International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

Crowd Density Estimation and Anomaly Detection

Kaustubh Bhat[1], Shwetha BC[1], Sujan RS[1], Varun BV[1], Dr. Hema Jagadish[2]

[1]*Student, Dept. of Information Science and Engineering, Bangalore Institute of Technology, Bengaluru, Karnataka, India*
[2]*Assistant Professor, Dept. of Information Science and Engineering, Bangalore Institute of Technology, Bengaluru, Karnataka, India*

Abstract - *Crowd density estimation and anomaly detection have a variety of applications such as ensuring public safety, avoiding congestion, detection of crimes, observing any sort of crowd abnormalities and for urban planning. The main aim of crowd density analysis is to calculate the crowd density of a given area and thereby calculating the number of people present. Pattern recognition technique is used for face detection. The task of face detection is inherently complex and challenging due to the many differences present in a human face such as color, expression, position etc. One of the methods available for face detection is the detection based approach. However, it performs poorly in highly congested scenes since most of the targeted objects are obscured. In order to overcome this problem, density based approach is used. The images captured undergo pre-processing before being fed to the Convolutional Neural Network (CNN) model. The CNN model processes the image and produces a density map using which the head count is calculated. The count is compared with the pre-determined threshold value for a given area. If the count exceeds the threshold value, the authorities are notified. The data obtained is stored in a database for future references. The data stored can be used to organize an event better in the future.*

Key Words: **Crowd Counting, Crowd Density Estimation, Anomaly Detection, Image Processing, Violence Detection, Convolutional Neural Network**.

1. INTRODUCTION

Any area which consists of a large number of people grouped together in a confined space is considered to be 'crowded'. Crowded areas are usually associated with frequent and heavy occlusions. There are a number of technologies available for detection and tracking but are limited by the fact that they can be used only for sparse scenes and isn't effective in crowded scenes. This has led to a new field of research which concentrates more on crowded scenes. It covers a broad range of topics, including crowd segmentation and detection, crowd tracking, crowd counting, pedestrian travelling time estimation, crowd attribute recognition, crowd behaviour analysis, and crowd abnormality detection. Many existing works on crowd analysis are scene-specific, i.e., models trained from a particular scene can only be applied to the same scene. When the scene in which the model is to be used changes, it is required to train the model again which is not convenient. The ideal situation would be to train a generic crowd model once and be able to use in any given scene. This is referred to as scene-independent crowd analysis. Achieving this is a big challenge considering the inherently complex nature of crowd which varies vastly from scene to scene. Comparing and characterizing all the different dynamics of crowds is the main challenge.

1.1 Problem Statement

Over the years it has become increasingly difficult to manage a large group of people gathered at one place. There are a number of riots taking place almost on a daily basis. And in most cases, the responsible authorities will not receive the information in time and are often helpless. In order to overcome the problem, a solution has been designed using image processing to find out the crowd density of a given area and detect any anomaly in that same area. If the crowd density increases beyond a given threshold or an anomaly is detected, the concerned authorities are immediately notified thereby giving them time to take necessary precautions and preventing any mishaps from happening.

1.2 Related Work

The Shanghai Tech dataset which is used is developed specifically for crowd detection by Li Biaping, et al. [4]. At present, it is the best dataset available which can be used for crowd detection using density based approach. The 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 consists of 700 images collected directly from the streets of Shanghai and is of a standard resolution of 1280x720 pixels.

2. PROPOSED DESIGN

The model follows a density-based approach as it not only gives us the crowd count but also gives us the spatial information of the given image. The proposed design consists of two different sections. They are crowd density estimation and anomaly detection. The system architecture and the training process of the two is explained in this section.

2.1 Crowd Density Estimation

The model used for crowd estimation is a pure convolutional neural network as it does not contain any fully connected layer. It is made up of six convolution layers and two max pooling layers. All convolution layers have a kernel size of 3x3 but the number of kernels in each layer vary. The first two layers have 64 kernels followed by a maxpooling layer of 2x2 filter size. The other two convolution layers have 128 kernels again followed by a maxpooling layer of the same filter size. The model consists of two dilated convolution layers which have 128 and 64 kernels respectively and a dilation rate of two. By using two maxpooling layers the output image is reduced to 1/4th the size of the input image.

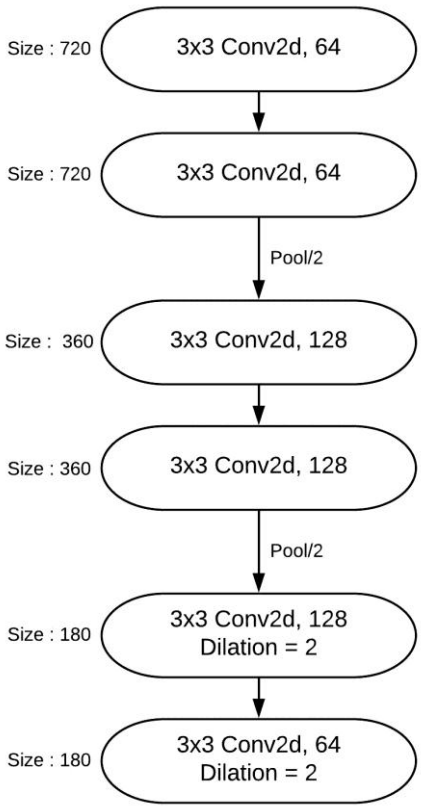

Fig -1: System Architecture of the crowd estimation model.

In order to preserve the further loss of resolution and spatial information of the image, dilation convolution layers are used. The activation function used is ReLU function as it is the most commonly used activation function in deep learning models.

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.

2.1.1 Implementation

The input is an image no conversion is needed. It is directly given as input to the model. 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. Only three frames are extracted from equal interval of the input video. The input image or frame is first taken as an RGB image and converted into a tensor (A tensor is a form n-dimensional array) for easier processing and then the tensor is normalized. The tensor is 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

2.2 Anomaly Detection

The system is able to detect two different types of anomalies. The first type of anomaly is based on the crowd count. If the crowd count is higher than the threshold value specified, it is considered as an anomaly. The other type of anomaly detected is based on the occurrence of any violent activities such as physical fights in a given image.

The model used to detect violence is a customized VGG16 model. The first 12 layers of the model are taken as it is. But the last 4 layers are modified. The last maxpooling layer is changed to an average pooling layer. The next two layers are dense layers with 64 nodes or neurons with an activation function as ReLu. But the last layer or the output layer has only two neurons with the softmax activation function. The first 12 layers in the model are pre-trained. The images are re-sized to 224x224 before being fed into the model.

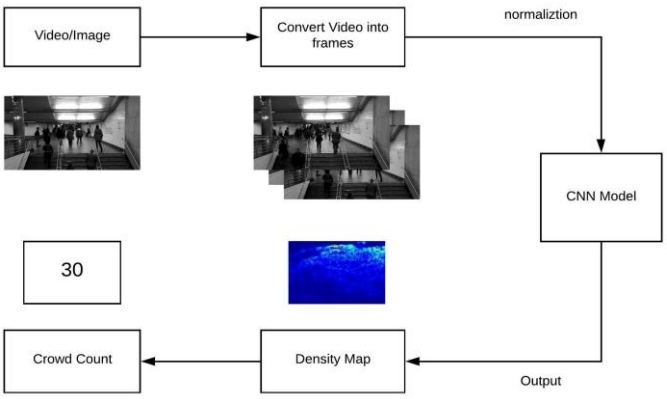

Fig -2: General Framework of the crowd estimation model.

The dataset used to train the model is the real time violence dataset which consists of over 1000 videos. This includes both classes of videos namely violent and non-violent. The videos are then converted into images before being fed into the model for training. The model is trained using ADAM optimization and the batch size is taken as 16. Categorical cross entropy is taken as the loss function with the learning rate fixed at 0.01.

2.2.1. Implementation

Video is given as input, which is split into frames and pre-processing is done on the frames. In pre-processing, the frame/image size is reduced to 224 * 224 * 3 (RGB image) and fed into the VGG16 model. The VGG16 model classifies and detects anomaly. It then gives the output frame with text printed on the top right corner if anomaly is present. The frames which have been given as the output are combined together to form a video of 5 frames per second and a .mp4 extension.

3. RESULTS AND ANALYSIS

The performance of the crowd estimation model is measured in terms of the Mean Absolute Error (MAE). MAE is preferred over accuracy because crowd estimation is an approximate estimate calculated and not there is no certainty of the answer being 100% accurate.The MAE calculated for part A and part B of the Shanghai tech dataset is 103.1 and 32.9 respectively.

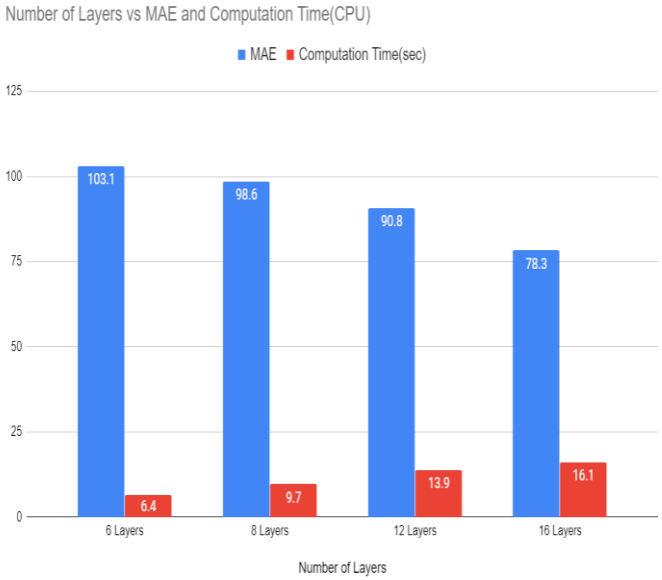

Chart -1: MAE and Computation Time for Different Number of Layers

As observed in the graph, it is clear that MAE of the model decreases as we increase the number of layers. However, the computation time increases as we increase the number of layers. The 6-layer model performs better than the other models as it has a significantly less computation time compared to the other models. The significantly less computation time makes up for the slightly larger MAE compared to other models. It should also be noted that lesser the number of layers of a model, easier it is to run on a system which has less computational power. When GPU is used, the difference in the computation time between all the models reduced to a few seconds. But if a powerful GPU is used then the difference in computation time would be in the matter of milliseconds.

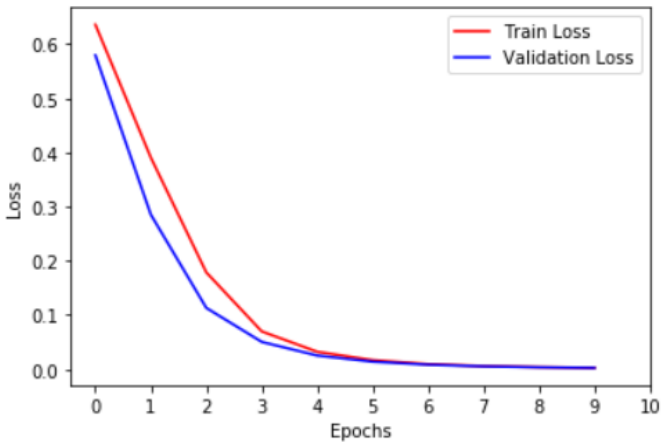

Chart -1: Loss for Training and validation phase for Anomaly Detection Model

The graph below represents the loss for training and validation phase of the anomaly detection model. It shows

how loss gradually decreases as the number of epochs increases. The x-axis represents the number of epochs and y-axis represents loss during training and testing phase. The accuracy of the anomaly detection model is 92.7%.

4. CONCLUSIONS

The project proposes a CNN model for estimating the crowd count and performs anomaly detection in a given area. The model takes the input in the form of an image or a video. If the input is a video, then three frames are captured in equal intervals and the crowd count is calculated for those 3 frames. After training the model for several values of epoch, the accuracy was found to be greatest for an epoch value of 300. With a MAE of 103.1 and an execution time of 6.4 seconds on CPU, it is performing better than many of the existing models when run on CPU. The challenges faced include identifying or recognizing people who are very close to the camera. The accuracy of the system depends on the angle at which the image is captured. The accuracy is maximum when the image is captured directly from above but it is difficult to install the cameras at such an angle. Training the model takes a long time and hence care should be taken to not interrupt the training process.

ACKNOWLEDGEMENT

The team are most humbled to acknowledge the enthusiastic influence that was provided by the guide Dr. Hema Jagadish for the ideas, insights that were provided, frequent advises and expertise that was given and the cooperation that was exhibited during the implementation of the proposed solution resulting in its success.

REFERENCES

[1] Raghad Jaza Alamri, Maha Suliman Alkhuriji, Malak Saed Alshamani, Omniyyah Yahya Ibrahim and Fazilah Haron, “Al-Masjid An-Nabawi Crowd Adviser Crowd Level Estimation Using Head Detection”, 2018

[2] Shiliang Pu, Tao Song, Yuan Zhang, Di Xie, “Estimation of crowd density in surveillance scenes based on deep convolutional neural network”, 2017

[3] Ryan Tan, Indriati Atmosukarto and Wee Han Lim, “Video Analytics for Indoor Crowd Estimation”, 2018

[4] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, Yi Ma, “Single-Image Crowd Counting via Multi Column Convolutional Neural Network”, 2016.

[5] Li Biaping, Han Xinyi, Wu Dongmei, “Real-time crowd density estimation based on convolutional neural networks”, 2018