Pedestrian Detection and the Effect of Diverse Benchmarks

Nagi OULD TALEB, Adil CHERGUI, Mohamed
Larbi BEN MAATI, Mohamedade Farouk NANNE, Mohamed O.M. Khelifa, Aicha Mint Aboubekrine

Abstract— Pedestrian detection is a popular research topic in computer vision community, with several applications including robotics, surveillance and automotive safety. It is a particularly difficult subject, in particular because of the great variability of appearances and possible situations. Much of the progress of the past few years has been driven by the availability of challenging public datasets. To solve these problems some recent researches has led to highlighting large databases. In our approach we will use the SSD method for detecting pedestrians in images using a single deep neural network.

The goal is to evaluate our model after fine-tuning with different datasets, and then analyze the performance gain from transfer learning.

Keywords— Pedestrian detection, Convolutional Neural Network, Best Benchmark, Deep Learning.

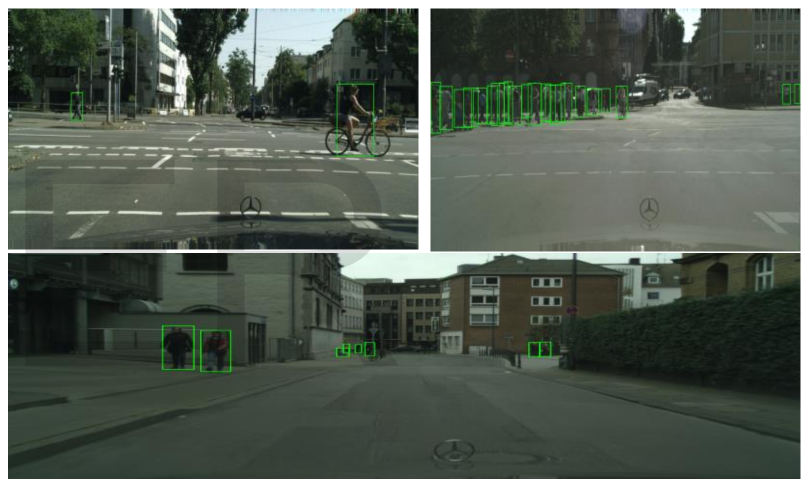
1. Introduction

The Pedestrian detection has been an important computer vision research topic over the years. This constitutes an important challenge because of the variety of scales, positions and lighting conditions [49]. So there are many problems that need to be solved in pedestrian detection as detecting objects with different sizes and in different locations. Also the problem of partial occlusion adds to the complexity of the task in question [47,48].

Now there are many searches proposing different strategies to solve this problem. They can be grouped into two large categories [13, 34, 35]: conventional approaches and deep learning approaches. In the conventional approaches, features are extracted, such as HOG-LBP [39], Haar [37], and HOG [38] from the images in order to train an SVM classifier [38] or a Boosting classifier [40]. The approaches based on Deep Learning have obtained very good results in different pedestrian detection topics [41, 42, 43, 44, 46]. This type of neural networks can learn discriminate features from raw image pixels.

A lot of progress has been made in recent years on object detection due to the use of convolutional neural networks (CNNs) [1, 2, 3, 4]. They have significantly improved image classification [5] and object detection [6, 7].

The main contributions of this work are summarized as follows: section 2, describes the general structure of two methods that appeared more important to us and the dataset used; in section 3, we will detail the model used; section 4 shows the experiments performed, together with the results obtained and finally, section 5 presents the conclusion.



2. Related Approaches

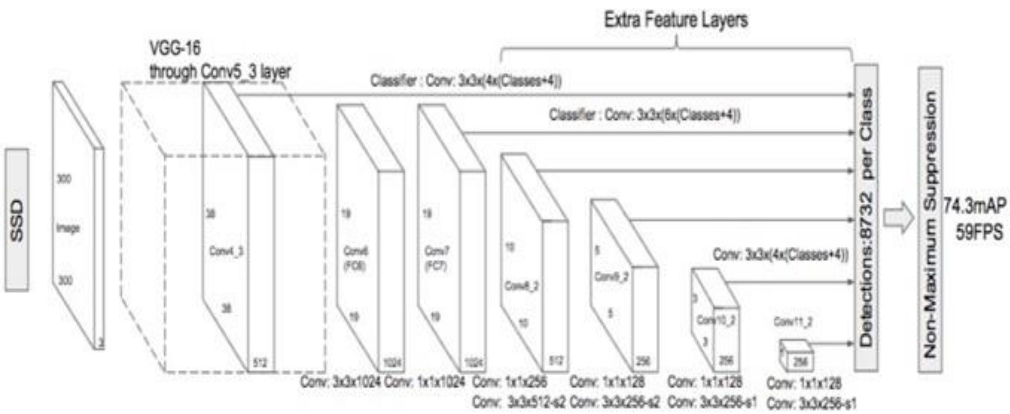
As mentioned before, we talked about two categories of topic, conventional approaches and deep learning approaches. Here we will be interested only in the deep learning approaches.

In the related works we notice that the current detectors of state-of-the-art can be divided into two categories [8]: (1) The two-stage approach [9, 10, 11, 12], and (2) the one-stage approach [2, 13]. In the two-stage approach, a number of candidate object boxes are generated and then they are classified and regressed. The one-stage approach detects objects by regular and dense sampling over locations, scales and aspect ratios. The two methods achieved top performances on many challenging benchmarks, such as PASCAL VOC [14] and MS COCO [15]. Among these methods we will be interested in Faster-RCNN and SSD.

Faster-RCNN:

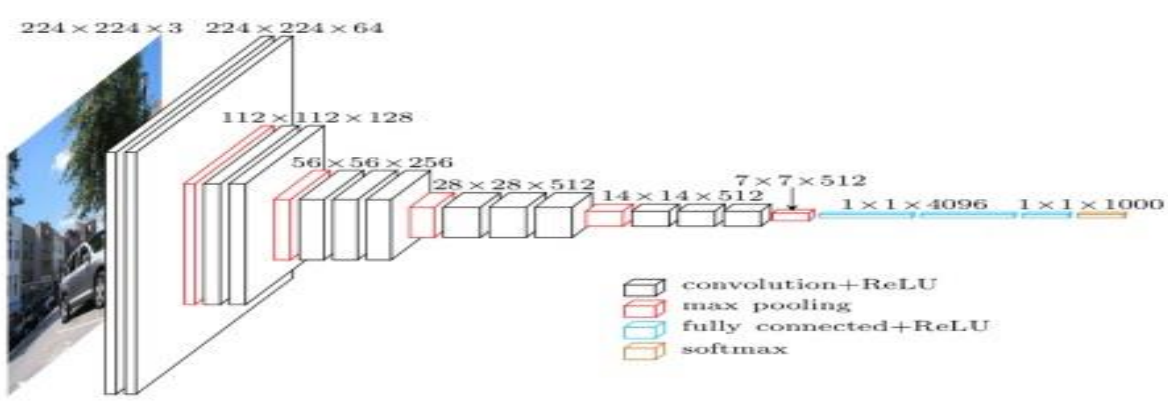
Since appearing in 2015, Faster R-CNN has been particularly influential, and has led to a number of follow-up works [12, 17, 18, 19, 20, 21, 22, 23].The Faster-RCNN object detection system is composed of two modules. One is a deep fully con-

International Journal of Scientific & Engineering Research, Volume 9, Issue ƝǑƐ2Ì×ÛÌÔÉÌÛıl ƳƱƝƜ
ISSN 2229-5518



volutional network that proposes regions. In the stage, called the region proposal network (RPN), images are processed by a feature extractor (VGG16). The other module is the Fast R-CNN detector [10] that uses the proposed regions. The entire system is a single, unified network for object detection [16]. The Faster RCNN is the successor of R-CNN [11] and Fast R-CNN [10]. For Faster R-CNN, we can also choose the number of region proposals to be sent to the box classifier at test time. Typically, this number is 300 in the setting. This method was evaluated on PASCAL VOC 2007 detection [14], on PASCAL VOC 2012 [1] and on Microsoft COCO benchmark [15].

SSD: Single Shot MultiBox Detector
SSD is a method for detecting objects in images (Figure1) using a single deep neural network [2].
The model we talked about before performed region proposal and region classification in two separate steps. First, they used a region proposal network to generate region of interest; next, they used either fully-connected layers or position-sensitive convolutional layers to classify those regions. SSD does the two in a "single shot", simultaneously predicting the bounding box and the class as it processes the image. SSD reached new records in terms of performance and precision for object detection tasks, scoring over 74% mAP (mean Average Precision) at 59 frames per second on standard datasets such as PASCAL VOC and COCO.



VGG architecture (input is 224x224x3)

Experimental results on the PASCAL VOC, COCO, and ILSVRC datasets confirm that SSD [2] has competitive accuracy to methods that utilize an additional object proposal step and is much faster, while providing a unified framework for both training and inference (Figure1).

Benchmarks:
The databases play a very important role in the field of pedestrian detection. They greatly influence the results of the detection according to the databases considered during the learning process. A several benchmarks have been created for this task [24, 25, 26]. These benchmarks have enabled great progress in this area [27]. The most popular publicly available benchmarks of them is the INRIA, KITTI [30], ETH [31], TUD-Brussels [32], Daimler, Caltech and CityPersons datasets [33]. The INRIA dataset [24] have contributed to spurring interest and progress in this area of machine vision. The Caltech Pedestrian Dataset is also very important compared to others benchmarks. The Caltech datasets contain richly annotated video, recorded from a moving vehicle, with challenging images of low resolution and frequently occluded people. Existing datasets may be grouped into two type: the first is "person" datasets containing people in unconstrained pose in a wide range of domains and the second is "pedestrian" datasets containing upright people (standing or walking). In this article we are limited to the Caltech and Citypersons databases.

International Journal of Scientific & Engineering Research, Volume 9, Issue ƝŦ2Ì×ÛÌÔÉÌÛıƖƔƕƚƛ
ISSN 2229-5518

3. Model Details

Nowadays, deep learning has become the go-to method for image recognition tasks, far surpassing more traditional computer vision methods used in the literature. Several methods that are based on this technique have been created to the goal of achieving real-time object detection. In our approach we will be interested in the

$$w= scale.\sqrt{ar}\ ;\ h= \frac{scale}{\sqrt{ar}}.$$

Training data	Aspect ratios	MR-Caltech
Caltech	$1, 2, \frac{1}{2}, 3, \frac{1}{3}$	32.18%
COCO+VOC+Caltech	$1, 2, \frac{1}{2}, 3, \frac{1}{3}$	25.52%
COCO+VOC+Caltech	AR=0.41	19.83%
COCO+VOC+Caltech	scale modified	11.98%

SSD method which is based on the VGG16 model. The original SSD 512x512 model uses many feature maps to represent different scales, and many default boxes with different aspect ratios and scales in each feature map. SSD defines a scale value for each feature map layer. Starting from the left, Conv4_3 detects objects at the smallest scale 0.2 and then increases linearly to the rightmost layer at a scale of 0.9.

SSD's architecture builds on the venerable VGG-16 architecture, but discards the fully connected layers. The reason VGG-16 was used as the base network is because of its strong performance in high quality image classification tasks and its popularity for problems where transfer learning helps in improving results.

We have modified some parameters at the level of the layers.

For COCO dataset, the authors use box aspect ratios (ar) from the set ar= {1, 2, 1/2, 3, 1/3}. Combining the scale value with the target aspect ratios, we compute the width and the height of the default anchor boxes as follows:

- Aspect ratios:

For Caltech dataset, the mean aspect ratio (width/height) is 0.41. Depending on this observation, we set the aspect ratio of the proposed anchor boxes to only 0.41. This helped to decrease the false positives

- Default boxes scales:

We used more scales F of small people.

Layer	Scales
conv4_3	0.03, 0.04, 0.055, 0.07, 0.085
fc7	0.1, 0.15
conv6_2	0.26
conv7_2	0.42
conv9_2	0.58
conv8_2	0.74
conv10_2	0.9

- Datasets: Caltech and Citypersons.
- Evaluation: Evaluation for "Reasonable" and "All" subsets in Caltech and CityPerson.

4. Main Results

A. Comparison between SSD and Faster R-CNN

We started by doing a small comparison between SSD and Faster R-CNN by making a detection on several benchmarks.

The evaluation metric is log average Miss Rate (MR) on False Positive Per Image (FPPI) in the range $[10^{-2} - 10^{0}]$. Convert both FPPI, and MR to log scales, then, in the range of FPPI $[10^{-2} - 10^{0}]$, average the corresponding miss rates. This is the evaluation metric used in Caltech paper. We used the following databases here: INRIA, Caltech, Daimler, ETHZ, and TUD Brussels.

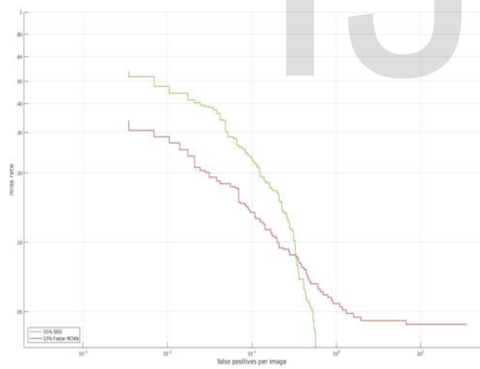
In this experience, SSD trained on MS-COCO, Faster-RCNN trained on Pascal VOC and the evaluation criteria is log average Miss Rate (MR) on False Positive Per Image (FPPI) in the range $[10^{-2} - 10^{0}]$.

International Journal of Scientific & Engineering Research, Volume 9, Issue ƝǑ2Ì×ûÌÔÉÌÛıl Ɣƀƀ
ISSN 2229-5518

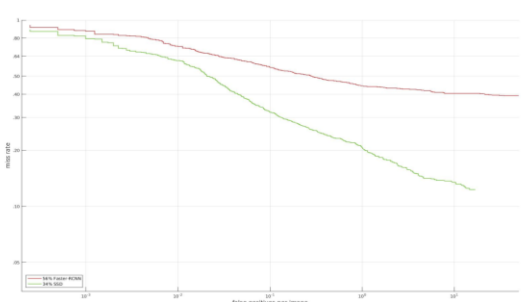
Dataset / Algorithm	Faster-RCNN	SSD
1- INRIA	**13%**	15%
2- Caltech	56%	**34%**
3- Daimler	44%	**28%**
4- ETHZ	58%	**53%**
5- TUD-Brussels	77%	**67%**

So we plotted the curves with each database lows:

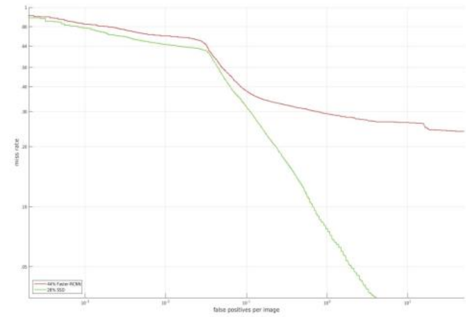
- INRIA Person dataset



- Caltech pedestrian detection benchmark

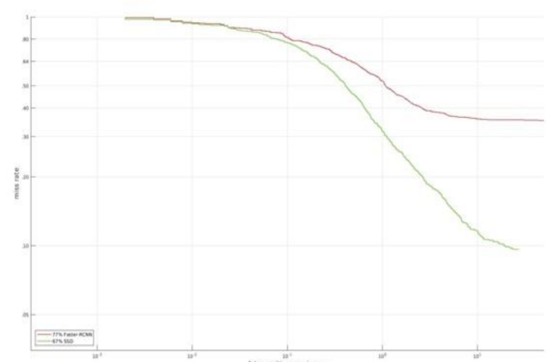


- Daimler pedestrian detection benchmark



- ETHZ pedestrian datasets

- TUD Brussels



International Journal of Scientific & Engineering Research, Volume 9, Issue ƝǞ2Ì×ÛÌÔÉÌÛılƔƞƜƝ
ISSN 2229-5518

From qualitative results, we can notice the problem of Faster-RCNN in detecting many false positives (trees, traffic signs ... etc), while detecting more people as well. However, SSD doesn't have the problem of false positives, but it misses many people (High miss rate). It is very obvious that Faster-RCNN is better if the qualities of images are high. We can say mainly that Faster-RCNN has more problems with hard negatives in low-res images, so it gives high False Positives. However, SSD can handle these hard negatives and small objects better, but it has higher miss rate.

In our opinion, these results are somehow misleading, because SSD is trained on Pascal VOC + COCO, but Faster R-CNN is trained only on Pascal VOC. To be fair, we need to train all of them on certain pedestrian datasets, and then evaluate.

B. Evolution with SSD

The first thing that we did here is the evaluation for the model trained with COCO and Pascal VOC. Firstly, we used the model trained with COCO and Pascal VOC. We then did the detection on Caltech and Citypersons benchmarks. Secondly we did the fine-tuning using both databases, Caltech and Citypersons. For Fine-tuning with Caltech, we use the improved 10x annotations [45]. To get the best out of the training/fine-tuning process, we used the videos set00/V014, set01/V005, and set02/V011 as validation set, which are chosen to be as general as possible. These three validation videos are removed from the training data. All training and fine-tuning are done using a batch size 32 and a learning rate starting from 0.0005, which decrease to 0.05 * previous learning rate each certain iterations.

However, for Citypersons, we used the training set of CityPerson except one video of the city "aachen" as validation. Note that for CityPerson we don't have the testing set annotations, so we use the validation set for testing.

After that, SSD is fine-tuned with Caltech and CityPerson training set, with the modifications in aspect ratio and bounding box scales as explained before. The results are evaluated with both datasets to measure the generalization of the fine-tuning process. In the following table we can notice that after doing the finetunig on Caltech and Citypersons, we find better results. We note that "Reasonable" is when the height of the person is greater than 50 pixels and without occlusion or occlusion that is less than 35% whereas "all" is when it comes to all other cases.

Model / Training data	Caltech benchmark	CityPerson
SSD512-VGGNet (COCO+VOC before fine-tuning)	Reasonable: 33.05% All: 73.02%	Reasonable: 69.40% All: 82.96%
SSD512-VGGNet (COCO+VOC+Caltech after fintunig)	Reasonable: 11.96% All: 55.18%	Reasonable: 70.13% All: 83.22%
SSD512-VGGNet (COCO+VOC+CityPerson after fintunig)	Reasonable: 22.15% All: 66.48%	Reasonable: 53.27% All: 75.61%

5. Conclusion

We have presented an approach that is based on SSD method using the Citypersons and Caltech databases. We have modified some parameters such as aspect ratios and also injected learning base to increase the detection performance. For future work, we plan to combine another system with this method to further improve this model at the level of miss rate.

References

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 39, Issue: 6, June 1 2017.

[2] Wei Liu1, Dragomir Anguelov2, Dumitru Erhan3, Christian Szegedy3, Scott Reed4, Cheng-Yang Fu1, Alexander C. Berg1, "SSD: Single Shot MultiBox Detector", Proceedings of the European Conference on Computer Vision (ECCV) , Dec 2016.

[3] Jifeng Dai, Yi Li*, Kaiming He, Jian Sun, "R-FCN : Object Detection via Region-based Fully Convolutional Networks",Neural Information Processing Systems (NIPS), Jun 2016.

[4] Christian Szegedy, Scott Reed, Dumitru Erhan, "Scalable High Quality Object Detection", Cornell University Library, Dec 2015.

[5] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems, Volume 1, Pages 1097-1105, Decembre 2012.

[6] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, Yann LeCun, "OverFeat : Integrated Recognition, Localization and Detection using Convolutional Networks",Cornell University Library, Feb 2014.

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, UC Berkeley, "Rich feature hierarchies for accurate object detection and semantic segmentation", ACM Digital Library, Oct 2014.

International Journal of Scientific & Engineering Research, Volume 9, Issue ƝǾ2Ì×ÛÌÔÉÌÙılɣhШ
ISSN 2229-5518

[8] Shifeng Zhang1,2, Longyin Wen3, Xiao Bian3, Zhen Lei1,2*, Stan Z. Li4,1,2, "Single-Shot Refinement Neural Network for Object Detection", Cornell University Library, Jan 2018.

[9] Zhaowei Cai1, Quanfu Fan2, Rogerio Feris2, and Nuno Vasconcelos1, "A Unified Multi-scaleDeep Convolutional NeuralNetwork for Fast Object Detection", Cornell University Library, Volume abs/1607.07155, Jul 2016.

[10] Ross Girshick, "Fast R-CNN", ACM Digital Library, ICCV'15 Proceedings of the 2015 IEEE International Conference on Computer Vision ((ICCV), Dec 2015.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", ACM Digital Library, Proceeding CVPR' 2014 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Pages 580-587, June 2014.

[12] Abhinav Shrivastava1, Abhinav Gupta1, Ross Girshick2, "Training Region-based Object Detectors with Online Hard Example Mining", IEEE Computer Society, pages= {761 - 769}, 2016.

[13] Joseph Redmon□, Joseph Redmon□, "YOLO9000 : Better, Faster, Stronger", Cornell University Library, Dec 2016.

[14] Mark Everingham, Luc Van Gool, Christopher K.I. Williams, "The PASCAL Visual Object Classes (VOC) Challenge", International Journal of Computer Vision manuscript, Volume 88, issue 2, pp 303-338, June 2010.

[15] Tsung-Yi Lin, Michael Maire and all, "Microsoft COCO: Common Objects in Context", Cornell University Library, Volume {abs/1405.0312}, 2014.

[16] Jan Chorowski, Dzmitry Bahdanau, "Attention- Based Models for Speech Recognition", Cornell University Library, Volume {abs/1506.07503}, 2015.

[17] SEAN BELL, KAVITA BALA and all, "Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks", dblp computer science bibliographi, Volume= {abs/1512.04143}, 2015.

[18] Abhinav Shrivastava and Abhinav Gupta, "Contextual Priming and Feedback for Faster R-CNN", Carnegie Mellon University, 2016.

[19] Sergey Zagoruyko*, and all, "A MultiPath Network for Object Detection", dblp computer science bibliographi, Pages {936-944}, 2017.

[20] Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun, "Deep Residual Learning for Image Recognition", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), ISSN: 1063-6919.

[21] Jifeng Dai Kaiming He Jian Sun, "Instance- aware Semantic Segmentation via Multi- task Network Cascades", Cornell University Library, Volume: {abs/1512-04412}, 2015.

[22] Kye- Hyeon Kim*_ and all, "PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection", dblp computer science bibliographi, Volume: {abs/1608.08021}, 2016.

[23] Bin Yang1, and all, "CRAFT Objects from Images", Cornell University Library, Volume= {abs/1604.03239}, 2016.

[24] Navneet Dalal and Bill Triggs, "Histograms of Oriented Gradients for Human Detection", CVPR'05 Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Volume 1, Pages 886-893, 2005.

[25] Piotr Doll ´ ar, and all, "Fast Feature Pyramids for Object Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 36 Issue 8, Aug 2014.

[26] Andreas Geiger and Philip Lenz, Raquel Urtasun, "Are we ready for Autonomous Driving?", CVPR'12 Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Pages 3354-3361, June 2012.

[27] Rodrigo Benenson Mohamed Omran Jan Hosang Bernt Schiele, "Ten Years of Pedestrian Detection, What Have We Learned?", ECCV 2014: ECCV 2014 Workshops, pp 613-627, 2014.

[28] Kaiming He Georgia Gkioxari Piotr Doll´ar Ross Girshick, "Mask R-CNN", the blue social bookmark and publication sharing system, 2018.

[29] Piotr Doll´ar1 and all., "Pedestrian Detection: A Benchmark", In CVPR, 2009.

[30] Andreas Geiger, Philip Lenz, Christoph Stiller and Raquel Urtasun, "Vision meets Robotics: The KITTI Dataset", International Journal of Robotics Research, Volume 32 Issue 11, September 2013.

[31] Ernest Cheung and all., "MixedPeds: Pedestrian in Inannotated Videos using Synthetically Generated Humman-agents for Training", Cornell University Library, Volume: {abs/1707.09100}, 2017.

[32] Miao He and al, "Pedestrian Detection with Semantic Regions of Interest", US National Library of Medicine, 2017.

[33] Shanshan Zhang and al., "CityPersons: A Diverse Dataset for Pedestrian Detection", Cornell University Library, Volume:{abs/1702.05693}, 2017.

[34] Carlos Ismael Orozco1 and al., "New Deep Convolutional Neural Network Architecture for Pedestrian Detection", Journa Image Communication, Volume 47 Issue C, Sep 2016.

[35] Rodrigo Benenson and al. "Ten Years of Pedestrian Detection, What Have We Learned", ECCV 2014: Computer Vision - ECCV 2014 Workshops, pp 613-627, 2014.

[36] Jan Hosang, "Taking a Deeper Look at Pedestrians", Cornell University Library, Volume: {abs/1501.05790}, 2015.

[37] PAUL VIOLA, "Robust Real-Time Face Detection", International Journal of Computer Vision, Volume 57, Issue 2, pp 137-154, May 2004.

[38] Navneet Dalal and Bill Triggs, "Histograms of Oriented Gradients for Human Detection", CVPR' 05 Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) – Volume1, 2005.

[39] Xiaoyu Wang* and al, "An HOG- LBP Human Detector with Partial Occlusion Handling", IEEE Computer Society, 2009.

[40] Piotr Doll´ar1, "Pedestrian Detection: A Benchmark", 2009 IEEE Conference on Computer Vision and Pattern Recognition, ISSN: 1063- 6919, 2009.

[41] Ping Luo and al., "Switchable Deep Network for Pedestrian Detection", IEEE Conference on Computer Vision and Pattern Recognition, ISBN: 978-1-4799-5118-5, 2014.

[42] Wanli Ouyang and Xiaogang Wang, "A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling", IEEE Conference on Computer Vision and Pattern Recognition, ISBN: 978-1- 4673- 1228-8, 2012.

[43] Wanli Ouyang and al, "Modeling Mutual Visibility Relationship in Pedestrian Detection", IEEE Conference on Computer Vision and Pattern Recognition, ISBN: 978-0- 7695- 4989-7, 2013.

[44] Pierre Sermanet and al, "Pedestrian Detection withUnsupervised Multi- Stage Feature Learning", IEEE Computer Society (CVPR), Pages: 3626- 3633, 2013.

[45] Shanshan Zhang and al, "How Far are We from Solving Pedestrian Detection?", Cornell University Library, Volume: {abs/1602-01237}, 2016.

[46] Weicheng and al, "Deep Learning Based pedestrian Detection", Chinese Control and Decision Conference (CCDC), ISSN: 1948-9447, 2018.

[47] Ming-Shi Wang* and and Zhe- Rong Zhang, "FPGA Implementation of HOG based Multi- Scale Pedestrian Detection", Proceedings of IEEE International Conference on Applied System Innovation, ISBN= {978-1-4503-5614-5}, 2018.

[48] JiaXiang Zhao and Jun Li, "RPN+ Fast Boosted Tree: Combining Deep Neural Network with Traditional Classifier for Pedestrian Detection", ACM Digital Library, Volume 47 Issue C, Sep 2016.

[49] Xiaowei Zhang and al, "Too Far to see? Not Really! Pedestrian Detection with Scale –Aware Localization Policy", University Library, Volume {abs/1709.00235}, 2017.