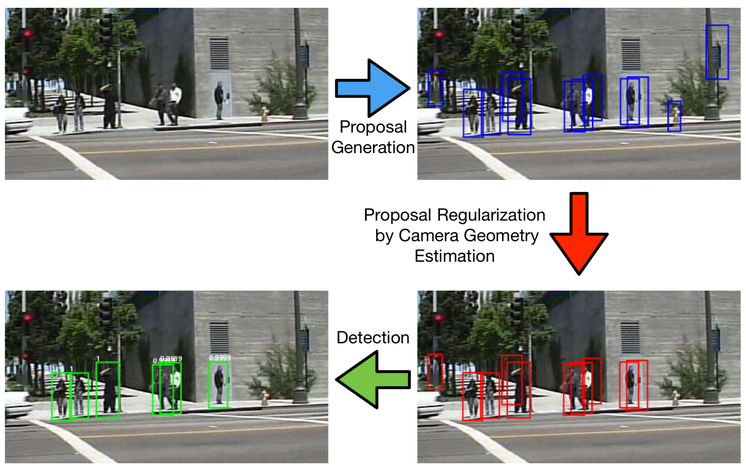
Geometrically Regularized Region Proposals for End-to-end Pedestrian Detection

Yinpeng Dong
donyp13@mails.tsinghua.edu.cn

Introduction

Pedestrian detection is a canonical sub-problem of general object detection which attracts much attention because of its direct applications in self-driving vehicles, surveillance, and robotics.

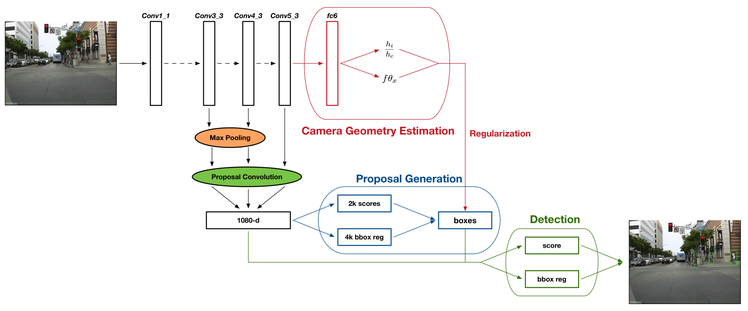
In this work, we propose a novel framework called ***CG-RPN (Region Proposal Networks with Camera Geometry Estimation)*** for high-quality region proposal generation and accurate pedestrian detection. It is comprised of three modules, namely **proposal generation** module, **camera geometry estimation** module and **detection** module.



Contributions

- We investigate a proper way to **fuse different convolutional layers** and produce **a coarse-to-fine feature map** for proposal generation and pedestrian detection.
- We **integrate camera geometry estimation and proposal generation network** together to get a reduced set of region proposals with high recall and accurate localization.
- Our framework **achieve the state-of-the-art performance** on Caltech pedestrian dataset.

Framework



- An entire image is forwarded through the network to generate feature maps and predict camera geometry parameters.
- Fuse the feature maps of different layers to produce region proposals which are then regularized by camera parameters.
- Proposals are classified by the detection module.

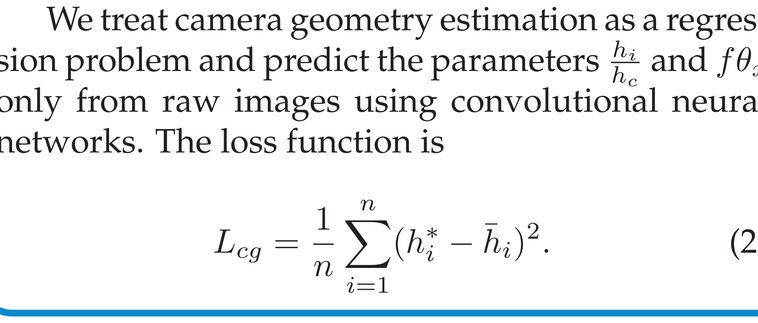
Feature Fusion

Deeper layers can **find object of interest with high recall** while earlier layers are good at **localizing objects**. To combine multi-level feature maps at the same space, we first add max pooling layers after high resolution feature maps and then use l_2 normalization layers to scale them.

Camera Geometry Estimation

Camera geometry estimation aims at regularizing region proposals. There are two monocular cues, one is that smaller object is farther from camera, the other is that farther object appears higher on image plane.

Adding camera geometry information can reject many hard negative backgrounds with inconsistent size and choose the proper one with better localization performance among a lot of candidates.



We use a simplified perspective camera model. The projection of a real world point $\mathbf{P} = (x, y, z)$ on image plane in homogeneous coordinate is given by

$$\begin{bmatrix} \tilde{u} \\ \tilde{v} \\ \tilde{w} \end{bmatrix} = \begin{bmatrix} f & 0 & c_u \\ 0 & f & c_v \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x \\ 0 & \sin\theta_x & \cos\theta_x \end{bmatrix} \begin{bmatrix} x \\ y + h_c \\ z \end{bmatrix}.$$

Using above equation, we find that the projection height of a pedestrian is a linear transformation of the vertical coordinate of the bottom point given by

$$\bar{h}_i = \frac{h_i}{h_c} \cdot (v_i - c_v + f\theta_x). \tag{1}$$

We treat camera geometry estimation as a regression problem and predict the parameters $\frac{h_i}{h_c}$ and $f\theta_x$ only from raw images using convolutional neural networks. The loss function is

$$L_{cg} = \frac{1}{n}\sum_{i=1}^{n}(h_i^* - \bar{h}_i)^2. \tag{2}$$

Jointly Training

We train proposal generation and camera geometry estimation jointly with a multi-task loss:

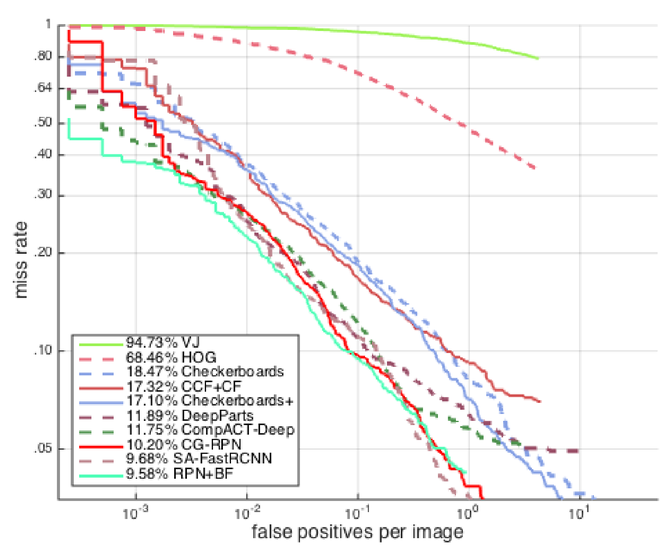
$$L = \frac{1}{N_b}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_p}\sum_i p_i^* L_{reg}(t_i, t_i^*) + \mu L_{cg}, \tag{3}$$

Detection

We warp each region to a fixed-size 80×32 and use the same architecture as proposal generation network to classify them. The framework runs much faster than RCNN and slower than Faster-RCNN within 2 times.

Results

Comparison with state-of-the-art



Fuse which layers?

Fusion layers	MR (%)
conv5_3 (baseline)	16.32
conv3_3, conv4_3, conv5_3	**14.01**
conv1_2, conv2_2, conv3_3, conv4_3, cnov5_3	15.48

Does camera geometry estimation work?