Pedestrian Detection aided by Deep Learning Semantic Tasks

Yonglong Tian[1], Ping Luo[3,1], Xiaogang Wang[2,3], Xiaoou Tang[1,3]
[1]Department of Information Engineering, The Chinese University of Hong Kong. [2]Department of Electronic Engineering, The Chinese University of Hong Kong. [3]Shenzhen Key Lab of CVPR, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.

State-of-the-art methods for pedestrian detection can be generally grouped into two categories, the models based on hand-crafted features and deep models. The first is to extract Haar [6], HOG [1], or HOG-LBP [7] features and train SVM [1] or AdaBoost classifiers [2], where the two stages cannot be jointly optimized to improve performance. In the second category, deep neural networks achieved promising results [3, 4, 5], owing to their capacity to learn discriminative features from raw pixels. While previous treated pedestrian detection as a single binary classification task, which are not able to capture rich pedestrian variations, this paper proposed a novel task-assistant CNN (TA-CNN) to jointly optimize detection with auxiliary semantic tasks, including pedestrian attributes and scene attributes. Fig.1 is an illustration for how this idea works. If only a single detector is used to classify all the positive and negative samples in Fig.1 (a), it is difficult to handle complex pedestrian variations. Therefore, the mixture models of multiple views were developed in Fig.1 (b), i.e. pedestrian images in different views are handled by different detectors. If views are treated as one type of semantic tasks, learning pedestrian representation by multiple attributes with deep models actually extends this idea to extreme.

All attributes are summarized in Fig.2. Given a pedestrian dataset **P**, for example Caltech, we manually label the positive patches with nine pedestrian attributes. For background, we transfer hard negative patches with attribute information from three public scene segmentation datasets to **P**, including CamVid ($\mathbf{B}^a$), Stanford Background ($\mathbf{B}^b$), and LM+SUN ($\mathbf{B}^c$). As shown in Fig.2, pedestrian attributes only present in **P**, shared attributes present in all **B**'s, and the unshared attributes present in one of them.

We construct a training set **D** by combing patches cropped from both **P** and **B**'s. Let $\mathbf{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ be a set of image patches and their labels, where each $\mathbf{y}_n = (y_n, \mathbf{o}_n^p, \mathbf{o}_n^s, \mathbf{o}_n^u)$ is a four-tuple. Specifically, y_n denotes a binary label, indicating whether an image patch is pedestrian or not. $\mathbf{o}_n^p = \{o_n^{pi}\}_{i=1}^9$, $\mathbf{o}_n^s = \{o_n^{si}\}_{i=1}^4$, and $\mathbf{o}_n^u = \{o_n^{ui}\}_{i=1}^4$ are three sets of binary labels, representing the pedestrian, shared scene, and unshared scene attributes, respectively. Then TA-CNN can be formulated as minimizing the log posterior probability with respect to a set of network parameters $\mathcal{W}$

$$\mathcal{W}^* = \arg\min_{\mathcal{W}} -\sum_{n=1}^{N} \log p(y_n, \mathbf{o}_n^p, \mathbf{o}_n^s, \mathbf{o}_n^u | \mathbf{x}_n, \mathbf{z}_n; \mathcal{W}), \qquad (1)$$

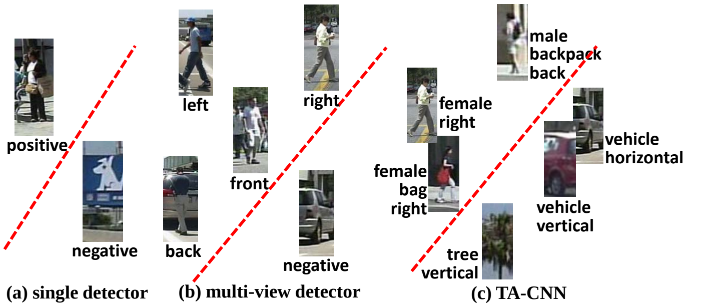
where $\mathbf{z}_n$ is named as structural projection vector and designed within a tree clustering structure to bridge the visual gaps between datasets **P** and **B**'s.

To learn network parameters $\mathcal{W}$, we reformulate Eqn.(1) as optimizing a single multivariate cross-entropy loss,

$$E = -\mathbf{y}^{\mathsf{T}} \mathrm{diag}(\lambda) \log p(\mathbf{y}|\mathbf{x}, \mathbf{z}) - (\mathbf{1} - \mathbf{y})^{\mathsf{T}} \mathrm{diag}(\lambda) (\log \mathbf{1} - p(\mathbf{y}|\mathbf{x}, \mathbf{z})), \qquad (2)$$

where λ denotes a vector of tasks' importance coefficients and $\mathrm{diag}(\cdot)$ represents a diagonal matrix. Here, $\mathbf{y} = (y, \mathbf{o}^p, \mathbf{o}^s, \mathbf{o}^u)$ is a vector of binary labels, concatenating the pedestrian label and all attribute labels. The network parameters are updated by minimizing Eqn.(2) using stochastic gradient descent and back-propagation (BP). We fix the important coefficient $\lambda_1 \in \lambda$ of the main task y, i.e. $\lambda_1 = 1$. As the auxiliary tasks are independent, their coefficients can be obtained by greedy search between zero and one. To simplify the learning procedure, we have $\forall \lambda_i \in \lambda, \lambda_i = 0.1, i = 2, 3, ..., 18$ and found that this setting provides stable and reasonable good results.

We analyze the performance of different components of TA-CNN by training on Caltech-Train and evaluating on Caltech-Test reasonable subset. The performances show clear increasing patterns when gradually adding more components. For examples, TA-CNN (main task) cascades on ACF and reduces the miss rate of it by more than 5 percent. TA-CNN (PedAttr.+SharedScene) reduces the result of TA-CNN (PedAttr.) by 2.2 percent, because it can bridge the gaps among multiple scene datasets. After modeling the unshared attributes, the miss rate is further decreased by 1.5 percent, since more attribute information is incorporated. The final result of 20.86 miss rate is obtained by using the structure projection vector as input to TA-CNN.



(a) single detector (b) multi-view detector (c) TA-CNN

Figure 1: Comparisons between different detectors.

	Pedestrian Attributes									Scene Attributes							
										Shared				Unshared			
	$p1$	$p2$	$p3$	$p4$	$p5$	$p6$	$p7$	$p8$	$p9$	$s1$	$s2$	$s3$	$s4$	$u1$	$u2$	$u3$	$u4$
	Backpack	Dark-Trousers	Hat	Bag	Gender	Occlusion	Riding	Viewpoint	White-Clothes	Sky	Tree	Building	Road	Traffic-light	Horizontal	Vertical	Vehicle
Caltech (**P**)	√	√	√	√	√	√	√	√	√								
CamVid ($\mathbf{B}^a$)										√	√	√	√	√			
Stanford ($\mathbf{B}^b$)										√	√	√	√		√	√	
LM+SUN ($\mathbf{B}^c$)										√	√	√	√				√

Figure 2: Attribute summarization.

Details for extensive experiments, such as detection effectiveness of hard negative mining, pedestrian attributes, and background attributes, are described in this paper. Our conclusion is that auxiliary attribute tasks are favorable for enriching the learned features to account for combinatorial pedestrian variations. Future work tends to explore more attribute configurations. The proposed approach also has potential for attribute prediction and background scene understanding.

[1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*. 2005.

[2] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009.

[3] Ping Luo, Yonglong Tian, Xiaogang Wang, and Xiaoou Tang. Switchable deep network for pedestrian detection. In *CVPR*, 2014.

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

[5] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *CVPR*. 2013.

[6] Paul Viola, Michael J Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 63(2):153–161, 2005.

[7] Xiaoyu Wang, Tony X. Han, and Shuicheng Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, 2009.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.