Human Activities Detection using DeepLearning Technique- YOLOv8

Nilesh Parmanand Motwani*, Soumya* S
School of Robotics, Defence Institute of Advanced Technology (DU),

Abstract. Using a mask during the pandemic has occasionally been crucial and difficult. The use of universal masks can greatly lower and possibly even stop the spread of viruses within communities. So, mask detection has become avery critical task for security agencies in all the buildings, Government offices & other places. With the advent of GPUs, high computing machines, and Deep Convolution Neural Networks (DCCN), automatic Face & Mask Detection is possible by considering the image processing feature of extracting, 3-dimensional shapes from 2- dimensional images. This paper discuss about the YOLOv8 model to confirm its overall applicability, on two datasets namely FDDB & MASK. This helps to examine the behavior of the feature from the Mask dataset, which is intendedfor COVID-19 Mask Detection alone. Mask is the main dataset in this experiment. Above this, the ImageNet datasetis utilized for pretraining and FDDB (Face Detection Dataset & Benchmarks) datasets for recognizing face of a human being. The precision of models on FDDB is 58.9 % & on MASK dataset is 66.5%.

Keywords. *Object Recognition, Human activity, Intersection over union, Deep Learning, YOLOv8, IOU.*

1 Introduction

The inception of computer vision as a field was made possible by a better understanding of how biological neurons in the visual cortex of the brain work. This understanding revealed that visual input processing begins with the recognition of basic shapes like edges before progressing to more complex structures. The Neocognitron was invented by Kunihiko Fukushima in 1980 [4]. Yann LeCun applied a backpropagation algorithm to Neocognitron and released the first convolutional neural network, LeNet-5, in 1989. Similarly, Krizhevsky et al. introduced AlexNet in 2012, which is an 8-layer convolutional neural network with non- saturating ReLU activation neurons [5]. The network won the 2012 ImageNet Large Scale Visual Recognition Challenge [6] achieving a top-5 error of 15.3%. AlexNet architecture was considered the most influential neuralnetwork architecture.

In computer vision, face detection is related to the detection of human faces in images. Some are discussed below: While deep learning face detectors provide substantially improved accuracy and robustness, OpenCV'sHaar cascades continue to serve a useful purpose as they are light and secondly they are Superfast and even runwith low-resource devices. The model size is limited to around 930 KB. Apart from this, Haar cascades have anumber of flaws, including the fact that they are more prone to false-positive detections and are less reliable than their HOG + Linear SVM, SSD, YOLO, and other equivalents.

The most popular algorithm was first presented by Viola and Jones [7]. The method creates rectangular partsfrom the supplied image. Following that, each portion is run through a series of weak classifiers that scan for straightforward features that like those in the haar. A Haar-like feature is the variation in the sum of pixel intensities in several nearby locations. If the section successfully navigates all levels of the cascade, it is deemed to include a face; otherwise, it is rejected. Repeating this process results in rectangular pieces of varied sizes. This classifier is trained using AdaBoost. The fundamental advantage of the method is that since integral pictures are employed, the amount of time needed to calculate a haar-like characteristic does not change. A summed- area table, also known as an integral picture, is a type of data structure where each cell's value is determined byadding the values of the cells to its left and right. The summed-area table can be calculated in a single step. Many well-known libraries for computer vision use Haar cascade face detection, including Open CV.

Another alternative approach for this task is a feature descriptor called a histogram of oriented gradients (HOG). Authors [8] first proposed it as a way to recognise faces. The algorithm determines the gradient of pixelintensities for each individual pixel in the image. The image is then cut up into smaller pieces after that. Each section creates a histogram of the gradients.

After then, the most striking gradient is stored. Finally, the HOG picture is classified using Support Vector Machines (SVM). The library package Dilib has been using face detection based on HOG.

Various(Multi) Tasks Cascaded A deep learning technique for face detection called Convolutional Neural Networks (MTCNN) was first presented [9]. The technology recognises the face and facial landmarks in a singlepass. It makes use of a cascade of convolutional neural networks where, an image pyramid is initially created by resizing the provided image. An image pyramid is a representation of a photograph in various scales. It enables humans to identify objects in images of varying scales. This is frequently paired with a sliding window that can place things in different places. Compared to other facial landmark identification systems, the MTCNNsimply locates five landmarks: the tip of the nose, the corners of the mouth, and two eyes.

In CNN, a centered detection method is used where a region proposal component for creating a lot of promising regions consisting of one type of object which is trailed by the CNN classifier for classification. Themain idea after this model was to change multiple object classification concept to single object classification concept. Due to the slow region proposals methods, classification part is delayed. The major drawback of this system is that both the factors accuracy and speed cannot be simultaneously achieved during time critical situations [26].

1.1 Proposed Architecture based on MobileNet & on YOLOv8

In this paper, the foundation is based on the MobileNets neural network architecture [22]. This decision was made because the architecture is suitable for software that needs to balance processing speed and accuracyon embedded or mobile platforms. A convolutional layer can be divided into "depthwise" and "pointwise" operation while still keeping a significant portion of the network's representational power, according to MobileNets' creators. Due to this division, 3 x 3 convolutions require a lot less operations and parameters. A separable convolution is preceded by a further pointwise layer with linear activation to produce a "bottleneck,"and the linear bottleneck layers of the MobileNets architecture are built from separable ones. This bottleneck multiplies feature maps before lowering them and spreads the input in a higher-dimensional space to make use of ReLU activation's nonlinear power without sacrificing information. To enhance backpropagation and enhance computation graph execution, the MobileNets authors also inserted leftover connections from earlier work. Since each residual block's input and output tensor sizes are lower than the enlarged tensors processed between the bottlenecks, the presence of these skip connections requires an order

of execution where memoryusage is primarily determined by this.

Epochs:
100 epochs were run to check the performance in the iterations.

One crucial feature of YOLOv8 is its extensibility. It is designed as a framework that is compatible withall earlier iterations of YOLO, making it simple to move between them and evaluate their effectiveness. For individuals who want to take advantage of the most recent YOLO technology while maintaining the functionality of their existing YOLO models, YOLOv8 is the ideal choice as a result. The same model is then used to conductan inclusive investigation over the Mask Dataset. The conclusion acknowledges that YOLOv8 is an outstanding or the best model for detecting objects, human activity, or masks.

2. Design of System and Methodology

A. Parameter Setting For Neural Network

The GoogLeNet network, which has 24 convolutional layers and two layers which are FC, served as the only inspiration for this network. This system has the ability to take photos of any size as input and reshape them to 448*448 prior to sending them to this network. 7 * 7 grids are used to disperse the input photos. Each grid predicts three bounding boxes, each with four box coordinates, one class probability, and twenty class-specific sums. Thus, the result is a 7 * 7 (3 * 5 + 30) tensor.

B. Cost Function

A cost function indicates how quickly a neural network learns for oneself. The network predicts class- specific scores and further bounding boxes during the item recognition exercise. As a result, the overall cost$Cost_{total}$, which is denoted by Equation 3.1, is the outcome of adding the class cost $Cost_{class}$ with the boundingbox cost $Cost_{IOU}$. The change in the class probability and desired values (0 or 1) is what is referred to as the cost function.

$$Cost_{total} = Cost_{class} + Cost_{IOU} \quad 3.1$$

The cost of calculation in the real scenario is very difficult. In our test, the cost is calculated by the Equation 3.2. There are factors which are used to modify weights of various costs are noobject scale, object scale and class scale. The addition of the S x S grids costs is known as the total cost. Every grid consists of Nboxes having cost for presence of object, C class cost and cost for assuming the box, i.e. IOU cost and are calculated by square error function. When the object is existing in the grid cell (isObj = 1), the cost function will impose a charge by summing cost of class and cost of best IOU.

$$\text{Cost} = \sum_{m=0}^{S*S} \{ \sum_{n=0}^{N} n \text{ oobject\_scale} * P^2 \text{ object,n} + \text{isObj} * \{-\text{noobjectscale} * P^2 \text{ object,bestprediction} + \text{objectscale} * (1 - P_{\text{object,bestpredict}})^2 + (1 - IOU_{\text{bestpredict}})^2 + \sum_{i=0}^{C} (\text{classscale} * (P_{\text{class,i}} - P_{\text{truth,i}})^2) \} \} \quad 3.2$$

Here S = grid size, N = in every grid how many bounding boxes are predicted. C = no. of classes. The noobject scale, object scale, and class scale are some of the variables that are used to change the weights of different expenses. $P_{object,n}$ is the likelihood that an object is present in box n of grid m. According to the intendedresult, isObj indicates if an object is present in the current box. The P_{object} with the best IOU prediction out of allthe n boxes is the P_{Object}, bestpredict. The intersection of the ground truth bounding box with the best prediction bounding box in grid m is known as $IOU_{bestpredict}$. $P_{class,i}$ is the class i class probability for prediction and $P_{truth,i}$ is the class i ground truth class probability in grid m, both of which are either 0 or 1.

C. Evaluation Criteria

To authorize and confirm results of detecting objects, I have used two standards, viz. overall precision & overall recall. For authenticating object detecting results, we used overall precision and overall recall. For direction estimation, direction accuracy is used. Schematicdiagram Figure 1, gives the understanding of overallprecision, overall recall.

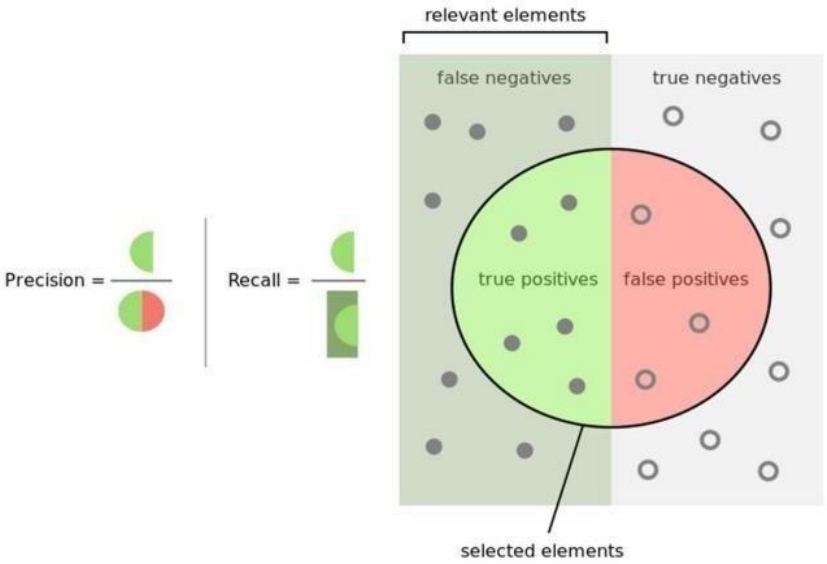

Fig. 1. Precision and Recall [15]

The calculation of overall precision and overall recall is specified in the below Equation Overall precision is defined as how many predicting objects are the related objects. Similarly, Overall Recall is defined as how many predicted elements are relevant.

$$Precision = \frac{tp}{tp+fp} \quad Recall = \frac{tp}{tp+fn}$$

D. Datasets:

In this study, mainly three major datasets are involved for experimental purpose. One is Mask dataset- specifically for COVID-19 mask detection, other is ImageNet [4] and Face Detection Dataset & Benchmark[FDDB].

E. Mask Dataset

MASK Dataset is the best suited for the mask detection system. The collection of data was done at the gates of big companies, Airports & other various sensitive premises by installing various sensors. The Maskdataset includes total 3 classes having 5237 labelled images. Single image size for this dataset is around 750 kb to 850kb, hence the approximate total size becomes 5.5 GB. These 3 classes are persons with Mask, without mask and Improper Mask. The total of 5237 images are divided into 80% % 20% which consists of 4190 number of images and 1047 number of images respectively for training and testing respectively. Speedof detection is: 1.8ms pre-process, 9.0ms inference, 0.0ms loss, 2.0ms post-process per image.

F. ImageNet Dataset:

This is one of the largest dataset having around 1000 classes and total images of around 1.40 million.

These all images are used for pre-training means before training the model, it is used.

G. Face Dataset [FDDB]

This dataset consists of one and only one class, which is known as “face”. Total number of faces available in2845 images are 5171. The data set size is around 523 MB.

3 Results Analysis

A. Dataset FDDB

Fig. 2. Detection of Face Results on Face detection Datasets

Table 1. Results after training dataset-FDDB

Input Image	**Test**		**Results**		
	Testing	***Training***	***Precision***	***Recall***	***Detecting Speed***
2845	569	2276	58.9%	54.2%	0.031 sec/image

B. Mask Dataset Results

Fig. 3. Detection of mask on Mask Detection Datasets

Table 2. Results after training dataset-MASK

Input Image	**Test**		**Results**		
	Training	**Testing**	**Precision**	**Recall**	**Detecting Speed**
5237	4190	1047	66.5.%	62.2%	0.02 Sec/image

YOLOv8 for MASK dataset: The values of Precision, Recall, mAP50 are shown below:

Table 3. Precision, Recall of YOLO v8 model on MASK Dataset

Class	Images	Instances	Box(P	R	mAP50	mAP50-95): 100%
all	294	1473	0.665	0.622	0.615	0.387
no_mask	294	225	0.774	0.68	0.715	0.416
mask	294	1235	0.827	0.8	0.814	0.502
improper_mask	294	13	0.394	0.385	0.315	0.243

4 Conclusion

In this paper, YOLOv8 model got trained on various datasets i.e. FDDB & ImageNet. The resultsfound from the experimentation depicts that the YOLOv8 performed extremely well for all the FDDB Dataset compared to the benchmark.

Eventually, this work helped us come to the conclusion that YOLOv8 has achieved 84.6% accuracy with 66.5% recall at 0.030 sec. per picture while employing the model over the Mask benchmark.These results gave us hope and highlighted that the YOLO model is a better tool for identifying different things needed in the field of self-driven vehicles.

If we look at the exponential rise from version 5 to version 7, it is clear that training time was a major problem, however version 8 is taking less time in delivering results with a greater mean average precision. The problem of lengthy training is partially addressed here. Yolov8 more successfully balances precision with training. Completely developed network-backbone, an anchor-free detection head, and anovel reducing loss(cost) function have made it much faster.

References

[1] J. D. S. G. R. A. F. A. REDMoN, “You only look once: Unified, real- time object detection.,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016

[2] M. A. W. J. EvEriNghAM, “The pascal visual object classes challenge 2011 (voc2011) development kit,” Pattern Analysis, Statistical Modelling and Computational Learning, Tech, 2016.

[3] A. L. P. S. C. A. U. R. GEigEr, “Vision meets robotics: The kitti dataset,” The International Journal of Robotics Research, 2013.

[4] J. B. A. S. S. S. H. K. A. A. F.-F. L. DENg, “Imagenet large scale visual recognition competition 2012 (ilsvrc2012),” 2012.

[5] V. A. L.-M. E. G. F. JAiN, “A benchmark for face detection in unconstrained settings,” UMass Amherst Technical Report, 2010.

[6] A. S. I. A. H. G. E. KrizhEVsKy, “Imagenet classification with deep convolutional neural networks,” in In Advances in neural information, 2012, pp. 1097-1105.

[7] K. A. Z. A. SiMoNyAN, Very deep convolutional networks for large- scale image recognition., arXiv preprint arXiv:1409.1556, 2014.

[8] R. D. J. D. T. A. M. J. GirshicK, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014

[9] P. A. J. M. VioLA, “Rapid object detection using a boosted cascade of simple features.,” in Computer Vision and Pattern Recognition, Proceedings of the 2001 IEEE Computer Society Conference, 2001.

[10] C. L. W. J. Y. S. P. R. S. A. D. E. D. V. V. A. R. A. SzEgEDy, “Going deeper with convolutions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[11] M. A. A. F. D. SADEghi, “30hz object detection with dpm v5.,” in European Conference on Computer Vision, 2014.

[12] P. A. J. M. VioLA, “Rapid object detection using a boosted cascade of simple features.,” in Computer Vision and Pattern Recognition, Proceedings of the 2001 IEEE Computer Society Conference, 2001.

[13] S. Xu, X. Wang, W. Lv, Q. Chang, C. Cui, K. Deng, G. Wang, Q. Dang, S. Wei, Y. Du,

et al., “Pp-yoloe: An evolved version of yolo,” arXiv preprint arXiv:2203.16250, 2022.

[14] X. Huang, X. Wang, W. Lv, X. Bai, X. Long, K. Deng, Q. Dang, S. Han, Q. Liu, X. Hu, et al., “Pp-yolov2: A practical object detector,” arXiv preprint arXiv:2104.10419, 2021.

[15] R. Liu, J. Lehman, P. Molino, F. Petroski Such, E. Frank, A. Sergeev, and J. Yosinski, “An intriguing failing of convolutional neural networks and the coordconv solution,” Advances in neural information processing systems, vol. 31, 2018.

[16] G. Jocher, A. Chaurasia, and J. Qiu, “YOLO by Ultralytics.” https://github.com/ultralytics/ ultralytics, 2023. Accessed: February 30, 2023.

[17] Alibaba, “TinyNAS.” https://github.com/alibaba/lightweight-neural-architecture-search, 2023. Accessed: March 18, 2023.

[18] C.-Y. Wang, H.-Y. M. Liao, and I.-H. Yeh, “Designing network design strategies through gradient path analysis,” arXiv preprint arXiv:2211.04800, 2022.

[19] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 658–666, 2019.

[20] Z. Gevorgyan, “Siou loss: More powerful learning for bounding box regression,” arXiv preprint arXiv:2205.12740, 2022.

[21] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, “Tood: Task-aligned one-stage object detection,” in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3490–3499, IEEE Computer Society, 2021.

[22] Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li, and Y. Fu, “Rethinking classification and localization for object detection,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10186–10195, 2020.

[23] S. Wang, J. Zhao, N. Ta, X. Zhao, M. Xiao, and H. Wei, “A real-time deep learning forest fire monitoring algorithm based on an improved pruned+ kd model,” Journal of Real-Time Image Processing, vol. 18, no. 6, pp. 2319–2329, 2021.

[24] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, “Distance-iou loss: Faster and better learning for bounding box regression,” in Proceedings of the AAAI conference on artificial intelligence, vol. 34, pp. 12993–13000, 2020.

[25] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, “Cspnet: A new backbone that can enhance learning capability of cnn,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 390–391, 2020

[26] N. P. Motwani, S. S and U. Singh, "Object Detection and Tracking for Autonomous Vehicles using Deep Learning Technique- YOLO," 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2022, pp. 1-6, doi: 10.1109/SMARTGENCON56628.2022.10083703.