Learning Scene-Specific Pedestrian Detectors without Real Data
— Supplementary Material —

Hironori Hattori
Sony Corporation
Hironori.Hattori@jp.sony.com

Vishnu Naresh Boddeti, Kris Kitani, Takeo Kanade
The Robotics Institute, Carnegie Mellon University
naresh@cmu.edu, kkitani@cs.cmu.edu, tk@cs.cmu.edu

1. Data Simulation

Our scene-and-location specific pedestrian detection system is trained entirely using synthetic computer generated pedestrian images. Therefore, the detection performance of our system is critically dependent on the quality of pedestrian rendering. Quality however does not necessarily mean the how realistic the graphics is but the factors which effect the feature extraction and the classifier performance. We now enumerate the different factors that we took care of while rendering pedestrians.

1.1. Pose and Appearance

The generalization capability of our appearance model is dependent on the variety of rendered pedestrians. To span a large space of pedestrian appearance, we render pedestrians spanning a wide range of shape and size, gender, height and clothing over the entire 360° pose. Figure 1 shows a few examples of the renders that we use.

Figure 1. A few examples of the pedestrian renderings used for training our pedestrian detectors. We have a total of 36 different pedestrian models and for each location we simulate pedestrians with 3 different walking poses and 12 (every 30°) different orientations.

1.2. Training Data

Figure 2 shows a few examples of rendered positive and negative training samples for the CMUSRD dataset.

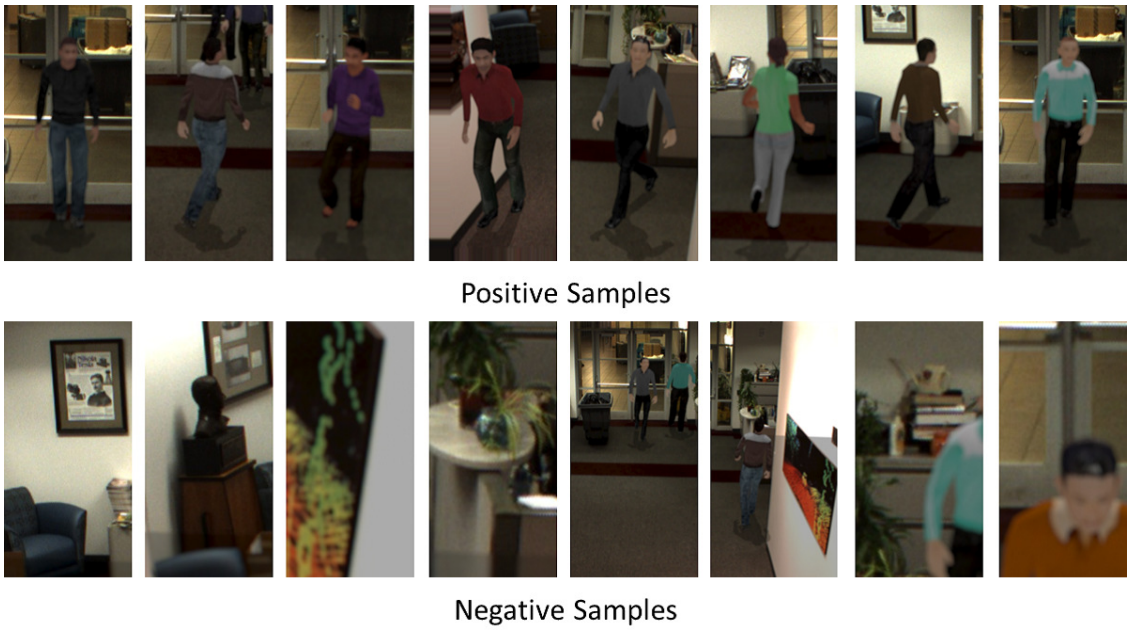

Figure 2. The positive samples have variations in pedestrian pose, appearance, height, gender etc. On the other hand the negative samples consist of many variations of the background including samples with partial occluded pedestrians and pedestrians at very different scale.

1.3. Blending

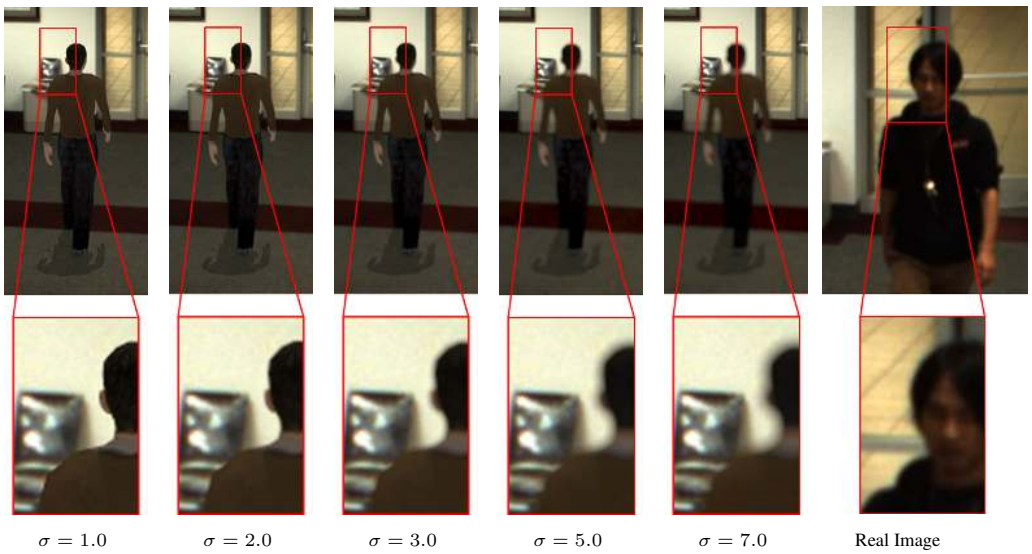

Figure 3. Pedestrian renderings for different amounts of blurring. Note that the best Gaussian smoothing parameter that matches the real data is about $\sigma = 5.0$.

The feature extraction and the classifier performance is also heavily dependent on the quality of blending (sharpness) of the rendered pedestrians. While the rendering engine can generate very sharp renders the performance of the system, unsurprisingly, is best when the rendering sharpness matches the real pedestrians in the scene. Figure 3 shows pedestrian renders with different amounts of blurring along with the appearance of a real pedestrian in the scene.

2. Experimental Evaluation

In this section, we report additional experimental results that were omitted from the main paper due to space constraints.

2.1. 2D Bounding Box Evaluation

Most object detection systems use a criterion of 50% overlap ratio to determine a correct detection. However, one of our goal in this work is precise 2D localization of pedestrians in scenes. Therefore, we evaluate pedestrian detection performance under increasingly stringent overlap ratios of 50%, 60%, 70% and 80%.

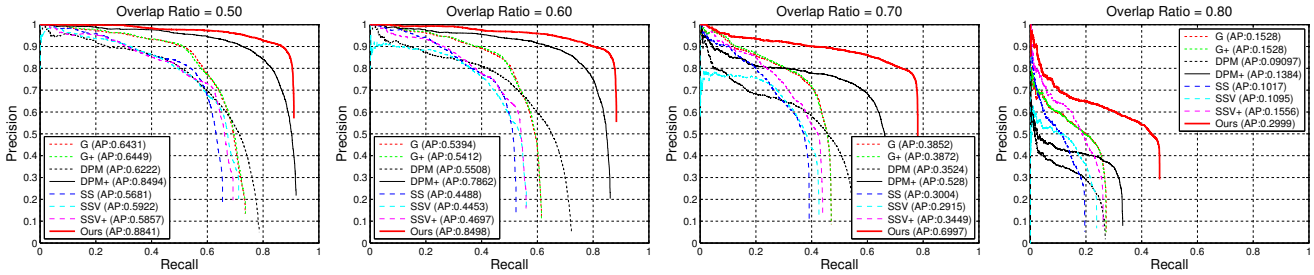

Figure 4. Precision-recall curves for differing overlap ratio criteria on TownCenter.

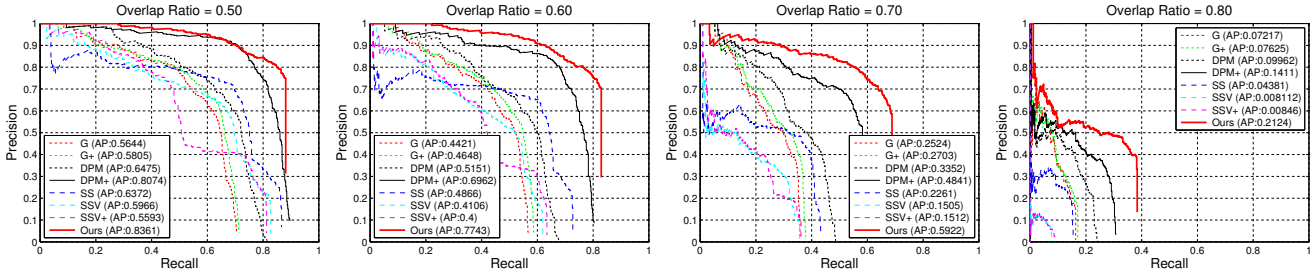

Figure 5. Precision-recall curves for differing overlap ratio criteria on PETS2006.

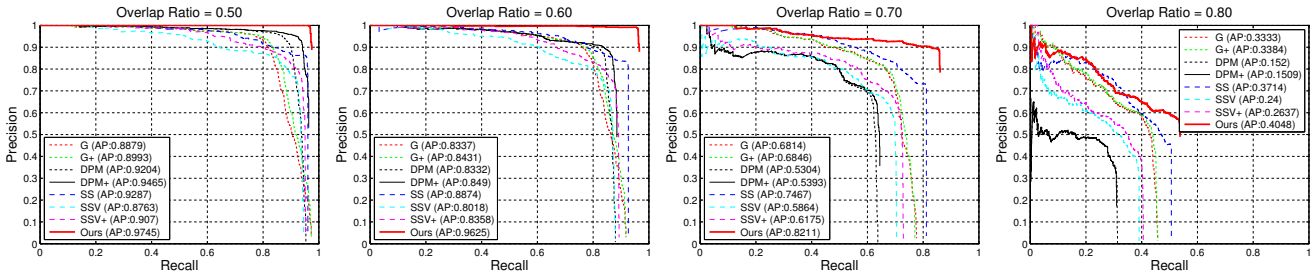

Figure 6. Precision-recall curves for differing overlap ratio criteria on CMUSRD.

2.2. Effect of grid-size resolution

Building upon our observation that a single generic detector is not flexible enough to cover the entire scene, we would like to understand how many detectors are needed to effectively cover all appearance variations. We evaluated the effect of the grid-size on system performance using small portion of the scenes to understand how appearance is affected by location. Table 2.2 shows how AP performance changes with respect to the grid size (number of learned detectors). The results indicate that a smaller grid size of 8×8 patches perform better which means that pedestrian appearance is in fact varying significantly by location. Our results show a plateau effect starting at 16×16 so we use this setting for all our experiments.

Table 1. Town Center

Patch Size	# of Detectors	AP
8×8	371	0.802
16×16	102	0.798
32×32	30	0.764
64×64	8	0.755
128×128	4	0.693

Table 2. PETS 2006

Patch Size	# of Detectors	AP
16×16	879	-
32×32	238	0.912
64×64	63	0.906
128×128	15	0.847

Table 3. CMUSRD

Patch Size	# of Detectors	AP
16×16	640	0.971
32×32	102	0.966
64×64	32	0.947
128×128	8	0.927

2.3. Localization in 3D

The primary goal of our work is accurate 3D localization of pedestrians on the ground plane in the scenes. Therefore, we evaluate 3D pedestrian localization performance under increasingly stringent distances (90cm, 70cm, 50cm, 30cm) from the ground truth on the ground plane.

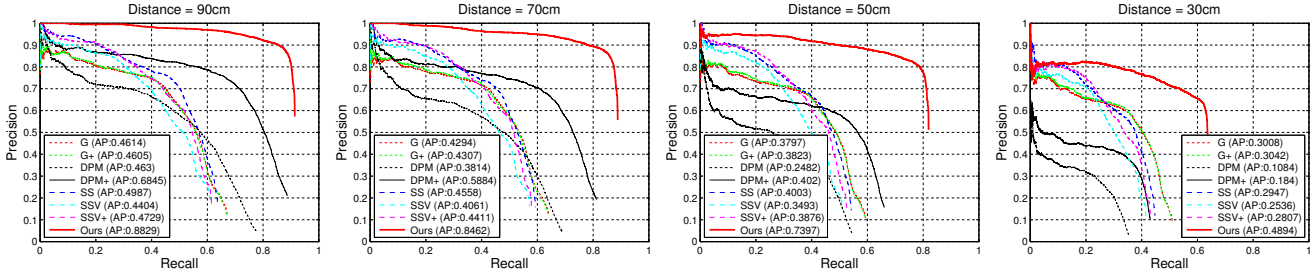

Figure 7. Precision-recall curves for different amounts of distance on TownCenter.

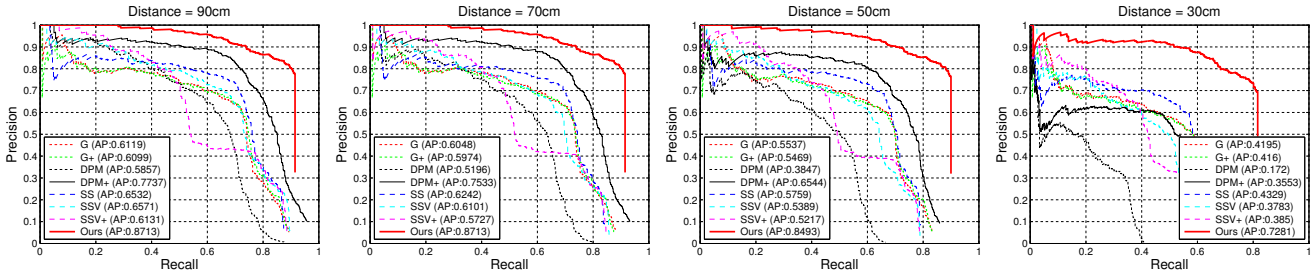

Figure 8. Precision-recall curves for different amounts of distance on PETS2006.

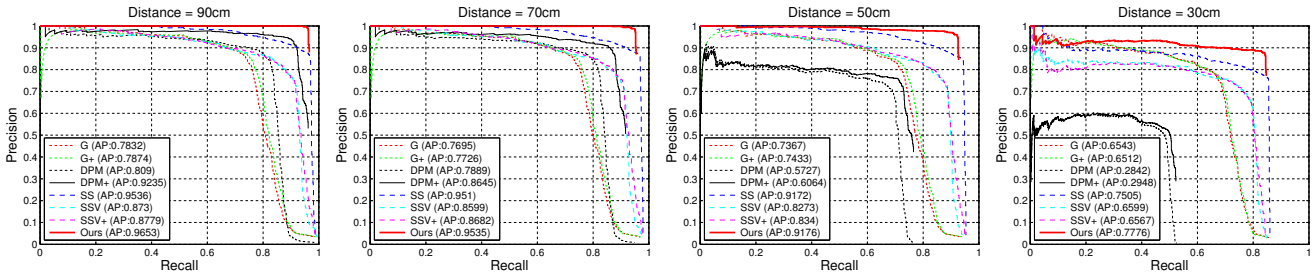

Figure 9. Precision-recall curves for different amounts of distance on CMUSRD.