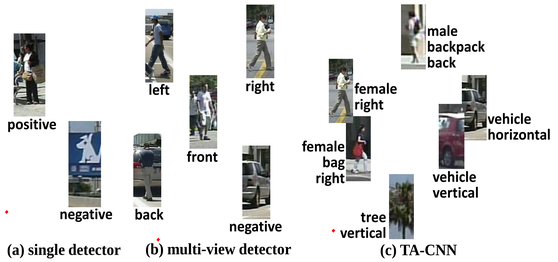
Pedestrian Detection Aided by Deep Learning Semantic Tasks

Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang
The Chinese University of Hong Kong
{ty014, pluo, xtang}@ie.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk

IEEE 2015 Conference on Computer Vision and Pattern Recognition
CVPR 2015 BOSTON JUNE 7-12

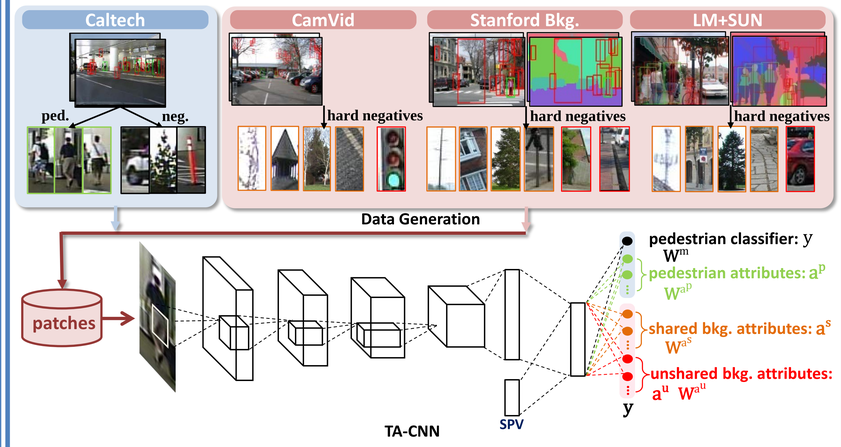
1. Motivation

- **Existing deep models**: treated pedestrian detection as a single binary classification task, which can not capture rich pedestrian variations
- **Our idea**: jointly optimizes pedestrian detection with auxiliary semantic tasks, including pedestrian attributes (*e.g.* 'backpack', 'gender', and 'views') and scene attributes (*e.g.* 'vehicle', 'tree', and 'vertical')

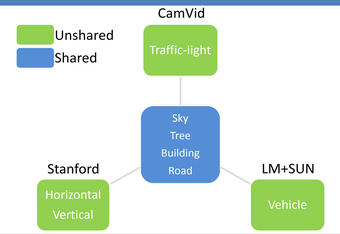


(a) Single Detector: difficult to handle complex pedestrian variations
(b) Multi-View Detector: pedestrian images in different views are handled by different detectors
(c) TA-CNN: treat view as one type of semantic attribute and extend the idea by involve more pedestrian and background attributes

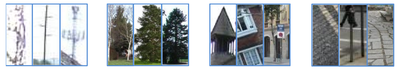
2. Overall Pipeline



2.1 Scene Attributes

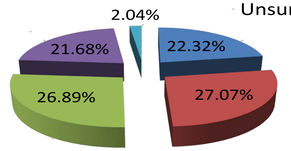


- **Shared**: enable the learning of shared representation across multiple datasets



- **Unshared**: Enhance the diversity of attributes

2.2 Pedestrian Attributes

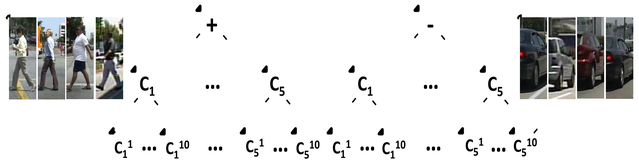


Other Attributes:

Backpack	Gender
Dark-trousers	Occlusion
Hat	White-cloth
CarryingBag	Riding

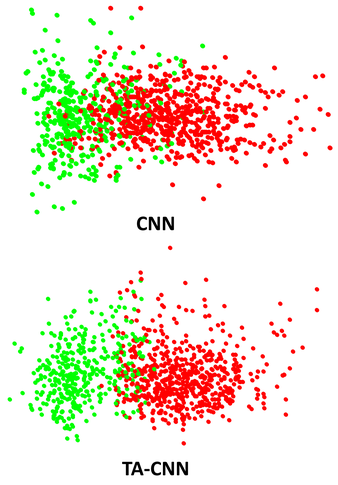
2.3 Structural Projection Vector

- **Objective**: as additive information to bridge the visual gaps between the background datasets and the pedestrian dataset



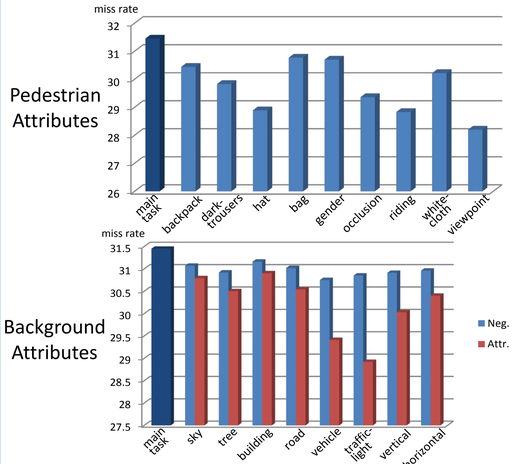
- **Computation**
  (1) Organizing the positive and negative patches of pedestrian dataset into two tree structures by **HOG** feature clustering. Specifically, each tree partitions patches top-down, and have 3 layers and 50 leaf nodes
  (2) SPV of each sample is obtained by concatenating the distances between its HOG feature and the mean HOG feature of each leaf node

3. Feature Visualization

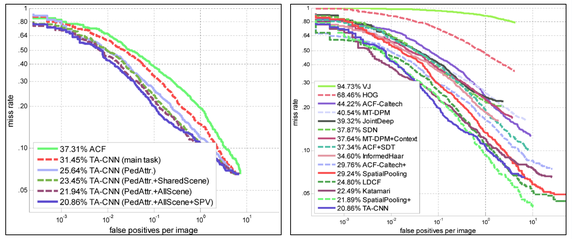


4 Experimental Results

(1) Effectiveness of Auxiliary Tasks



(2) Overall Results on Caltech



(3) Attribute Prediction (mean accuracy 75+%)

Viewpoint Estimation

		Predict State			
		Frontal	Back	Left	Right
True State	Frontal	226	32	15	10
	Back	24	232	12	8
	Left	22	13	164	21
	Right	5	15	40	96
Accuracy		0.816	0.796	0.701	0.711

5. Discussions

(1) Improvement Potentials

- Pre-train models on ImageNet
- Explore more attribute configurations
- Learn sharing representations between multiple pedestrian datasets
- ……

(1) Future Work

- Simultaneously detection and scene parsing