Mask-Streaming CNN for Pedestrian Detection

Peilei Dong, Wenmin Wang *, Mengdi Fan, Ronggang Wang, Ge Li

School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University
Lishui Road 2199, Nanshan District, Shenzhen, Guangdong Province, China 518055
pldong@pku.edu.cn, wangwm@ece.pku.edu.cn, fanmengdi@sz.pku.edu.cn, {rgwang,gli}@pkusz.edu.cn

***Abstract*—Inspired by humans recognition of pedestrians, we propose a mask-streaming convolutional neural network (CNN) for pedestrian detection. The mask stream consists of one original region proposal and six masked regions. These masked regions, which aim at highlighting significantly discriminative semantic characteristic of head, body parts and contextual information, are generated through wiping off a part of the original region by 'masks'. We feed the mask stream into the network to generate features of each masked region, then a concatenate layer integrates these features as the final representation of pedestrians. We evaluate the proposed model on the challenging Caltech and Eth datasets. Our approach performs better than the baseline, and achieves competitive performance compared to numerous pedestrian detection methods.**

***Index Terms*—Pedestrian Detection, Mask Stream, CNN, Masked Region, Semantic Characters**

I. Introduction

Pedestrian detection, the aim of which is to detect pedestrian instances and locate their positions in images, has attracted widespread attention in the past few years. Accurate pedestrian detection could be applied to many fields, including automotive safety, video surveillance and robotics. Despite the considerable efforts of computer vision researchers, it is still a problem to be further solved due to the complexity of pedestrian detection in reality.

Recently, deep learning has been widely used for object detection. The most notable model is the work of Girshick [1] with the R-CNN framework. Girshick's framework relies on two critical steps: It designs an algorithm to generate a set of region proposals and then the candidate proposals are sent to the convolutional neural network (CNN). Different from general object detection, pedestrian detection is far more challenging on account of the complex backgrounds. In the process of humans recognition for the pedestrian, we always see the whole pedestrian at first glance. Then we observe the characters of head, body parts, legs and the surroundings. The characters of these parts have significantly different features from other objects. Therefore not only consider the feature of the entire pedestrian, we should also highlight the sematic characters of clearly distinguishable parts of a pedestrian.

In this paper, we propose a novel mask-streaming CNN to achieve the above assumptions. We take the pipeline of [1] as the baseline of our work. Moreover our model generates multiple masked regions through wiping off a part of the original regions by "masks". Each masked region represents one peculiar characteristic of pedestrians or contextual information. A sequence of them originated from the same region proposal and made available over time in our framework is defined as one *mask stream*. The architecture of our model is shown in Fig. 1. Firstly we use the region proposal algorithm to generate a set of region proposals. Then our model generates other six masked regions for each region. We combine the original region and six masked regions as one mask stream, and feed the raw images with a set of mask streams to the convolutional neural network. Each masked region is projected in the convolutional feature maps. The features of each masked region in a mask stream are pooled into fixed-length vectors by region of interest (RoI) pooling, and then they are passed into two fully connected layers one by one. A concatenate layer is utilized to integrate these features to the final representations of this mask stream. Finally, there are two sibling output layers followed the concatenate layer. One of them is softmax probability estimates over the pedestrian class, and the other is four real-valued numbers representing the bounding-box positions of pedestrians.

We evaluate the proposed model on the challenging Caltech and Eth datasets. Our method yields better result than [1], and achieves competitive performance compared to numerous pedestrian detection methods. It achieves 23.7% miss-rate on the Caltech dataset and 37.4% miss-rate on the Eth dataset, respectively.

The rest paper is organized as follows. Section II briefly introduces related works. Section III describes details of the mask-streaming CNN model. The following section presents the results of our experiments. Finally, section V concludes this paper.

II. RELATED WORK

Current methods for pedestrian detection can be generally grouped into two categories. The first category is known as conventional approaches, which extract hand-crafted features from images to train SVM or boosting classifiers. Dalal et al. proposed the Histograms of Oriented Gradients (HOG) descriptor in [2]. Wang et al. [3] combined HOG and Local Binary Pattern (LBP) as the feature set to handle partial

This project was supported by Shenzhen Peacock Plan (20130408-183003656), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (ZDSYS201703031405467), and Guangdong Science and Technology Project (2014B010117007).

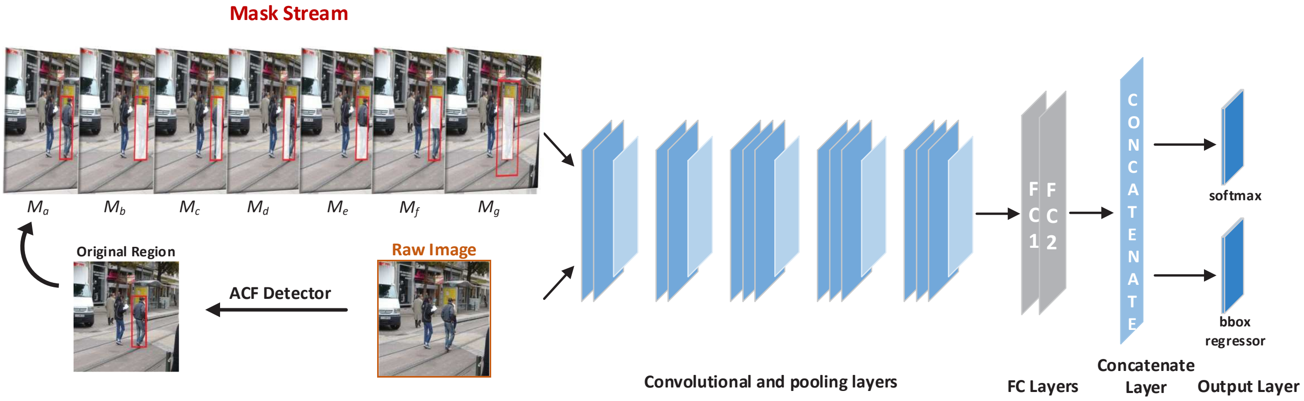


Fig. 1. Our mask-streaming CNN architecture. The red bounding box is the original region, and the gray mask is wiped off. We use the ACF detector to generate the region proposals for the image. We generate six masked regions (M_b~M_g) and combine them with the original region (M_a) as the mask stream of one region proposal. The ratio of each masked region is explained in section 3.2. The raw image with the mask stream of each region proposal are input into the CNN. We get the features of each masked region by convolutional and pooling layers. The features of a masked region are pooled into fixed-length vectors by RoI pooling and fed into fully connected (FC) layers one by one. The concatenate layer integrates the features of seven masked regions as final representation of the mask stream. The network has two output vectors: softmax probabilities and bounding-box regression offsets.

occlusion. Felzenszwalb et al. proposed the deformable part-based models (DPM) [4] to consider the appearance of each part and the deformation among parts for detection. Dollár et al. proposed Integral Channel Features (ICF) in [5] and Aggregated Channel Features (ACF) [6], which can be effective used to detection.

Recently, deep models are also applied to pedestrian detection. Deep learning methods improve the performance of pedestrian detection owing to learning features from raw images. Ouyang et al. [7] learned features and the visibility of different body parts through a discriminative deep model to handle occlusion. ConvNet [8] used an unsupervised method based on convolutional sparse coding to unsupervised pre-train CNN for pedestrian detection. The DeepParts method [9] improved the detection performance by handling occlusion with an extensive part pool. Different from the above deep methods using only one original region type for a pedestrian, we also consider other semantic parts of the pedestrian as significantly discriminative characters and highlight them through the mask stream in the network.

III. PROPOSED METHOD

A. Region Proposal Algorithm

We adopt the ACF detector [6] to generate the original regions. The ACF detector is a fast and effective sliding window detector. It extracts 10 channels of hand-crafted features and trains an AdaBoost classifier. Different from other region proposal methods [10], [11], [12] for detecting generic objects, the ACF detector can be trained to detect objects of a specific category, which can be used for extracting and mining the high-quality region proposals. In this work, we train 2048 depth-two trees combining the region detector for the pedestrian class. And we low the detection threshold to generate sufficient region proposals.

B. Mask Stream

We take the original region proposals as the base regions and generate other six masked regions. We combine the original region and six masked regions as one mask stream. All the masked region types we employ are described as follows.

The original region is the candidate detection box being used in [1]. We define this original region as:

$$M_a = (x, y, w, h) \tag{1}$$

where (x, y) represents the top-left position of the region in the image and (w, h) represents its width and height. The original region is used to lead the network to capture the appearance information of the entire pedestrian.

The masked region of head is the head part of one pedestrian. According to the average ratio between head and body of humans, we define this mask as :

$$M_b = (x, y, w, 0.125h) \tag{2}$$

This masked region type is adopted to highlight the semantic character of pedestrian head.

The semi masked regions are the left, right, up and bottom half parts of the base region. They are defined as:

$$M_c = (x, y, 0.5w, h) \tag{3}$$

$$M_d = (x + 0.5w, y, 0.5w, h) \tag{4}$$

$$M_e = (x, y, w, 0.5h) \tag{5}$$

$$M_f = (x, y + 0.5h, w, 0.5h) \tag{6}$$

M_c and M_d guide the network to learn the appearance characteristics present in left and right half parts of the pedestrian, aiming to make the representation more robust with respect to occlusions. M_e and M_f teach the network to learn the semantic information of top half body, and highlight the legs of pedestrian which is also one of the clearly distinguishing features of pedestrians.

There is one type of contextual masked region that has rectangular ring shape. We enlarge the original region 1.8 times to get the contextual part, which is defined as:

$$M_g = (x - 0.4w, y - 0.4h, 1.8w, 1.8h) - (x, y, w, h) \quad (7)$$

It drive the network to focus on the contextual appearance that surrounds the pedestrian such as the appearance of background or of other objects next to it.

C. Architecture of the Network

In this section we first present the network architecture, then we specify the loss function of this network.

The network has five convolutional phases. The first two phases consist of two convolutional layers and the others are three convolutional layers. We utilize a rectified linear unit (ReLU) as the nonlinear activation function for each convolutional layer. Each convolutional phases is followed by a space max pooling layer. We put the mask streams into the network and map each masked region on the feature maps. The features of each masked region in a mask stream are reshaped to 4096-length vectors by RoI pooling, and then they are passed to fully connected layers one by one. The fully connected layers consisting of 4096 nodes are initialized from zero-mean Gaussian distributions with standard deviations 0.01 and 0.001, respectively. Biases are initialized to 0. The concatenate layer integrates the features of seven masked regions to 28672-length vector end to end.

The concatenate layer is followed by two sibling output layers. The first outputs probability values, $p = (p_0, p_1)$, over "background" and pedestrian classes. p_0 and p_1 denote the probability values of "background" and pedestrian classes respectively. As usual, p is computed by a softmax over the 2 outputs of a fully connected layer. The output of second sibling layer is bounding-box regression offsets, $t = (t_x, t_y, t_w, t_h)$ for the pedestrian class. Each training region proposal is labeled with a ground-truth class u and a ground-truth bounding-box regression target v. We use a multi-task loss L on each labeled proposal to jointly train for classification and bounding-box regression:

$$L(p, u, t^u, v) = L_{cls}(p, u) + \lambda[u \geq 1] L_{loc}(t^u, v) \quad (8)$$

in which $L_{cls} = -logp_u$ is log loss for true class u. The second task loss, L_{loc}, is defined over the true bounding-box regression targets for class u, $v = (t_x, t_y, t_w, t_h)$, and a predicted tuple (formula) again for class u. The Iverson bracket indicator function $[u \geq 1]$ evaluates to 1 when $u \geq 1$ and 0 otherwise. By convention the catch-all background class is labeled $u = 0$. For background RoIs, there is no notion of a ground-truth bounding box and hence L_{loc} is ignored. For bounding-box regression, we use the loss:

$$L_{loc}(t^u, v) = \sum_{i \in \{x,y,w,h\}} smooth_L(t_i^u - v_i) \quad (9)$$

where the robust loss $smooth_L(\cdot)$ is defined as:

$$smooth_L(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \quad (10)$$

The parameter λ controls the balance between the two task loses. The ground-truth regression targets v is normalized to have zero mean and unit variance. And in all experiments, we set $\lambda = 1$. The minimization of loss function is achieved with the stochastic gradient descent (SGD) algorithm.

IV. EXPERIMENTS

The experiments are conducted on two public pedestrian datasets: the Caltech benchmark and the Eth dataset.

A. Datasets and Evaluation protocol

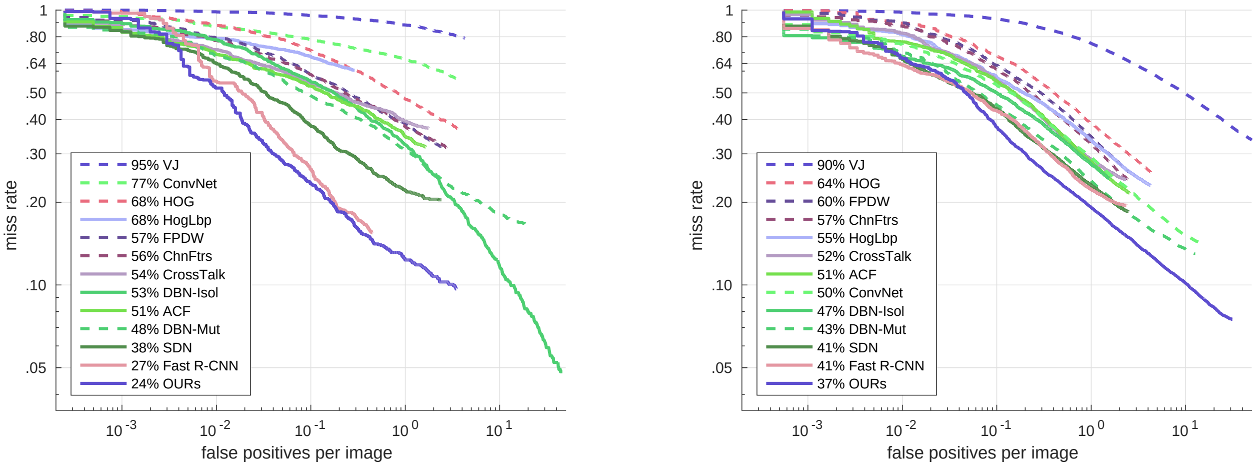
Datasets The Caltech dataset is one of the most popular pedestrian detection datasets. It consists of about 10 hours of 30 Hz videos collected from a vehicle driving through urban traffic. The 10 hours data consists of 11 videos with the first 6 videos are used for training and the last 5 videos for testing. To enlarge the number of training samples, we sample a frame from every 3 frames instead of every 30 frames. The Eth dataset was recorded using a pair of AVT Marlins F033C mounted on a chariot respectively a car, with a resolution of 640×480. The total number of images is 1804 in three testing sets. We follow the commonly-used methods [13] to use the Inria dataset to train our model and test on the Eth dataset.

Evaluation protocol To evaluate the detection results, we use the bounding boxes labels and the evaluation software (version 3.2.1) provided by Dollar et al. on the website. The per-image evaluation methodology is adopted, i.e. all the detection results are compared using miss rate vs. False-Positive-Per-Image (FPPI) curves. The log-average miss rate is also used to summarize the detection performance, and is computed by averaging the miss rate at nine FPPI points[2] that are evenly spaced in the log-space in the range from 10^{-2} to 10^0. In order to validate the effectiveness of the proposed approach, the following experiments will be mainly conducted on the most popular Reasonable subset (pedestrians of ≥ 50 pixels high, fully visible or less than 35% occluded).

B. Implementation Details

We adopt the pre-trained ImageNet VGG16 model [14] to initialize the weights of the convolutional and pooling layers. Regarding one masked region that is rectangular ring, both the inner and the outer box are projected on the feature maps and then the activations that lay inside the inner box are masked out by setting them to zero. Following the guidelines of Fast R-CNN [1], the network is trained with SGD with momentum of 0.9, and weight decay of 0.0005. And all layers use a per-layer learning rate of 1 for weights and 2 for bias and a global learning rate of 0.001. Each of the two fully-connected layers is followed by a drop-out operation with a ratio of 0.5 to combat over fitting. The minibatch has 128 samples. The positive samples are defined as the region proposals that overlap a ground-truth bounding box by at least 0.5. As negative samples we take the region proposals that overlap with a ground-truth bounding box on the range [0.1, 0.3). The labelling of the training samples is relative to the original

www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/



(a) Comparison on the Caltech dataset

(b) Comparison on the Eth dataset

Fig. 2. The comparison of ours with the state-of-the-art methods for pedestrian detection on (a) Caltech dataset and (b) Eth dataset respectively. The Mask-Streaming CNN outperforms most of other methods including Fast R-CNN.

candidate boxes and is the same across all the masked region types. The average amount of region proposals of each image in our model is 500.

The network is implemented based on the publicly available Caffe platform [15]. The whole network is trained on NVIDIA Tesla GPU K80 with 12GB memory.

C. Comparison with the State-of-the-arts

We compare our framework with hand-crafted models such as VJ, ConvNet, HOG, HogLbp, FPDW, ChnFtrs, CrossTalk, ACF. And we also compare with deep models including DBN-lsol, DBN-Mut and SDN. The overall experimental results are reported in Fig. 2. Our method significantly outperforms most of pedestrian methods, achieving 23.7% miss-rate on the Caltech dataset (Fig. 2a). We also obtain a competitive result on the challenging Eth dataset, where the miss-rate is reduced to 37.4% (Fig. 2b).

Moreover, we compare our model with the baseline framework Fast R-CNN [1], which only consists of the original region. We take the region proposals generating from the ACF detector instead of selective search as input for consistency's sake. It can be observed in Fig. 2 that the proposed model achieves lower log-average miss-rate than Fast R-CNN both on the Caltech dataset and Eth dataset. The experimental results demonstrate that the semantic information of mask stream facilitate the outstanding performance in pedestrian detection.

V. Conclusion

In this paper, we present a novel mask-streaming CNN for pedestrian detection. The mask stream is composed of one original region with six masked regions to highlight the discriminative semantic characters of pedestrian parts and context. Our model extracts the features of each masked region, and they are integrated by the concatenate layer. We illustrate the details of each masked region and the architecture of our model. The experimental results demonstrate that the proposed model achieves competitive results compared to plentiful pedestrian detection methods, and performs better than the baseline convolutional neural network (Fast R-CNN) in pedestrian detection task. In the future, we will continue to expand our model, and discuss the effect of each semantic part for pedestrian detection in details.

References

[1] R. Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. IEEE, 2005, pp. 886–893.

[3] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *ICCV*, 2009, pp. 32–39.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[5] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *BMVC*, 2009.

[6] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *TPAMI*, vol. 36, no. 8, pp. 1532–1545, 2014.

[7] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *CVPR*. IEEE, 2012, pp. 3258–3265.

[8] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *CVPR*. IEEE, 2013, pp. 3626–3633.

[9] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *CVPR*. IEEE, 2015, pp. 1904–1912.

[10] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

[11] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*. IEEE, 2014, pp. 3286–3293.

[12] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014, pp. 391–405.

[13] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *ICCV*, 2013, pp. 2056–2063.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM MM*, 2014, pp. 675–678.