2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) | 978-1-6654-2867-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICESC51422.2021.9532683

Crowd Detection and Analysis for Surveillance Videos using Deep Learning

Aman Ahmed
Student, Department of Computer Science and Engineering,
G H Raisoni College of Engineering,
Nagpur, India
ahmed_aman.cs@ghrce.raisoni.net

Prateek Bansal
Student, Department of Electronics and Telecommunication Engineering,
G H Raisoni College of Engineering, Nagpur, India.
bansal_prateek.et@ghrce.raisoni.net

Atiya Khan
Faculty, Department of Computer Science and Engineering,
G H Raisoni College of Engineering, Nagpur, India
atiya.khan@raisoni.net

Neha Purohit
Faculty, Department of Computer Science and Engineering,
G H Raisoni College of Engineering, Nagpur, India.
neha.purohit@raisoni.net

Abstract*—Crowd identification and analysis has drawn a lot of attention recently, owing to a wide variety of video surveillance applications. We present a detailed review of crowd analysis and management, focusing on state-of-the-art methods for both controlled and unconstrained conditions. The paper illustrates both the advantages as well as disadvantages of state-of-the-art methods. Mass or crowd gathering can be seen at a lot of places like airports, sports stadiums, at various religious, educational, and entertainment-related events, etc. When tens of thousands of people gather in limited space, a tragedy is probably bound to happen. Automated video surveillance has become the need of the day and supports the analysis and management of data on a massive scale. It is very important to identify the presence of a crowd and detect thenumber of people in the gathering. This can prove very useful for the detection of sudden troupe build-up to avoid riots. Moreover, it can also be very useful in the Covid-19 pandemic situation to avoid people gathering at a place. This paper presents a system to detect the presence of a crowd by counting unique people and then performing crowd analysis. The crowd is analyzed by detecting the gender and age of people in the crowd.

***Keywords*—Deep Learning, Crowd Density Estimation, CNN, MobileNets, Neural networks**

I. INTRODUCTION

With the expanding population and several problems arising due to crowded scenarios in the cities, there is a necessity for crowd detection. Automatic crowd detection involves assessing the number of individuals in the video or an image. Further, the estimation of such crowd density can be done from the images of the crowded scene extracted from the surveillance video.

Crowd Density Estimation (CDE) is a challenging problem that can assist in solving various real-life problems. CDE is essential for disaster management and for maintaining the maximum people count during the COVID-19 pandemic.

Moreover, in the recent past, analysis of facial attributes has attained much credit in the field of computer vision. Various features of the human face, such, emotions, age, gender and ethnicity, can be used for categorization.

Some of these features can be used for Security, and video monitoring, electronic customer relationship management, biometrics, cosmetology, and forensic art are only a few of the real-world applications where age and gender classification is essentially helpful. However, the results obtained are not up to the mark. Various age and gender classification problems remain persistent complications.

Even with the advancement that has been made in the computer vision community with the continuous onward motion of modern techniques that improves state of the art, age, and gender predictions of the raw, real-life face are ineffective to meet the demands of commercial and real-world applications. Over many years, a lot of effort has gone into solving the classification problem. Many of these custom methods perform poorly when it comes to determining the age and gender of unconstrained in the wild pictures. These traditional tailor-made methods rely on discrepancies in attributes of facial feature and face descriptors, which are unable to cope with the various unpredictable variations encountered in these difficult unconstrained imaging conditions. There is a variation of images in these categories, such as Noise, posture, and lighting can all affect the ability of those manually developed computer vision methods to accurately classify the age and gender of the images.

Deep learning-based approaches have recently shown promising results in the age and gender classification of unfiltered face images. Further elaboration of existing works in age and gender prediction, as well as evidence of deep learning and CNN advances, we propose a VGG16 architecture model based on deep learning that predicts age group and gender of unfiltered in-the-wild facial images generated using a crowd detection model based on object tracking algorithm. It works by calculating the Euclidean distance between existing object centroids and new object centroids between subsequent frames in a video. We build an object tracker for each of our detected objects to track it as it moves around the frame. We monitor until we hit the Nth frame, then rerun our object detector. This complete process then repeats.

The remaining part of the paper is organized as follows. We present related work in crowd identification and age and gender classification in Section 2. The background of the models used in the method is presented in Section 3. The proposed method is described in Section 4. Section 5 presents the datasets used for training and experimentation. In section 6, the results are reported, accompanied by conclusions in section 7.

II. LITERATURE REVIEW

In this section, we present a review of various crowd detection approaches followed by a review of gender and age prediction methods.

A. *Crowd Detection*

There are various approaches for person detection. Mostly the crowd detection methods proceed by finding individual persons and counting them. People with a clear vision are involved in detecting, recognizing, and tracking items. Clustering-based methods,regression-based methods, and detection-based methods are the three types of methods. The clustering method[1-2], is used to detect different objects, and their trajectories are clustered to count the objects. Regression-based methods [3-4] first find low-level information such as foreground features, edge, and texture features. Scene level information is extracted from local and global properties such as Histogram of Oriented Gradients (HOG), Local Binary Pattern (LBP), etc. Finally, a regression function is exploited for counting. Detection-based methods involve person detection, target localization, tracking, and trajectory classification. A comprehensive study on person detection can be found in [5]. In this section we briefly review is the detection-based methods, as we utilize the detection-based method.

The early methods are based on low-level features. In [6],both appearance and motion features are utilized for the detection of a pedestrian. Haar filter, absolute difference Haar filter, and shifted difference filter are used to detect the objects with motion. Moreover, eight pedestrian detectors are trained using the Adaboost algorithm. Salim, et al. [7] presented method that can detect all the people passing through the field of view of the camera, with an average efficiency of 83.14%. It uses the Kalman filter [8] for predicting the tracked person's location in each frame. PETS2009 dataset was utilized for experimentation. Frame differencing is used in [9] to segment the crowd for people and then counting. Features are described for individual patterns, and counting is performed. Various methods [10,11] are also proposed that utilize Kinect camera to capture the depth information along with the low-level features. The most recent methods rely on deep learning. A crowd counting model is presented in [12] that uses a compact convolutional neural network to save computational resources and at the same time achieve great real-time speed, which is superior to the existing lightweight models. [13] proposes a deep model focused on Convolutional Neural Networks (CNN) and Spatio-Temporal context. The CNN model is used to detect people, and STC is used to monitor moving people's heads. [14] describes another supervised approach that uses Spatio-temporal features and their fusion.

B. *Gender and Age Prediction*

Face images are often used in gender and age detection methods. After extracting facial features, the images are classified into age and gender groups using classification and regression methods. The classification method [15- 16] used the Support Vector Machine based method. Various classification methods like Support Vector Machines, Radial Basis Function Networks, and the classical Discriminant methods were compared in [17], in which SVMs were able to succeed to give an acceptable error rate with storage of 20 per cent the training set. In [18], two competing hyper bf networks, one for male and one for female, were trained on geometrical shapes. Standard regression methods for age and gender classification include linear regression [19], Reinforce Vector Regression (SVR) [20], and Partial Least Squares (PLS) [21].

In [22], a gender and age prediction system is proposed. Face images are used for the task. First, the quality of the face image is improved using the histogram equalization method called Brightness Preserving Dynamic Fuzzy Histogram Equalization (BPDFHE). For detection of a face in the given image, Image segmentation and image filling are applied. For age estimation, Eigen face is used. More recently, Weber law descriptor was used in [23] for gender recognition which demonstrated outstanding performance on the FERET benchmark[24].The best results in [23] were obtained with the block size of 12X12 and T, M, and S values of 8, 4, and 4 respectively. In [25], features like intensity, shape, and texture were used with mutual information which again resulted in almost perfect results on the FERET benchmark.

In recent history, numerous methods have been introduced to solve classification problems leveraging deep learning techniques. However, the early neural network methods utilized small datasets. In [26], A neural network was trained on a minimal collection of 90 images (45 male and 45 female), with an error rate of 8.1%. Ranjan et al. [27] presented a model utilizing CNN for gender recognition and age estimation. It's an end-to-end network that shares CNN's lower layers' parameters. In [28], a robust estimations solution (CNN2ELM) is proposed. It uses CNN and ELM.

For age classification, authors in [29] brought the importance of deep neural networks how adding or subtracting a layer could change the output of the model. In [30] the measurements of the face were utilized for age detection. It was shown that using readily available dense building blocks to approximate the expected optimal sparse structure can be a viable method for improving neural networks [31].

III. BACKGROUND

A. *Single Shot Detectors and MobileNet*

Faster R-CNNs (Girshick et al.) [32], You Only Look Once (YOLO) (Redmon and Farhadi) [33], and Single Shot Detectors (Liu et al.) [34] are the most common methods for object detection. The R-CNN technique is difficult to train and is based on the Region proposal.The YOLO method is the fastest algorithm, capable of processing between 40-90 frames per second. SSDs, on the other hand, were developed by Google as a balance between the two. Depending on which network version is used, a faster FPS throughput can be achieved. SSDs are even more precise than YOLO.

For image classification and mobile vision, MobileNet is the CNN architecture model used for classification. Figure 1 depicts the structure of MobileNet. The Common Objects in Context dataset was used to train MobileNet. MobileNets are different from conventional CNNs because they construct lightweight deep neural networks using depth wise separable convolution.

Depth wise separable convolution, which was introduced in [33], is a combination of depth wise convolution and point wise convolution.

Convolution is split into two stages by depth wise separable convolution:

1. A 3×3 depth wise convolution
2. Followed by a 1×1 point wise convolution

This allows reducing the number of parameters in the network.

In comparison to other existing models, the MobileNet architecture needs very little computing power to run or apply transfer learning. They provide better performance for resource - constrained devices.

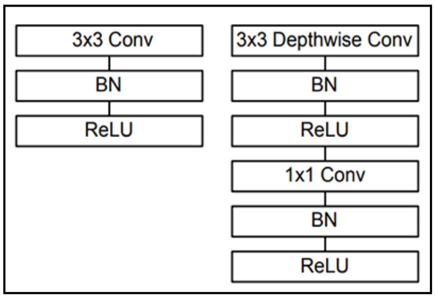


Figure 1- Architecture of MobileNet

B. *VGG-16 Model*

VGG-16 [35] is a pre-trained CNN proposed by Zisserman and Simonyan. VGG Net is trained on ImageNet dataset and is capable of performing the classification task with good accuracy. Images from 1000 classes are divided into three sets: 1.3 million training images, 100,000 testing images, and 50,000 validation images in the ImageNet dataset. There are 16 filtering layers in VGG-16.

Figure 2 shows the VGG-16architecture. The convolution layers with a non-linear activation function, which is a ReLU(rectified linear unit), are represented by the blue rectangles, 13 convolution layers and 5 max-pooling layers are included. The network has completely connected layers represented by three green rectangles.

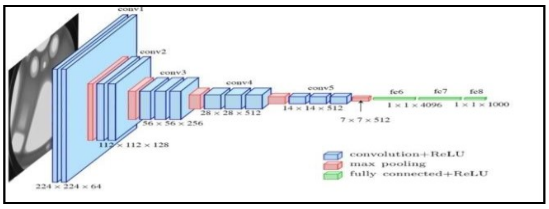


Figure 2. VGG-16 Architecture (from [35])

RGB images with an input size of 224 x 224 x 3 are accepted by the input layer. The images are passed through the convolution layers, which each have a small receptive field of 3x3 and a stride of 1. After the convolution, the resolution remains unchanged. A maxpool window with a size of 2x2 and a stride equal to 2 is used. The maxpool windows are non-overlapping in this case. Moreover, the maxpool layer does not follow the convolution layers. Without the having maxpool layer in between, a few convolution layers are followed by another convolution layer. The output layer has 1000 channels, one for each group of images in the dataset, and the first two completely connected layers have 4096 channels each. The activation function of the hidden layers is ReLU.

IV.PROPOSED APPROACH

A crowd detection approach is presented in this section. The people detected in the crowd are also analyzed by finding the gender and age of the people.

The overview of the proposed approach is presented in Figure 3. A video is accepted as input. Frames from the video are extracted by performing the uniform sampling. The frames are considered after every 30 frames. These video frames are passed as input to the Person detection and tracking module. This module detects the presence of persons in the image and extracts the regions from an image containing persons. It also counts the unique people in the given frames that assists in counting the number of people in the presented video. There by, one or more sub- images are created by the extraction of the regions of interest from a single frame. All the sub -images containing unique people in them are passed on to the Recognition module. The extracted person images passed to the recognition module are then analyzed to predict the gender and age of the person. The proposed approach is divided into two parts- crowd detection, and crowd analysis. We discuss these modules next.

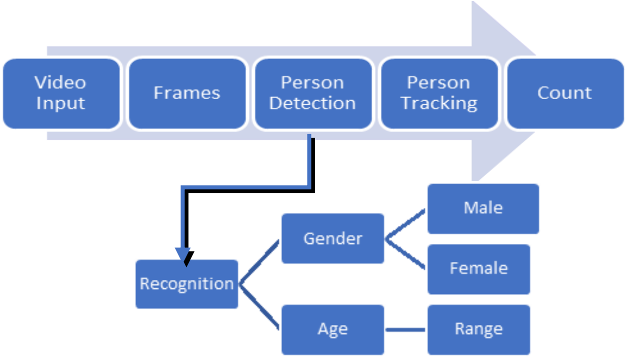


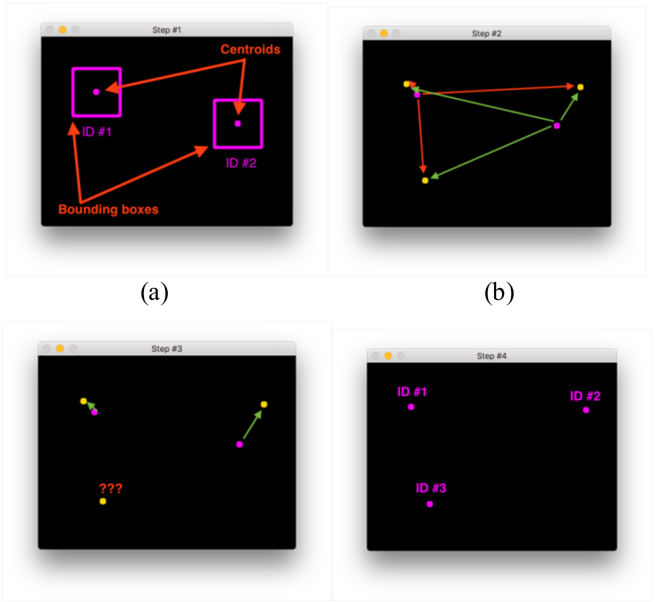
Figure 3. Overview of the Proposed Approach

A. CROWD DETECTION AND CROWD DENSITY ESTIMATION (CDE)

For detection of the crowd, Person detection is performed and the detected persons are tracked. The presence of humans is found using MobileNet SSDs Tracking is performed using Unique Person Detection Algorithmand it also displays the count of detected people.

For efficient detection of humans, we combine MobileNets and Single Shot Detectors. Specifically, MobileNets + SSD is used along with Opencv 3.3's DNN module (Deep Neural Network) to detect humans in images. The centroid tracking algorithm works as follows.

Crowd Density Estimation (CDE) Algorithm used in this paper is inspired by the centroid tracking algorithm[13]. Figure 4 shows the diagrammatic explanation of the algorithm and the algorithm is presented as algorithm1.First, the bounding boxes having coordinates (x, y) for each detected human in each frame is found. The bounding boxes can be generated using some of the common object detectors such as Haar cascades, RNN, color thresholding, etc. After detection of bounding boxes, the centroid is computed for each bounding box utilizing their respective (x,y) coordinates. Each bounding box is allotted a unique ID.



(a) (b)

(c) (d)

Figure 4. Unique Person Detection Procedure

(a) Compute centroids and find bounding box coordinates
(b) Calculate the distance between new bounding boxes and existing objects in Euclidean.
(c) Existing object coordinates are updated.
(d) Allot unique ids to objects

Crowd Density Estimation (CDE) Algorithm used in this paper is inspired by the centroid tracking algorithm[13]. Figure 4 shows the diagrammatic explanation of the algorithm and the algorithm is presented as algorithm1.First, the bounding boxes having coordinates (x, y) for each detected human in each frame is found. The bounding boxes can be generated using some of the common object detectors such as Haar cascades, RNN, color thresholding, etc. After detection of bounding boxes, the centroid is computed for each bounding box utilizing their respective (x,y) coordinates. Each bounding box is allotted a unique ID.

However, if a new unique ID is allotted to objects in each new incoming frame, it may cause a problem in the motive of object tracking. To alleviate this, we relate the centroid of the new object to that of the already existing human proposal and calculate the distance between them.

A list of tracked humans (TH) is maintained to detect the presence of unique people. To detect if any new humans are present in the new frame compared to the previous frame, the number of objects is counted. If the count of human proposals in the new frame is more compared to the previous frame then those newly detected objects are added to the list TH and a unique ID is allotted. Again the bounding box and centroid of the new proposal is computed. Following that, the path of the human proposal is detected by calculating the minimum distance using the Euclidean distance formula.

Furthermore, for any given video it is important to consider the fact that a human will move out of the field of view after some time. To address this issue, we deregister the object by removing the unique ID. When there is no match of the human proposal with the other existing object for a certain number of frames, say for N frames then we consider that the human proposal is lost and has moved out of the view, and hence it is deregistered. This assists in counting unique people and to develop the crowd counter.

Algorithm 1 Crowd Density Estimation

(CDE) Algorithm Input: Video Frames

Output: Crowd Count

Step 1: Bounding box and Centroid Detection

Step 2: Compute Euclidean distance between the two human proposals, using the following formula:

$$d\,(x, y) = \sqrt{(x1 - y1)^2 + (x2 - y2)^2}$$

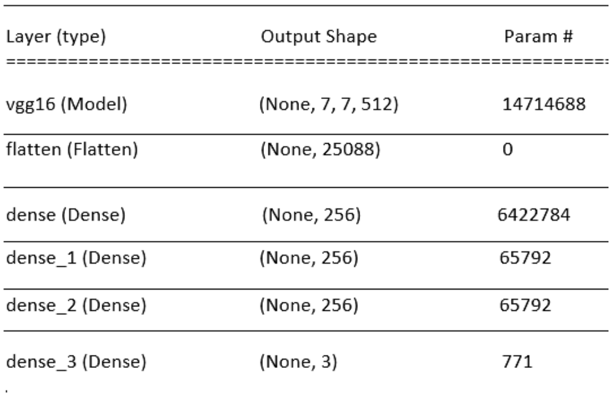
where (x1,y1) and (x2,y2) are the coordinates of the bounding boxes for which distance is computed.

Step 3: Register new human proposals
(a) Create a list of tracked humans called TH.
(b) Consider consecutive frames and find if any new human proposal is detected. If yes then add to the TH and assign a new unique ID to it.
Step 4: Deregister the human proposals that are out of the field of view

Step 5: Count all the unique human proposals detected in the video frames and output the crowd count.

B. CROWD ANALYSIS

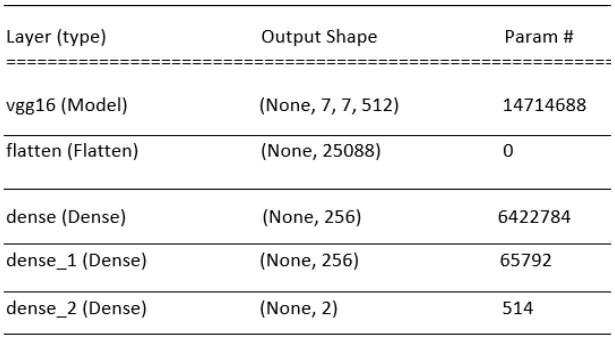
For analysis of unique people detected in the previous module, age and gender are predicted. We cater to the task of age and gender prediction as a classification task. For both the tasks we use a pre-trained CNN model for feature extraction as done in [23]. A two-level CNN architecture is leveraged to perform age and gender prediction. First, the features are extracted from the pre-trained VGG-16 [34], and then classification is performed. We classify the humans into three age groups (1-30), (30-60), and 60+. The age prediction model has 4 dense layers after feature extraction from the VGG-16. The last layer of Softmax has 3 nodes as we have three classification classes. The model summary is shown in figure 5. Test accuracy of the age prediction model is 69%.



```
Layer (type)                 Output Shape              Param #
==========================================================
vgg16 (Model)                (None, 7, 7, 512)         14714688
flatten (Flatten)            (None, 25088)             0
dense (Dense)                (None, 256)               6422784
dense_1 (Dense)              (None, 256)               65792
dense_2 (Dense)              (None, 256)               65792
dense_3 (Dense)              (None, 3)                 771
```

Figure 5. Age prediction model summary

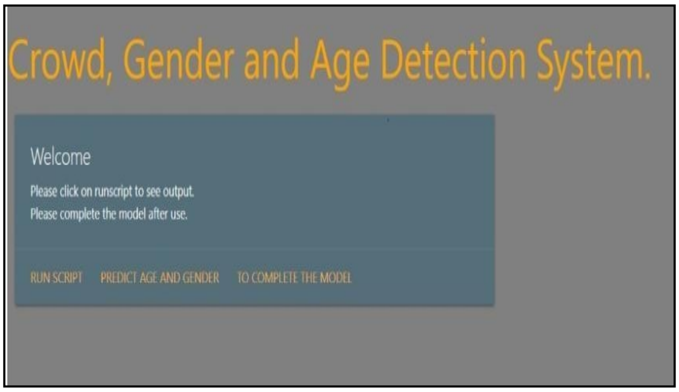
The same images of persons, passed to the age prediction module are passed through the gender prediction module also. The test accuracy of 90% was achieved by the model shown in figure 6. The gender prediction network's output layer is Softmax, which has two nodes that represent the two classes "Male" and "Female.



```
Layer (type)                 Output Shape              Param #
==========================================================
vgg16 (Model)                (None, 7, 7, 512)         14714688
flatten (Flatten)            (None, 25088)             0
dense (Dense)                (None, 256)               6422784
dense_1 (Dense)              (None, 256)               65792
dense_2 (Dense)              (None, 2)                 514
```

Figure 6. Gender prediction model summary

V. DATASETS



Crowd, Gender and Age Detection System.

Welcome

Please click on runscript to see output.
Please complete the model after use.

Figure 7. Home page of the UI

The dataset for age prediction as shown in the table-1 was taken from Kaggle, they were divided into 80% training, 20% validation, and testing was done manually. The dataset for age prediction contained JPG and PNG images for three age categories and was equally divided into directories and cleaning of data was done manually. The total number of images in the dataset and distribution of images into labels can be seen inTable-1.Thedataset was divided into three broad categories of age group 1-30,30-60 and 60+ to obtain better accuracy.

The dataset for gender prediction as shown in table-2 was taken from Kaggle, they were divided into 80% training, 20% validation, and testing was done manually. The dataset contained JPG images for men and women and was equally divided into directories and cleaning of data was done manually. The total number of images in the dataset and distribution of images into labels can be seen in Table-2.

	1-30	30-60	60+	Total
Train	5316 Images	4927 Images	3016 Images	13259 Images
Validation	855 Images	885 Images	911 Images	2651 Images

Table 1. Dataset for Age prediction

	Male	Female	Total
Train	1418 Images	1912 Images	3330 Images
Validation	312 Images	324 Images	636 Images

Table 2. Dataset for Gender prediction

VI. RESULTS AND DISCUSSION

Experimentations were performed on various videos from CCTV available over the internet. Figure 7 shows the home screen of our system. Following functionalities are available. Run Script: To run the MobileNet SSD model to find the unique people count in the given video. Predict age and gender: To pass the frames obtained from videos to the models created for age and gender prediction.

A crowd summary is then presented by displaying the extracted person images, their age, and gender. Complete the model: To rerun the model for another input and to erase the captured frames from the specified folder.

Figure 8 shows the result obtained for a sample video where the people are clearly visible and the video quality is good. Whereas figure 9 shows the results obtained on a night video. Although human proposals were detected correctly, the analysis was not completely correct. The age and gender were not correctly classified. Based on the experimentations performed, it was observed that the images obtained from videos that are of high resolution gave accurate results for both age and gender whereas the images obtained from videos that had low resolution or in which a person's face was not clearly visible did not give accurate results.

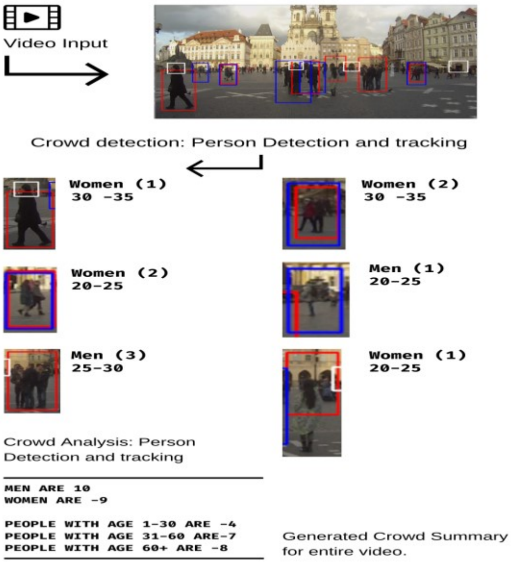

Video Input
Crowd detection: Person Detection and tracking
Women (1)
30 -35
Women (2)
30 -35
Women (2)
20-25
Men (1)
20-25
Men (3)
25-30
Women (1)
20-25
Crowd Analysis: Person Detection and tracking
MEN ARE 10
WOMEN ARE -9
PEOPLE WITH AGE 1-30 ARE -4
PEOPLE WITH AGE 31-60 ARE-7
PEOPLE WITH AGE 60+ ARE -8
Generated Crowd Summary for entire video.

Figure 8. Result obtained on a sample input video

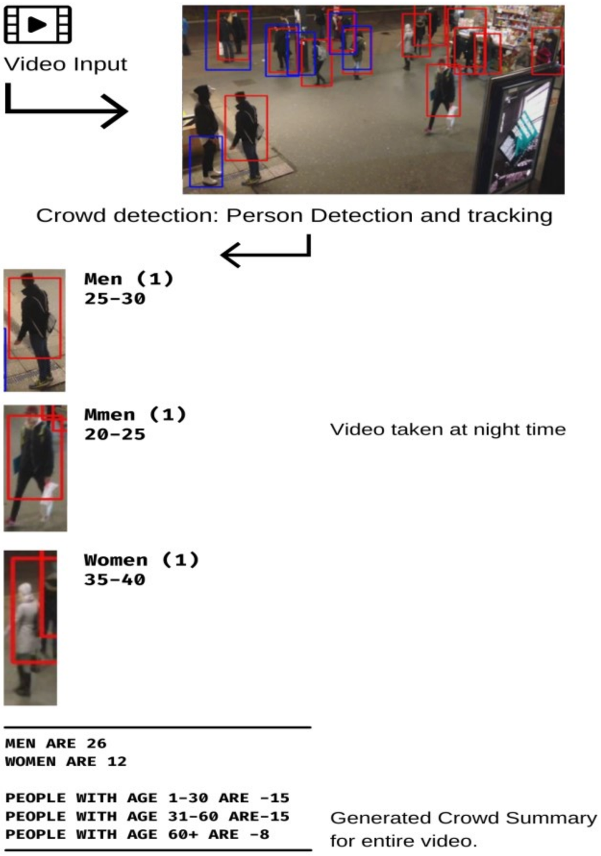

Video Input
Crowd detection: Person Detection and tracking
Men (1)
25-30
Mmen (1)
20-25
Video taken at night time
Women (1)
35-40
MEN ARE 26
WOMEN ARE 12
PEOPLE WITH AGE 1-30 ARE -15
PEOPLE WITH AGE 31-60 ARE-15
PEOPLE WITH AGE 60+ ARE -8
Generated Crowd Summary for entire video.

Figure 9. Results obtained on a sample input video captured during nighttime.

VII.CONCLUSION AND FUTURE WORK

The proposed system is capable of detecting the presence of a huge number of people in surveillance videos. It successfully detects the gender and age of people and gives the summary of people belonging to different age groups along with the gender for further analysis by the authorities. The system accepts input in the form of a video which further gets divided into frames using MobilNet SSD. The person detected in the frame is tracked, and the images of the person detected are sent to the age and gender prediction module. From the analysis of experiments performed on different types of surveillance videos, it can be concluded that the system can be very effective for solving many security and surveillance problems and can be used for data generation for analysis and further ruse. This system can be easily plugged into many other systems. Further, we intend to handle the night vision videos for the summarization of the crowd.

By far the most difficult portion of this project was setting up the training infrastructure to properly divide the data into folds, train each classifier, cross-validate, and combine the resulting classifiers into a test-ready classifier. I foresee future directions building off of this work to include using gender and age classification to aid face recognition, improve experiences with photos on social media, and much more. Finally, As a part of future work, a survey on other classifcation methods for age estimation can be done. Moreover, gender and ethnicity estimation and various other demographic features can be tested for their performance using Neural Networks classifer. I hope that additional training data will become available with time for the task of age and gender classification, which will allow successful techniques from other types of classification with huge datasets to be applied to this area as well.

References

[1] G. Antonini and J. P. Thiran, "Counting pedestrians in video sequences using trajectory clustering," IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 8, pp. 1008–1020, 2006.

[2] I. S. Topkaya, H. Erdogan, and F. Porikli, "Counting people by clustering person detector outputs," in 11th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2014, 2014, pp.313–318.

[3] Gould, Stephen, Tianshi Gao, and Daphne Koller. "Region-based segmentation and object detection." *Advances in neural information processing systems* 22 (2009): 655-663.

[4] Idrees, H., Saleemi, I., Seibert, C., Shah, M., 2013. "Multi-source multiscale counting in extremely dense crowd images", in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2547-2554.

[5] Raghavachari, Chakravartula, V. Aparna, S. Chithira, and Vidhya Balasubramanian. "A comparative study of vision based human detection techniques in people counting applications." *Procedia Computer Science* 58 (2015): 461-469.

[6] Jones, M.J., Snow, D., 2008. "Pedestrian detection using boosted features over many frames". In: 19th International Conference on Pattern Recognition, 2008. ICPR 2008. IEEE, pp. 14, http://dx.doi.org/10.1109 /ICPR.2008.4761703.

[7] Salim, Sohail, et al. "Crowd Detection And Tracking In Surveillance Video Sequences." *2019 IEEE International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)*. IEEE, 2019.

[8] Q. Wan and Y. Wang, "Multiple moving objects tracking under complex scenes," in The Sixth World Congress on Intelligent Control and Automation, Proc. IEEE 2, pp. 9871–9875, 2006.

[9] C. Chen, T. Chen, D. Wang, and T. Chen, "A cost-effectivepeople-counter for a crowd of moving people based on two-stagesegmentation," J Inform Hiding Multimedia . . ., vol. 3, no. 1, pp.12–23, 2012.

[10] L. Del Pizzo, P. Foggia, A. Greco, G. Percannella, and M. Vento, "Counting people by RGB or depth overhead cameras," PatternRecognition Letters, vol. 81, pp. 41–50, 2016.

[11] D. Pizzo, P. Foggia, A. Greco, G. Percannella, and M. Vento, "Aversatile and effective method for counting people on either RGB ordepth overhead cameras," in2015 IEEE International Conference onMultimedia Expo Workshops (ICMEW), 2015, pp. 1–6.

[12] Nascimento, Jacinto C., Arnaldo J. Abrantes, and Jorge S. Marques. "An algorithm for centroid-based tracking of moving objects." In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing.*

Proceedings. ICASSP99, vol. 6, pp. 3305-3308. IEEE Computer Society, 1999.

[13] G. Liu, Z. Yin, Y. Jia, and Y. Xie, "Passenger flow estimation basedon convolutional neural network in public transportation system," Knowledge-Based Systems, vol. 123, pp. 102–115, 2017.

[14] X. Wei, J. Du, M. Liang, and L. Ye, "Boosting Deep Attribute Learning via Support Vector Regression for Fast Moving Crowd Counting" Pattern Recognition Letters, vol. 47, pp. 178–193,2017.

[15] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[16] M. A. Beheshti-nia and Z. Mousavi, "A new classification method based on pairwise support vector machine (SVM) for facial age estimation," *Journal of Industrial and Systems Engineering*, vol. 10, no. 1, pp. 91–107, 2017.

[17] Moghaddam, Baback, and Ming-Hsuan Yang. "Learning gender with support faces." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, no. 5 (2002): 707-711.

[18] Poggio, Brunelli, R. Brunelli, and T. Poggio. "HyberBF networks for gender classification." (1992).

[19] A. Demontis, B. Biggio, G. Fumera, and F. Roli, "Super-sparse regression for fast age estimation from faces at test time," *Image Analysis and Processing—ICIAP*, Springer, Berlin, Germany, 2015.

[20] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1178–1188, 2008.

[21] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 657–664, Colorado Springs, CO, USA, June 2011.

[22] Kumar, S., Singh, S. and Kumar, J., 2019, January. Gender classification using machine learning with multi-feature method. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0648-0653). IEEE.

[23] Ullah, Ihsan, Muhammad Hussain, Ghulam Muhammad, and Anwar Mirza. "Gender Recognition From Face Images With Spatial WLD Descriptor."

[24] Phillips, P. Jonathon, Harry Wechsler, Jeffery Huang, and Patrick J. Rauss. "The FERET database and evaluation procedure for face-recognition algorithms." *Image and vision computing* 16, no. 5 (1998): 295-306.

[25] Perez, Claudio, Juan Tapia, Pablo Estévez, and Claudio Held. "Gender classification from face images using mutual information and feature fusion." *International Journal of Optomechatronics* 6, no. 1 (2012): 92-119

[26] Golomb, Beatrice A., David T. Lawrence, and Terrence J. Sejnowski. "Sexnet: A neural network identifies sex from human faces." In *NIPS*, vol. 1, p. 2. 1990.

[27] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 17–24, Biometrics Wild, Bwild, Washington, DC, USA, June 2017.

[28] M. Duan, K. Li, and K. Li, "An ensemble CNN2ELM for age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 758–772, 2018.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[30] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," IEEE Transactions on pattern analysis and machine intelligence, vol. 29, pp. 2234-2240, 2007.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9.

[32] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." *IEEE transactions on pattern analysis and machine intelligence* 39, no. 6 (2016): 1137-1149.

[33] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779-788. 2016.

[34] Fu, Cheng-Yang, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C. Berg. "Dssd: Deconvolutional single shot detector." *arXiv preprint arXiv:1701.06659* (2017).

[35] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

[36] Shakya, Subarna. "Collaboration of Smart City Services with Appropriate Resource Managementand Privacy Protection." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 3, no. 01 (2021).

[37] Ranganathan, G. "Real Life Human Movement Realization in Multimodal Group Communication Using Depth Map Information and Machine Learning." Journal of Innovative Image Processing (JIIP) 2, no. 02 (2020): 93-101.