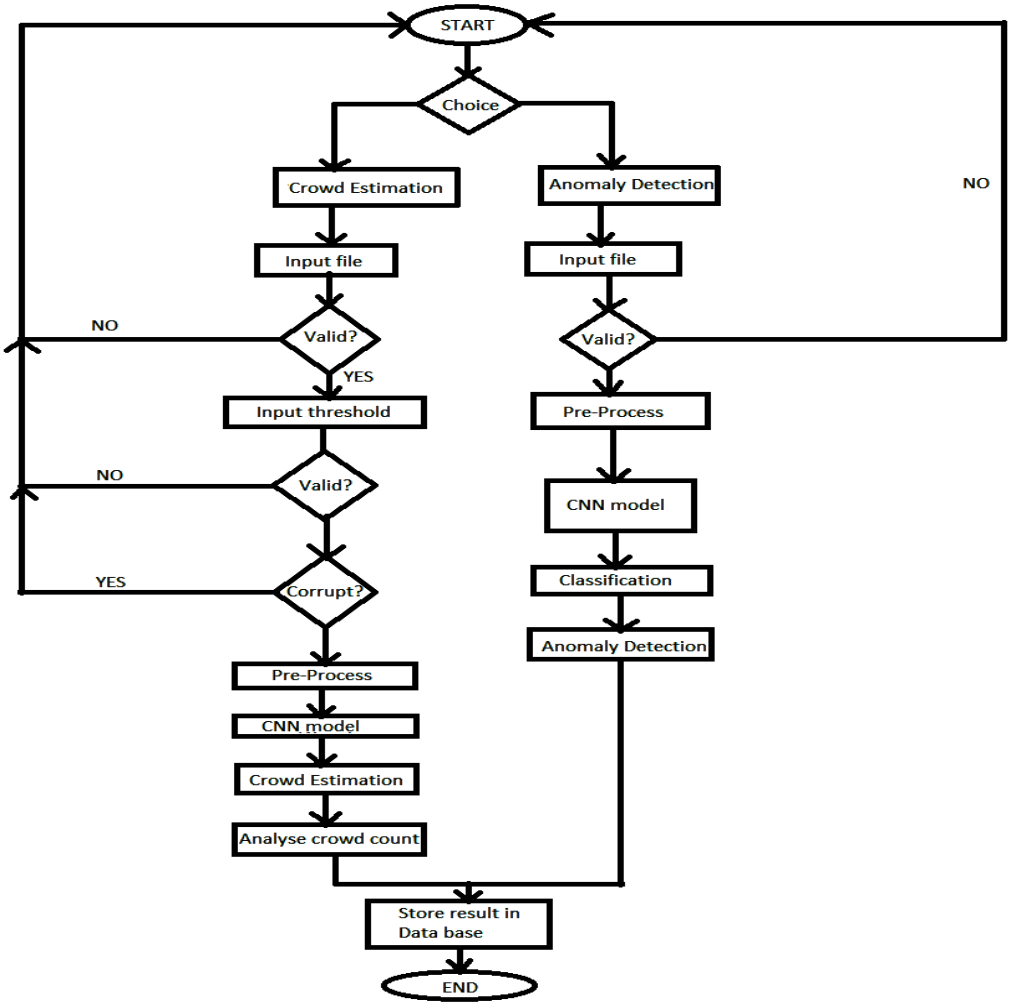

Fig. 6 : The Process flow chart

The crowd estimation system in order to provide the density of crowd accepts threshold or limit number from the user, which determines the sparse or dense crowd gathering to produce an appropriate output. Along with the overall threshold number, horizontal sectional threshold number is also taken for 4 divisions which is an idea based on camera sight, where farther the distance from the camera more the number of people can gather within safety limits but not near the camera because the area covered by the camera coverage narrows down limiting less number of people. The fig. 6, explains the flow of events of the system where the user inputs a file or data for either crowd estimation or violence detection, The system performs file format validation. The process continues only if the format is suitable; if not, an error is displayed and user is redirected to the home page. In case of crowd estimation the validation of thresholds is also done as the fields must not be left empty or given a alphanumeric value or a negative integer or float value. The system also detects corrupt file and if the file is corrupt it redirects it to home page by revoking output request. Once the validation checks are cleared the input videos are converted into frames and pre processed and fed into the model. If the input is a image it directly be subjected to pre process stage. The crowd estimation output consists of original image, its horizontal quadrant images and their heatmap in a table followed by the count, with appropriate font color. The results can be downloaded in the portable device format (PDF) file. For the violence detection system the output is of the same file format as it's input. This is further discussed in results and analysis section of this paper.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 703-714 www.ijaem.net ISSN: 2395-5252

V. RESULTS AND ANALYSYS

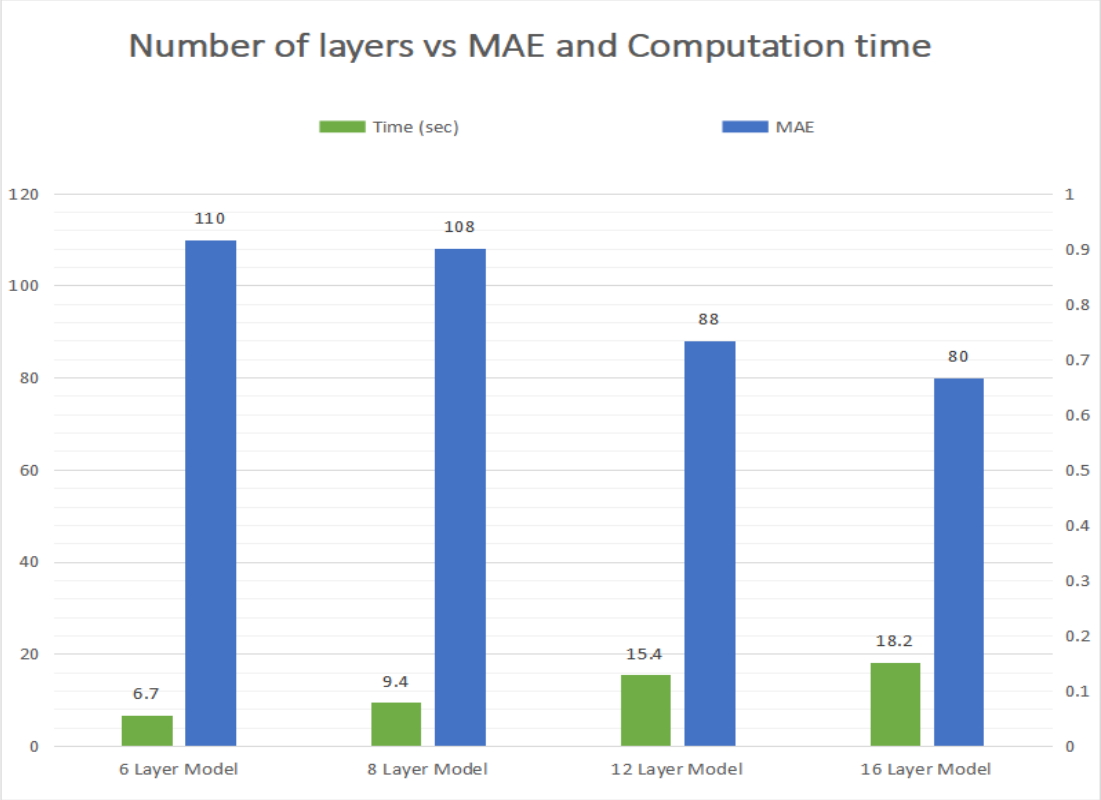

Fig. 7 : Comparison between number of layers, Mean Absolute Error (MAE) and computation time in the crowd estimation model.

In the Fig. 7, we have compared the trade - off between 3 parameters namely computation time (CPU), number of layers in the model and Mean absolute error (MAE). From this experimental result it can be notice that if the infrastructure is basic and cannot wait for the result for a long time, then a lesser a layered custom model can be deployed with the increase in MAE. Theses are the systems which do not require highly accurate count. If the infrastructure is advanced with more graphics processing unit (GPU) or Tensor Processing Unit (TPU) cores one can use the VGG based model for higher accuracy. Henceforth this study provides a view over the trade - offs for the system.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 703-714 www.ijaem.net ISSN: 2395-5252

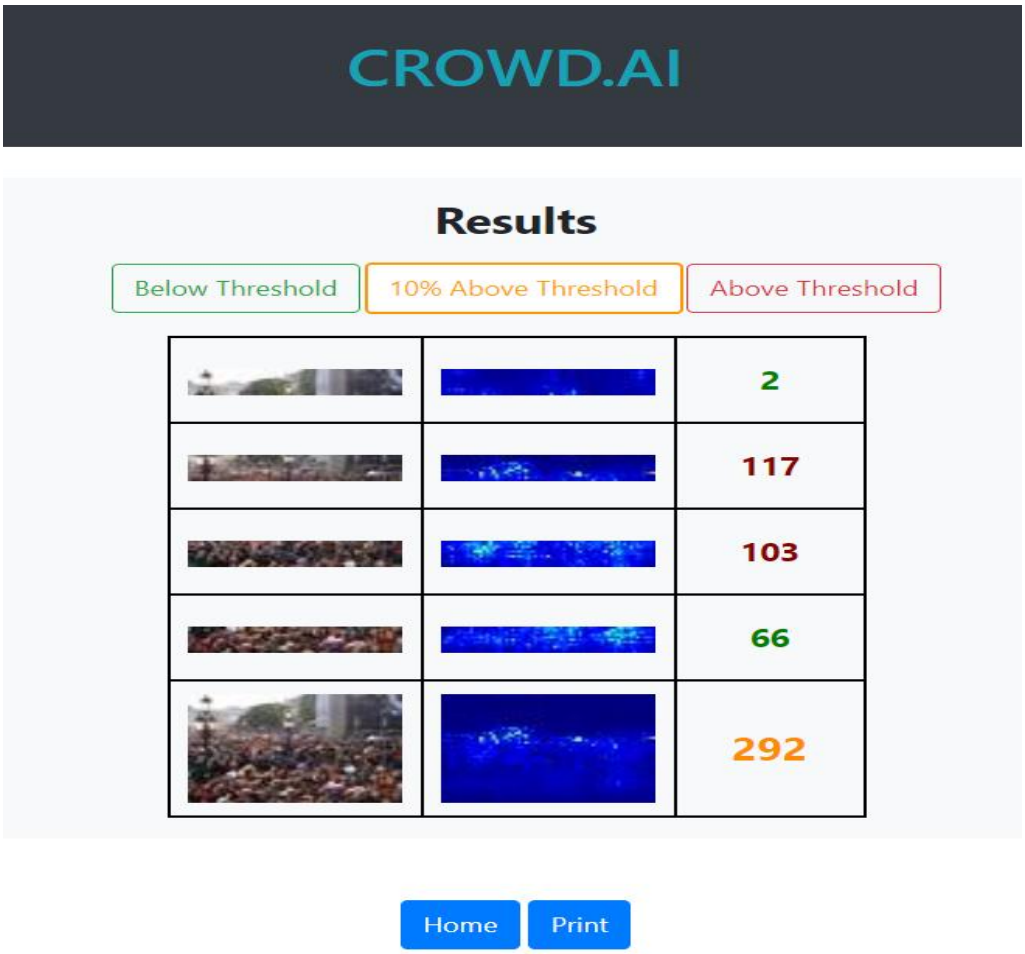

Fig. 8 : An output of the crowd estimation model.

The output produced by the crowd estimation model is shown in the Fig. 8, it can be observed that there is a 5x3 table. In the table first column consists of the original input given by the user split into horizontal quadrants and also the original image in the last row. The second column consists of heat map generated with respect to the image. The last column provides us the actual count which can be of 3 colors given by the color coding. The green color represents safe level, orange represents to be alert and red means danger. The coloring is based on the thresholds provided by the user if the count is below the threshold number then green, if above by 10% then orange and beyond that is red; these can be seen in the Fig. 9. This system provides both the count of the crowd and also the alert for the user. Hence can be used for strategic and statistical purposes and for the crowd control by security.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 703-714 www.ijaem.net ISSN: 2395-5252

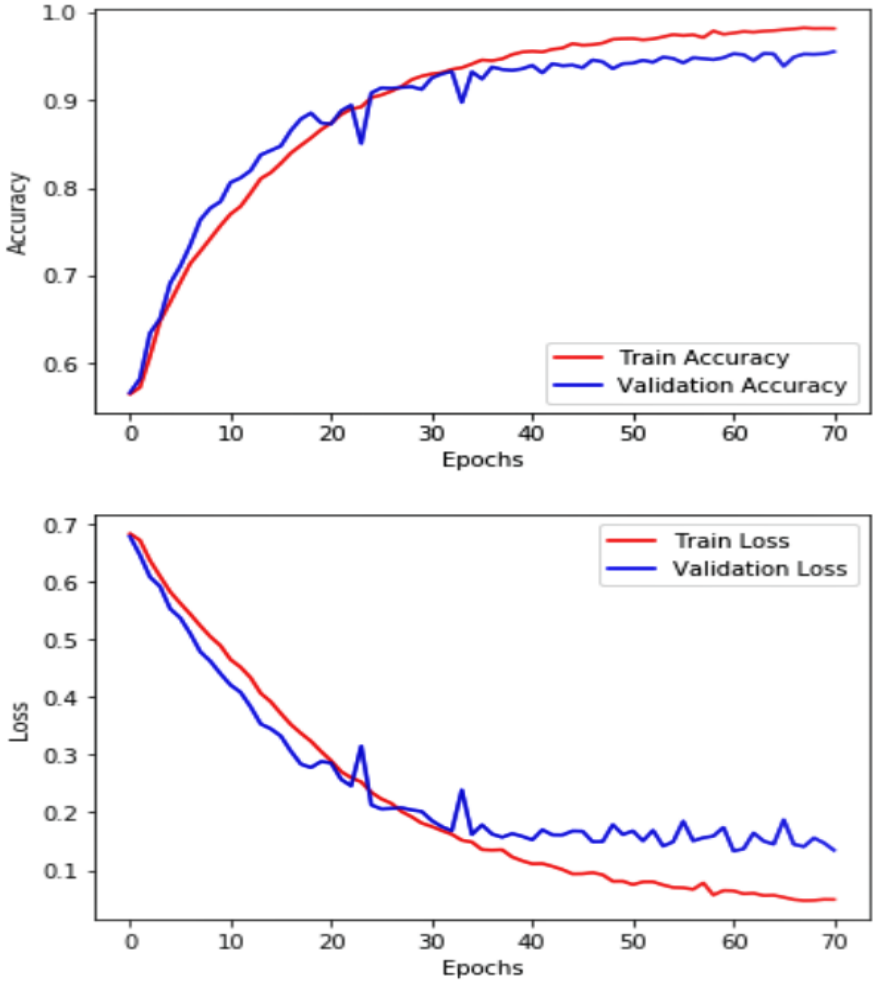

Fig. 9 : The accuracy vs validation accuracy(Top) and the loss vs validation loss(Bottom) graph for violence detection model.

The violence detection model had a early stopping criterion and a model checkpoint function was used to save the best possible model weights before the divergence of the curve. The model achieved training accuracy of 98% and 95.55% validation accuracy with training loss of 04% and validation loss of 11%. The model performs very well given that it is a binary classification model. The model classified different videos from different sources with an average accuracy of 95% and it also classified graphical videos and low resolution images with the same accuracy. The testing was done with different set of frames ranging from cartoons, video games with lower graphics to higher graphics, pictures from different textbooks and novels, street fights, protests, riots, fights between 2 groups of sport team supporters and political campaign fights, scenes from the movies etc,. and can also differentiate between close movement of people, hug against physical violence; few of them can be seen in Fig. 10. The location of these videos were also of different varieties indoor, streets, parks, fields, stadiums,etc,. The violence detected in videos or images would result in message stating violence detected in the top left of the frame as shown in the Fig. 11.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 2, pp: 703-714 www.ijaem.net ISSN: 2395-5252

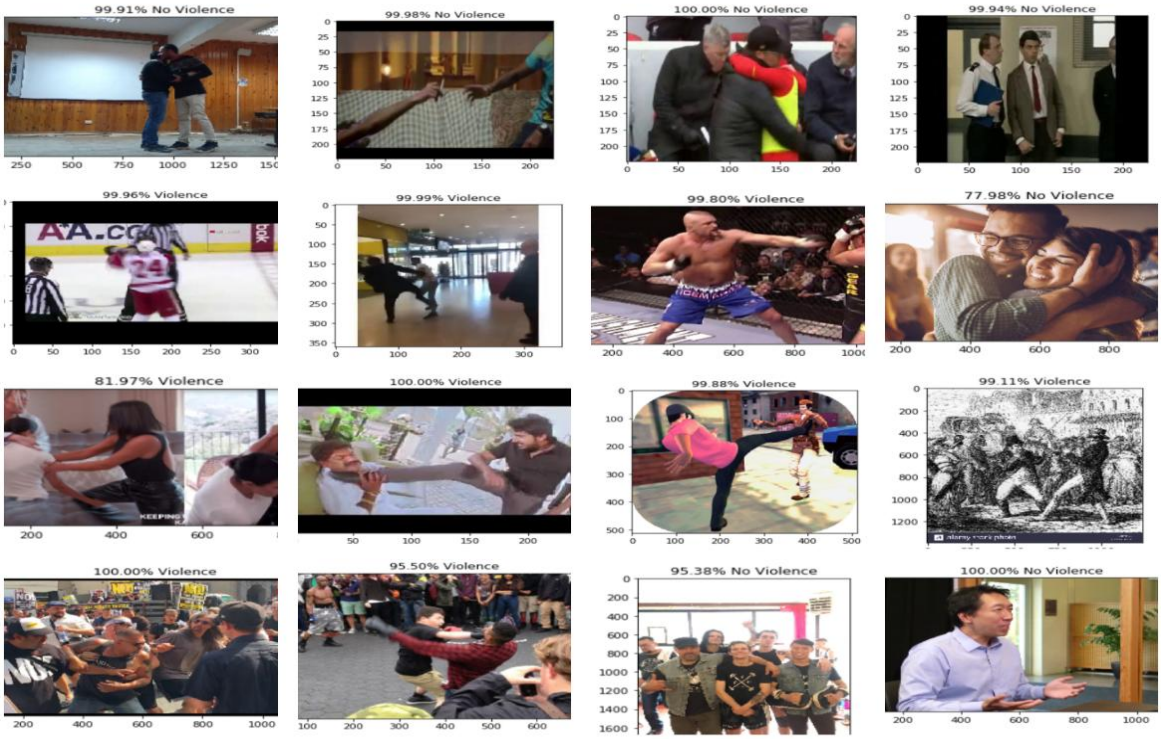

Fig. 10 : Few results of the violence detection model.

Fig. 11 : An output frame of the violence detection model.

VI. FUTURE ENHANCEMENT

For the crowd estimation in future, attempts shall be made to minimize the MAE and provide the count with much higher accuracy using different image processing techniques and advancements in the neural networks. MCNN and LSTM are other promising methods to estimate crowd and shall be further studied. For the violence detection various other image processing models can be used and experimented along with LSTM or few other methods. The proposed model can also be subjected for different hyper parameters and a locally built dataset for each surveillance system would helpful in building a system with higher classification accuracy. A self learning system can be built to learn from the new varieties of actions captured by the surveillance camera. Different

varieties of anomalies can also be included for detection in future.

VII. CONCLUSION

A crowd counting model and a violence detection model is built using Convolutional Neural Network. The violence is detected at an average accuracy of 95%. The crowd estimation model is built with Dilated Convolutional Neural Network using different custom layers and the MAE and the time taken for each is mentioned in the results section of this paper. This system is a prototype which can be further improvised and deployed as a real time violence detection and alert system along with crowd density estimation. The GUI of this system makes it easier for the user to obtain the desired result. Overall this paper is a study on usage of a system built upon image processing technique in public areas to monitor and control the crowd.

ACKNOWLEDGEMENT

We are indeed grateful to our Mentor Dr. Sharath Kumar Y H, Professor and The Head of the department of Information Science And Engineering, Maharaja Institute of Technology Mysore, India, for guiding us throughout this project and giving us the opportunity to present this paper "A System To Estimate Crowd And Detect Violence".

REFERENCES

[1]. Direction Estimation of Crowd Flow in Surveillance Videos, Muhammed Anees V, G. Santhosh Kumar, 2017 IEEE.
[2]. "Developing an Intelligent System for Crowd Density Estimation", International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 4, April 2016, Dr. Salem Saleh Ahmed Alamri, Dr. Ali Salem Ali Bin-Sama, .
[3]. "Al-Masjid An-Nabawi Crowd Advisor Crowd Level Estimation Using Head Detection",Raghad jaza alamri, 978-1-5386-4427-0/18/$31.00 ©2018 IEEE.
[4]. "Crowd Reckoning towards Preventing the Repeat of '2015 Hajj Pilgrims Stampede", 978-1-5386-2303-9/17/$31.00 ©2017 IEEE, A. Musa, M. M. Rahman, M. S. Sadi, M. S. Rahman.
[5]. "Deep Learning Framework For Density Estimation of Crowd Videos", Muhammed Anees V, 2018 Eighth International Symposium on Embedded Computing and System Design.
[6]. "A CNN-RNN Neural Network Join Long Short-Term Memory For Crowd Counting and Density Estimation", Jingnan Fu, Hongbo Yang, Ping Liu, Yuzhen Hu, Y. Wang et al. (Eds.): IGTA 2017, CCIS 757, pp. 85–95, 2018.
[7]. Single-Image Crowd Counting via Multi - Column Convolutional Neural Network Yingying Zhang Desen Zhou Siqin Chen Shenghua Gao Yi Ma, Shanghaitech University, 2016.
[8]. CSRNet : Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Yuhong Li ,Xiaofan Zhang, Deming Chen.
[9]. "Video analytics for indoor crowd estimation",Ryan tan, Indriyati atmosukarto, Wee han lim, 978-1-5386-4987-0/18/$31.00 ©2018 IEEE.
[10]. "Spatio - Temporal Anomaly Detection in Crowd Movement Using SIFT", Nitish Ojha, 978-1-5386-0807-4/18/$31.00 ©2018 IEEE.
[11]. "Abnormal Crowd Behavior Detection Using Speed and Direction Models", CHIBLOUN,Sanaa EL FKIHI, Haza MLIKI,978-1-5386-8173-2/18/$31.00 ©2018 IEEE.
[12]. "Holistic Features For Real-Time Crowd Behaviour Anomaly Detection", Kevin McGuinness, Suzanne Little, Mark Marsden, Noel E. O'Connor, IEEE 2016.
[13]. "Robust Real - Time Violence Detection in Video Using CNN And LSTM", Al-Maamoon R. Abdali, Rana F. Al-Tuma, 2019, 2nd Scientific Conference of Computer Sciences (SCCS), University of Technology - Iraq, 978 - 1 - 7281 - 0761 - 5/19/$31.00 ©2019 IEEE.
[14]. http://visal.cs.cityu.edu.hk/downloads/
[15]. http://academictorrents.com/details/38d9ed996a5a75a039b84cf8a137be794e7cee89
[16]. http://academictorrents.com/details/70e0794e2292fc051a13f05ea6f5b6c16f3d3635
[17]. https://www.kaggle.com/mohamedmustafa/real-life-violence-situations-dataset

International Journal of Advances in Engineering and Management
ISSN: 2395-5252

IJAEM

Volume: 02 Issue: 01 DOI: 10.35629/5252

www.ijaem.net
Email id: ijaem.paper@gmail.com