Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

People detection and counting using YOLOv3 and SSD models

Pooja Gupta [a,*], Varsha Sharma [a], Sunita Varma [b]

[a] School of Information Technology, RGPV, Bhopal, India
[b] Department of Information Technology, SGSITS, Indore, India

ARTICLE INFO

Article history:
Received 11 November 2020
Received in revised form 11 November 2020
Accepted 17 November 2020
Available online xxxx

Keywords:
You only look once
Single shot multibox detector
Object detection
Object counting

ABSTRACT

Object detection has become a crucial task for the various applications used in the real world such as surveillance, security, and automated vehicle system. The counting of the numbers of peoples at any junction also having various applications to provide integrity to any task. To count the number of peoples at any junction, we have various methods. Among the present methods, we analyzed the two algorithms that are You Only Look Once (YOLOv3) and Single Shot multi-box Detector (SSD). Two tasks were performed independently; one is for object detection by using the image dataset, and the other one is the counting of objects by using the video dataset. In this research, these two methods are analyzed for counting as well as detection efficiency, and comparison is presented. The results have shown that the precision, recall, and F1 measure achieved for SSD is higher than YOLOV3 v3.

Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

1. Introduction

Object counting has been used in a variety of fields for improved service providing as well as for the detection of different perspectives of the objects [1]. It is one of the important work that is seen in nowadays and is used for intelligent systems, driverless cars also in various fields. This field has achieved a large attraction from the researchers. The convolutional neural networks have been used earlier for object detection. As there are various object detection techniques available, the YOLOv3 and the SSD, have attracted more researchers for object detection. Both methods can work for the image as well as for the video detection of the objects.

Here in this research, we did tasks like object detection as well as the counting of the objects. The primary task in our algorithm is to count the number of humans present in the video. Initially, the object is detected as to whether it is human or any other thing in the image followed by counting the numbers of humans present in that video. The number of people counting is performed vertically and horizontally.

In this work, YOLOv3 and SSD algorithms are used. The performance of both the algorithms in object detection as well in counting the number of peoples in the video is studied. The working method of both algorithms makes them useful for the task of object detection as well for object counting.

2. Object detection

Object detection is one of the major techniques that is used for locating the objects present in the given image or the video. The task of object detection uses mainly machine learning and deep learning methods for detecting accurate objects [2]. Humans do the detection of objects by looking at the image or video; we want the same detection capability by the computer by using its intelligence. Object detection is used in video surveillance systems and many more fields.

Some of the deep learning-based algorithms for object detection are Region convolutional neural network(R-CNN) and YOLO v3 [3]. And in machine learning the algorithms used are Support Vector Machine (SVM) etc. Object detection is a two-step procedure. First to train the model to be used and secondly, the original data set is used to perform object detection [3]. Also, we can train the system as well we can use the already trained system. The selection of training methods between deep learning and machine learning depends on the purpose of the model as shown in Fig. 1.

3. Object counting

Counting the humans present in the videos has various applications in intelligent systems [4]. A system that has a counting capa-

* Corresponding author.
E-mail address: pooja1porwal@gmail.com (P. Gupta).

https://doi.org/10.1016/j.matpr.2020.11.562
2214-7853/
Selection and peer-review under responsibility of the scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.

Fig. 1. Object Detection for Humans in an image.

bility of the humans present in the video is necessary for multiple tasks. Vision-based object counting has multiple tasks involved in it like detection of a particular object, recognizing the object, and also tracking that object.

Three categories can help in performing such tasks that are regression methods, clustering methods, and detection-based methods [5]. The regression method uses a regression function by making the use of the regions that are used for detection, and this is used for the counting. The next method clustering does the counting by tracking certain features for the discrete objects, and their trajectories are clustered and used for counting. The third method is the detection based that uses tracking, trajectory, extraction for counting.

In this research work, we have used the deep learning methods that make the use of the convolutional neural network. This method has achieved great attention recently for object detection and the counting of objects [6]. The object counting is performed bi-directional, where the counting for both the incoming and outgoing peoples is done as shown in Fig. 2.

4. Literature review

Wang et al. have worked for vehicle trajectory data in the traffic conditions. They analyzed their work for the unmanned aerial vehicles (UAV) and used the UAV videos. They proposed a vehicle trajectory model in which first the CNN model and then YOLO v3 is used for the detection of vehicles. The results obtained were accurate and wise. [4]

Stahl et al. overcome the limitations observed in normal CNN, the authors have designed a model that is based on the YOLOv3. The YOLOv3 based model they implemented for the fast detection on the PASCAL VOC dataset. They have worked for the reduction of power consumption in the use of DRAM. [5]

Dai et al.have worked on human detection in the given dataset. They used the PDCS dataset, which has more than 4500 videos. The detection of humans in the given video is done. The presented work calculates the point cloud from the depth video. They counted the human beings in the video having 45fps on a 1.7 GHz processor.[6]

R. Feng et al, performed the teat detection for image recognition and the detection of objects. They have worked on the feature like HAAR. Also, the YOLO v3 algorithm was used for comparison. They performed their detection on the fake teat model made from silicon for the indoor environment. [7]

D. T. Nguyen et al. have analyzed the layers used in the SSD for detection and stated that there is a need for correction in that. They have presented a balanced feature fusion SSD model for the proper detection of objects. This algorithm has improved the detection of small objects in the traditional SSD. [8]

S. sun et al. have worked for the detection of vehicles and wheels. The limitations on CNN for this type of detection have been removed in this paper. They have presented a novel optimized SSD algorithm. The work was carried out on the dataset of PASCAL VOC 2007. They achieved a 1.5% high mAP that traditional SSD, [9]

A. Rastogi et al. presents a real-time behavior detection for the egg breeders. For this, they have used the YOLO v3 for the detection

Fig. 2. Counting Peoples in Both Direction in a video.

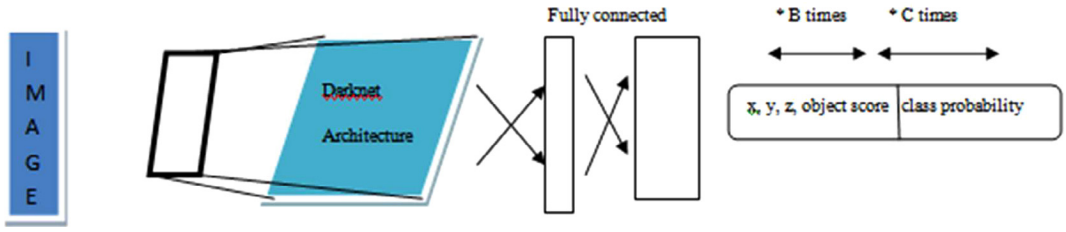

Fig. 3. YOLO v3 Architecture [14].

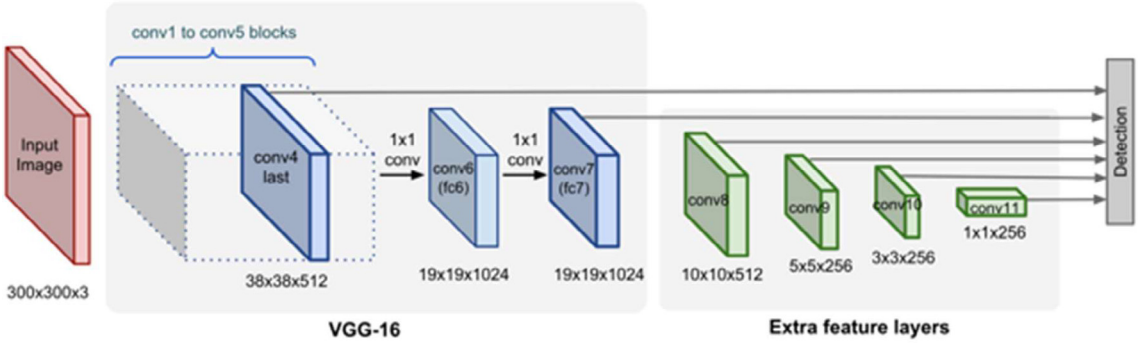

Fig. 4. SSD Architecture [15].

of the object. The manual training of the model was done. All the dataset was collected manually. The obtained result was satisfying [10].

H. Zhao et al. stated that face detection is a crucial task, and its proper detection is a must. The faster RCNN algorithms work well as they use 2 stages for detection. The drawbacks in YOLOv3 was removed by using the presented approach that is YOLOv3 face based on YOLO v3. The result was more accurate for face detection in the presented method [11].

J.Fu et al. have worked on object detection from the perspective of blind peoples. To guide a blind person, they have designed a prototype that can be used for any device creation. The object and distance should be provided to a blind person through that device. They have used SSD for this task, and the object detection was carried out [12].

J. Wang, has worked on the detection of gates for the drones used. The cameras used in the drone should have the capability to detect the gate. For this, they have presented a CNN based SSD for drone racing circuits. Various parameters were analyzed and provided efficient results [13].

5. Analytics performed

5.1. YOLO v3

It is an algorithm that works on the detection of objects for the real-time environment. This algorithm works fast and straightforward as it does not consume time for the generation of the region proposals. This algorithm focus on the recognition of the objects and the speed of the detection without perfection in locating the objects. Some previous algorithms, like faster RCNN, are seen to be more accurate, but complex implementation. Also, it has many outputs that surely result in errors. After the training of this model also, they are not worthy in the real-time environment.

For our proposed work, we have used the algorithm in two steps as shown in Fig. 3:

Object detection in YOLO v3:

1. The pre-training of the CNN for image classification.

2. Divide the image into small cells. In case the center of the object is in the cell, then this cell is responsible for the detection of that particular object.

3. Three tasks are done by every cell. Location of the box, b. confidence score, c. I am finding the probability of classification of the object in the bounding box.

i. The bounding box is defined by (x, y, width, height), width, and height vary between (0,1).

ii. The confidence score of the cell is the matching of the cell with the object.

P (containing an object) * IoU (pred, truth); here the P is a probability, and IoU is interaction under union.

In case the cell contains the actual object, then the prediction is made for every belonging class

P (Object belongs to class Ci | containing an object). Here single class probability is calculated in each cell, without noticing the number of bounding boxes.

The last layer of CNN (Pre-trained) is modified for outputting the prediction tensor of size.

Object counting in YOLO v3:

1. Set the threshold value 0.2 of Detection.

2. Detect people with re-training Region-based Convolutional Network (R-CNN) using YOLO v3 Model

3. Detect people and count based on appropriate boundary Selection.

4. Process the Count Information of the Previous Step

5. The output of Count in People in the form of In and Out the boundary.

5.2. Single shot multibox detector (SSD)

The full form for the SSD is a Single Shot Detector; the name itself defines the work done in this technique. This method a single shot for the detection of multiple objects present in the given image. At the same time, the Regional proposed network (RPN) based methods like R-CNN requires at least 2 shots for the detection of the objects. The two shots are used as the first for the generation of the regional proposals and the second for the detection

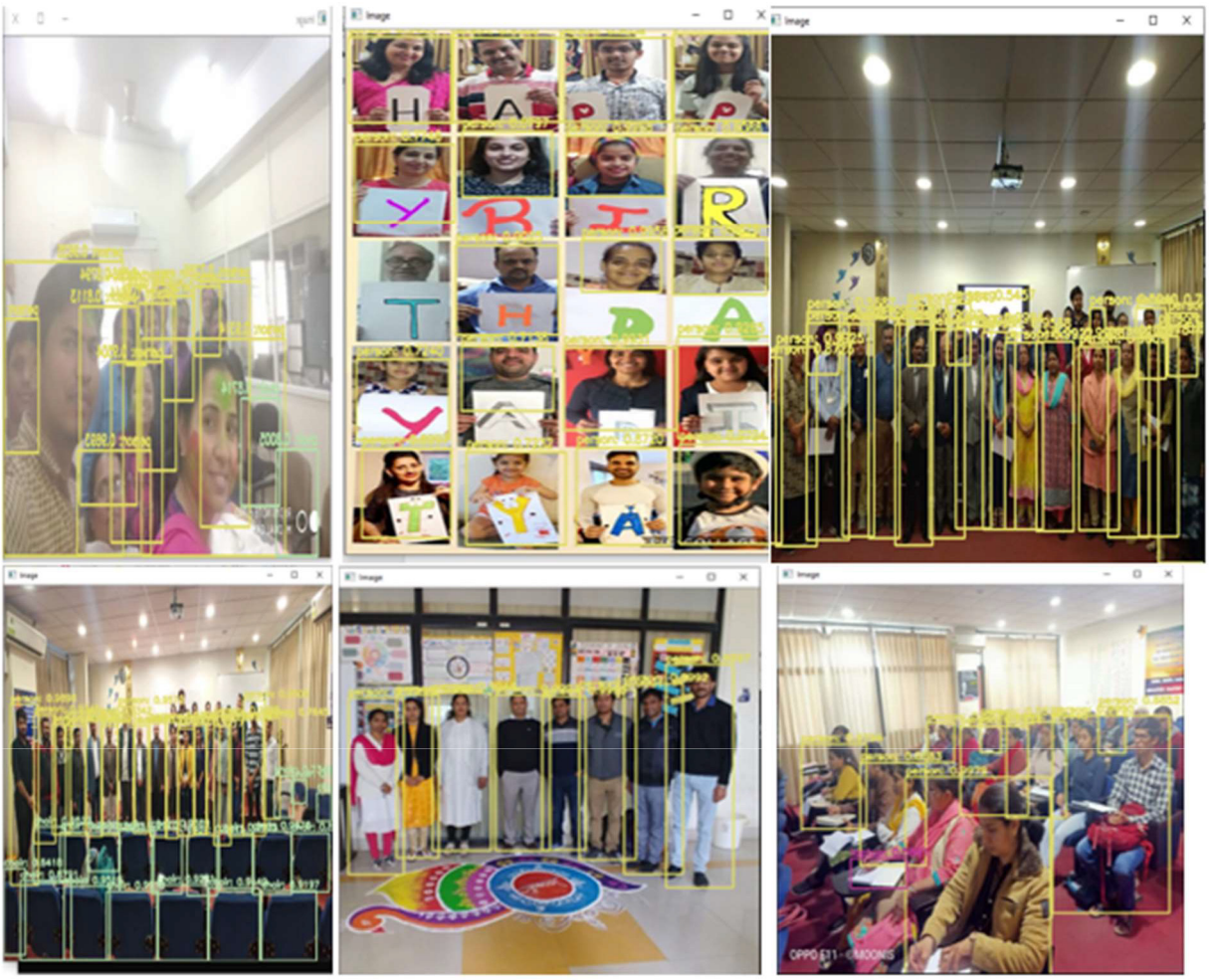

Fig. 5. Object Detection using SSD and YOLO v3 Method in images.

Fig. 6. Object Counting using SSD and YOLO v3 Method in videos.

of the objects in each proposal. So it is seen that the SSD is very fast as compared with the RPN based methods at the same time is works much faster and accurate than the YOLO v3.

The SSD here uses the pre-trained model of VGG-16. The pre-training of this model is done on our dataset. Using this adds some feature layers which decrease in sizes. Deep layers for fine granularity helps in accurate detection of even smaller objects. Here the detection is done at every layer for the detection of all the objects that are varying in sizes.

Table1
Results for the Training dataset.

DATA	PRECISION	RECALL	F1
Image 1	0.99	0.94	0.95
Image 2	0.99	0.91	0.92
Image 3	0.99	0.96	0.97
Video 1	0.99	0.90	0.92
Video 2	0.99	0.93	0.96
Video 3	0.99	0.93	0.96

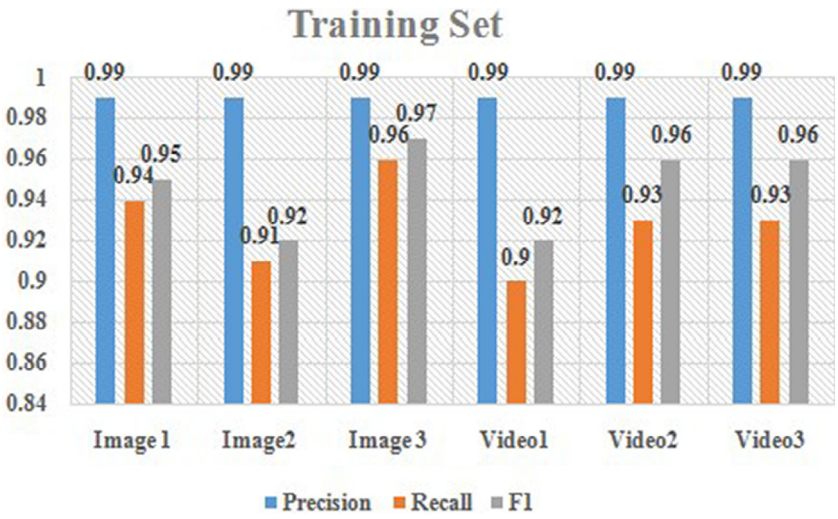

Fig. 7. Graph for the Training Dataset Results.

The algorithm architecture as shown Fig. 4 and steps for SSD implementation is as below:

Object detection in SSD:

1 The first task done is the category prediction

Let q is object category, where anchor boxes are q + 1, h and w are height and width of the feature map, and ×, y are coordinates of the feature map.

2 The bounding box prediction

This is the same as step one, but we need to define 4 offsets for every anchor box, except the categories, i.e. q + 1

3 Concatenating predictions for multiple scales

Here prediction of batches for different scales are done as

P (batch size, number of channels, height, width)

4 Height and Width downsampling

The h and w, here is reduced by 50% approx.

5 Base network block: here the extraction of features from the actual images is done.

Object counting in SSD:

1. Set the threshold value 0.2 of Detection.

2. Detect people with re-training Fast Region-based Convolutional Network (Fast R-CNN) using SSD Model

3. Detect people and count based on appropriate boundary Selection.

4. Process the Count Information of the Previous Step

5. The output of Count in People in the form of In and Out the boundary.

6. Experimental analysis

6.1. Hardware used for implementation

The following configuration is used for the implementation of the algorithms.

For the design Python Programming Language, 15.6 in HD WLED touch screen (1366 × 768), 10-finger multi-touch support. 10th Generation Intel Core i7-1065G7 1.3 GHz up to 3.9 GHz. 8 GB DDR4 SDRAM 2666 MHz, 512 GB SSD, No Optical Drive. Intel Iris Plus Graphics, HD Audio with stereo speakers. HP True Vision HD camera. Realtek RTL8821CE 802.11b/g/n/ac, Bluetooth 4.2, 1 HDMI 1.4, 1 USB 3.1 Gen 1 Type-C, 2 USB 3.1 Gen 1 Type-A. The Python Programming was run on Windows 10 64 bit Operating System platform. The python library was used during implementation like NumPy, Pandas, Matplotlib, SciPy, Scikit-Learn, PyTorch, Seaborn, XG Boost, Plotly, Tensor Flow, Keras, Seaborn.

6.2. Dataset for analysis

We have used two data set, the first is the image dataset, and the second is the video dataset. Image data set is used for object detection like humans and video datasets for object counting for humans.

Image data sets collected from the SGSITS college event, some family events. Video data sets were collected from SGSITS College MPPSC exam center and Independence Day celebration in SGSITS, College.

6.3. Object detection

The dataset used for the detection of human beings has the following images, and the detection is done in Fig. 5. The humans are detected in the following images using both the YOLO v3 and SSD methods. The size of the feature map varies for every object, and the algorithms obtained accurate detection.

6.4. Object counting

The task of object counting is shown in Fig. 6. The humans are counted in particular videos. The object counting was performed for bidirectional traffic that is incoming humans and outgoing humans. They are stated as up for incoming and down for outgoing humans. The following are some screenshots that show the object counting task done for both the algorithms.

7. Result analysis

The overall implementation was satisfactory, and the required results are obtained. In the comparison made between the YOLOv3 and SSD, SSD results show higher accuracy than YOLO. We have

Table 2
Table for the Test Dataset Results.

DATA	PRECISION	RECALL	F1
IMAGE 1	0.90	0.86	0.89
IMAGE 2	0.87	0.82	0.85
IMAGE 3	0.88	0.83	0.86
VIDEO 1	0.89	0.81	0.85
VIDEO 2	0.91	0.85	0.88
VIDEO 3	0.87	0.81	0.84

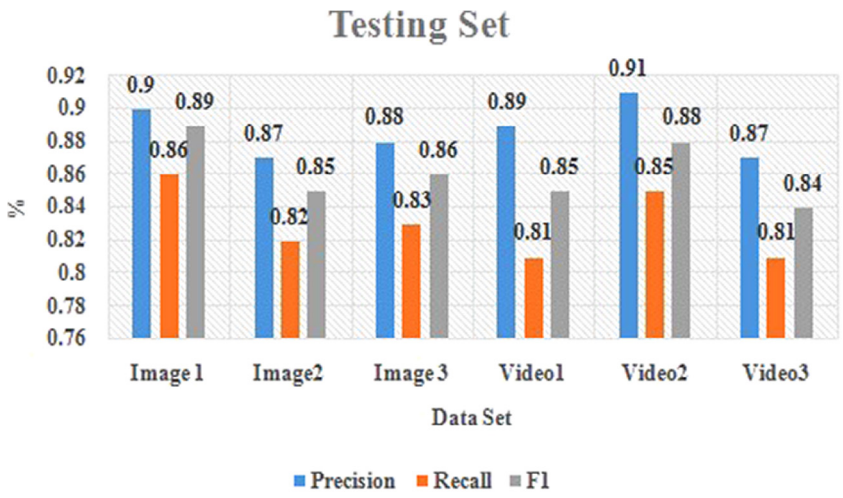

Fig. 8. Graph for the Test Dataset Results.

used 3 images and videos for the analysis. The parameters used for the analysis are precision, recall, and F1 measure. The result for both the methods have shown below:

7.1. YOLO v3 model gave the following results, which are stated below

The analysis was done for both the training and testing datasets.

The Table 1 shows the results obtained for the training model. The precision found for this is 0.99, the recall obtained was high for all the images and the video. As well the F1 measure founded is shown in the table.

The graphical representation is shown in Fig. 7. All three parameters have shown in the graph. The precision is found at the highest of all. The three datasets for both the video and images have been compared here.

Now the testing dataset has been analyzed for the YOLO v3 model. Below is the tabular representation of the testing dataset is stated.

The Table 2 shows the results observed for our testing dataset. Here the precision observed is an average of 90, and the recall is of average 84 also the F1 measure is obtained of average 0.86.

The graphical representation of the values obtained is shown in Fig. 8. The numerical values for both the image and video datasets are given here. The three parameters have been defined in the graph with their obtained numerical values.

Table 3
Results for training dataset in SSD.

DATA	PRECISION	RECALL	F1
IMAGE 1	0.98	0.93	0.96
IMAGE 2	0.99	0.91	0.95
IMAGE 3	0.99	0.88	0.92
VIDEO 1	0.98	0.89	0.93
VIDEO 2	0.98	0.84	0.94
VIDEO 3	0.99	0.87	0.95

Table 4
Results for Testing Dataset in SSD.

DATA	PRECISION	RECALL	F1
IMAGE 1	0.96	0.82	0.92
IMAGE 2	0.96	0.86	0.91
IMAGE 3	0.95	0.88	0.89
VIDEO 1	0.95	0.84	0.93
VIDEO 2	0.96	0.85	0.92
VIDEO 3	0.95	0.83	0.91

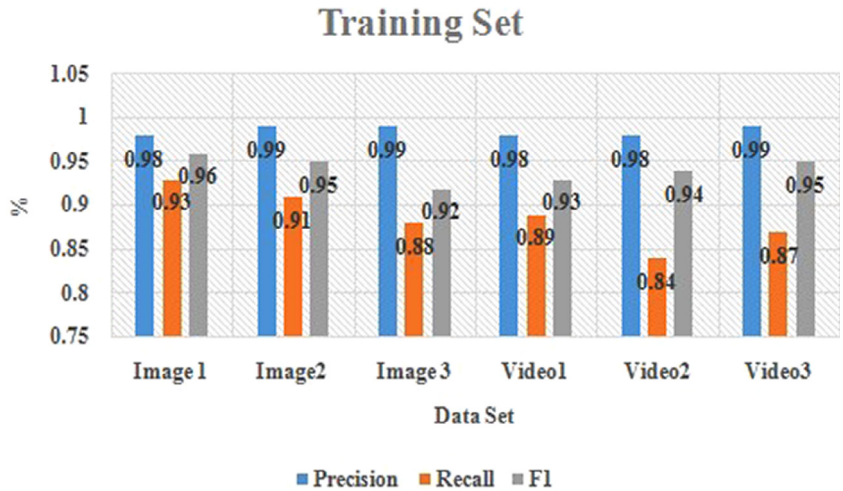

Fig. 9. Graph for Training Dataset in SSD.

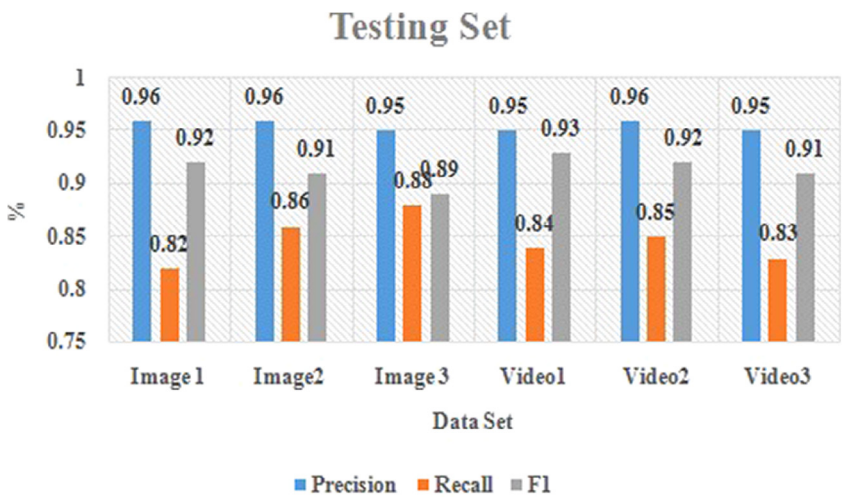

Fig. 10. Graph for training dataset in SSD.

7.2. SSD model gave the following results, which are stated below

The table 3 gives the numerical values obtained for the image and video dataset that are trained for the SSD model. The obtained values are satisfying. The results show that the model is well trained for any incoming inputs.

The training dataset observations for the SSD model have been stated in the graph Fig. 9. The results obtained are more accurate as compared to the YOLO v3 model. The average precision is 98, the average recall obtained of 0.87, and the average F1 measure is 0.94.

Table 4 shows results obtained for the testing dataset used for the SSD implementation have achieved higher results as compared to the YOLO v3 model. The image ad the video datasets were used here for the evaluation. The average value obtained is 95, 84, and 90 for precision, recall, and F1 measure respectively.

The graphical representation shows in Fig. 10 the numerical values obtained for all three parameters, i.e. precision, recall, and F1 measure. The values obtained on an average is far better in SSD as compared with the YOLO v3 model

8. Conclusion

The object detection field has been in trend for various applications. At the same time, we also need the number count of the objects for a particular application. In this research, a comparison of the two existing models for object detection as well as the object counting of humans is discussed. The YOLO v3 and SSD model has been analyzed. Image datasets are used for object detection and videos for object counting. The results have been carried out on three parameters, precision, recall, and F1. The obtained average values for SSD are 95, 84, 90 for precision, recall, and F1 measure, respectively. That is give better than the obtained values for the YOLO v3 model. For object detection as well as object counting the SSD model can be used. The results have shown that the precision, recall, and F1 measure achieved for SSD is higher than YOLO v3.

CRediT authorship contribution statement

Pooja Gupta Role: Conceptualization, Data Curation, Formal analysis, Writing -original draft, review & editing. **Varsha Sharma Role:** Supervision. **Sunita Varma Role:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Leng, Y. Liu, An enhanced SSD with feature fusion and visual reasoning for object detection, Neural Comput. Appl. 31 (10) (2019) 6549–6558, https://doi.org/10.1007/s00521-018-3486-1.

[2] L. Fang, X. Zhao, S. Zhang, Small-objectness sensitive detection based on shifted single shot detector, Multimed. Tools Appl. 78 (10) (2019) 13227–13245, https://doi.org/10.1007/s11042-018-6227-7.

[3] W. Fang, L. Wang, P. Ren, Tinier-YOLO: A Real-Time Object Detection Method for Constrained Environments, IEEE Access 8 (2020) 1935–1944, https://doi.org/10.1109/Access.628763910.1109/ACCESS.2019.2961959.

[4] Y. Wang, Y. Zou, W. Wang, Manifold-Based Visual Object Counting, IEEE Trans. Image Process. 27 (7) (2018) 3248–3263, https://doi.org/10.1109/TIP.2018.2799328.

[5] T. Stahl, S.L. Pintea, J.C. Van Gemert, Divide and Count: Generic Object Counting by Image Divisions, IEEE Trans. Image Process. 28 (2) (2019) 1035–1044, https://doi.org/10.1109/TIP.2018.2875353.

[6] Z. Dai, H. Song, X. Wang, Y. Fang, X.u. Yun, Z. Zhang, H. Li, Video-based vehicle counting framework, IEEE Access 7 (2019) 64460–64470, https://doi.org/10.1109/Access.628763910.1109/ACCESS.2019.2914254.

[7] R. Feng, C. Fan, Z. Li, X. Chen, Mixed Road User Trajectory Extraction from Moving Aerial Videos Based on Convolution Neural Network Detection, IEEE Access 8 (2020) 43508–43519, https://doi.org/10.1109/Access.628763910.1109/ACCESS.2020.2976890.

[8] D.T. Nguyen, T.N. Nguyen, H. Kim, H.-J. Lee, "A high-throughput and power-efficient FPGA implementation of YOLO CNN for object detection", *IEEE Trans*, Very Large Scale Integr. Syst. 27 (8) (2019) 1861–1873, https://doi.org/10.1109/TVLSI.9210.1109/TVLSI.2019.2905242.

[9] S. Sun, N. Akhtar, H. Song, C. Zhang, J. Li, A. Mian, Benchmark Data and Method for Real-Time People Counting in Cluttered Scenes Using Depth Sensors, IEEE Trans. Intell. Transp. Syst. 20 (10) (2019) 3599–3612, https://doi.org/10.1109/TITS.697910.1109/TITS.2019.2911128.

[10] A. Rastogi, B.S. Ryuh, Teat detection algorithm: YOLO vs Haar-cascade, J. Mech. Sci. Technol. 33 (4) (2019) 1869–1874, https://doi.org/10.1007/s12206-019-0339-5.

[11] H. Zhao, Z. Li, L. Fang, T. Zhang, A Balanced Feature Fusion SSD for Object Detection, Neural Process. Lett. 51 (3) (2020) 2789–2806, https://doi.org/10.1007/s11063-020-10228-5.

[12] J. Fu, C. Zhao, Y.e. Xia, W. Liu, Vehicle and wheel detection: a novel SSD-based approach and associated large-scale benchmark dataset, Multimed. Tools Appl. 79 (17-18) (2020) 12615–12634, https://doi.org/10.1007/s11042-019-08523-y.

[13] J. Wang, N. Wang, L. Li, Z. Ren, Real-time behavior detection and judgment of egg breeders, based on YOLO v3, Neural Comput. Appl. 32 (10) (2020) 5471–5481, https://doi.org/10.1007/s00521-019-04645-4.

[14] Manjari, K., Verma, M. and Singal, G., 2019, November. CREATION: Computational ConstRained Travel Aid for Object Detection in Outdoor eNvironment. In 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (pp. 247-254). IEEE.

[15] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y. and Berg, A.C., 2016, October. Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

Further Reading

[1] W. Chen, H. Huang, S. Peng, C. Zhou, C. Zhang, YOLO-face: a real-time face detector, Vis. Comput. (2020), https://doi.org/10.1007/s00371-020-01831-7.

[2] A. Arora, A. Grover, R. Chugh, and S. S. Reka, "Real-Time Multi-Object Detection for Blind Using Single Shot Multibox Detector," Wirel. Pers. Commun., no. 0123456789, 2019, DOI: 10.1007/s11277-019-06294-1.

[3] A.A. Cabrera-Ponce, L.O. Rojas-Perez, J.A. Carrasco-Ochoa, J.F. Martinez-Trinidad, J. Martinez-Carranza, Gate Detection for Micro Aerial Vehicles using a Single Shot Detector, IEEE Lat. Am. Trans. 17 (12) (2019) 2045–2052, https://doi.org/10.1109/TLA.990710.1109/TLA.2019.9011550.