Effective Deep Learning Based Architecture for Pedestrian Detection from Digital Images

Khushaboo Gill, Veenu Mangat

Abstract. ***This paper is to present an efficient and fast deep learning algorithm based on neural networks for object detection and pedestrian detection. The technique, called MobileNet Single Shot Detector, is an extension to Convolution Neural Networks. This technique is based on depth-wise distinguishable convolutions in order to build a lightweighted deep convolution network. A single filter is applied to each input and outputs are combined by using pointwise convolution. Single Shot Multibox Detector is a feed forward convolution network that is combined with MobileNets to give efficient and accurate results. MobileNets combined with SSD and Multibox Technique makes it much faster than SSD alone can work. The accuracy for this technique is calculated over colored (RGB images) and also on infrared images and its results are compared with the results of shallow machine learning based feature extraction plus classification technique viz. HOG plus SVM technique. The comparison of performance between proposed deep learning and shallow learning techniques has been conducted over benchmark dataset and validation testing over own dataset in order measure efficiency of both algorithms and find an effective algorithm that can work with speed and accurately to be applied for object detection in real world pedestrian detection application.***

Keywords. Convolution network, Deep Learning, Histogram of Oriented Gradients, Object Detection, Pedestrian Detection, Multibox Detector.

I. INTRODUCTION

Object detection is widely used application of image processing that helps in identifying specifications of object in an image. Recent trends in object detection is based on deep learning algorithms included in machine learning which includes algorithms inspired by artificial neural network. Convolution Neural Network (CNN) is a popular technique in deep learning for object detection. Many techniques exist that are based on CNN such as RCNN, Fast RCNN, and Mask RCNN. MobileNets SSD is a technique that is based on convolution neural network in deep learning and uses depth wise separable convolution. This technique is really fast and more accurate than RCNN and YOLO technique (You Only Look Once) [34].

Important application of object detection is pedestrian detection which is also a progressing research field. Major applications of pedestrian detection includes, security systems, traffic management, CCTV footages in bad weather conditions, automated driving vehicles and robotics [36]. Pedestrian detection system breaks down an image into small sections that are processed by a classifier which detects the presence or absence of a pedestrian/s in a given image. This paper presents a deep learning algorithm combining MobileNets and Single Shot Detector and its comparison with combined shallow learning technique using feature extraction focused on Support Vector Machine (SVM) and Histogram of Oriented Gradients (HOG) technique [1]. HOG is based on occurrences of gradient orientation of cells in an image and SVM is based on supervised learning techniques that can be used in object detection. Section 2 contains a survey based on relevant literature in the area of object detection. Section 3 describes HOG plus SVM technique. Section 4 elaborates the concept of MobileNets, depthwise separable convolutions, network structure and training and also describes Single Shot multibox Detector (SSD). Section 5 provides experimental settings, results and comparative analysis of Mobilenet SSD technique versus HOG plus SVM technique, as applied to the benchmark dataset and own dataset of images (RGB and thermal). Conclusion and future work of the paper is discussed in Section 6.

II. PRIOR WORK

Many techniques have been used for object detection in literature. The HOG technique [1] counts instance of gradient direction. It works on confined area of the image which is similar to SIFT feature descriptors, and shape contexts [2]. Another method related to this is the CHOG-Filter (Circular HOG)[3] which densely processes HOG. This filter can be prepared by a discriminative way to identify discretionary molded structures. Since Fourier HOG are identified with SHOG [4] which can be utilized for 3D protest discovery in volumetric pictures, this system is called the Circular HOG-Filter. The pyramid of histogram of orientation gradients (PHOG) features [5] are utilized to represent local shape descriptor. This filter was proposed by Bosch et al. [6] and it is utilized as a part of object classification disintegrating an image into sequence of increasingly fine sub-areas named as cells at several pyramid levels.

Centre-symmetric local binary pattern (CS-LBP) technique has been utilized in case of separating features for every pixel of the area [7]. In this method, to create stronger binary arrangements, center symmetric sets of pixels are recognized. Scale Invariant Feature Transformation (SIFT) [8] is a method which is used to categorize and describe bounded features of an image. SIFT features [9] of the objects are firstly extricated from an arrangement of related images to store it in database.

Revised Manuscript Received on February 15, 2020.

* Correspondence Author

***Ms. Khushaboo Gill,** M.E. (I.T.), U.I.E.T., Panjab University, Chandigarh, India. Email: khushaboogill@gmail.com

Dr. Veenu Mangat, Associate Professor (I.T.), U.I.E.T., Panjab University, Chandigarh, India. Email: veenumangat@yahoo.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Objects are identified from other images by separately contrasting elements from the images in this database and discovering hopeful coordinating features in light of Euclidean separation based on component vectors. Another feature extraction method which is inspired by this is Speeded up Robust Features (SURF). SURF method can be utilized for object recognition, images registration, grouping or 3D recreation [10]. Its component descriptor depends on the entirety of the Haar wavelet reaction around the region of interest.

Rotation Invariant Fast Feature (RIFF) descriptor is considered as pipeline [11] which leads to upto-date image recognition in an extensive informational collection. This algorithm is quicker than SURF. This type of tracker is quick and precise. By this method, current computational bottleneck for extensive scale picture handling can be reduced.

Another technique, Wavelet Transformation, introduced in [12] is utilized for strong element extraction from images or time arrangement information for its favorable position of local analysis in the spatial and recurrence areas. The 2-D double tree complex wavelet changes (2-D DT CWT) is executed by utilizing four 2-D discrete wavelet changes in parallel with various filter banks in lines and segments independently. Twelve wavelets are procured by taking the sum and difference of each pair of sub groups. The DD-DT CWT is finished by iteratively changing the low pass sub groups with the decided levels.

Deep learning is a widely used technique nowadays for object detection. Recurrent CNN (RCNN) approach [13] is a deep learning technique in object detection used to obtain a specific amount of possible object areas [14] and calculate convolutional networks [15] on each region. Faster R-CNN [16] exceeding RCNN with a Region Proposal Network. It is simplest and popular that is used for updates [17]. A deep learning based convolutional neural network is presented in [18] that is used for pedestrian detection. It includes different approaches on both learned and handcrafted features at specified computational complexity. L. Cai and J. Zhu described HOG features with deep learning module for pedestrian gender recognition in [19]. A multi modal structure for object detection in RGB-D format in deep feature learning is discussed in [20], which depicts process associated element portrayal. A salient objects detection model is introduced in [21] that is based on image simplification, feature correction and residual network in order to solve complex environmental inference problems. A survey is given in [22] on pedestrian detection and tracking system that show how modern approaches based on deep learning and CNN work and compares results of optimal ANNs topologies with SVM. In recent studies, an automatic traffic density estimation technique [39] uses Mobilenet SSD for car counting and performed quantitative analysis between Mobilenet SSD and SSD. Another recent technique that is based on convolutional neural network for field object detection [40] used for disaster response and recovery.

By analysing all these techniques, we observed that deep learning features are better in object detection and mostly techniques are based on convolutional neural network. In 2016, W. Liu and D. Anguelov [23] presented a method for that is based on deep learning neural network that is used in case of object detection which is easy to train and consolidate with the systems that needs component detection. They calculated experimental results on various datasets and over 4952 images. And in 2017, A. G. Howard and M. Zhu presented a small, very fast and efficient class of neural networks called MobileNets [24] which is based on a sleek structure using intensity distiguishable convolutions [25]. MobileNets is very effective in object detection and large scale geo-localization. COCO dataset has been used in that paper for preliminary outputs.

Thus, in following paper we have presented MobileNets SSD technique combining MobileNets and Single Shot Multibox Detector to build an efficient deep learning based model for object detection and particularly, for pedestrian detection and depicted its comparison with HOG plus SVM technique.

III. HOG WITH SVM

In HOG technique, pixels of an image are grouped into small cells. It is a very popular algorithm in object detection and recognition. It is carried out by breaking the image aperture into blocks and each block is composed of cells. It counts occurrences of gradient orientation for each cell [1]. Each feature in feature pool can be figured out at any position by the help of an array of integral images. We treat each bin of histogram as feature and it is utilized as a fundamental building component of the cascade classifier. Figure 1 shows working steps for HOG plus SVM technique.

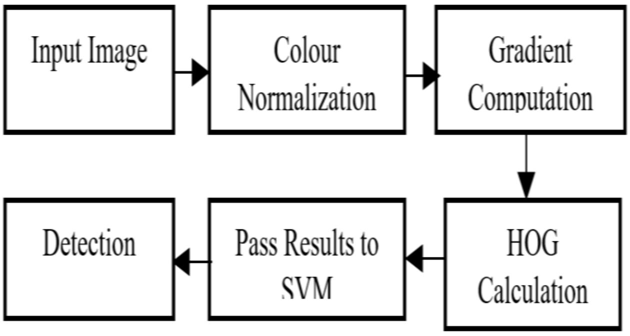


Fig. 1. Working steps for HOG plus SVM technique

The gradient calculation is done in two stages: the initial step of gradient calculation is the calculation of centered mask. This is done to smooth the shading or intensity data on the image. Next step of gradient calculation is to discover the angle and magnitude for every pixel in a cell. Features are separated from every cell, and cells are connected to each other to build a block descriptor. The last descriptor is acquired by the link of the considerable number of features included in the window.

The magnitude and direction of gradient is computed as:

$$mg = \sqrt{mg_x^2 + mg_y^2} \qquad (1)$$

$$\Theta = \arctan \frac{mg_y}{mg_x} \qquad (2)$$

Support vector machines are supervised learning methods that are used for classification, outlier detection and regression. They are useful in multidimensional spaces and are memory efficient and versatile. Linear Support Vector Machines trained on HOG features [26] helps in detecting objects. Downside of HOG method is, it generates numerous element arrangements and this is computationally expensive.

IV. PROPOSED METHODOLGY

A. MobileNet Technology

MobileNet technology is the latest in field of convolution neural network. These are small, very fast and very accurate neural networks. They are based on streamlined architecture that is useful in building light weight deep neural networks using depth wise separable convolutions [25]. MobileNets focuses on building a small network for optimising on latency.

Depthwise Separable Convolution

Depthwise Seperable Convolution is a type of factorised convolution [27] which factorizes an original convolution into separable convolution. A single filter is applied to every input signal in this convolution for MobileNets. The outputs are then combined by applying 1*1 convolution which is called as point wise convolution. In MobileNets, this is initial step to utilize distinguishable convolution that divides the interface into output signals and kernel size.

In a standard convolution layer, input is taken as feature map I i.e. (C_I * C_I * P) which produces output feature map O i.e. (C_O * C_O * Q). Here, C_I is spatial width along with the height of a square input component feature, P shows number of input channels, C_O shows spatial width along with the height of a square output component feature, Q shows number of output channels. Then the output standard layer is parameterized by convolution kernel K which has size equivalent to C_K * C_K * P * Q. Here, D_K is the spatial coordinate of the kernel which is assumed to be a square. The output feature [28] is computed as:

$$O_{k,l,n} = \sum_{i,j,m} K_{i,j,m,n} \cdot I_{k+i-1,l+j-1,m} \qquad (3)$$

Basic convolutions have the computational cost [29] of:

$$C_K \, . \, C_K \, .P \, . \, Q \, . \, C_I \, . \, C_I \qquad (4)$$

The basic convolution procedure has the impact of separating components in view of the convolutional pieces and consolidating components so as to deliver another portrayal. The shifting and mixing steps may be divided in two stages by means of the divided convolutions called depthwise seperable convolutions for significant downfall in calculational cost.

Intensity Distinguishable convolution are comprised of two layers, those are Depth-wise convolutions and point-wise convolutions. MobileNets utilises both batch norm and ReLU nonlinearities for the two layers. Point-wise convolution is often used to make a direct mix of the yield of the depth-wise layer. Figure 2 shows how a standard convolution 2(a) can be reconstructed into a depth-wise convolution 2(b) and a point-wise convolution 2(c). For one feature on each input channel, depth-wise convolution [30] can be shown by the equation:

$$\hat{O}_{k,l,n} = \sum_{i,j} \hat{K}_{i,j,m} \cdot I_{k+i-1,l+j-1,m} \qquad (5)$$

here, $\hat{K}$ is the depth-wise convolutional kernel having size C_K . C_K . P where the m^th feature in $\hat{K}$ is enforced to the p^th layer in I to induce the p^th layer of the filtered resultant feature map $\hat{O}$.

Depth-wise seperable convolution was initially presented in [25]. It is the mutation of depth-wise convolution and point-wise convolution. Its computational cost [31] can be formulated as:

$$C_K \, . \, C_K \, .P \, . \, C_I \, . \, C_I + P. \, Q \, . \, C_I \, . \, C_I \qquad (6)$$

MobileNet utilises 3X3 intensity distinguishable convolutions which comprises of minimal calculation time when compared to the standard convolutions.

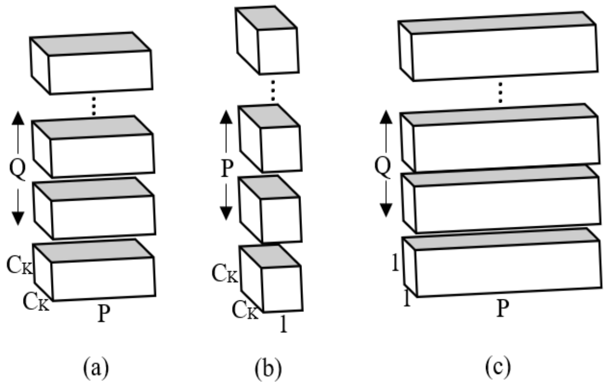


Fig. 2. Standard Convolution Filters shown in (a) are reconstructed by two layers consisting depthwise convolution filters shown in (b) and pointwise convolution (1*1 Convolution filters) shown in (c) in context of depthwise separable filter.

Network Structure and Training

The MobileNet structure is based on intensity distinguishable convolutions with the exception of the main layer that it is a full convolution. All the layers are trailed by batchnorm [32] and ReLU (Rectified Linear Unit) nonlinearity apart from the last, completely associated, layer that has no nonlinearity and inputs into a softmax layer for arrangement.

Figure 3 describes a layer from general convolutions, which has batchnorm and ReLU nonlinearity to the factorised layer along with depth-wise convolution, point-wise convolution, also highlighting batch norm and ReLU afterwards of each convolutional layer. Downsizing is taken care of with stridden convolution of the depth-wise convolutions with the primary layer. In the end, normal pooling diminishes the spatial resolution to 1 preceding the completely associated layer. Counting depth-wise and point-wise convolutions as particular layers, the proposed MobileNet has, in total, 28layers.

B. SSD: Single Shot MultiBox Detector

The SSD approach is dependent on a supervised convolutional neural network that gives a rigid size gathering of bounding boxes. It marks for the closeness of object class instances in those boxes, followed by a padding value. The main characteristics are:

Multi-scale feature maps

Convolutional feature layers are added to end of the truncated base system. These types of layers diminish in estimate continuity and permit forecasts of discoveries at numerous scales. The convolutional model for foreseeing identifications is diverse for each component layer. Overfeat [33] and YOLO [34] which work on a solitary scale component attribute are some of the many ideas which provide similar information.

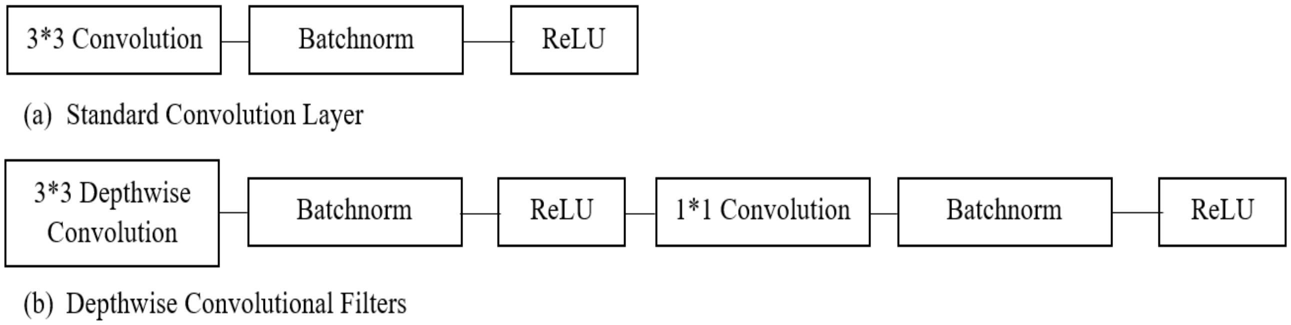

3*3 Convolution
Batchnorm
ReLU
(a) Standard Convolution Layer
3*3 Depthwise Convolution
Batchnorm
ReLU
1*1 Convolution
Batchnorm
ReLU
(b) Depthwise Convolutional Filters

Fig. 3. (a): Standard convolution layer with batchnorm and ReLu (b): Depthwise convolution filters with pointwise layers followed by batchnorm and ReLu

Convolutional predictors for detection

Every additional component tier (or alternatively a current component tier of the ground system) is able to deliver a rigid arrangement of exposure forecasts utilizing an arrangement of convolutional channels. These are shown over the SSD method design in Fig.2. Considering a component layer of size p * q with m carriers, fundamental component to calculate specifications is a 3 * 3 * *m* small kernel which generates mark for classification, and can also generate a frame counterbalance in respect to the default boxes. At every *p**q* areas on which kernel is connected, it delivers a potential object. This is similar to the design of YOLO [34] that uses a middle associated layer rather than convolutional channel based on this progression.

Aspect Ratios and Default Boxes

An arrangement of default bounding boxes having every component attribute is related, to various component attributes on the highest point of system. The default encloses the component attributes on a convolutional way, therefore the location of every box with respect to the related pixel is settled. For every component outline, offsets corresponding to default box configurations on that cell are predicted, and also the values for every class which show the nearness of the class case for each box is predicted.

Data Augmentation

In order to build a structure more vigorous for multiple input object dimensions, every image that is to be trained is inconstantly sampled by using any of these options:

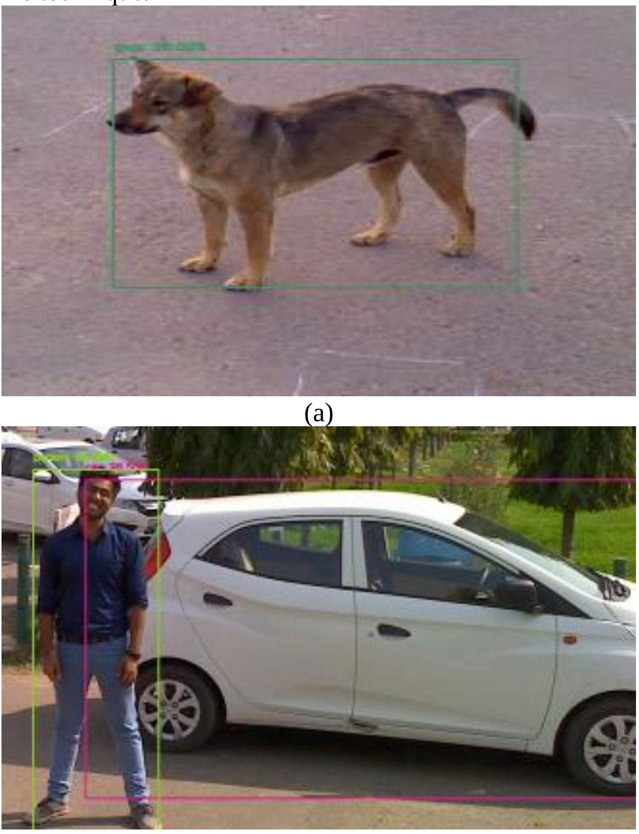
- Adopt complete input fed image.
- Fragment bit in order to make the Jaccard overlap minimum for the objects that is within range of 0.1-0.9.
- Irregular sampling of bits.

The extent of every fragment bit is [0.1, 1] for the input image dimension, and the perspective proportion is within half or two. The covered piece is kept of basic truth box in case if its focal point is in the examined fix. After the previously mentioned testing step, each examined fix is resized to settled size and is on a level plane changed with possibility of half [35].

V. EXPERIMENTAL RESULTS AND ANALYSIS OF COMBINED MOBILENETS AND SSD

When MobileNets and SSD are taken along with each other, they can be utilized for very fast, efficient and real time object detection. Then we can use deep learning methodology to process a supervised object detection method. That will make us to process input images from the model and to attain the resultant bounding box (x,y) coordinates for every object in image. At last the outputs of obtained by MobileNet SSD MultiBox Detector will be analyzed.

We comprehensively evaluate the method on benchmark as well as own data set. By default, a threshold of 0.2 is used for determining accuracy (mAP) in these datasets. It contains test images (206 images) with object classes as person, car, bicycle, chair, dog and bottle. By running the algorithm on dataset, the accuracy has been shown in the various images. Accuracy is the calculated confidence of objects in the observed image. Figure 4 shows results of object detection of this technique.


(a)

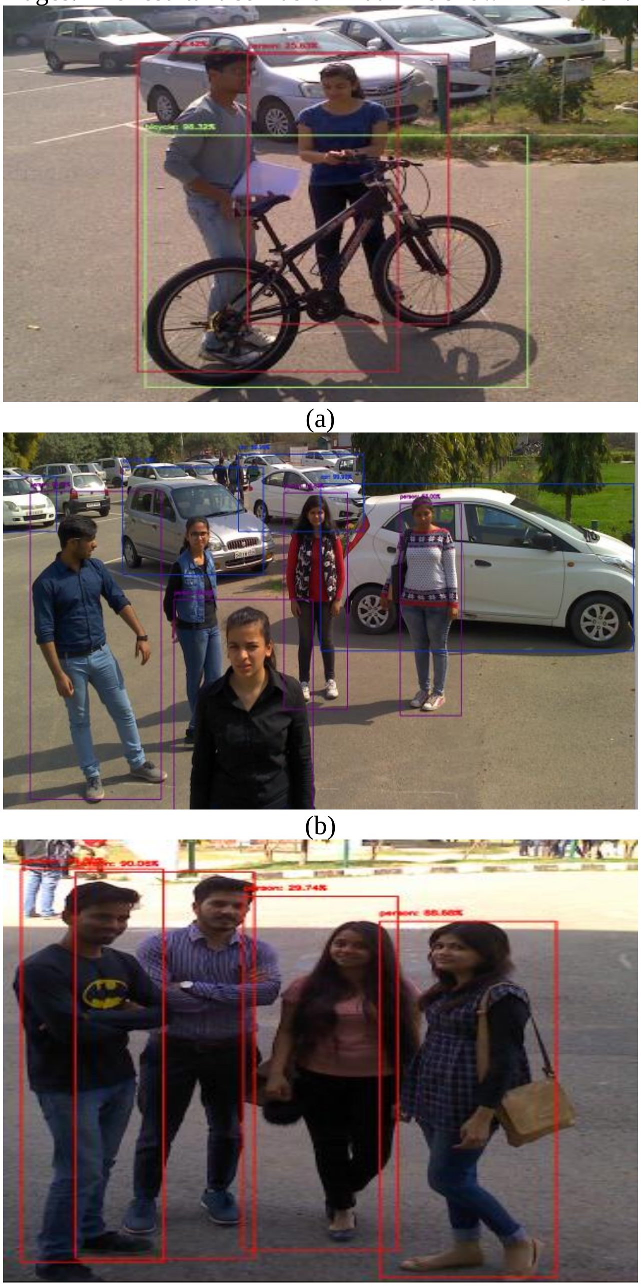
(b)

Fig. 4. (a) A dog is detected (b) a person and a car are detected

A. Pedestrian Detection on own dataset

We have discussed about efficiency of Mobilenets SSD to find objects in this paper and now we will see its results as compared to HOG plus SVM technique, which is a popular algorithm for pedestrian detection. We have used 290 images from our dataset to train and test the model. And various datasets have been taken.

When Mobilenets SSD is applied to an image, it shows detected persons in bounding boxes. Bounding box boundary is coloured according to detected accuracy of objects, along with it, it shows the calculated confidence of the persons detected in the image. Figure 5 shows the output images when this technique is applied to personal dataset on RGB images. The resultant confusion matrix is shown in Table I.



(a)

(b)

(c)

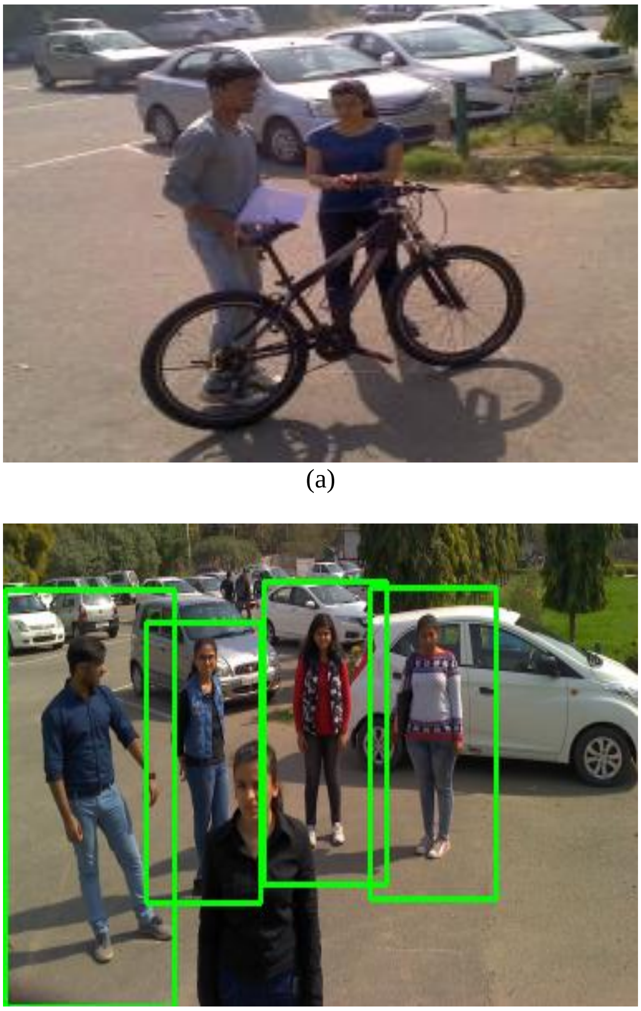
Fig. 5. (a) Two persons are detected (calculated confidence is as follows: bicycle: 98.32%, person: 76.42% and person: 25.63%) (b) Four persons are detected (calculated confidence is as follows: person:99.53%, person:98.79%, person:63.63%, person:63.00%)

(c) Four persons are detected (calculated confidence is as follows: person:49.80%, person:90.06%, person:29.74%, person:88.68%)

Table I Confusion Matrix for MobileNets SSD

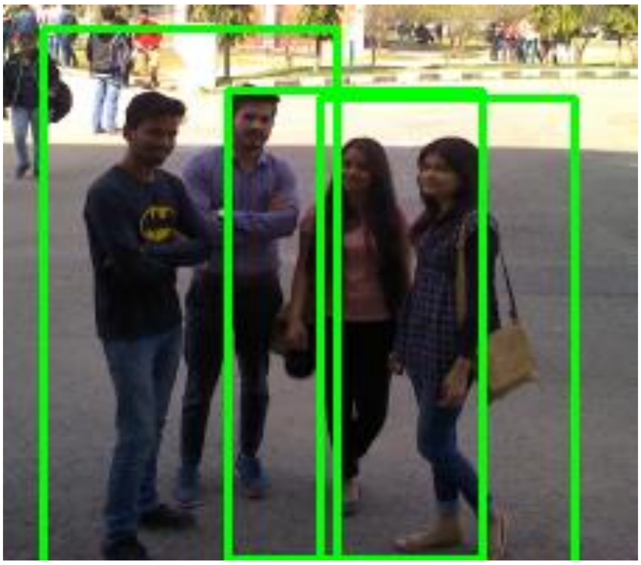
108 (TP)	13 (TN)
23 (FN)	0 (FP)

Now, when combined HOG with SVM is applied to an image, it shows detected objects in bounding boxes. Bounding box boundary is coloured green on detected objects. Figure 6 shows the output images when this technique is applied to own dataset on RGB images. The resultant confusion matrix is shown in Table II.



(a)

(b)



(c)

Fig. 6. (a) No objects are detected. (b) Four persons are detected amongst five persons. (c) Three persons are detected amongst four.

Table II Confusion Matrix for HOG Plus SVM

58 (TP)	13 (TN)
69 (FN)	5 (FP)

Based on the results obtained by implementing the methods given above, the Precision- Recall curve is plotted which is shown in Figure 7.

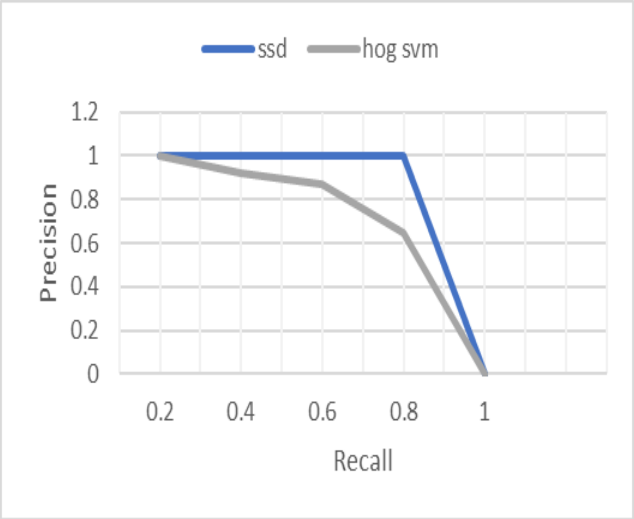
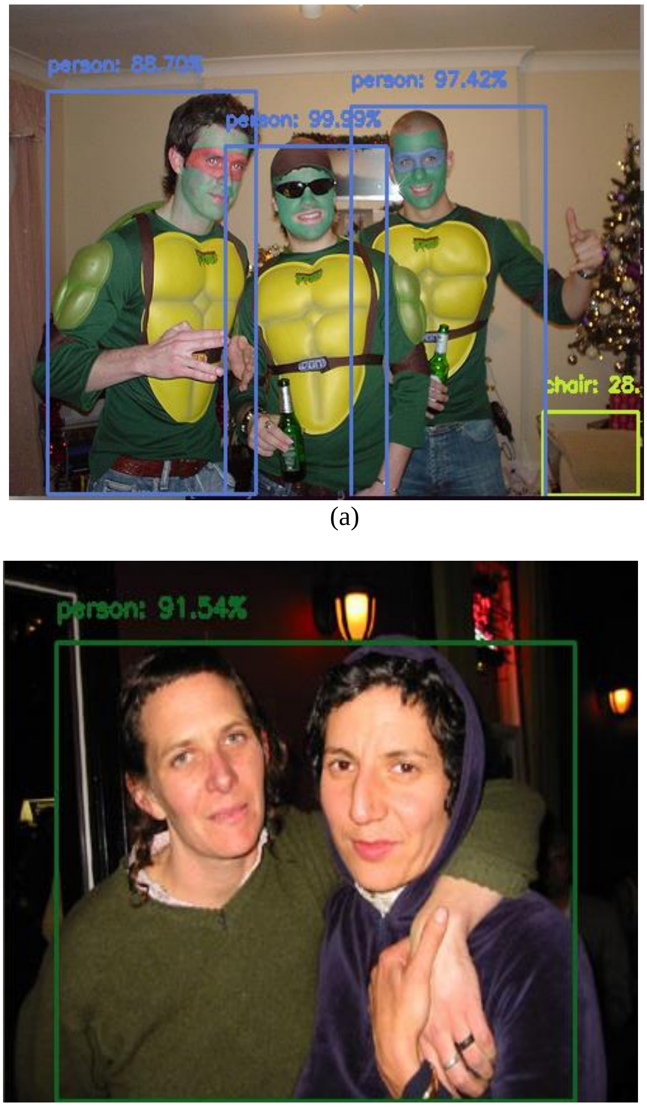


Figure 7. Precision Recall Curve for Mobilenet SSD versus HOG plus SVM

B. Pedestrian Detection on PASCAL VOC 2012 dataset

PASCAL VOC dataset provides standardised images for object detection and also provides a general set of tools for datasets and their annotation that allows to evaluate and compare multiple methods. PASCAL VOC 2012 has 20 classes containing 11,530 images having 6,929 segmentations and 27,450 ROI annotated objects. We have taken pedestrian images from this dataset for our results.

Applying MobileNets SSD to this dataset, the obtained results are depicted in Figure 8 and resultant confusion matrix is shown in Table III.



(a)

(b)

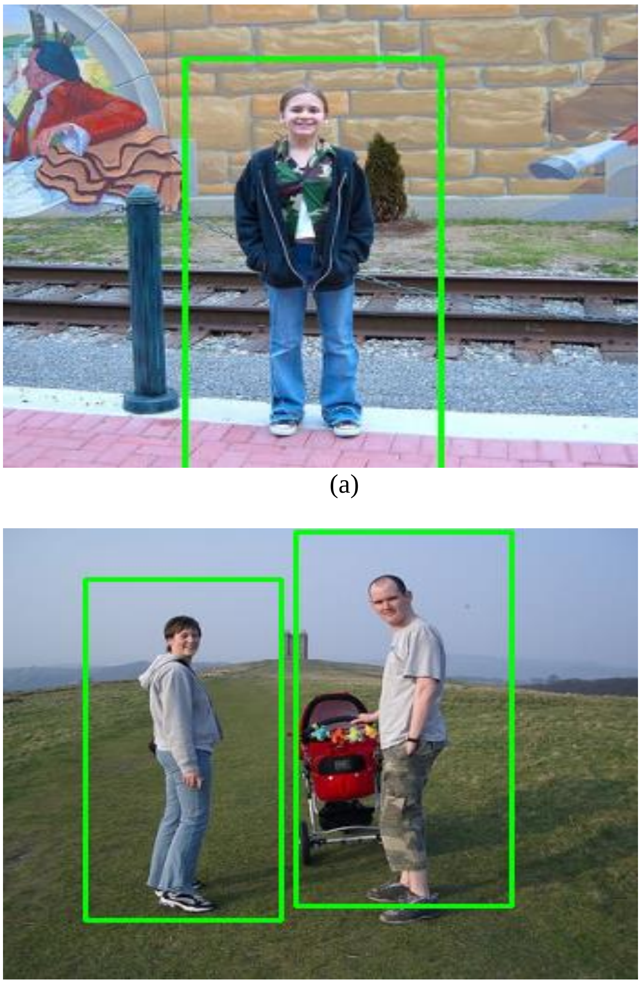
Fig. 8. (a) Three persons are present and are detected with 99.9%, 97.4% and 88.7% accuracy. (b) Two persons present and single is detected with 91.5% accuracy.

Table III depicts the confusion matrix for this category.

Table III Confusion Matrix for MobileNets SSD

126 (TP)	10 (TN)
60 (FN)	4 (FP)

Now by applying HOG plus SVM technique on this dataset, results are as shown in Figure 9. Resultant confusion matrix is depicted in Table IV.



(a)

(b)

Fig. 9. (a) single person is present and detected. (b) two persons present and are detected.

Table IV depicts the confusion matrix for this category.

Table IV Confusion Matrix for HOG Plus SVM

60 (TP)	10 (TN)
110 (FN)	20 (FP)

Based on the results obtained by implementing the methods given above, the Precision-Recall curve is plotted and shown in Figure 10.

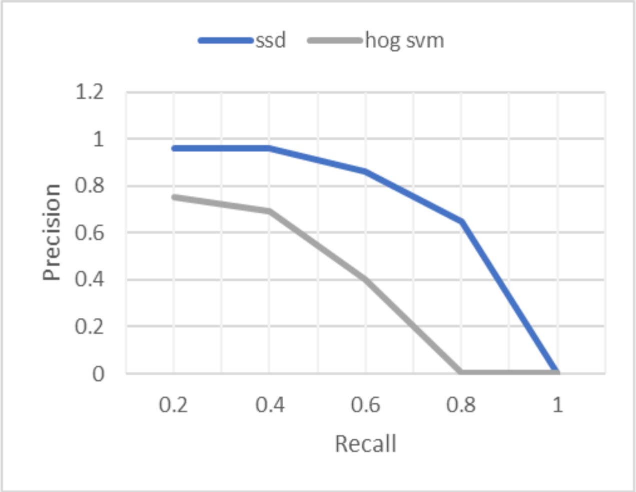
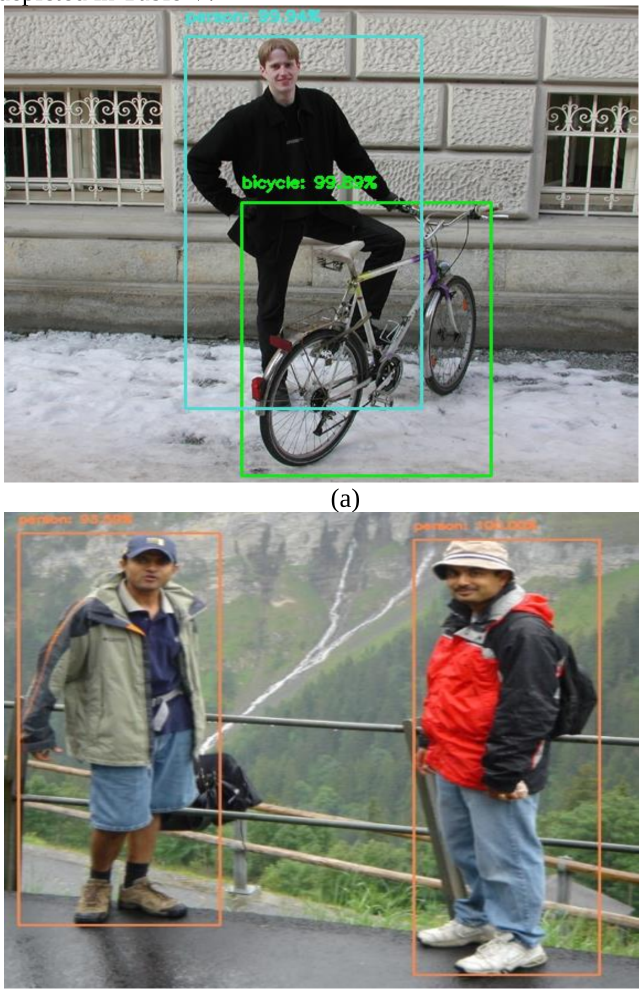


Fig. 10. Precision Recall Curve for Mobilenet SSD and HOG with SVM

C.Pedestrian Detection on INRIA Person dataset

INRIA Person Dataset is described in [37] and was used for research work on people detection in images and video. This dataset is available in two formats: first format contains original images with corresponding annotation file and second contains positive images with negative images normalized n 64×129 pixel format. For our experiment, we have considered 288 images from Test Positive images part. By applying MobileNets SSD technique, the obtained results are as shown in Figure 11. Resultant confusion matrix is depicted in Table V.



(a)

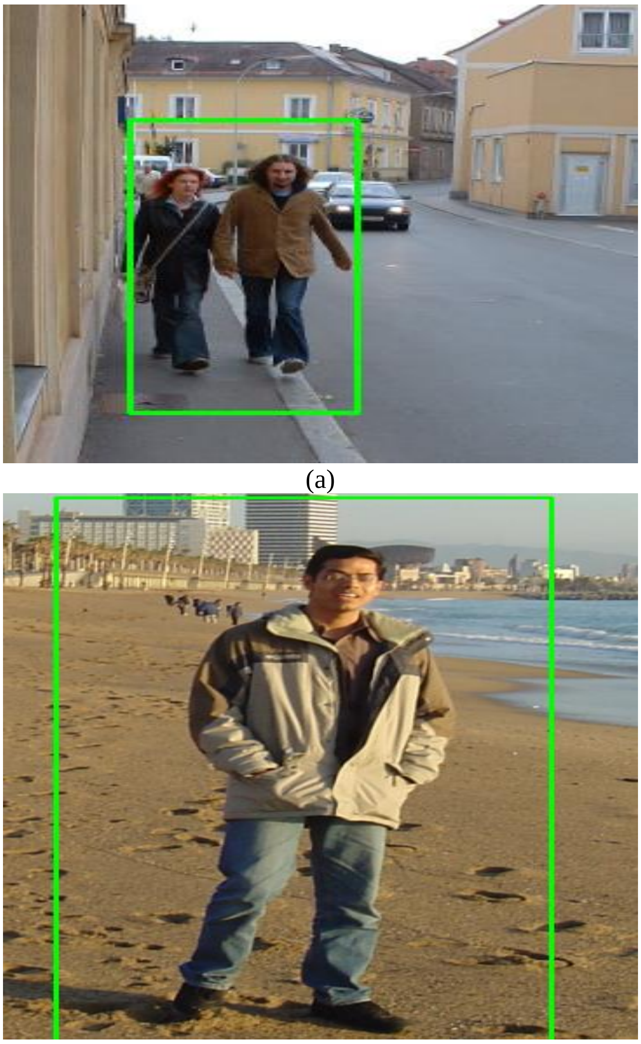
(b)

Fig. 11. (a) a person is detected with 99.9% accuracy. (b) two persons present and are detected with 99.8% and 99.3% accuracy respectively.

Table V Confusion Matrix for MobileNets SSD

172 (TP)	0 (TN)
110 (FN)	6 (FP)

By applying HOG plus SVM technique on this dataset, results obtained are as shown here in Figure 12 and resultant confusion matrix is depicted in Table VI.



(a)

(b)

Fig. 12. (a) two persons are present and are detected. (b) single person is present and detected accurately.

Table VI Confusion Matrix for HOG Plus SVM

208 (TP)	0 (TN)
41 (FN)	39 (FP)

Based on the results obtained by implementing the methods given above, the Precision-Recall curve is plotted and shown in Figure 13.

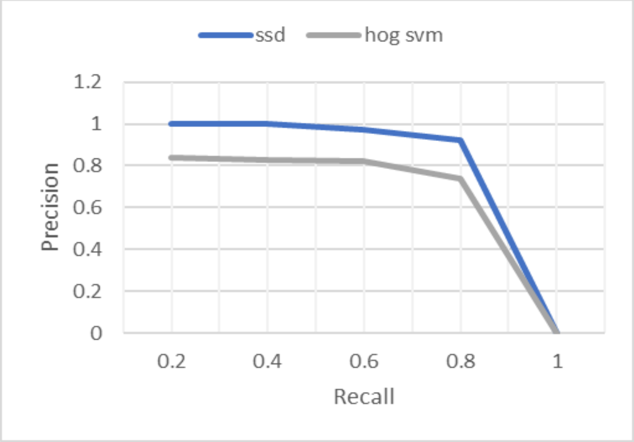
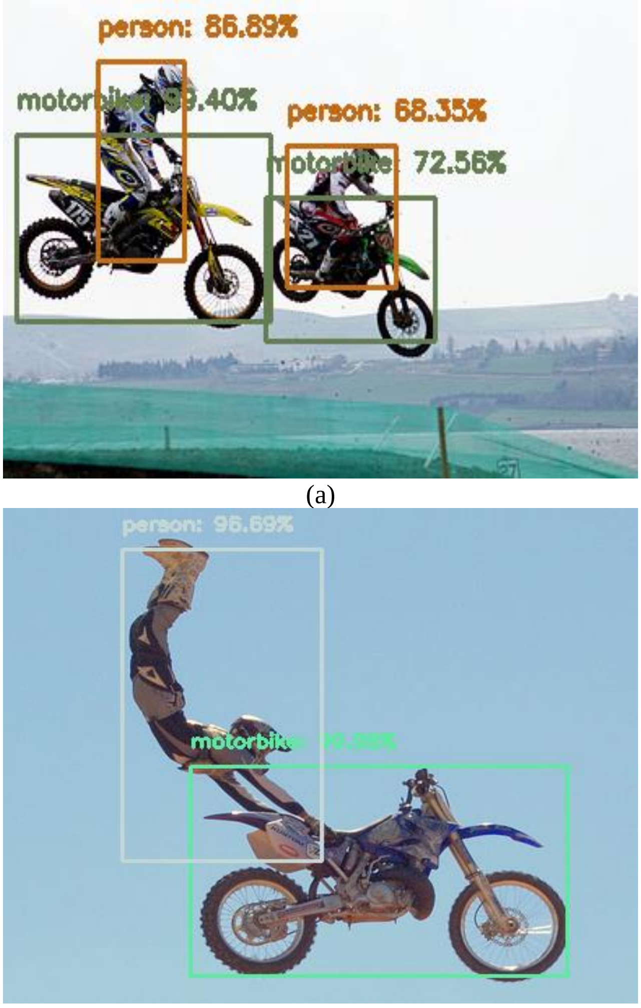


Fig. 13. Precision Recall Curve for Mobilenet SSD and HOG with SVM

D.Pedestrian Detection on Freestyle Motocross dataset

The Freestyle motocross dataset is mentioned in [38]. This dataset contains motorbikes along with person in difficult conditions such as illumination and partial occlusions. It contains two sets: one that contains bikes without in-plane rotations and another with multiple rotations in plane. We considered this dataset for person detection in extreme rotation conditions.

By applying MobileNets SSD technique we got following results in Figure 14. Table VII depicts the confusion matrix for this category.



(a)

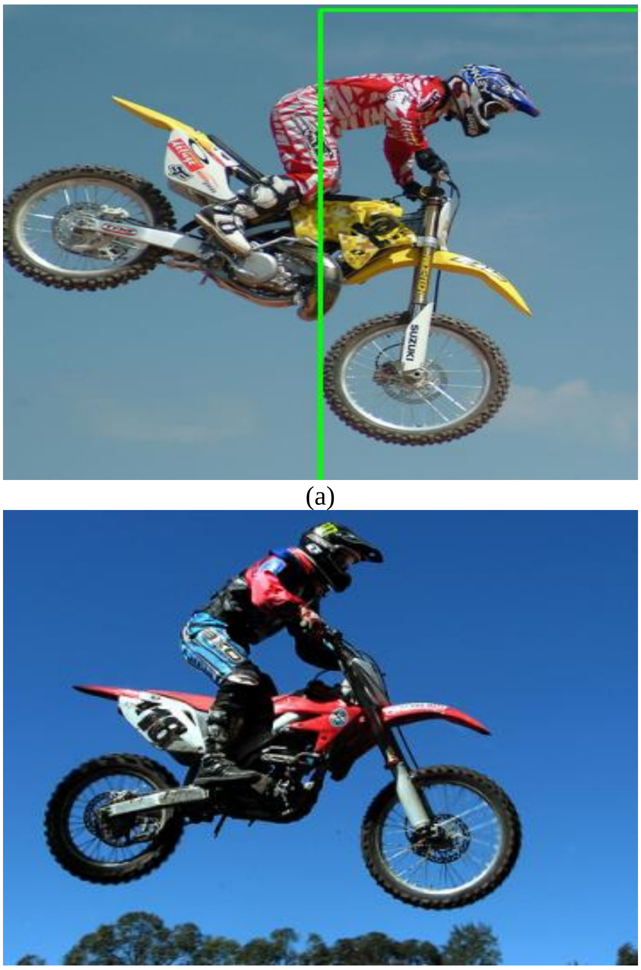
(b)

Fig. 14. (a) two persons present and are detected with 86.8% and 68.3% accuracy respectively. (b) a person is detected with 96.6% accuracy.

Table VII Confusion Matrix for MobileNets SSD

96 (TP)	4 (TN)
69 (FN)	0 (FP)

By applying HOG with SVM technique on this dataset, obtained results are shown in Figure 15. Table VIII depicts the confusion matrix for this category.



(a)

(b)

Fig. 15. (a) a person is present and detected. (b) single person is present but not detected.

Table VIII depicts the confusion matrix for this category.

Table VIII Confusion Matrix for HOG Plus SVM

22 (TP)	4 (TN)
129 (FN)	14 (FP)

Based on the results obtained by implementing the methods given above, the Precision-Recall curve is plotted and shown in Figure 16.

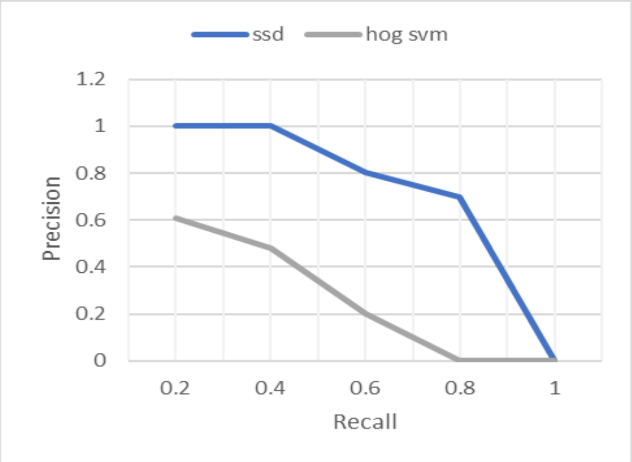
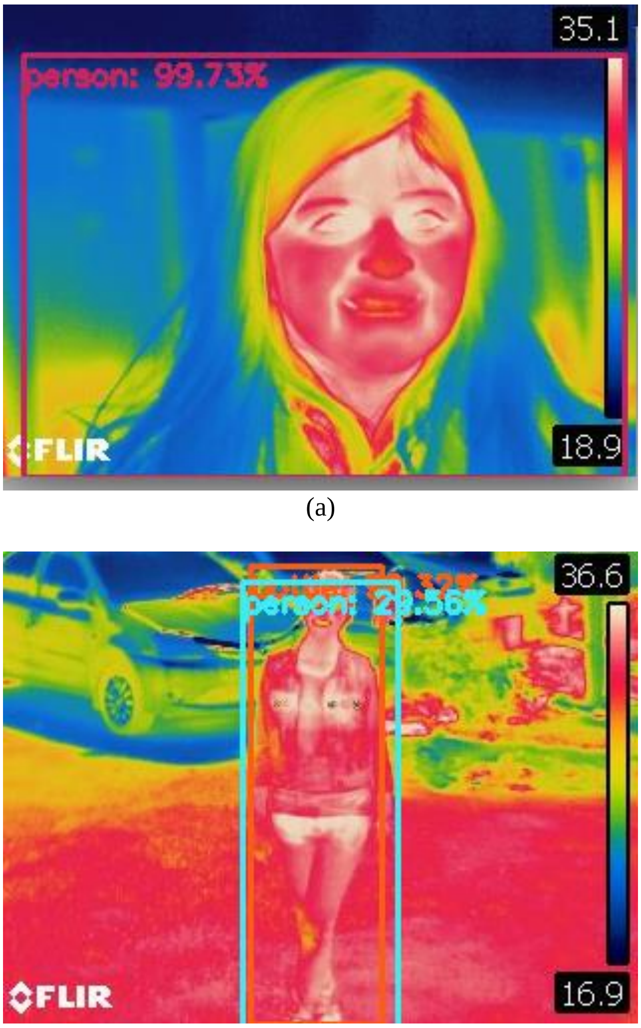


Fig. 16. Precision Recall Curve for Mobilenet SSD and HOG with SVM

E. Comparing Mobilenets SSD versus HOG Plus SVM on Thermal images (Pedestrian Detection)

Now we provide a comparison of these techniques on dataset of infrared images. When Mobilenets SSD is applied to infrared images of own data set, it is able to detect efficiently when the objects are near to camera but not able to detect much efficiently when objects are far and not clear. As shown in Figure 17, this technique is able to detect person efficiently in first image but in second image, the algorithm is reading the person as bottle more accurately than detecting it as a person. Table IX depicts the confusion matrix for this category.



(a)

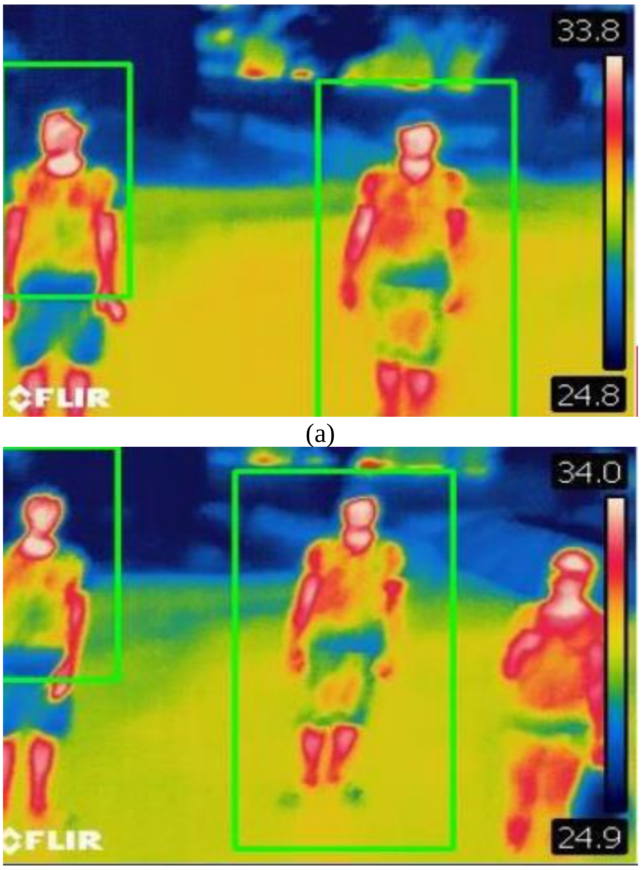
(b)

Fig. 17. (a) person detected with 99.73% accuracy (b) person detected with 29.56% accuracy, also the system takes person as bottle with increased accuracy of 99.22%

Table IX Confusion Matrix for MobileNets SSD

84 (TP)	13 (TN)
43 (FN)	5 (FP)

Now we will apply HOG plus SVM technique on thermal images for pedestrian detection. Figure 18 shows results for this and Table X depicts the confusion matrix for this category. Based on the results calculated upon different throughputs, we can draw Precision-Recall curve in Figure 19.



(a)

(b)

Fig. 18. (a) two persons present and detected (b) three person present and two detected

Table X Confusion Matrix for MobileNets SSD

52 (TP)	7 (TN)
83 (FN)	3 (FP)

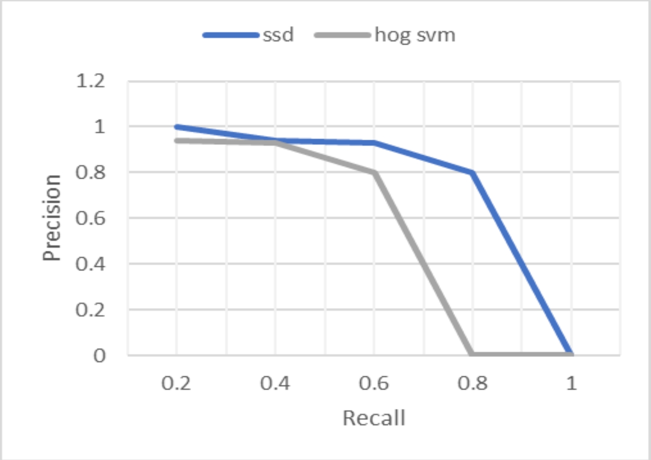


Fig. 19. Precision Recall Curve for Mobilenet SSD and HOG Plus SVM

By comparing the results from confusion matrix and precision recall curves, we can conclude that MobileNets SSD is an accurate and effective technique for pedestrian detection. It works accurately in case of object detection particularly, pedestrian detection, in case of coloured images. Although in case of thermal image, the results of proposed deep learning using MobileNets SSD need to be further improved, yet the results are better than shallow learning based on combined HOG plus SVM technique.

VI. CONCLUSION AND FUTURE WORK

In this research paper, we have discussed about the effectiveness of combined deep learning Mobilenets SSD technique in object detection, that is established on depthwise separable convolution and is an efficient algorithm in deep learning for object detection and pedestrian detection. We applied this technique on multiple benchmark and own datasets and compared its results with HOG plus SVM technique. Experimental results indicate that combined MobileNets SSD technique performs better in terms of accuracy and precision-recall tradeoff as compared to HOG plus SVM technique. In case of thermal images, both methods did not give good results, but deep learning approach is relatively more promising in case of accuracy and speed. In the future, work can be done to acquire large number of images under varying conditions in order to better train the deep learning architecture. Additionally, optimisation of the number of layers and weights of the deep network, can be done to improve the performance.

REFERENCES

1. S.K. Uma, B.J. Srujana, “Feature Extraction for Human Detection using HOG and CS-LBP methods”, International Journal of Computer Applications (0975 – 8887), National Conference Electronics, Signals, Communication and Optimization, Vol. NCESCO 2015, Issue 2, pp. 11-14.,2015.
2. Liu, H Wu, W Su,j. Sun, “Sector-ring HOG for rotation-invariant human detection”, Elsevier journal Signal Processing: Image Communication, vol. 54, pp. 1-10, china, 2017
3. H. Skibbe and M. Reisert, “Circular fourier-hog features for rotation invariant object detection in biomedical images”, presented at the 9th IEEE International Symposium on Biomedical Imaging (ISBI), Barcelona, Spain, May 2-5, 2012, Accession Number: 12864365.
4. M.Reisertand, H.Burkhardt, “Equivariant holomorphic filters for contour denoising and rapid object detection”, IEEE Trans. Image Processing, vol. 17(2), pp. 190– 203, Feb 2008.
5. P. P. Sarangi1, B.S.P. Mishra1 and S. Dehuri, “Pyramid Histogram of Oriented Gradients based Human Ear Identification”, International science press International journal of control theory and applications, vol 10(15), pp. 125-133, April 2017.
6. A. Bosch, A. Zisserman, and X. Munoz, “Representing shape with a spatial pyramid kernel”, in Proceedings of the 6th ACM International Conference on Image and Video Retrieval, July 2007, pp. 401-408.
7. J. Shen, W. Yang, C. Sun, “Real-time human detection based on gentle MILBoost with variable granularity HOG-CSLBP”, Neural computing and applications, vol 23(7-8), pp. 1937-1948, Dec 2013.
8. D. Lowe, “Distinctive image features from scale-invariant”, International Journal of Computer Vision, vol 60 (2), pp. 91–110, Jan 2004.
9. S. Zhong, J. Wang, L.Yan, L. Kang, Z. Cao, “A real-time embedded architecture for SIFT”, Journal of Systems Architecture vol 59(1), pp 16-29, Jan 2013.
10. H. Bay, T. Tuytelaars, L.V. Gool, “SURF: speeded up robust features”, Springer Proceedings of the European Conference on Computer Vision(ECCV), Austria, pp. 404-417, Dec 2006.
11. G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, B. Girod, “Rotation-invariant fast features for large-scale recognition and real-time tracking”, Elsevier journal Signal Processing: Image Communication, vol 28, pp.334–344, Dec 2013.

12. B K Alsberg, A M Woodward,D B Kell, “An introduction to wavelet transforms for chemometricians: A time-frequency approach”, Elsevier journal Chemometrics and Intelligent Laboratory Systems, vol 37(2) , pp. 215-239, June 1997.
13. R. Girshick, J. Donahue, T. Darrell, J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation”, Computer Vision and Pattern Recognition (CVPR), DOI: 10.1109/CVPR.2014.81, 2014.
14. J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, “Selective search for object recognition”, International Journal of Computer Vision, vol 104(2), pp. 154–171, Sept 2013.
15. A.G Howard, “Some improvements on deep convolutional neural network-based image classification”, arXiv:1312.5402 [cs.CV], Dec 2013.
16. S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 39(6), pp. 1137-1149, June 2017.
17. A. Shrivastava, A. Gupta, and R. Girshick, “Training region-based object detectors with online hard example mining”, Computer Vision and Pattern Recognition, arXiv:1604.03540 [cs.CV], April 2016.
18. D. Tome, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, S.Tubaro, “Deep Convolutional Neural Networks for pedestrian detection”, Elsevier journal Signal Processing: Image Communication vol 47, pp. 482–489, Sept 2016.
19. L. Cai, J. Zhu, H. Zeng, J. Chen, C. Cai, K.K. Ma, “HOG-assisted deep feature learning for pedestrian gender recognition”, Elsevier Journal of the Franklin Institute vol 355, pp.1991–2008, March 2018.
20. X. Xu,Y. Li, G. Wu, J. Luo., “Multi-modal deep feature learning for RGB-D object detection”, Elsevier journal Pattern Recognition, vol 72, pp. 300–313, Dec 2017.
21. H. Wanga, L. Dai, Y. Cai, X. Sun, L. Chen, “Salient object detection based on multi-scale contrast”, Elsevier journal Neural Networks, vol 101, pp. 47–56, May 2018.
22. A. Brunetti, D. Buongiorno, G. F. Trotta, V. Bevilacqua, “Computer vision and deep learning techniques for pedestrian detection and tracking: A survey”, Elsevier journal Neurocomputing, submitted for publication, Vol. 300, pp. 17-33, July 2018.
23. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu1, A. C. Berg, “SSD: Single Shot MultiBox Detector”, presented at the European Conference on Computer Vision, Sept 2016, pp. 21-37.
24. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”, Computer Vision and pattern Recognition, arXiv:1704.04861[cs.CV], April 2017.
25. L. Sifre. “Rigid-motion scattering for image classification”. PhD thesis, Ph. D. thesis, 2014.
26. [26] H. Bristow, S. Lucey, “Why do linear SVMs trained on HOG features perform so well?”, CVPR, arXiv:1406.2419[cs.CV], Jun 2014.
27. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, “Imagenet large scale visual recognition challenge”, International Journal of Computer Vision, vol 115(3), pp. 211–252, Dec 2015.
28. [28] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, “Tensorflow: Large-scale machine learning on heterogeneous systems”, Distributed, Parallel, and Cluster Computing, arXiv:1603.04467[cs.DC], Mar 2015.
29. W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen, “Compressing neural networks with the hashing trick”, Proceedings of the 32nd International Conference on Machine Learning, Lille, France, JMLR: W&CP, vol 37, 2015.
30. F. Chollet, “Deep learning with depthwise separable convolutions”, Computer vision and Pattern recognition, arXiv:1610.02357[cs.CV], Oct 2016.
31. S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding”, Computer vision and Pattern recognition, arXiv:1510.00149[cs.CV], Oct 2015.
32. S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift”, Learning, arXiv:1502.03167[cs.LG], Feb 2015.
33. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks”, Computer vision and Pattern recognition, arXiv:1312.6229[cs.CV], Feb 2014.
34. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, “You only look once: Unified, real-time object detection”, Computer vision and Pattern recognition, arXiv:1506.02640 [cs.CV], May 2016.
35. L. Zhang, L. Lin, X. Liang, K. He, “Is Faster R-CNN Doing Well for Pedestrian Detection?”, presented at the European conference on computer vision, Sept 2016, pp. 443-457.
36. C.F. Lin, C.S. Chen, W.J. Hwang, C.Y. Chen, C.H. Hwang, C.L. Chang, “Novel outline features for pedestrian detection system with thermal images”, Pattern Recognition, vol 48(11), pp. 3440-3450, Nov 2015.
37. N. Dalal, B. Triggs, “Histograms of oriented gradients for human detection”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol 1(01), pp. 886-893, July 2005.
38. M. Villamizar, F. Moreno-Noguer, J. Andrade-Cetto, A. Sanfeliu, “Efficient rotation invariant object detection using boosted random ferns”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, June 2010 pp. 1038–1045.
39. D. Biswas, H. Su, C. Wang, A. Stevanovic, “An automatic traffic density estimation using Single Shot Detection (SSD) and MobileNet-SSD”, Elsevier Journal Physics and chemistry of Earth, vol 110, pp 176-184, April 2019.
40. Y. Pi, N. D. Nath, A. H. Behzadan, “Convolutional neural networks for object detection in aerial imagery for disaster response and recovery”, Elsevier Journal Advanced Engineering Informatics, vol 43, 101009, Jan 2020.

AUTHORS PROFILE

Ms. Khushaboo Gill is pursuing her Masters of Engineering in Information Technology at Panjab University, India. Her research interests include image processing, machine learning and object detection.

Dr. Veenu Mangat is an Associate Professor of Engineering in Information Technology at Panjab University. She holds a Ph.D. in Engineering and Technology in Computer Science in the area of data mining. She has successfully guided 20 Masters' dissertations and is supervising 6 Ph.D. students. Her areas of interest include machine learning, cloud computing, pattern recognition, data privacy. She is a member of several technical societies like IEEE, ACM and IAENG.