PART-LEVEL FULLY CONVOLUTIONAL NETWORKS FOR PEDESTRIAN DETECTION

Xinran Wang, Cheolkon Jung

Xidian University
School of Electronic Engineering
Xian, Shaanxi, China

Alfred O Hero

University of Michigan
Electrical Engineering and Computer Science
Ann Arbor, Michigan, USA

ABSTRACT

Since pedestrians in videos have a wide range of appearances such as body poses, occlusions, and complex backgrounds, pedestrian detection is a challengeable task. In this paper, we propose part-level fully convolutional networks (FCN) for pedestrian detection. We adopt deep learning to deal with the proposal shifting problem in pedestrian detection. First, we combine convolutional neural networks (CNN) and FCN to align bounding boxes for pedestrians. Then, we perform part-level pedestrian detection based on CNN to recall the lost body parts. Experimental results demonstrate that the proposed method achieves 6.83% performance improvement in log-average miss rate over CifarNet.

Index Terms— Bounding box alignment, convolutional neural network, deep learning, fully convolutional networks, part detection, pedestrian detection

1. INTRODUCTION

Pedestrian detection is a key problem for visual surveillance, automotive safety, and robotics applications. Pedestrians in videos have a wide variety of appearances: Body poses, occlusions, clothing, lighting, and complex backgrounds. Pedestrian detection in videos has received much attention by researchers in the computer vision field [1, 2, 3, 4, 5]. Part-based detection is able to deal with the occlusion problem in pedestrian detection. Felzenszwalb *et al.* [6, 7] proposed a star model to search the whole image for detecting body parts. These works [8, 9, 10, 11, 12] inspired researchers to consider part detection in pedestrian detection. Deep learning has a relatively short history in pedestrian detection. The first work was an unsupervised model for limited labeled training data proposed by Sermanet *et al.* [13]. Then, a series of research works [8, 9, 10, 11] considered part detection in their deep model. DBN-Isol [8] extended the deformable parts model (DPM) [6, 7] with a deep belief network to estimate visibility of a pedestrian. JointDeep [9] was a deep model to generate feature extraction, occlusion handling, deformation and classification in single network. SDN [11] used a switchable Restricted Boltzmann Machines (RBMs) variant to extract high-level features for body parts. Some successful general object detectors [3] were also applied to pedestrian detection tasks. Hosang *et al.* [14] analyzed the feasibility of an R-CNN framework for pedestrian detection. Considering part detection, Tian *et al.* [12] proposed a body part pool that the detector could map templates onto various occluded samples. However, it is still hard to apply common object detection methods to the pedestrian detection task. [3] obtained detection proposals by semantic segmentation [15], but required high computational costs for pedestrian detection. It produced thousands of detection proposals per image, and thus made pedestrian detection difficult. Thus, current convolutional neural network (CNN) detectors for pedestrian detection [8, 9, 10, 11, 16, 17, 12, 18, 14] have employed detection proposals to guide pedestrian detection, thereby preventing redundant exhaustive search on images. JointDeep [9] and SDN [11] used "HOG+CSS" as features and a Linear SVM as the classifier to generate detection proposals (HOG: Histogram of oriented gradient, CSS: Color-self-similarity). The "HOG+CSS+SVM" detector recalled most pedestrian candidates from images. Also, the performance of the CNN detector was improved by hard negatives because the false positives are simultaneously hard negative samples for training. Other detection proposals were generated by ACF [19], LDCF [20], SquaresChnFtrs [21], and checkerboards [22]. Based on the careful observations on detection proposals, we have found that there exists the proposal shifting problem

Fig. 1. Proposal shifting examples. Colored boxes are detection proposals, while black boundaries represent ground truth.

This work was supported by the National Natural Science Foundation of China (No. 61271298) and the International S&T Cooperation Program of China (No. 2014DFG12780).

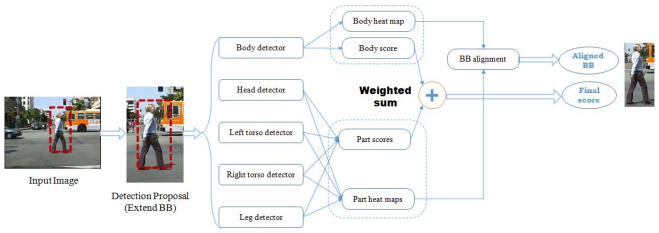

Fig. 2. Whole framework of the proposed method.

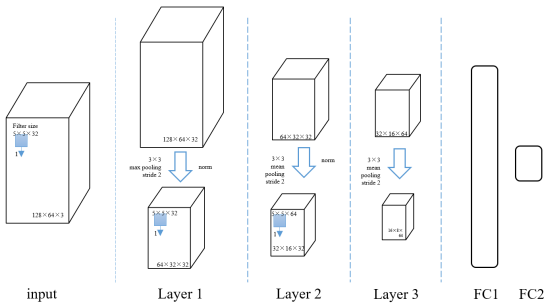

Fig. 3. Architecture of the CNN.

(PSP) in detection proposals which causes the loss of body parts. As shown in Fig. 1, the detection proposal has various types of PSP, and sometimes contains no head or leg. Thus, part-based proposal alignment is needed for accurate pedestrian detection. In this paper, we propose part-level FCNs for pedestrian detection. We combine CNN and FCN to obtain the confidence map and align detection proposal. Our pedestrian detector is composed of one body detector and four part detectors. We combine confidence maps from body and part detectors to estimate the pedestrian location by their FCNs. Based on the combined confidence map, we perform bounding box (BB) alignment to obtain accurate pedestrian locations from detection proposals. Finally, we combine body and part scores to produce the detection score. Fig. 2 illustrates the whole framework of the proposed method. Compared with existing methods, our main contributions are as follows: (1) We design a part-level detection framework (one body detector and four part detectors) to recall the lost body parts in detection proposals; (2) We utilize FCN to produce the confidence map from each detector and combine CNN with FCN for BB alignment. Therefore, we achieve 6.83% performance improvement in log-average miss rate over CifarNet.

2. PROPOSED METHOD

2.1. Training Process: CifarNet

CifarNet is originally designed to solve the CIFAR-10 classification problem [23] which has 60000 32×32 color images in 10 classes. According to Hosang *et al.*'s work [14], this network has a fair performance on Caltech dataset. We employ the vanilla convolutional network as the base model for our part-level detector. We modify the network by adding one fully connected layer `FC1` under the top layer. The input window size varies with different human body parts. The input size of the head detector is 32×32 pixels, while that of the body detector is 128×64 pixels. According to [6, 7], all pedestrian windows are divided into four parts, and thus we get body part labels. We crop small windows of 32×32 for head, left torso, right torso, and 64×64 for leg from the pedestrian window. As shown in Fig. 3, we train five CifarNets for five part detectors using CNN architecture. The CNN architecture consists of 3 convolutional layers, 3 pooling layers, 1 fully connected layer, and 1 output layer. The input size of each CifarNet is various by the cropped window size. Raw images are used as input for our CifarNets. On the top of CifarNet, we set a softmax layer to get a confidence score, which indicates the probability of a pedestrian/body part in BB.

2.2. Bounding Box Alignment

Pedestrian detectors suffer from PSP which loses some of the body parts in detection proposals and causes low confidence score and bad pedestrian localization. In this work, we perform BB alignment based on FCN to deal with PSP. For BB alignment, we first extend the detection proposal size to contain more regions. We enlarge the detection proposal from 128×64 pixels to 160×80 pixels, i.e. 32 and 16 pixels in height and width, respectively. We combine CifarNet with FCN to get the confidence map, named as CifarNet-FCN. The size of input images is supposed to be fixed for CifarNet. Based on the trained CifarNet, we change the shape and dimension of the weights between `pool3` and `fc1` to make these weight matrix convolute with a larger feature map. In CifarNet for body detector, i.e. body-CifarNet, we obtain the convolution weights by rolling the inner product weights (8192×16 for `fc1` and 16×2 for `fc2`) into $16 \times 8 \times 64 \times 16$ (height $\times$ width $\times$ input channel $\times$ out channel) for `fc1-conv` and $1 \times 1 \times 16 \times 2$ for `fc2-conv`. Since we do not change the magnitude of the weights, the perceptual field for body-FCN is not changed (128×64 pixels). That is, 1 pixel in `fc1-conv` maps 2^3 pixels in CifarNet-FCN on the input image window and there are total $s = 8$ pixels between every stride. The output of FCN is a heat map with confidence scores with a size of 5×3. A heat map with a higher resolution is needed to localize the pedestrian in the enlarged regions. We shift the proposal by f steps on the horizontal and vertical axis uniformly to make total distance no more than 8 pixels. That is, shift distance is s/f. Also, body-CifarNet generates a 5×3 heat map by every step, and we interlace all f^2 outputs together according to the relative direction of ev-

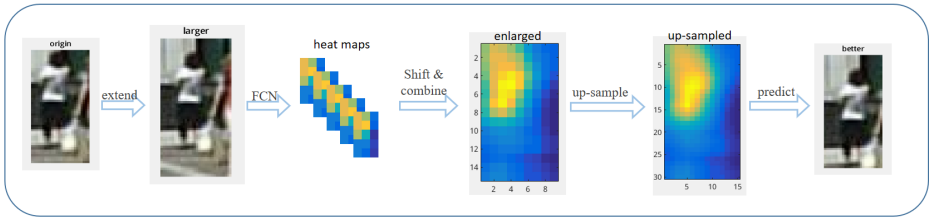

Fig. 4. Pipeline of BB alignment. Warmer color indicates higher score for a pedestrian. Origin: the original BB. The person lays on the top left corner of the BB. Larger: an enlarged BB. Heat map: the output of the FCN. Enlarged: shift the input image by f (=3) pixels on 2 directions. Up-sampled: up-sample the heat map into a corresponding size. Better: a better aligned bounding box is predicted.

ery shift-and-stitch. As a result, a $(5 \cdot f) \times (3 \cdot f)$ heat map is populated. Once we get a heat map with a higher resolution, we apply a simple up-sampling method to produce the heat map with nice aspect ratio and determine the shifting direction easily. Shifting direction is calculated without a stretch operation. We use a enlarging ratio parameter L to determine the size of the target rectangle. Width/height of the rectangle w/h is obtained by multiplying L with the width/height of the input region W/H as follows:

$$w/h = L \cdot W/H \tag{1}$$

Define the coarse position in the input region as (x_p, y_p) and the original position as (x_o, y_o). Then, we update x as follows:

$$\Delta x = \frac{2 \times \sum_{i=1}^{n} (score_i^t - score_i^o)^2}{\sum_{i=1}^{n} score_i^{t\,2} + \sum_{i=1}^{n} score_i^{o\,2}} * (x_p - x_o) \tag{2}$$

where $score_i^t$ is the value of the ith element in the target rectangle in the score heat map, $score_i^o$ is the value of the ith element in the original rectangle, n is the total number of elements in the rectangles. We obtain Δy in the same way, and the position of the detection proposal is updated by

$$x_a = x_o + \Delta x \tag{3}$$

y_a is also updated, and the updated position of the detection proposal (x_a, y_a), i.e. anchor position. Based on (x_a, y_a), the proposed part-level detector is operated to yield part scores and part positions.

2.3. Score Merging

We perform detection on the aligned BB. We obtain four part scores from part detectors, and determine the final detection score as follows:

$$score = score_{root} + \sum_{i=\{parts\}} w_i * (score_i + P_i) \tag{4}$$

where $score_{root}$ is the output score of the body detector; $score_i$ is the output score of four part detectors; w_i is the weight that indicates the importance of part scores, and we set $\sum_{i=\{parts\}} w_i = 1$; and P_i is the penalty term of the spatial distance between anchor position and part position:

$$P = a*(|x_p - x_a| + |y_p - y_a|) + b*(|x_p - x_a|^2 - |y_p - y_a|^2) \tag{5}$$

where a and b are weights which balances the orientation and geometrical shifting distance; and (x_a, y_a) is the anchor position which is the position of an aligned detection proposal.

3. EXPERIMENTAL RESULTS

We evaluate the proposed method on the Caltech pedestrian detection dataset [24]. Caltech dataset consists of approximately 10 hours of 640×480 30Hz videos taken from a vehicle driving through regular traffic in an urban environment. About 250,000 frames (in 137 approximately minute long segments) with a total of 350,000 BBs and 2,300 unique pedestrians were annotated. We use an Nvidia GTX-980 GPU with 4GB graphics memory and the deep learning framework caffe with matcaffe (MATLAB-caffe) interface in our experiments. Because all samples are annotated in every frame, we use every 3rd frame instead of every 30th frame to extract training data [14, 20]. We use SquaresChnFtrs [21] for proposal detection. There are total $2 \cdot 10^4$ annotated pedestrians in 42,782 frames, i.e. positive samples, while the number of negative samples is the same as positive ones. In Caltech pedestrian dataset [24], every frame has two BBs: One BB indicates the full extent of the entire body (BB-full), and the other BB is for visible region (BB-vis). For part detectors, we only select BB-vis for part division to avoid collecting background regions into positives. More than 70% pedestrians are occluded in at least one frame in Caltech dataset. In the part training samples, we select visible area in detection proposals whose IoU is higher than 0.5 as positive samples, while we select occluded areas in detection proposals whose IoU is lower than 0.5 as negative samples. Thus, partially occluded samples are fully used to increase the number of training data. The detection system returns a BB and a confidence score for each detection proposal. We use the

Table 1. Performance comparison between CifarNet, CifarNet-FCN, and Part-FCNs

Method	Avg. miss rate (%)	Improvement(%)
CifarNet	29.35	-
CifarNet-FCN	26.27	3.08
Part-FCNs	*22.52*	3.75

Fig. 5. BB alignment results. (a)(c)(e)(g) Detection proposals from SquaresChnFtrs. (b)(d)(f)(h) Their aligned proposals by the proposed method.

standard setting in Caltech dataset. Log-average miss rate is to summarize the detector's performance, which is computed by averaging miss rate at nine false positive per image(FPPI) rates evenly spaced in log-space in the range 10^{-2} to 10^{0}. In our experiments, we follow most of the settings on CifarNet in [14]. Compared with other deep models, CifarNet has a relatively small parameter set compared with AlexNet. To evaluate every stage of the part-level FCN which consist of a pipeline of body detection, BB alignment, and part-level detection. We evaluate the performance of body-CifarNet by a BB that is not aligned by FCN and a body score. Then, BB alignment improves the performance of CifarNet-FCN. Thanks to the BB alignment, CifarNet-FCN produces more accurate BB and a higher response score for each pedestrian. We apply part-level detection to larger detection regions with aligned BBs. Thus, we achieve more than 3% increase than CifarNet-FCN as shown in "Part-FCNs" of Table 1. The performances of our 3-stage pipeline are shown in Table 1. We provide some examples of BB alignment in Fig. 5.
We compare our part-level FCN with state-of-the-art deep learning methods including DBN-Isol [8], DBN-Mut [10], MultiSDP [25], JointDeep [9], SDN [11], CifarNet [14] and AlexNet [14]. Fig. 6 shows the performance comparison between the proposed method and other deep learning ones.

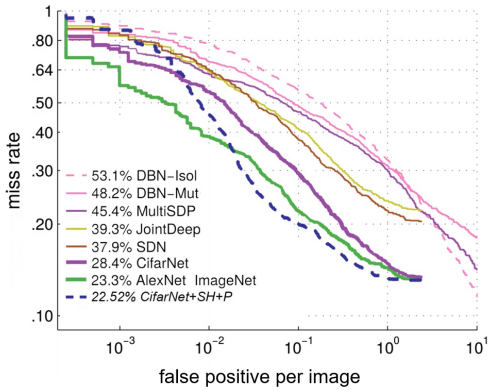

Fig. 6. Performance comparison between the proposed method and other deep learning ones.

Table 2. Performance comparison between different detection proposal methods on Caltech dataset. MS: Log-average miss-rate. CSS: Color-self-similarity.

Method	Detection Proposal	MS(%)
"HOG+CSS+SVM"	-	45.5
Proposed	"HOG+CSS+SVM"	36.5
SquaresChnFtrs	-	34.8
Proposed	SquaresChnFtrs	30.7

Under the small number of layers, i.e. 3 layers [14], the proposed method achieves the best performance due to CifarNet, part detection, and BB alignment. We also provide the performance comparison between different detection proposal methods on Caltech dataset in Table 2. We evaluate two detection proposal methods of "HOG+CSS+SVM" [9] and SquaresChnFtrs [21]. As listed in the table, SquaresChnFtrs outperforms "HOG+CSS+SVM" in pedestrian detection. Moreover, the proposed method improves the performance of both detection proposal methods about 9% and 3.9% over "HOG+CSS+SVM" and SquaresChnFtrs, respectively. Therefore, it can be safely concluded that the proposed method can be effectively applied to the pedestrian detection task.

4. CONCLUSION

In this paper, we have proposed an FCN for pedestrian detection. We have utilized BB alignment to recall the lost body parts. We have generated the confidence map using FCN, and estimated accurate position of pedestrians based on the map. Thanks to the BB alignment and part-level detection, we have achieved 6.83% performance improvement in the average miss rate over CifarNet.

5. REFERENCES

[1] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2005, vol. 1, pp. 886–893.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[4] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[6] P. F. Felzenszwalb, D. Mcallester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," 2008.

[7] P. F. Felzenszwalb, R. Girshick, D. Mcallester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[8] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3258–3265.

[9] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2056–2063.

[10] W. Ouyang, X. Zeng, and X. Wang, "Modeling mutual visibility relationship in pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3222–3229.

[11] P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable deep network for pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 899–906.

[12] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1904–1912.

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

[14] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4073–4082.

[15] J. R. R. Uijlings, Koen EA K. E. A. Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[16] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5079–5087.

[17] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, and D. Ferguson, "Real-time pedestrian detection with deep network cascades," 2015.

[18] J. Li, X. Liang, S. Shen, T. Xu, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *arXiv preprint arXiv:1510.08160*, 2015.

[19] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[20] W. Nam, P. Dollar, and J. H. Han, "Local decorrelation for improved detection," *arXiv preprint arXiv:1406.1134*, 2014.

[21] R. Benenson, M. Mathias, T. Tuytelaars, and L. V. Gool, "Seeking the strongest rigid detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3666–3673.

[22] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 1751–1760.

[23] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[24] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[25] X. Zeng, W. Ouyang, and X. Wang, "Multi-stage contextual deep learning for pedestrian detection," pp. 121–128, 2013.