QALAAI ZANIST JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil – Kurdistan – Iraq
Vol. (2), No. (2), April 2017
Speical Issue : The 1st International Conference on Information Technology (ICoIT'17)
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

Fast and Accurate Real Time Pedestrian Detection Using Convolutional Neural Network

Hayder Albehadili
Department of Software , Kadhum College for Islamic Science University, Iraq
Sadqmo4@gmail.com
Laith Alzubaidi
Department of Systems and Applications, University of Information Technology & Communications Baghdad, Iraq
Laithicci@gmail.com
Jabbar Rashed
Department of Electrical Engineering , Engineering College, University of Misan- Iraq
Dr.jabar72@gmail.com
Murtadha Al-Imam
Department of Software , Kadhum College for Islamic Science University, Iraq
Muam1987_computer@yahoo.com
Haider A. Alwzwazy
Department of Mathematics, Mathematics College, University of Misan- Iraq
Haiderabd83@gmail.com

ARTICLE INFO

Article History:

Received: 20 March 2017

Accepted: 1 April 2017

Published: 10 April 2017

DOI:

10.25212/lfu.qzj.2.2.29

Keywords:
Convolutional Neural Network, Pedestrian Detection, Multiscale input images

ABSTRACT

Recently, pedestrian detection has become an important problem of interest. Our work primarily depends on robust and fast deep neural network architectures. This paper used very efficient and recent methods for pedestrian detection. Recently, pedestrian detection has become an important problem of interest. This paper suggests robust convolutional neural network models to solve this problem. We primarily evaluate accuracy and speed. Our work primarily depends on robust and fast deep neural network architectures; substantial changes to those models achieve results that are competitive with prior state-of-the-art methods. As a result, we outperformed all the prior state-of-the-art pedestrian detection methods. We also overtook other models that use extra information during testing and training. All experiments used three pedestrian detection challenge benchmarks: Caltech-USA, INRIA, and ETH.

1. INTRODUCTION

Object detection is one of the most challenging tasks in computer vision. Recently, pedestrian detection has become a topic of interest for many researchers [1, 2, 3, 4, 5, 6] because it is widely used in many real world applications such as safe driving systems, security surveillance, and robotic navigation. A large number of difficulties such as object appearance make pedestrian detection a challenging task. Different appearances can result from different poses and environmental factors such as illumination. Pedestrian

QALAAI ZANIST JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil – Kurdistan – Iraq
Vol. (2), No. (2), April 2017
Speical Issue : The 1st International Conference on Information Technology (ICoIT'17)
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

detection becomes an even more challenging task when multiple persons are adjacent to each other, leading to a bounding box that covers more than one pedestrian. Prior methods of pedestrian detection can be categorized into two groups: hand-crafted or end-to-end feature learning systems. For the first type, researchers first proposed different approaches such as Integral Channel Features [7] and HoG [3] followed by trainable classifiers such as Support Vector Machine SVM [8, 9], boosted classifiers [10], or random forests [11]. In contrast, a convolutional neural network (CNN) is an example of an end-to-end feature extraction system. Recently, CNNs have become a very effective and popular approach for pedestrian detection [2, 12, 13]. CNNs have been used in variety of machine learning tasks including image recognition [14,15, 16, 17, 18]. We summarize our contributions as follows: we adapt contemporary CNN object detection architectures to pedestrian detection. We extensively analyse and investigate two contemporary CNN architectures: the SPPnet architecture in [19] and Fast Regions with CNN (F-RCNN) architecture. Our work principally depends on prior work; however, we comprehensively study the suggested models and carefully choose their parameters. Consequently, we consider the advantages of some prior existing methods and propose a unified deep model. We achieve the best current performance on the Caltech-USA, ETH, and INRIA benchmarks. We not only show that this model outperforms all existing models but also demonstrate its performance as a real-time pedestrian detector. This proposed CNN achieves results that are superior to all existing methods. Note that we always follow the best parameter settings, as investigated in [20] (e.g. methods for generating proposals such as Squares ChnFtrs, the size of input windows, and so on).

2. RELATED WORK

Because CNNs are used in our study, in this section, we briefly review all deep neural network models used for pedestrian detection. A very efficient model was proposed in [23] that achieves particularly good results. This model efficiently joins detection components such as feature extraction, deformation, occlusion, and the final stage of classification. All these parts are combined into one unified CNN and trained using back-propagation. The model presented in [2] essentially focused on using ConvNet (proposed by [27]) with slight changes, and the model is entirely trained in an unsupervised fashion. The main contribution of ConvNet [12] is a classifier with hierarchical features. In this case, the classifier is fed from two successive preceding layers simultaneously. This allows the classifier to use both local and global features. One interesting work is JointDeep [23], which combines several learning-stage pedestrian detection models to produce a unified deep model. All the stages are trained simultaneously as a synergistic system. The system joins four main stages used in pedestrian detection into one unified CNN: feature extraction, a PDM, occlusion, and finally classification. A third implementation of the model was demonstrated in [16]. The authors reviewed prior work and showed the effect of different factors on detection. In addition, they showed the effect of combining different methods and achieved results that were superior to previous work. In [15], a pedestrian detection model was proposed called Switchable Deep Network [15], which achieves state-of-the-art results on the Caltech-USA dataset with a 37.87% miss rate (MR). The main contribution of this system is the hierarchical extraction of features. In other words, the model extracts features from different parts of the body such as the head or shoulder. In [24], two methods to enhance the performance of a CNN were proposed. The first method is a technique called Random Dropout. In this method, the authors replaced the conventional Dropout method used by CNN. Instead of using a fixed rate of dropout, they used a random dropout. Their second

QALAAI ZANIST JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil – Kurdistan – Iraq
Vol. (2), No. (2), April 2017
Speical Issue : The 1st International Conference on Information Technology (ICoIT'17)
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

method was named Ensemble Inference Network, which uses several fully connect layers that each have different network topologies. Both methods were evaluated on the Caltech Pedestrian Dataset and Daimler Mono datasets, achieving reasonable results. Recently, the authors of [25] proposed different methods to generate foreground and background windows to feed a CNN using Aggregated Channel Features. In [26], a multistage deep network classifier called MultiSDP was proposed. This model has more than one stage of classifier. Each stage is fed with contextual information. The model requires unsupervised pre-training to avoid overfitting resulting from the multi-stage classifiers. Finally, the most recent work of interest was proposed in [20]. The authors introduced a type of ConvNet called a 'vanilla' network, and performed extensive experiments and tested different parameters that can influence the performance of a CNN. The authors showed that big ConvNets such as AlexNet or small ConvNets such as CifarNet can achieve results that are competitive with prior methods on the widely used Caltech-USA dataset. Although both SpatialPooling+ [21] and Katamari [22] represent the state-of-the-art, these methods require additional information at testing and training time, and the vanilla network achieved competitive results. Fig. 1 shows the results of all abovementioned methods on both the ETH and Caltech-USA datasets. It is clear that the vanilla network achieves results that are superior to all existing deep learning methods on the Caltech-USA dataset. In addition, JointDeep [23] surpasses the other methods on the ETH dataset.

Vanilla networks [20] achieve the state-of-the-art results over all existing deep neural network approaches. Consequently, in this work, we enhance the performance of ConvNet [12] and address the weakness inherited from vanilla networks by using a very recent powerful deep neural network model. Different network architectures are demonstrated to enhance pedestrian detection performance.

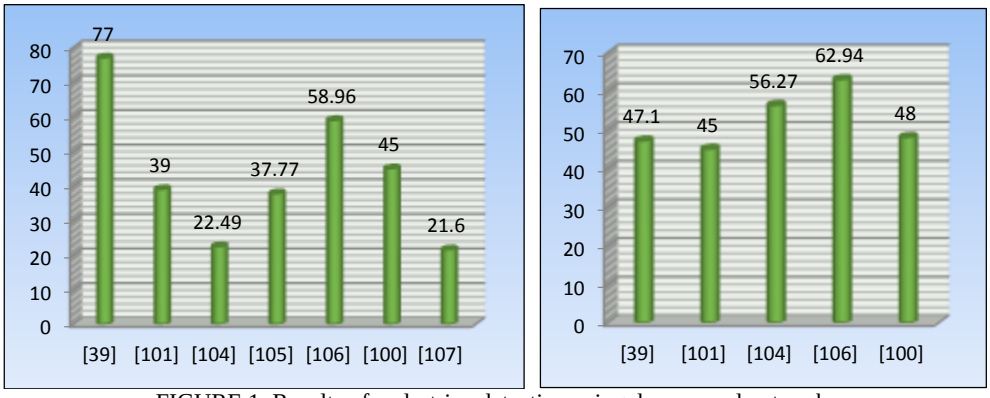


FIGURE 1. Results of pedestrian detection using deep neural networks

In addition, we further show the MR results of all other prior strategies in Fig. 2. Note that a lower MR is better. Optical flow is used by pedestrian detection strategies as additional information during training and testing [21, 22].

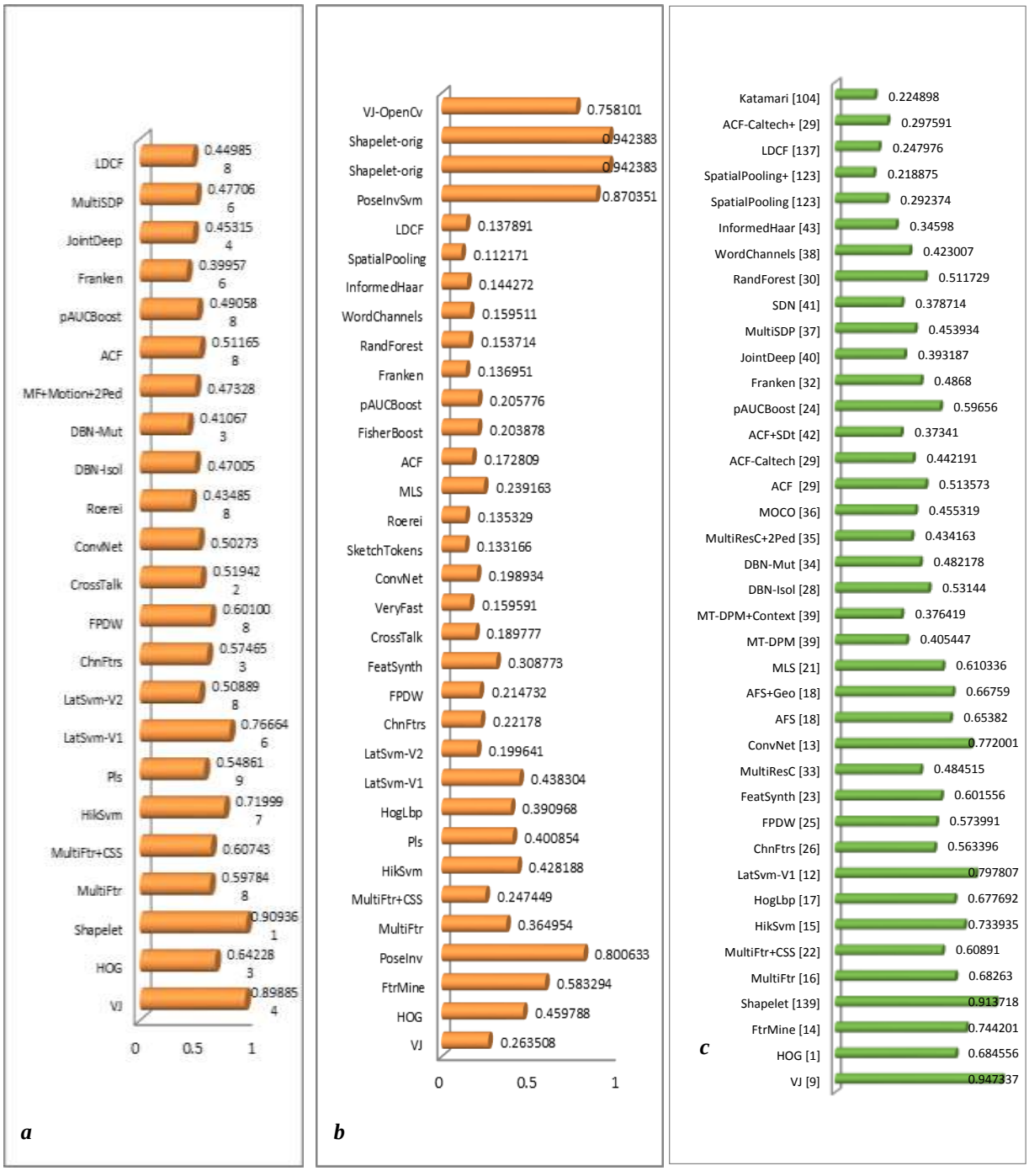
The CNN state-of-the-art pedestrian detector [20] and most existing models that use CNN follow the same training procedure, which is similar that of [18], which was originally used for object detection. The steps for detection can be summarized as follows:

- Warp or crop the proposals generated by the first step and feed them into the CNN because CNNs require a fixed input size, as demonstrated in [18].
- Train the CNN using candidate windows.

QALAAI ZANIST JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil – Kurdistan – Iraq
Vol. (2), No. (2), April 2017
Speical Issue : The 1st International Conference on Information Technology (ICoIT'17)
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

- After training, use the features extracted from one of the last layers of the CNN to train the SVM, which is used later for classifying extracted features from the last or preceding layers of the CNN.

Note that the best performance achieved using a CNN follows the procedure above [20]. This procedure has some drawbacks, as demonstrated in [27]. First, the training and testing phases are very slow because it is necessary to extract features for both training and testing. Second, because the input to the CNN is fixed, the model is trained with only a single input image size [20]. Finally, the systems still have three dependent steps: proposal creation, training the deep model, and classification using SVM, as shown in Fig. 3. These steps reduce both the performance and speed. In addition, the CNNs used in [20, 18] accept a fixed image size of 227 × 227. Therefore, all candidate proposal windows must be cropped before they are fed to the CNN. To summarize the limitations of the vanilla networks described in



QALAAI ZANIST JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil – Kurdistan – Iraq
Vol. (2), No. (2), April 2017
Speical Issue : The 1st International Conference on Information Technology (ICoIT'17)
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

FIGURE.2 Results of pedestrian detection for all prior works

1. Only a fixed input image size is allowed.
2. Several exhaustive separate stages are required: proposal generation, deep model training, and classification using SVM.
3. Only a single resolution is allowed for input images.
4. A large amount of time is consumed because of limitations (1−3).

We endeavor to enhance the performance and overcome the limitations inherent in vanilla networks. Very efficient and fast deep models proposed in [19, 27] are used, and the potential parameters that can influence performance are further investigated.

3. DEPLOYMENT OF DIFFERENT CNN ARCHITECTURES FOR PEDESTRIAN DETECTION

Because our experiments use SPPnet and F-RCNN proposed in both [19] and [27], respectively, they have crucial intrinsic differences. SPPnet is very similar to the methods used in [20, 18]; however, SPPnet accepts different image sizes and keeps the same training and testing procedures as in [20]. F-RCNN has more qualifications than SPPnet because it has a unified CNN that shortens all three main steps mentioned earlier further details are in Section 4.1

3.1. PIPELINE STEPS OF SPPNET FOR PEDESTRIAN DETECTION

As described earlier, we essentially use two predominant CNN paradigms. First, we use SPPnet as described in [19], which is similar to the vanilla networks in [20]. The only difference is that SPPnet is flexible and can take multi-scale input images for training or testing. In this section, we explore the leverage of multi-scale input images. The pedestrian detection pipeline shown in Fig. 3 consists of three main parts. In the first part, multi-scale input images are prepared. In the next step, a large CNN is used to extract the features from the input proposals. Finally, SVM is used on top of the CNN for classification. These steps match most pedestrian detection CNN approaches, specifically the vanilla network, although here, multi-scale input images are used.

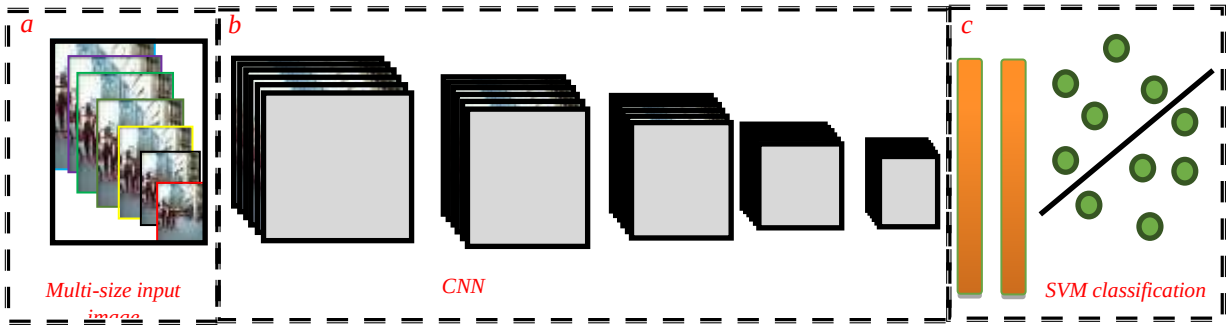


FIGURE 3. Pedestrian detection pipeline: a) multi-scale input images, b) the CNN, and c) final, fully connected layer and SVM.

It is clear that this model is similar to [20] except for the first stage of multi-scale images. Note that SPPnet can process whole input images or we can generate candidate proposals. In this work, we use the second option. The following steps describe Fig. 3.

Step 1. Generated candidate windows. There are a variety of methods for generating proposal windows. For fair comparison with [20], we followed the methodsuggested by the authors for generating proposals.

QALAAI ZANIST JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil – Kurdistan – Iraq
Vol. (2), No. (2), April 2017
Speical Issue : The 1st International Conference on Information Technology (ICoIT'17)
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

Step 2. The generated category-independent windows are passed into the CNN for training and testing purposes. Using the robust CNN architecture described by Krizhevsky et al. [12], 4096-dimensional feature vectors are extracted from the last layers using Caffe [28]. In contrast to [20], which uses RCNN [18], our CNN can accept any size of input image because we use the SPPnet architecture [19]. In addition, the need for crop and warp proposal windows is eliminated using the SPPnet approach [19]. Cropping or warping images is considered one of the main drawbacks of RCNN because it takes a long time to warp images and loses information. However, SPPnet can process entire images in a single step.

Step 3. Features are extracted from the CNN layers and the SVM is trained for use later for testing.

4. EXPERIMENTAL SETUP

The model above was evaluated on three challenge benchmarks, the INRIA, Caltech-USA, and ETH datasets. The next section briefly describes these datasets. For the evaluation, we follow the conventions proposed by Dollar and Zitnick. [1].

4.1. CALTECH-USA DATASET

One of the most challenging datasets that we used is the Caltech-USA Pedestrian dataset [1]. It contains numerous video clips (‘set00’ to ‘set11’). The dataset has 250,000 frames. All image frames are accompanied by annotations. The annotations number 350,000 bounding boxes for 2,300 unique pedestrians. All videos were taken by a driven vehicle. All images have the same resolution of 640 × 480 pixels. The image quality is very poor compared with the INIRA dataset. The general setting for the dataset is that the first six datasets are used for the training and the other five sets are used for testing. In the standard setting, sampling is chosen as one of thirty frames in both the training and testing sets. However, it is possible to increase the training and testing samples by choosing different numbers of frames to skip. We increased the training and testing sample rate by tenfold, as described in [20].

4.2. INRIA DATASET

The INRIA dataset [3] contains very high resolution images. However, it is a small dataset and has some incorrectly labeled images. The dataset has 614 positive frames and 1,218 negative frames for training. In addition, it has 288 negative images and 453 positive images for testing. Overall, it has around 1,774 pedestrian annotations.

4.3. ETH dataset

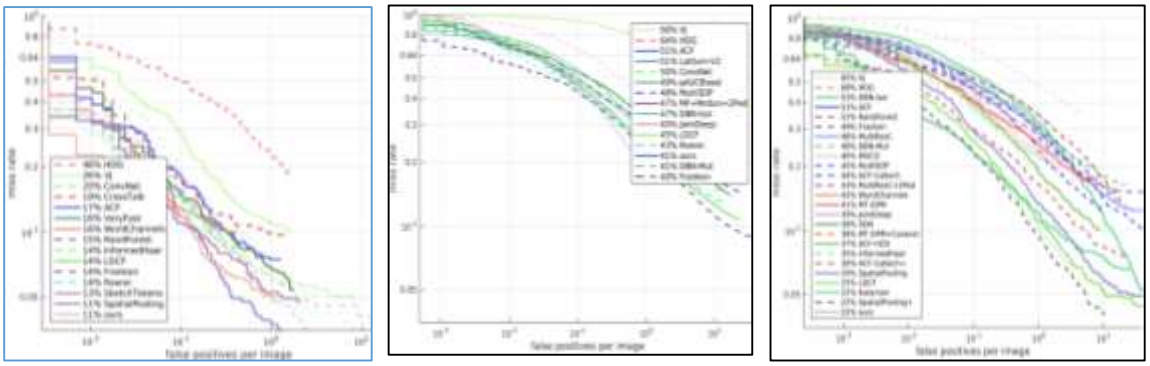
In this dataset, all the data frames have a resolution of 640 × 480 pixels. This dataset has a different density than the other datasets. We followed the author-recommended setting in [28].

5. COMPARISON OF THE ACCURACY OF THE PROPOSED METHOD AND VANILLA NETWORKS

To evaluate the advantages of using multi-scale images for the deep model over the traditional method used in vanilla networks [20], we used the AlexNet [12] model, which is the same model as used in [20]. However, a SPP-net [19] was exploited to use multi-scale input images. The experiments were conducted using the three benchmarks listed above. After building the model described in 5.1, we conducted our experiments without substantial changes to the model. We not only achieved results that were competitive with all former CNN models, but the results were reasonable with respect to other approaches that use additional information during testing such as optical flow.

QALAAI ZANIST JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil – Kurdistan – Iraq
Vol. (2), No. (2), April 2017
Speical Issue : The 1st International Conference on Information Technology (ICoIT'17)
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

Figs. 4(a), (b), and (c) show the results on the Caltech-USA, ETH, and INRIA datasets, respectively. It is clear that SpatialPooling+ [21] and Katamari [22] are state-of-the-art



and outperform all the prior models. Furthermore, SpatialPooling and Katamari overtake the vanilla network, which is the state-of-the-rt for pedestrian detection using deep neural networks. The highest reported result on the Caltech-USA dataset using the vanilla network is 23.3% MR [20].

FIGURE 4.SPPnet results using multi-scale input images: a) Caltech-USA, b) ETH, and c) INRIA

Using the SPPnet [19] technique without any significant changes, we achieved accuracies of 22.21%, 40.49%, and 11.1% MR on the Caltech-USA, ETH, and INRIA datasets, respectively. Note that we significantly outperformed vanilla networks. Thus, we next investigated a more powerful CNN model called F-RCNN [18] for pedestrian detection.

6. WEAKNESSES OF SPPNET

Although SPPnet achieves competitive results to the state-of-the-art pedestrian detection using multi-scale input images, it still has the following shortcomings: 1) we still need to perform the four training and testing stages described in Section 4 and 2) fine-tuning can be accomplished only for the last two fully connected layers [27]. Thus, to alleviate these limitations, a more sophisticated model is used. In this section, we use the F-RCNN model proposed in [27] for pedestrian detection. As stated earlier, F-RCNN has a more powerful architecture than all the CNNs mentioned in [18, 19, 20] for the following reasons:

- It is a unified deep model that enfolds more steps in a single CNN and shortens all four stages of CNN training and testing. For instance, feature extraction, cropping candidates, and SVM training and testing are all encapsulated in a single unified CNN. Moreover, it is much faster than all previously mentioned models because F-RCNN does not extract features and store them in a database; instead, all proposals are fed into the CNN with the original images so they can project the proposals into the images. F-RCNN can also propagate error back-to-back, in contrast to SPPnet, which can propagate error only to the last max-pooling layer. Therefore, all layers of the CNN are tuneable, which gives it a greater ability to tune the model for pedestrian detection.
- There are more features recognized in F-RCNN than in other models.

QALAAI ZANIST JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil – Kurdistan – Iraq
Vol. (2), No. (2), April 2017
Speical Issue : The 1st International Conference on Information Technology (ICoIT'17)
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

Experiments were also conducted on the three pedestrian detection datasets. Again without any substantial changes, F-RCNN achieved 21.83%, 38.18%, and 10.9% MR on the Caltech-USA, ETH, and INRIA datasets, respectively as shown in figure 5,6,7. The results are shown in Fig. 5. Intuitively, F-RCNN is clearly superior to SPPnet. Thus, we designed a more effective model based on F-RCNN.

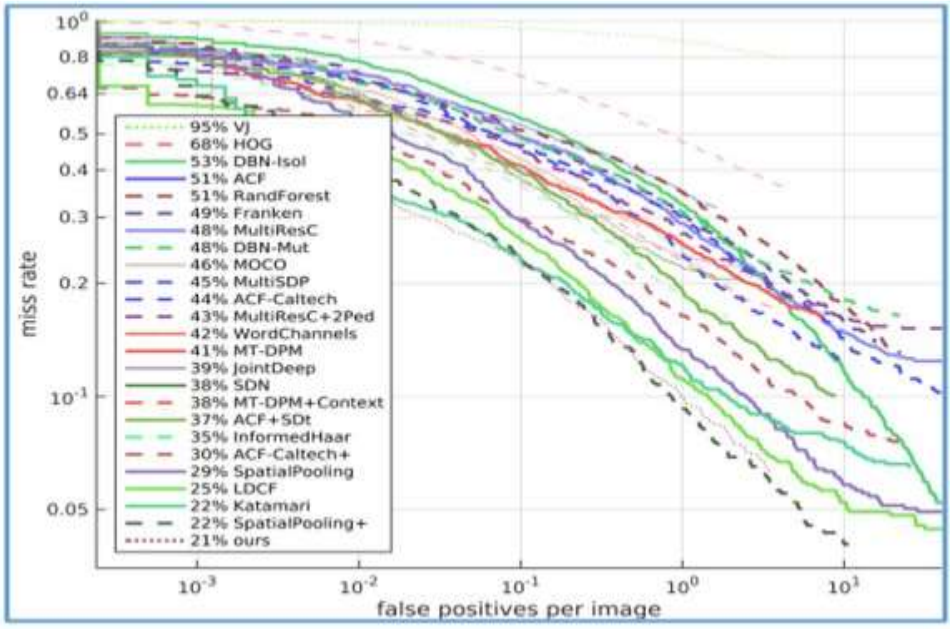


FIGURE5: F-RCCN results using multi-scale input images,Caltech-USA dataset

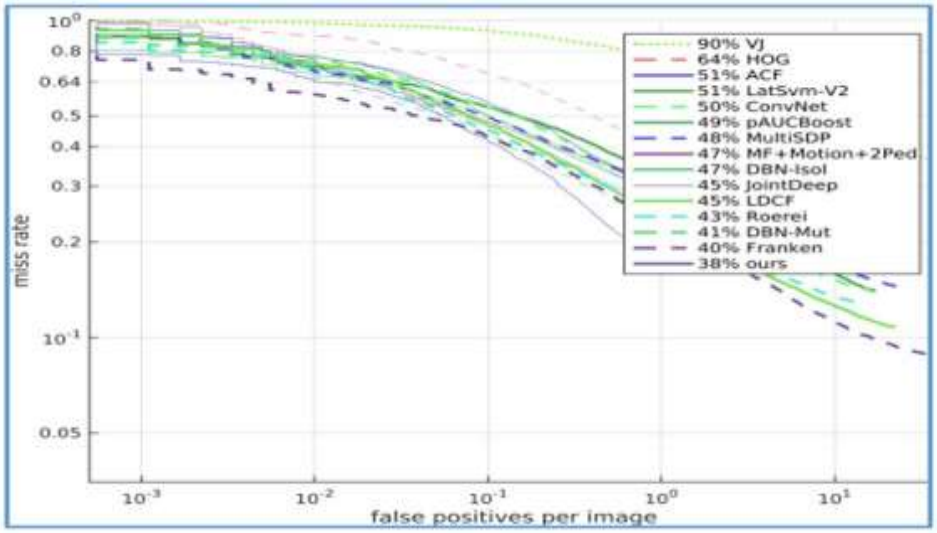


FIGURE 6: F-RCCN results using multi-scale input images,ETH dataset

QALAAI ZANIST JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil – Kurdistan – Iraq
Vol. (2), No. (2), April 2017
Speical Issue : The 1st International Conference on Information Technology (ICoIT'17)
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

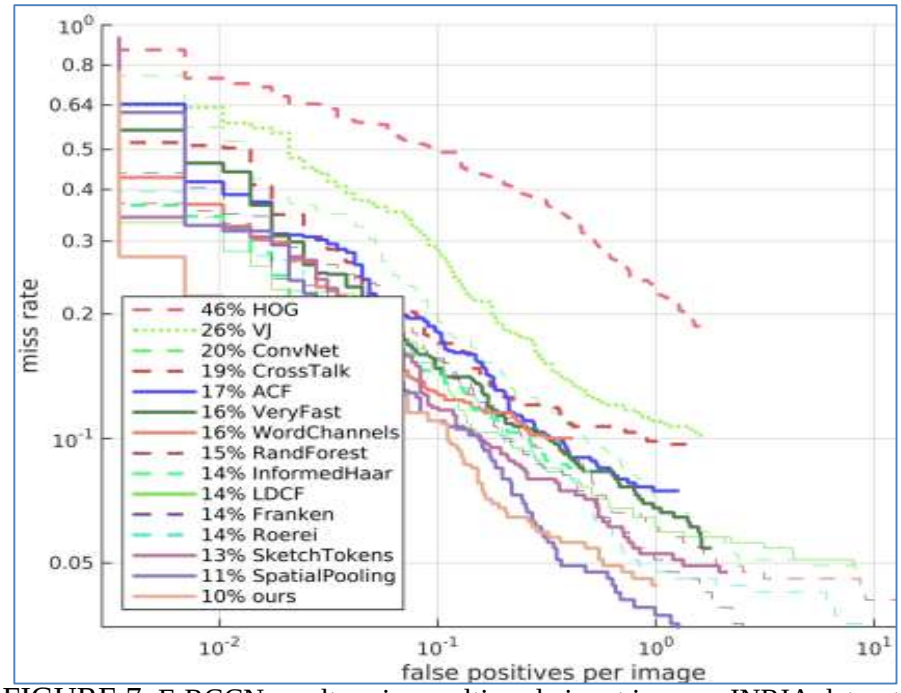


FIGURE 7: F-RCCN results using multi-scale input images, INRIA dataset

7. CONCLUSION

In this work, we used very efficient and recent methods for pedestrian detection. We evaluated both SPPnet and F-RCNN, which were originally used for object detection. We showed that SPPnet performs better than conventional CNNs, which are used in prior pedestrian detection methods [20]. Using SPPnet without substantial changes, we achieved better accuracy than the best-designed model using deep neutral networks. In addition, we accomplished results that are competitive with other models that do not use CNN. Motivated by the results from SPPnet, we used a leveraging CNN called F-RCNN that is more efficient than SPPnet. As a result, we outperformed all the prior state-of-the-art pedestrian detection methods. We also overtook other models that use extra information during testing and training.

8. REFERENCES

[1] [1] P. Doll´ar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: an evaluation of the state of the art. TPAMI, 34(4):743–761, 2012.

[2] [2] P. Sermanet, K. Kavukcuoglu, S. C. Pedestrian, and Y. Le-Cun.Pedestrian detection with unsupervised multi-stage feature learning.CVPR, 2013.

[3] [3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection.In CVPR, 2005. 1, 2, 5, 7

[4] [4] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classifcation on riemannian manifolds. IEEE Trans. PAMI, 30(10):1713–1727, Oct. 2008. 1, 2

[5] [5] C. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In ECCV, 2004.

[6] [6] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. In PAMI 2010. 1

[7] [7] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral Channel Features" Proc. British Machine Vision Conf., 2009.

[8] [8] M. Bertozzi, A. Broggi, M. Del Rose, M. Felisa, A. Rakotomamonjy and F. Suard "A Pedestrian Detector Using Histograms of Oriented Gradients and a Support Vector Machine Classifier" Proceedings of the 2007 IEEE Intelligent Transportation Systems Conference Seattle, WA, USA, Sept. 30 - Oct. 3, 2007

[9] [9] Reecha P. Yadav, Vinuchackravarthy, Sunita P. Ugale "Implementation of Robust HOG-SVM based Pedestrian Classification" International Journal of Computer Applications (0975 – 8887) Volume 114 – No. 19, March 2015

[10] [10] JiaolongXu, David Vazquez, Sebastian Ramos, Antonio M. Lopez , and Daniel Ponsa "Adapting a Pedestrian Detector by Boosting LDA Exemplar Classifiers" CVPR 2013

[11] [11] [1] J.Gall and V.Lempitsky. Class-specific hough forests for object detection. In CVPR, 2009.

[12] [12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks.In NIPS 2012: Neural Information Processing Systems. 1

[13] [13] P. Luo, Y. Tian, X. Wang, and X. Tang. Switchable deep network for pedestrian detection.In CVPR, 2014. 2, 5

[14] [14] 61. Yin, X.C., Yin, X., Huang, K.: Robust text detection in natural scene images. CoRR abs/1301.2628 (2013)

[15] [15] Min Lin, Qiang Chen, and Shuicheng Yan "Network In Network" arXiv 1312.4400v3,4 Mar 2014

[16] [16] Dan Cires¸an, Ueli Meier and JurgenSchmidhuber, "Multi-column Deep Neural Networks for Image Classification" CVPR 2012

[17] [17] JulienMairal, PiotrKoniusz, ZaidHarchaoui, and CordeliaSchmid, "Convolutional Kernel Networks" arXiv 14 Nov 2014 .

[18] [18] Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik "Rich feature hierarchies for accurate object detection and semantic segmentation" arXiv:1311.2524v5 [cs.CV] 22 Oct 2014

[19] [19] Kaiming, He and Xiangyu, Zhang and Shaoqing, Ren and Jian Sun "Spatial pyramid pooling in deep convolutional networks for visual recognition" European Conference on Computer Vision, 2014

[20] [20] Jan Hosang Mohamed Omran Rodrigo BenensonBerntSchiele, "Taking a Deeper Look at Pedestrians" arXiv:1501.05790v1 [cs.CV] 23 Jan 2015

[21] [21] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Strengthening the effectiveness of pedestrian detection with spatially pooled features.In ECCV, 2014. 1, 2, 4, 7

[22] [22] Rodrigo Benenson Mohamed Omran Jan HosangBerntSchiele, "Ten Years of Pedestrian Detection,What Have We Learned?"

[23] [23] W. Ouyang and X. Wang.Joint deep learning for pedestrian detection.In ICCV, 2013. 2, 4, 5

QALAAI ZANIST JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil – Kurdistan – Iraq
Vol. (2), No. (2), April 2017
Speical Issue : The 1st International Conference on Information Technology (ICoIT'17)
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

[24] [24] Hiroshi Fukui Takayoshi Yamashita, Yuji Yamauchi, Hironobu Fujiyoshi, Hiroshi Murase2 "Pedestrian Detection Based on Deep Convolutional Neural Network with Ensemble Inference Network"

[25] [25] Xiaogang Chen, Pengxu Wei, Wei Ke, Qixiang Ye, Jianbin Jiao "Pedestrian Detection with Deep Convolutional Neural Network"

[26] [26] X. Zeng, W. Ouyang, and X. Wang.Multi-stage contextual deep learning for pedestrian detection.In ICCV, 2013. 2

[27] [27] Ross Girshick, "Fast R-CNN "arXiv preprint arXiv:1504.08083, 2015

[28] Y. Jia, "Caffe: An open source convolutional architecture for fast feature embedding," http://caffe.berkeleyvision.org/, 2013.