Journal of Theoretical and Applied Information Technology
15th November 2019. Vol.97. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN EFFICIENT DEEP LEARNING FRAMEWORK FOR PEDESTRIAN DETECTION

HOANH NGUYEN
Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam
E-mail: nguyenhoanh@iuh.edu.vn

ABSTRACT

Vision-based pedestrian detection has achieved a large successful with the fast development of deep convolutional neural network (CNN). However, due to difficult environments such as large-scale variation, heavy occlusion and small scale of pedestrian, recent deep CNN-based approaches for pedestrian detection still do not achieve very good accuracy over public benchmark dataset. In this paper, an efficient framework based on deep CNN for pedestrian detection is proposed, especially for small-scale pedestrians. The proposed method first uses the reduced ResNet-34 architecture for generating convolution feature maps. Then, deconvolutional modules is used after the base convolution layers to bring additional context information which is more effective to detect small-scale pedestrians. In the region of interest pooling process, different feature maps at different scales are used to produce high quality region proposals. Furthermore, a modified loss function based on cross entropy loss is used to increase the loss contribution from hard-to-detect small-sized pedestrians. Experiment results demonstrate the effectiveness of the proposed approach with good detection performance over the Caltech pedestrian dataset.

Keywords: *Pedestrian Detection, Convolutional Neural Network, Intelligent Transportation Systems, Object Detection, Deep Learning*

1. INTRODUCTION

Pedestrian detection plays an essential role in many intelligent transportation systems, such as advanced driving assistance systems, autonomous driving and intelligent traffic management systems. Although many existing approaches for pedestrian detection have achieved high detection accuracy over public benchmark dataset, the detection of the presence of small-scale pedestrians who are relatively far from the camera is still very challenging for researchers. Vision-based pedestrian detection methods can be divided into two groups: traditional methods and deep CNN-based methods. Traditional methods are usually based on hand-craft features to detect pedestrian, including MultiSDP [9], Haar cascade classifier [12], Haar-like features [13], HOG [14], ACF [15] and so on. Traditional methods have limited ability of feature representation, which is difficult to handle complex scenarios. In recent years, deep convolutional neural networks (CNNs) have achieved incredible success on object detection tasks as well as pedestrian detection [16], [20]. Thus, many deep learning-based methods have also applied to pedestrian detection, including SDS-RCNN [7], SAF R-CNN [8], MultiSDP [9], RPN+BF [10], TA-CNN [11] and so on. Deep CNN-based methods achieved good accuracy on large-sized pedestrian detection, but they did not perform well on small-sized pedestrian detection because of obscure appearances, blurred boundaries and distortion. Furthermore, the features extracted by the existing methods contain less useful information for the small objects. Many recent methods have improved the feature extraction of small objects by using additional context information and increasing the spatial resolution of feature maps. DSSD [17] used deconvolution layers in combination with existing multiple layers to reflect the large-scale context. MS-CNN [18] applied deconvolution on shallow layers to increase the feature map resolution before using the layers to extract region proposals and pool features. Recently, Long et al. [19] introduced the fully convolution network (FCN), which demonstrated impressive performance in semantic segmentation and object detection. However, these kinds of frameworks are less explored in pedestrian detection area.

In view of the above research challenges, this paper proposes an improved framework based on

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Faster R-CNN [6] for pedestrian detection. First, ResNet-34 architecture [21] is adopted for generating convolution feature maps. Then, deconvolutional modules are added after the base convolution layers, so the semantics from higher layers can be conveyed into lower layers to increase the representation capacity. Furthermore, different convolution layers of the base network are forwarded to RoI pooling layer to match proposals obtained from proposal generation. Finally, a modified loss function is adopted to increase the loss contribution from hard-to-detect examples.

This paper is organized as follows: an overview of previous methods is presented in Section 2. Section 3 describes the detail of the proposed method. Section 4 demonstrates experimental results. Finally, the conclusion is made in Section 5.

2. RELATED WORK

Vision-based pedestrian detection methods can be divided into two groups: traditional methods and deep CNN-based methods. Traditional methods mainly devoted to extract hand-craft features. In [14], the Histogram of Gradients is adopted to describe the image local variance and linear Support Vector Machine is then used to detect pedestrians. Felzenszwalb et al. [24] designed the Deformable Part Model based on Histogram of Oriented Gradient features. In [25], the Integral Channel Features coupled with a standard boosting algorithm. Wang et al. [26] proposed to combine Histograms of Oriented Gradients and Local Binary Pattern as the feature set to handle partial occlusion in pedestrian detection. In [27], the authors conducted a detailed investigation of the different factors affecting the integral channel features detectors. In [12], a Haar cascade classifier and validated candidates through a novel part-based HOG filter are used to generate possible pedestrian candidates. Benenson et al. [28] proposed a decision forest detector named Katamari by combining multiple published strategies. In [9], Zeng et al. proposed to jointly train multi-stage classifiers through several stages of back propagation. Liu et al. [29] used a linear kernel function to combine heterogeneous features, which enhanced the description power of the heterogeneous features. To incorporate with rich information from image data, the authors in [30] extracted Haar-like features by convolution between shape templates and features of colors and gradients. In [31], Baekthe et al. proposed cascade implementation of an additive kernel support vector machine (AKSVM) and the AKSVM was sub-optimized by a genetic algorithm. Zhang et al. [32] extracted patches which were potential to include objects-of-interest and composed the extracted patches into an optimal number of sub-frames. Traditional methods have limited ability of feature representation, which is difficult to handle complex scenarios.

With fast development of deep learning [36], many methods based on deep learning for pedestrian detection have been proposed and achieved better results compared to traditional methods. Pierre et al. [33] proposed an unsupervised method based on convolutional sparse coding to pretrain the filters at each stage. In [34], the authors used a derivation of the Faster R-CNN, which adopted multiple parallel classifiers with soft-rejection-based network fusion. In [34], the authors used a derivation of the Faster R-CNN, which adopted multiple parallel classifiers with soft-rejection-based network fusion. In [8], the authors introduced the Scale-Aware Fast RCNN by incorporating a large-size subnetwork and a small-size sub-network into a unified architecture. The method selected weights of large-size sub-network or small-size sub-network according to the height of proposal, but it is not accurate if a big proposal contains small objects. In [10], Zhang et al. proposed the region proposal network followed by boosted forests on shared and high-resolution convolutional feature maps. In [11], task-assistant CNN is proposed to jointly optimize detection with semantic tasks. Many approaches used multiple-scale feature maps and cascaded deep learning have also been proposed. Li et al. [35] proposed to generate multiple potential regions around an upper body for fast R-CNN. This method only used convolution features for multiple potential regions and did not make full use of low latitude convolution features. In [37], the authors investigated the scale-dependent pooling and layer-wise cascaded rejection classifiers. The scale-dependent pooling improved detection accuracy for object detection with scale-specific branches attached convolutional layers and the layer-wise cascaded rejection classifiers speeded up the detection.

3. PROPOSED APPROACH

Figure 1 shows the overall framework of the proposed approach. As shown in this Figure, the proposed method uses ResNet-34 as the base network for generating convolution feature maps at first step. Deconvolutional modules are then used after the base convolution layers to provide richer context for pedestrian detection at individual feature output scale. Next, the region proposal network uses different feature maps for generating proposals, and the region of interest (RoI) polling layer adjusts the

Journal of Theoretical and Applied Information Technology
15th November 2019. Vol.97. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

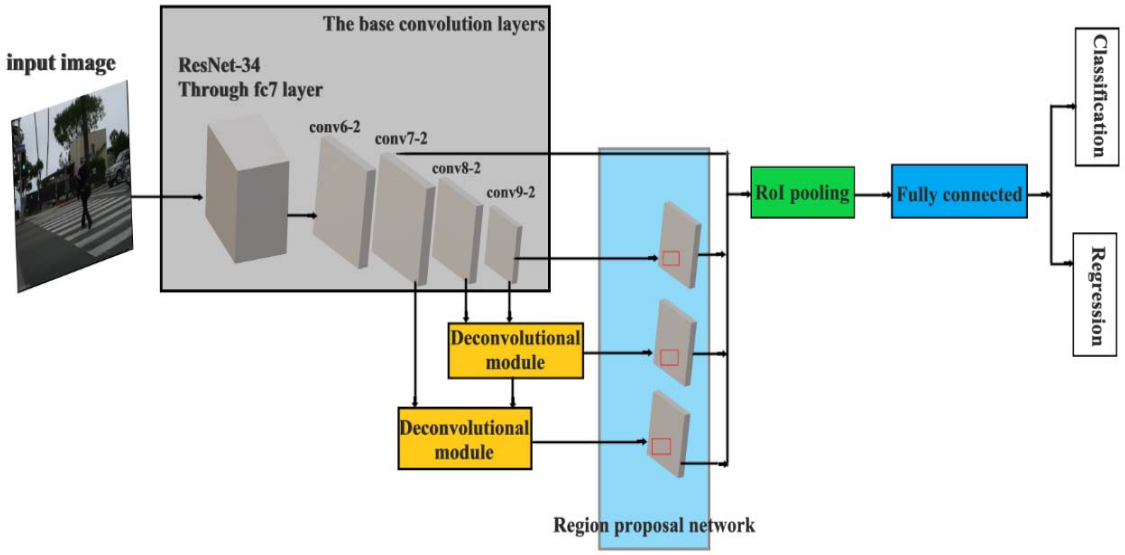

Figure 1: The Overall Framework of The Proposed Approach.

proposals to the specified size without sacrificing important contextual information. Finally, the classifier with two fully connected layers is used to classify proposals into pedestrian and background class and regress the bounding box of each detected pedestrian. Details of the proposed approach are explained in next sections.

3.1 The Base Network

In this paper, ResNet [21] is used as the base network instead of the VGG-16 [1] in the original Faster R-CNN framework. ResNet is an efficient network which adopted residual learning to every few stacked layers, so the training of networks can be eased and substantially deeper than other networks. ResNet-50 and ResNet-101 have high precision, but they are slower than VGG-16 [23]. Thus, they are not suitable for real time processing. Table 1 shows the comparison of computational cost of ResNet-34 and VGG-16. As shown, ResNet-34 significantly reduces the amount of computation compare with VGG-16. Thus, ResNet-34 is not only more accurate than VGG-16 but also faster than VGG-16. To improve the accuracy, the layer after res4f of the ResNet-34 is removed to retain the original FC6 layer and the FC7 layer. Then, extra layers after the ResNet-34 are added as in [22]. Figure 2 shows the architecture of the base network. The first number in the labels such as 4 and 6 represents the associated hidden layer in ResNet-34 architecture, and the second number represents the ID of the convolution layer in a hidden layer. For large-scale pedestrians, this paper fuses the Conv9-2 layer of the ResNet-34 network to generate many proposals which could better describe the characteristics of the large-scale pedestrians. In the fusion process, the deconvolution layer is used after Conv9-2 layer and the first deconvolutional module to enlarge the feature map size in order to match the size of the lower-level feature maps.

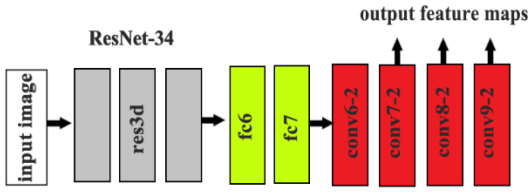

Figure 2: The Architecture of The Base Network. Conv7-2, Conv8-2 and Conv9-2 Layers Are Used as Output Feature Maps.

Table 1: The Comparison of Computational Cost of ResNet-34 and VGG-16.

Architecture	Computational Cost	Top-1 accuracy on ImageNet
VGG-16	15.3 billion FLOPs	0.715
ResNet-34	3.6 billion FLOPs	0.733

3.2 Region Proposal Network

The region proposal network (RPN) in Faster R-CNN is a fully convolutional detector which predicts many bounding boxes for the further object

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

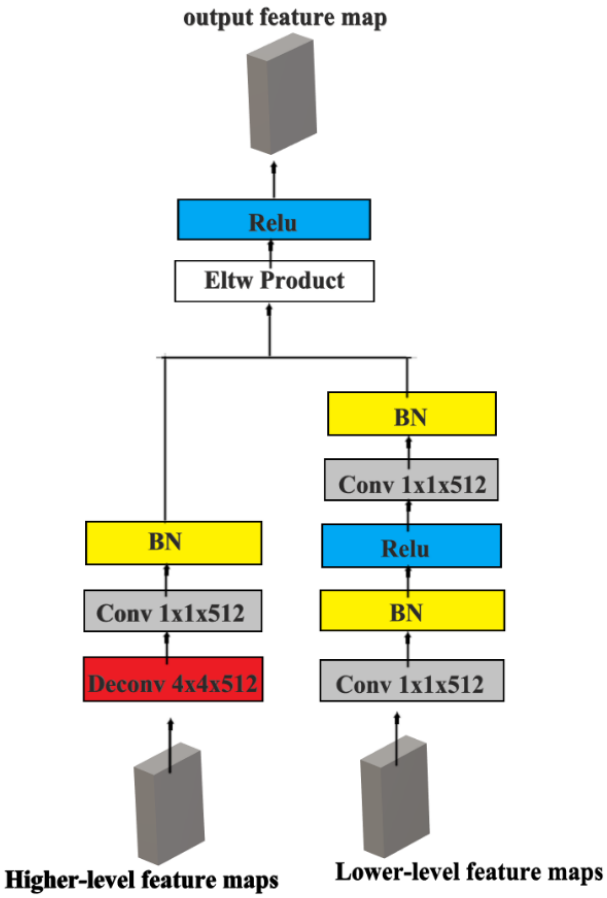

Figure 3: The Architecture of The Deconvolutional Module. Two Deconvolutional Module Are Used in Proposed Framework.

detection. The RPN includes two fully convolution layer: classification layer and box regression layer. The objectness classification layer distinguishes the objects and background. The box regression layer regresses the coordinates of the proposal location. However, the RPN in original Faster R-CNN framework is built on the Conv5-3 layer of the VGG network, which cannot provide enough detailed features because of many pooling layers in VGG network. Current deep CNN-based object detectors exploit multi-scale features to produce predictions of different scales, which showed improved object detection performance over Faster-CNN and SSD. However, shallow feature maps from the low layers of feature pyramid inherently lack fine semantic information for object recognition. Thus, this paper adds two deconvolutional modules after the base convolution layers. With this enhancement, the semantics from higher layers can be conveyed into lower layers to increase the representation capacity. Figure 3 illustrates the architecture of the deconvolutional module. As shown in this Figure, a 1x1 convolution layer and rectified linear activation are used. For the deconvolution branch, the encoder-decoder structure with 4x4 deconvolution is used followed by a 1x1 convolution. A batch normalization layer (BN) is added after each convolution layer. Higher-level feature maps are extracted after conv7-2, conv8-2 and conv9-2 layers. Then, the deconvolution layer is added to enlarge the feature map size in order to match the size of the lower-level feature maps. Finally, element-wise product is performed as a combination method, which is followed by rectified linear activation to generate the new output feature layer.

The region proposal network (RPN) receives output feature maps from the deconvolutional modules and Conv9-2 layer to produce high quality region proposals. The RPN first takes feature maps to generate a set of anchor boxes. An anchor is centered at the sliding window and is associated with a scale and aspect ratio. Because pedestrian is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

usually in rectangular shape, this paper uses one scale and one aspect ratio for each anchor, yielding 1 anchor at each sliding position in a feature map as shown in Figure 4. More specific, the aspect ratio is set at 0.5 in this paper. Next, the RPN takes all the anchor boxes and outputs two different outputs for each of the anchor. The first one is objectness score, which means the probability that an anchor is an object. The second output is the bounding box regression for adjusting the anchors to better fit the object. The anchors with estimated classification scores and the bounding box for each feature map location then are processed to form good quality proposals. Since anchors usually overlap, proposals end up also overlapping over the same object, Non-Maximum Suppression (NMS) is used to solve the issue of duplicate proposals. The proposal whose region overlaps a ground truth region more than 70% is regarded as a positive proposal. Otherwise, it is regarded as a negative proposal. After applying NMS, this paper keeps the top 256 proposals sorted by score.

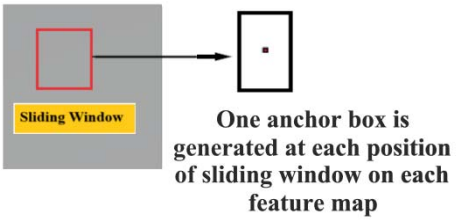

Figure 4: Anchor Boxes Generated by the RPN

3.3 Detection Network

The RoI pooling layer [6] followed by proposal generation step maps proposals to fixed-size feature maps as shown in Figure 5. In the RoI pooling process, this paper fuses different convolution layers of ResNet-34 network to match proposals obtained from proposal generation, including Conv7-2 layer, Conv8-2 layer and Conv9-2 layer.

After applying ROI pooling process, a list of regions with different sizes are transformed into a list of corresponding feature maps with a fixed size. Fixed size feature maps are needed for the classifier at final stage in order to classify them into a fixed number of classes. The classifier has two different goals: Classify proposals into pedestrian and background class and adjust the bounding box for each of detected pedestrian. The proposed classifier has two fully connected (FC) layers, a box classification layer and a box regression layer as shown in Figure 5. The first FC layer has two outputs, which are fed into the softmax layer to compute the confidence probabilities of being pedestrian and background. The second FC layer with linear activation functions regresses the bounding boxes of detected pedestrian. All convolutional layers are followed by a batch normalization layer and a ReLU layer.

3.4 Loss Function

In Faster R-CNN, the classification loss is defined as the following equation:

$$L_{cls}(p_i, p_i^*) = \begin{cases} -\log(p_i), & if\ p_i^* = 1 \\ -\log(1 - p_i), & otherwise \end{cases} \quad (1)$$

where i denotes the index of region proposals; p_i denotes the probability that the proposal i is an object. The ground-truth label p_i^* is set to 1 if the proposal is an object, and 0 if the proposal is not an object. For differentiating between easy instances and hard instances [2], this paper uses a loss function to up-weight hard instances. The loss function used in this paper is defined as the following equation:

$$L_{cls}(p_i, p_i^*) = \begin{cases} -(2 - p_i) \times \log(p_i), & if\ p_i^* = 1 \\ -\log(1 - p_i), & otherwise \end{cases} \quad (2)$$

If an instance is easy to classify correctly, the predicted probability p_i is close to 1. In this case, the modulation factor is near to 1 and the loss function is not affected by the modulation factor. If an instance is hard to classify correctly, p_i is closer to 0. In this case, the modulation factor is close to 2 and the weight from hard-to-detect samples increases. Thus, the model can be better trained for hard-to-detect samples. Thus, the proposed loss function increases the loss contribution from hard-to-detect examples to better detect the hard-to-detect targets, such as small-sized pedestrians.

4. EXPERIMENTAL RESULTS

In this section, this paper analyses and compares the performance of the proposed approach with other state-of-the-art approaches. The proposed method is implemented on a machine with Intel Core i5 9400 processor, 8GB of RAM, NVIDIA GeForce GTX 1660Ti. TensorFlow is used for implementing deep CNN frameworks.

4.1 Dataset

The Caltech pedestrian dataset [3] is one of the most popular datasets for pedestrian detection. It contains 250k frames captured from 10 hours of urban traffic videos. The training data (set00-set05) consists of six training sets, each with 6–13 one-

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

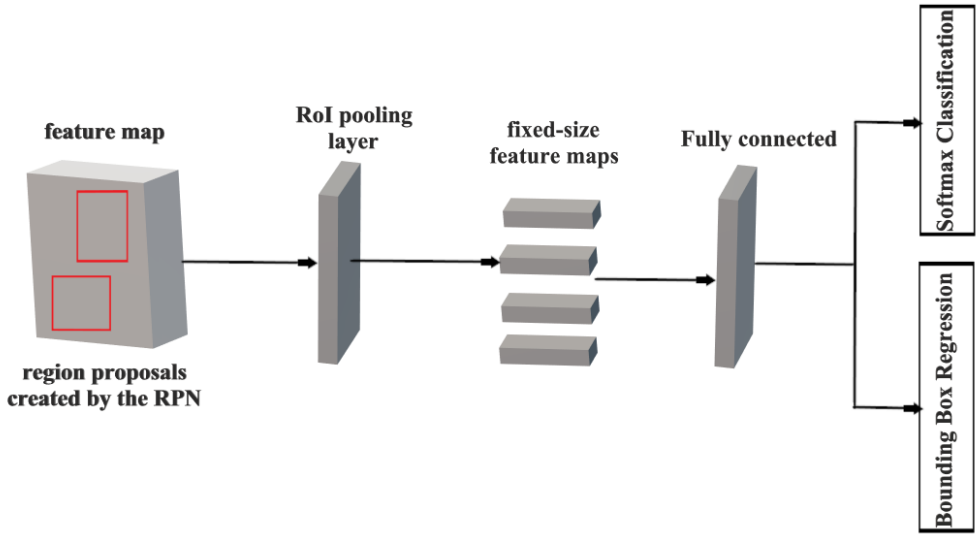

Figure 5: The Architecture of The Detection Network.

minute long sequence files, along with all annotation information. The testing data (set06-set10) consists of five sets, again along with all annotation information. The training and testing dataset have different video sequences with respect to the difficulty of pedestrian height, visibility, and aspect ratio. In proposed experiments, the training images are extracted with one out of every frame. There are 128,419 images for training and 4024 images for testing. Figure 6 shows some examples of extracted images of testing data.

4.2 Evaluation Metrics

The average precision (AP) [4], [5] is used as evaluation metric in the fields of pedestrian detection and object detection. To calculate the AP, intersection-over-union (IoU) [6] is used as an important parameter. The value of IoU is defined as:

$$IoU(b_1, b_2) = \frac{area(b_1 \cap b_2)}{area(b_1 \cup b_2)} \quad (3)$$

where b1 and b2 are the two proposal bounding boxes. The IoU is set to 0.7 in this paper, which means only the overlap between the detected bounding box and the ground truth bounding box greater than or equal to 70% is considered as a correct detection.

4.3 Performance Results

In this section, this paper examines and compares the performance of the proposed approach with other state-of-the-art approaches on Caltech pedestrian dataset, including SDS-RCNN [7], SAF R-CNN [8], MultiSDP [9], RPN+BF [10] and TA-CNN [11]. SDS-RCNN [7] proposed a segmentation infusion network to enable joint supervision on semantic segmentation and pedestrian detection. SAF R-CNN [8] introduced the Scale-Aware Fast RCNN by incorporating a large-size subnetwork and a small-size sub-network into a unified architecture. MultiSDP [9] proposed to jointly train multi-stage classifiers through several stages of back propagation. RPN+BF [10] proposed the region proposal network followed by boosted forests on shared and high-resolution convolutional feature maps. In TA-CNN [11], task-assistant CNN is proposed to jointly optimize detection with semantic tasks.

Figure 7 presents some examples of detection results of the proposed method. As shown in this Figure, the proposed method improves the detection performance from several aspects:

- The proposed method can detect correctly most pedestrians in images, including small-scale pedestrians who are relatively far from the camera (the first row and the second row of Figure 7).
- The proposed method can accurately detect pedestrians from a complex and diverse environment.
- The proposed method can eliminate false proposals and avoid producing multiple bounding boxes for one pedestrian.

Table 2 shows a comparison report between the proposed method and recent popular pedestrian detection methods in terms of detection accuracy. As shown from this table, the efficiency of the proposed approach surpasses the current state-of-the-art methods for pedestrian detection. More

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 6: Examples of Extracted Images of Testing Data in Caltech Pedestrian Dataset.

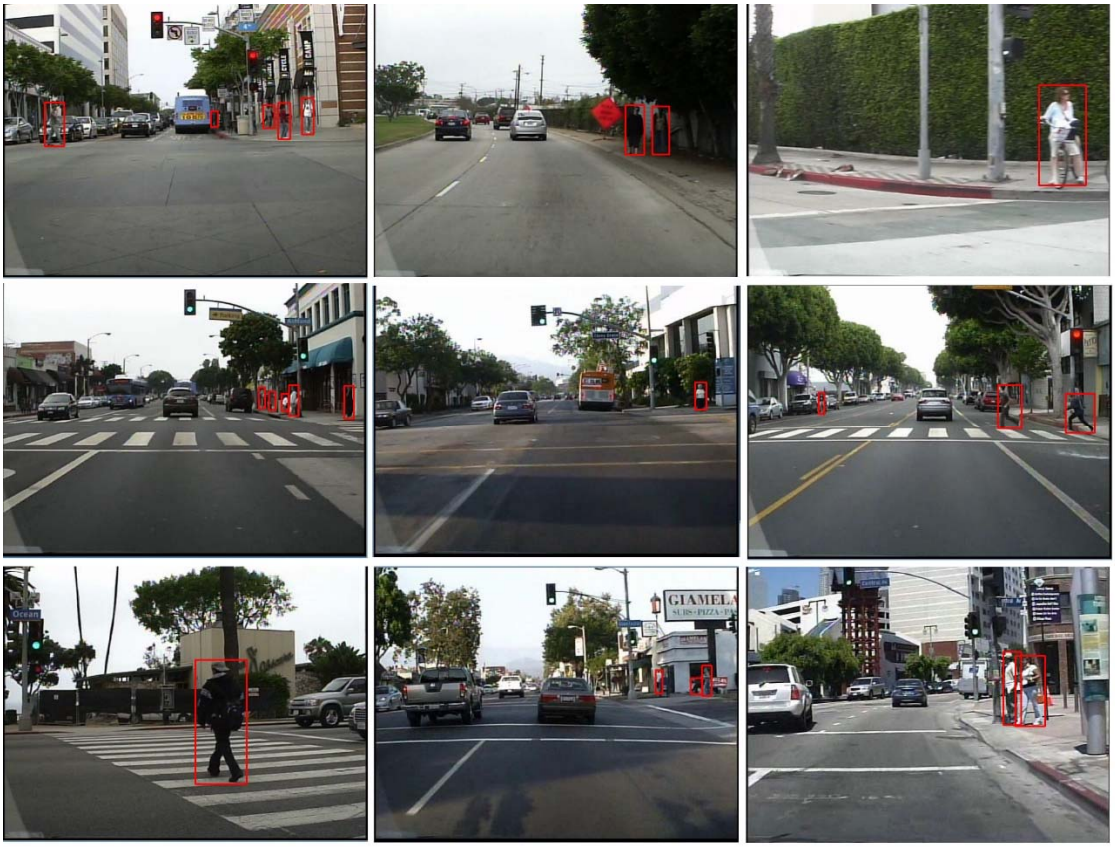

Figure 7: Detection Results of The Proposed Method on Caltech Pedestrian Dataset.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

specific, the performance of the proposed method is improved comparing with SDS-RCNN, SAF R-CNN, MultiSDP, RPN+BF, TA-CNN by 0.7%, 1.8%, 22.5%, 3.8%, 11.1% respectively. For the computational efficiency, the proposed method takes 0.4 second for processing an image with a low-end hardware machine.

Table 2: Comparison Experiment on Caltech Pedestrian Dataset.

Method	AP (%)
SDS-RCNN [7]	35.5
SAF R-CNN [8]	34.4
MultiSDP [9]	13.7
RPN+BF [10]	32.4
TA-CNN [11]	25.1
Proposed Approach	36.2

5. CONCLUSIONS

Vision-based pedestrian detection plays an essential role in many intelligent transportation systems. Due to difficult environments such as large-scale variation, heavy occlusion and small scale of pedestrian, existing deep CNN-based approaches for pedestrian detection still do not achieve very good accuracy over public benchmark dataset. Thus, this paper proposes an efficient framework based on Faster R-CNN for pedestrian detection, especially for small-scale pedestrians. First, ResNet-34 architecture is used for generating convolution feature maps. Then, deconvolutional modules are added after the base convolution layers, so the semantics from higher layers can be conveyed into lower layers to increase the representation capacity. Furthermore, different convolution layers of the base network are forwarded to RoI pooling layer to match proposals obtained from proposal generation. Finally, a modified loss function is adopted to increase the loss contribution from hard-to-detect examples. Experimental results on the Caltech pedestrian dataset show that the proposed approach achieved good detection performance compared to other state-of-the-art methods. In future, the author will extend the proposed method to detect other small objects in traffic scenes and explore a faster pedestrian detection algorithm.

REFRENCES:

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *NIPS*, 2015.

[2] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection", *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 2999-3007.

[3] Dollár P,Wojek C, Schiele B, Perona P, "Pedestrian detection: a benchmark", *2009 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2009, pp 304–311.

[4] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective", *Int. J. Comput. Vis.*, vol. 111, no. 1, 2015, pp. 98–136.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge", *Int. J. Comput. Vis.*, vol. 88, no. 2, 2010, pp. 303–338.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks", *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[7] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection segmentation", Jun. 2017, arXiv:1706.08564. [Online]. Available: https://arxiv.org/abs/1706.08564.

[8] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection", *IEEE Trans. Multimedia*, vol. 20, no. 4, 2018, pp. 985–996.

[9] X. Zeng, W. Ouyang, and X. Wang, "Multi-stage contextual deep learning for pedestrian detection", *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 121–128.

[10] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?", *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 443–457.

[11] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5079–5087.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[12] A. Prioletti, A. Møgelmose, P. Grisleri, M. M. Trivedi, A. Broggi, and T. B. Moeslund, "Part-based pedestrian detection and feature-based tracking for driver assistance: Real-time, robust algorithms, and evaluation", *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, 2013, pp. 1346–1359.

[13] S. Zhang, C. Bauckhage, and A. B. Cremers, "Efficient pedestrian detection via rectangular features based on a statistical shape model", *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, Apr. 2015, pp. 763–775.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[15] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, Aug. 2014, pp. 1532–1545.

[16] Huang, J., Rathod, V., Sun, C. *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors", *CVPR*, 2017.

[17] Fu C-Y, Liu W, Ranga A, Tyagi A, Berg AC, "DSSD: deconvolutional single shot detector", arXiv:1701.06659 [cs.CV]. http://arxiv.org/abs/1701.06659. Accessed 23 Jan 2017.

[18] Cai Z, Fan Q, Feris R S, Vasconcelos N, "A unified multi-scale deep convolutional neural network for fast object detection", *Lecture Notes in Computer Science*, vol 9908. Springer, Cham, 2016.

[19] Long J, Shelhamer E, Darrell T, "Fully convolutional networks for semantic segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 39(4), 2017, pp.640-651.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[21] He, K., Zhang, X., Ren, S., et al., "Deep residual learning for image recognition", *IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016, pp. 770–778.

[22] Liu, W., Anguelov, D., Erhan, D., et al., "SSD: Single Shot MultiBox Detector", *European Conf. on Computer Vision*, Amsterdam, Holland, 2016, pp. 21–37.

[23] L. Chen, Z. Zhang and L. Peng, "Fast single shot multibox detector and its application on vehicle counting system", *IET Intelligent Transport Systems*, vol. 12, no. 10, 2018, pp. 1406-1413.

[24] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[25] P. Dollr, Z. Tu, P. Perona, and S. Belongie, "Integral channel features", *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2009, pp. 7–10.

[26] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling", *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 32–39.

[27] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, "Seeking the strongest rigid detector", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3666–3673.

[28] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?", *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 613–627.

[29] W. Liu, B. Yu, C. Duan, L. Chai, H. Yuan, and H. Zhao, "A pedestrian-detection method based on heterogeneous features and ensemble of multi-view–pose parts", *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 813–824, Apr. 2015.

[30] S. Zhang, C. Bauckhage, and A. B. Cremers, "Efficient pedestrian detection via rectangular features based on a statistical shape model", *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 763–775, Apr. 2015.

[31] J. Baek, J. Kim, and E. Kim, "Fast and efficient pedestrian detection via the cascade implementation of an additive kernel support vector machine", *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 902–916, Apr. 2017.

[32] S. Zhang, W. Lin, P. Lu, W. Li, and S. Deng, "Kill two birds with one stone: Boosting both object detection accuracy and speed with adaptive patch-of-interest composition", *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 447–452.

[33] Sermanet P, Kavukcuoglu K, Chintala S, LeCun Y, "Pedestrian detection with unsupervised multi-stage feature learning" *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp 3626–3633.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[34] Du X, El-Khamy M, Lee J, Davis L, "Fused DNN: a deep neural network fusion approach to fast and robust pedestrian detection", *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp 953-961.

[35] X. Li et al., "A unified framework for concurrent pedestrian and cyclist detection", *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 2, pp. 269–281, Feb. 2017.

[36] Hanafi, Suryana, N., Basari, Abd Samad., "Deep learning for recommender system based on application domain classification perspective: A review", *Journal of Theoretical and Applied Information Technology*, vol. 96, 2018, pp. 4513-4529.

[37] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2129–2137.