Addressing the False Positives in Pedestrian Detection

N. J. Karthika and Saravanan Chandran

Abstract Pedestrian detection is a subfield of object detection that is necessary for several applications such as person tracking, intelligent surveillance system, abnormal scene detection, and intelligent cars. We prepared a dataset for addressing the false positives that occur during the person detection process. Some objects have very similar features to those of a person. If a model is trained using a dataset containing only persons, it leads to several false positives since it cannot differentiate such objects from that of a person. Our dataset includes person and person-like objects (PnPLO). Person-like objects that we introduce in our dataset are statues, mannequins, scarecrows, and robots. We used the SSD model to show that, on training a model using our dataset, we can significantly reduce the false positives during detection when compared to models trained on standard person datasets, thereby improving the precision. The dataset consists of 944 training images, 160 validation images, and 235 images for testing, with a total of 1626 person and 1368 nonhuman labelling.

Keywords Pedestrian detection · Nonhuman detection · Deep learning · SSD · Computer vision

1 Introduction

Humans can instantly recognise any object in an image. We can also simultaneously interpret the location of any object, as well as how the objects interact. The human visual system is very fast as well as accurate, helping us to perform even highly complex tasks such as driving a vehicle, with little conscious thought. Computer

N. J. Karthika (✉) · S. Chandran
Computer Science and Engineering, National Institute of Technology, Durgapur, West Bengal, India
e-mail: karthika.nj@gmail.com

S. Chandran
e-mail: cs@ieee.org

P. K. Mallick et al. (eds.), *Electronic Systems and Intelligent Computing*, Lecture Notes in Electrical Engineering 686, https://doi.org/10.1007/978-981-15-7031-5_103

vision is a research area, wherein the researchers try to make the computers work in the same way as to how the human visual system works. The recent trend in this field is the use of deep learning models, which is because various researches [1–5] show that deep learning models have made the computers much faster and accurate in object detection, classification, recognition, and various other computer vision problems. Until the year 2016, state-of-the-art object detection systems with the best accuracy were computationally intensive and too slow to run in real-time (e.g. faster RCNN [1]). Also, the models which ran on real-time were not accurate enough, especially for safety-critical applications. With the advent of SSD model [2], there was a significant improvement in the speed for detection with high accuracy.

Dataset plays a crucial role in problems of object classification, detection, recognition, segmentation, etc. There are many popular datasets widely accepted as benchmarks for object detection problem. ImageNet [6], PASCAL VOC [7], COCO [8], and SUN [9] datasets are few examples. Each of these differs in the type of images, number and type of object labels, and size of datasets.

Pedestrian detection is an object detection problem. It has several real-time applications such as person tracking, robotics, video surveillance, and driverless cars. Research works over the years have been using various approaches for the problem of pedestrian detection, such as part-based detection [10], holistic detection using features like HOG [11], motion-based detection [12], patch-based detection [13], and detection using multiple cameras. As in other object detection problems, deep learning is currently the most used approach in the research related to pedestrian detection [14].

All the research works related to pedestrian detection which use deep learning for training their models make use of person datasets as the benchmark [11, 15, 16]. These datasets consist of only the images of persons in various postures and under different lighting conditions. The popular object detection datasets [6–8] also have person as a class but no object classes to differentiate the person from objects having similar features as persons. Since there are objects such as mannequins and statues that have very similar features as that of a person, a model that is trained with datasets containing only person images will have higher false positives on encountering such objects. Considering this problem, we prepared a dataset (PnPLO) [17] containing persons as well as the objects having features very close to that of a person such as mannequins, statues, scarecrows, and robots. We train SSD model with 300*300 image input size (ssd_300) and show the improvements in precision on testing the newly trained model compared to the model trained on benchmark datasets, namely, COCO [8], INRIA [11], and PASCAL VOC [7]. We can observe considerable improvement in precision with the model trained on our dataset, on testing the models on PnPLO test dataset.

2 Recent Works

Computer vision is a field of study where extensive research is going on, especially with the use of deep learning models. Pedestrian detection or person detection, in general, is one of the topics of eminence in the field of computer vision, because of its wide variety of applications in real-world problems. Video surveillance, driverless cars, and person tracking are some of the applications. Deep learning gained popularity with the advent of AlexNet [3] in the year 2012, followed by many notable research works such as [4, 5, 18].

Most of the recent object detection research works use either faster RCNN [1] or single-shot multibox detector [2] as their backbone network because of their accuracy and speed of SSD. Before the advent of SSD, faster RCNN was widely used because of its excellent accuracy. This model is based on a region proposal network, which is class agnostic. This class agnostic nature of RPN networks leads to high time consumption, as the network needed two rounds of predictions—first, to predict the regions which may contain an object, and then to predict the class of the object present in that region. YOLO [19] considered this disadvantage and proposed to have only one round of prediction by making the region proposals to be class-specific so that the network needs to look at an image only once, thus, saving a great deal of time. As the image passes through the deep convolutional network only once, the model was speedy and could be run real time. Though YOLO worked in real time, it compromised the accuracy by a great deal when compared to the previous state-of-the-art model [1]. Problem with the first version of YOLO was that it could not capture scale variation and failed to detect very small objects. SSD provided a solution for this problem by proposing an auxiliary structure that can perform detections at multiple scales. SSD, therefore, can run in real time with an accuracy comparable to that of faster RCNN.

2.1 SSD Overview

SSD model was the first model that worked in real time with an accuracy as good as the previous state-of-the-art models in object detection. Before SSD, models such as the RCNN series [1, 20, 21] used RPN-based approach, which was time-consuming because of two stages—region proposal, followed by detecting objects in each proposal. The most significant advantage of SSD is its simplicity, with a single network encapsulating all the computations, eliminating the need for a proposal generation as well as the feature resampling stages. Figure 1 shows the architecture of the ssd model.

SSD takes an image and ground truth boxes as input. The model used VGG-16 network [5] as the base network. This network forms the first layers, following which, an auxiliary structure was added to the network to produce detections. For each location in feature maps of different scales, a small set of default boxes of

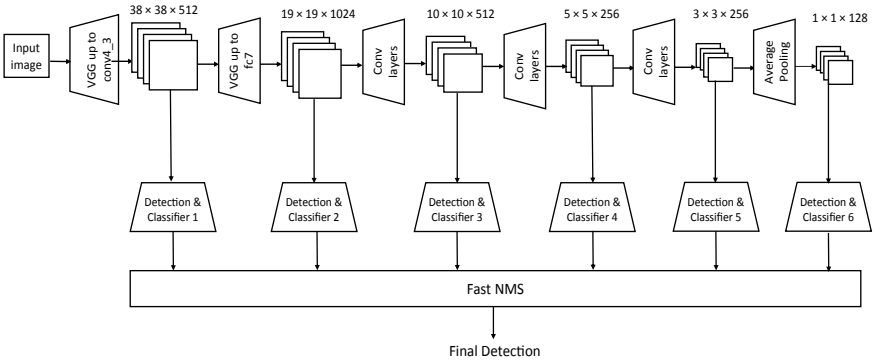

Fig. 1 SSD architecture

varying aspect ratios are evaluated in a convolutional manner. For each default box, confidence scores of all object categories, and the shape offsets are predicted. During training, the default boxes and the ground truth boxes are matched, and the model loss is calculated, which is the weighted sum of confidence loss and the localisation loss. The overall loss is given by :

$$L(x, c, l, g) = \frac{1}{N}(L_{\text{conf}}(x, c) + \alpha L_{\text{loc}}(x, l, g))$$

where L_{conf} and L_{loc} are confidence loss and localisation losses, respectively, and α is the weight term. Confidence loss is the Softmax loss over multiple class confidences (c). Localisation loss is the smooth $L1$ loss [21] between the parameters of ground truth box and the predicted box (l).

2.2 Datasets

Pedestrian detection is a subject of interest in various researches because of its widespread real-life applications. Hence, there are multiple standard datasets available, consisting of person as a class, used for these research works. We have considered three datasets used as benchmarks viz., COCO, INRIA, and PASCAL VOC datasets.

2.2.1 COCO

This dataset contains images of complex everyday scenes of common objects in their natural context. It is a large-scale dataset for object detection. It defines 91 classes, but only 80 classes are used by the data. Segmentations of 11 other classes were not

collected because of problems like too many instances, ambiguity, and difficulty in labelling, too few instances, etc. Compared to the previous datasets such as Imagenet and PASCAL VOC, COCO dataset has more object instances per image and also gives an additional focus on segmenting individual instances of different objects. Person is one among the 80 classes considered in the dataset.

2.2.2 INRIA

The INRIA person dataset has a training set constituting 1128 negative images and 614 positive images, and a testing set with 288 images. First created and used by Dalal and Triggs [11], the static person dataset comprised of people in various positions and orientations, taken in a variety of backgrounds and lighting conditions.

2.2.3 PASCAL VOC

The PASCAL visual object classes (VOC) challenge had been organised annually since the year 2005. The dataset associated with this challenge is publicly available and has been accepted as a benchmark in object detection. It constitutes images as well as annotations in XML format. The dataset consists of 20 classes which include 'person' as one among them.

3 PnPLO Dataset

Data plays a critical role in any deep learning research work, enabling the computers to work in the same way as to how the humans do. This is especially true in the field of object detection, where the number and the type of images used for training a deep learning model play a crucial role while applying the model to real-world problems.

All the datasets used as benchmarks for person detection problem contains only images labelled with person objects. Training with such a dataset leads to several false positives while testing, when the images include many objects having features close to that of a person. If an image contains a statue, then a model that was trained with only person images tends to identify the statue as a person, leading to a false positive. To address this problem of false positives, we prepared a dataset containing persons as well as the objects having features similar to a person—person and person-like objects (PnPLO) dataset [17]. We have labelled the person-like objects as 'nonhuman'.

3.1 Image Acquisition

Person images and their corresponding annotation files used for training are considered from the PASCAL VOC 2012 person training dataset, and images for testing are taken from PASCAL VOC 2007 person test set. The nonhuman images are taken from the Internet. These images are completely random, not taken for a specific purpose or a specific event, or from any particular angle. Because of this randomness, we get an unbiased dataset. Some of the nonhuman images also contain person objects in them.

3.2 Labelling

The dataset consists of a total of 944 images for training, 160 images in the validation set, and 235 for testing. In the training set, there are 1106 person and 960 nonhuman labellings. In the validation set, there are 203 person and 130 nonhuman labellings. The test set consists of 317 and 278 labellings of person and nonhuman, respectively.

We labelled the nonhuman images using the labelImg tool [22], which is a graphical tool for generating image annotations. This tool saves the annotations in the form of XML files in the format of PASCAL VOC dataset. We have labelled the images for two classes, person and nonhuman class. In the XML annotations of PASCAL VOC dataset, we have removed the annotations marked as difficult since such objects will have similar features for both the classes considered and are difficult for even the human eye to differentiate correctly. We have taken 526 person images for training from the PASCAL VOC 2012 dataset, and 125 images from PASCAL VOC 2007 test list for testing. The number of person and nonhuman objects in the dataset are comparable to avoid any over-fitting or under-fitting problems.

4 Experiment

We first tested the ssd_300 model trained on some standard datasets on the test data of our dataset. The ssd_300 model trained on COCO, INRIA, and PASCAL VOC datasets, respectively, are considered. Since the PASCAL VOC dataset includes person as a class, we used the SSD model trained on this dataset as the initial setting for the model to train on our dataset. This leads to a good initialisation for the model instead of any random weight initialisation methods. We trained this model on our train data for 50 epochs. We limited the number of epochs to 50 as further epochs did not give any considerable improvement in the loss. We have trained the model with a learning rate of 10^{-3} for 10 iterations, then with a learning rate of 10^{-4}, we trained the model up to 30 iterations, and for the final iterations, we used a learning rate of 10^{-5} to train the model. We then used the final trained model to evaluate the test data.

Fig. 2 Evaluation using model trained on PASCAL VOC 07+12 dataset: detecting robot, scarecrow, mannequins, and statues as person

Fig. 3 Evaluation using model trained on PnPLO dataset: robot, scarecrow, mannequins, and statues are correctly detected as nonhumans

We noticed a significant improvement in the precision of person detection after training on our dataset when compared to the precisions obtained on training on datasets containing only persons. This improvement is achieved with the help of PnPLO dataset [17] that considered the objects with features resembling those of a person. Figure 2 shows the evaluation on test images of PnPLO dataset using model trained on PASCAL VOC 07+12 dataset. We can see that the model wrongly detects the person-like images as persons. Figure 3 shows evaluation on the same four images using SSD model trained on PnPLO dataset. We can see that the model is able to differentiate person from other person-like objects.

4.1 Evaluation Metric

Average precision (AP) is the evaluation metric used to compare the performances of the SSD model trained on different datasets. Following metrics are involved in the calculation of average precision.

4.1.1 Intersection Over Union (IOU)

IoU is given by the following formula—the area of overlap over the area of the union of the predicted and ground truth bounding boxes.

$$\text{IoU} = \frac{\text{area of intersection}}{\text{area of union}}$$

IoU is used to measure whether the bounding box predicted by the model is true positive (TP), false positive (FP), or false negative (FN). If the IoU > 0.5, we consider the predicted box to be a true positive. The bounding box is considered to be FP either if IoU < 0.5, or if there are duplicate boxes predicted for the same object in an image. The predicted bounding box is an FN if IoU > 0.5, but a wrong prediction.

4.1.2 Precision

Precision is nothing but the measure of how accurate our predictions are.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

4.1.3 Recall

Recall measures how well the model finds all the true positives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Once the above values are obtained, the precision-recall curve (PR curve) is plotted, as shown in Fig. 4. Average precision is calculated by taking the area under the PR curve.

Table 1 shows the improvement we achieved after training the SSD model on PnPLO dataset. We have compared the performance of the model trained on PnPLO dataset with that trained on three standard datasets, namely, PASCAL VOC 07+12 [7], COCO [8], and INRIA [11] person datasets. Average precision is the metric used to compare the performances.

On evaluating the model trained on COCO, INRIA, and PASCAL VOC, on the test set of our dataset, the average precision obtained was 53.6%, 55.3%, and 61.6%, respectively. After training on our training set, the performance significantly improved to an average precision of 79.8%. Figure 4 shows the precision-recall curve on evaluating the SSD model on our test data.

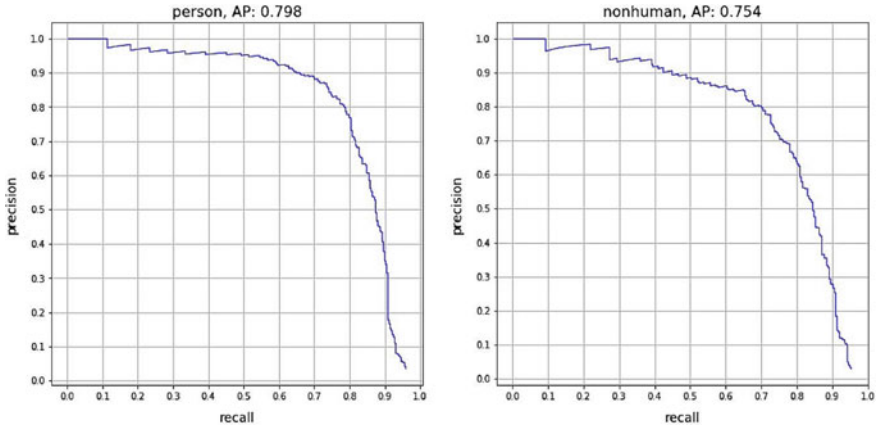

Fig. 4 Precision-recall curve for person and nonhuman on SSD model evaluation on our test data

Table 1 Comparing performance of SSD300 trained on different datasets

Dataset	Average precision (%)
COCO [8]	53.6
INRIA [11]	55.3
PASCAL VOC 07+12 [7]	61.6
PnPLO (ours) [17]	**79.8**

Bold represents the average precision of SSD300 model on the PnPLO dataset, which we created

5 Conclusion

Various research works are carried out focusing on detection of persons because of its widespread applications such as video surveillance, person tracking, and intelligent cars. These works use datasets comprising of only persons as benchmark dataset. Many objects have features similar to that of a person. A model trained on only persons fails to differentiate these objects from a person and person-like objects. Usage of only person datasets as a benchmark leads to many false positives, detecting person-like objects also to be persons. To overcome this problem, we prepared a person and person-like object (PnPLO) dataset consisting of persons as well as person-like objects such as statues, mannequins, scarecrows, and robots. We trained ssd_300 model on our dataset and tested on PnPLO test data. We show that the performance of the model trained on PnPLO dataset is better than performances of models trained on three standard datasets, namely, COCO, INRIA, and PASCAL VOC. The model trained on our dataset has an average precision of 79.8% compared to 53.3%, 55.6%, and 61.6% for SSD model trained on COCO, INRIA, and PASCAL VOC datasets, respectively.

References

1. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
2. Liu W et al (2016) SSD: Single shot multibox detector. In: Lecture notes in computer science, vol 9905. Springer, pp 21–37
3. Krizhevsky A, Sutskevar I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Commun ACM 60:84–90
4. Szegedy C et al (2015) Going deeper with convolutions. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, pp 1–9
5. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556
6. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. in: CVPR
7. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The PASCAL visual object classes (VOC) challenge. Int J Comput Vis 88:303–308
8. Lin T-Y, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollar P (2014) Microsoft COCO: common objects in context. In: European conference on computer vision (ECCV), Zurich
9. Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A (2010) SUN database: large-scale scene recognition from abbey to zoo. In: CVPR
10. Wang S, Cheng J, Liu H, Wang E, Zhou H (2018) Pedestrian detection via body part semantic and contextual information With DNN. IEEE Trans Multimedia 20(11):3148–3159
11. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 886–893
12. Pierard S, Lejeune A, Van Droogenbroeck M (2011) A probabilistic pixel-based approach to detect humans in video streams. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 921–924
13. Leibe B, Seemann E, Schiele B (2005) Pedestrian detection in crowded scenes. In: IEEE computer society conference on computer vision and pattern recognition (CVPR'05), San Diego, CA, USA, vol 1, pp 878–885
14. Karthika NJ, Chandran S (2019) Recent developments in pedestrian detection using deep learning. In: 2019 International conference on computing, communication, and intelligent systems (ICCCIS), Greater Noida, India, pp 353–358
15. Dollr P, Wojek C, Schiele B, Perona P (2012) Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell 34(4):743–761
16. Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? The kitti vision benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3354–3361
17. Karthika NJ, Chandran S. PnPLO Dataset. https://drive.google.com/open?id=1_HSXRasckZr8-LIZ9ms7LtBxzMwl-3Oa
18. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, pp 770–778
19. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, pp 779–788
20. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE conference on computer vision and pattern recognition, Columbus, OH, pp 580–587
21. Girshick R (2015) Fast R-CNN. In: 2015 IEEE international conference on computer vision (ICCV), Santiago, pp 1440–1448
22. Tzutalin L (2015) https://github.com/tzutalin/labelImg