Learning Scene-Specific Pedestrian Detectors without Real Data

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

Abstract

We consider the problem of designing a scene-specific pedestrian detector in a scenario where we have zero instances of real pedestrian data (i.e., no labeled real data or unsupervised real data). This scenario may arise when a new surveillance system is installed in a novel location and a scene-specific pedestrian detector must be trained prior to any observations of pedestrians. The key idea of our approach is to infer the potential appearance of pedestrians using geometric scene data and a customizable database of virtual simulations of pedestrian motion. We propose an efficient discriminative learning method that generates a spatially-varying pedestrian appearance model that takes into the account the perspective geometry of the scene. As a result, our method is able to learn a unique pedestrian classifier customized for every possible location in the scene. Our experimental results show that our proposed approach outperforms classical pedestrian detection models and hybrid synthetic-real models. Our results also yield a surprising result, that our method using purely synthetic data is able to outperform models trained on real scene-specific data when data is limited.

Figure 1. Scene-specific and location-specific pedestrian detection results using geometrically consistent computer generated data.

1. Introduction

Consider the scenario in which a new surveillance system is installed in a novel location and an image-based pedestrian detector must be trained without access to real scene-specific pedestrian data. A similar situation may arise when a new imaging system (*e.g.* a custom camera with unique lens distortion) has been designed and must be able to detect pedestrians without the expensive process of collecting data with the new imaging device. Both of these scenarios are ill posed zero-instance learning problems, where an image-based pedestrian detector must be created without having access to real data. Fortunately, in these scenarios we have access to two important pieces of information: (1) the camera's calibration parameters, and (2) scene geometry. In this work, we show that with this information, it is possible to generate synthetic training data (*i.e.* computer generated pedestrians) to act as a proxy for the real data. Moreover, we show that by using this 'data-free' technique (*i.e.*, does not require real pedestrian data), we are still able to train a scene-specific pedestrian detector that outperforms baseline techniques.

Our goal is to develop a method for training a 'data-free' scene-specific pedestrian detector which outperforms generic pedestrian detection algorithms (*i.e.* HOG-SVM, DPM). Generically trained pedestrian detectors are trained over large data sets of real data and thus work robustly across many scenes. However, generic models are not always best-suited for detection in specific scenes. In many surveillance scenarios, it is more important to have a customized pedestrian detection model that is optimized for a single scene. Optimizing for a single scene however often requires a labor intensive process of collecting labeled

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

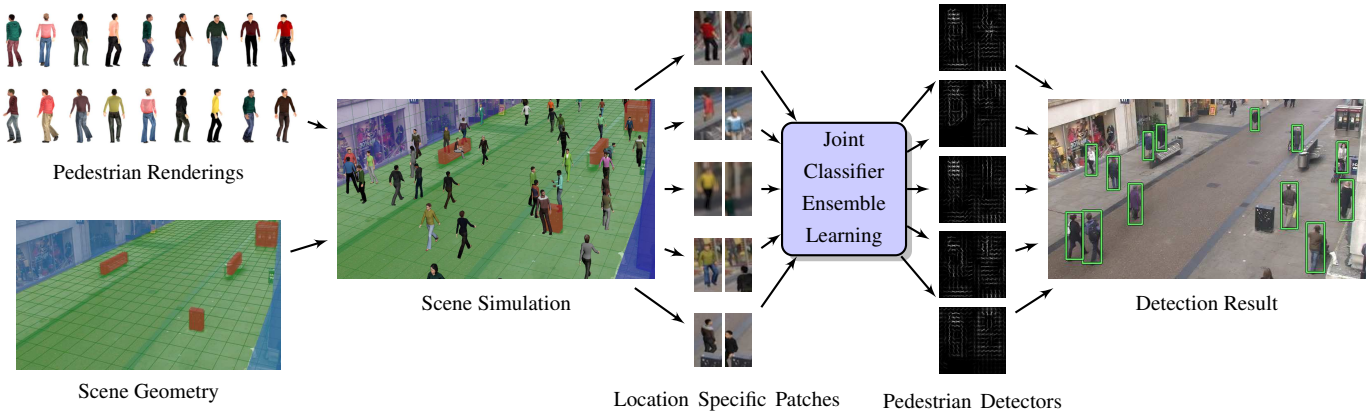


Figure 2. **Overview:** For every grid location, geometrically correct renderings of pedestrian are synthetically generated using known scene information such as camera calibration parameters, obstacles (red), walls (blue) and walkable areas (green). All location-specific pedestrian detectors are trained jointly to learn a smoothly varying appearance model. Multiple scene-and-location-specific detectors are run in parallel at every grid location.

data – drawing bounding boxes of pedestrians taken with a particular camera in a specific scene. The process also takes time, as recorded video data must be manually mined for various instances of clothing, size, pose and location to build a robust pedestrian appearance model. Creating a situation-specific pedestrian detector enables better performance but it is often costly to train. In this work, we provide an alternative technique for learning high-quality scene-specific pedestrian detector without the need for real pedestrian data. Instead we leverage the geometric information of the scene (*i.e*. static location of objects, ground plane and walls) and the parameters of the camera to generate geometrically accurate simulations of pedestrian appearance through the use of computer generated pedestrians. In this way, we are able to synthesize data as a proxy to the real data, allowing us to learn a highly accurate scene-specific pedestrian detector.

They key idea of our approach is to maximize the geometric information about the scene to compensate for the lack of real training data. A geometrically consistent method for synthetic data generation has the following advantages. (1) An image-based pedestrian detector can be trained on a wider range of pedestrian appearance. Instead of waiting and collecting real data, it is now possible to generate large amounts of simulated data over a wide range of pedestrian appearance (*e.g*. clothing, height, weight, gender) on demand. (2) Pedestrian data can be generated for any location in the scene. Taken to the extreme, a synthetic data-generation framework can be used learn a customized pedestrian appearance model for every possible location (pixel) in the scene. (3) The location of static objects in the scene can be incorporated into data synthesis to preemptively train for occlusion.

In our proposed approach, we simultaneously learn hundreds of pedestrian detectors for a single scene using millions of synthetic pedestrian images. Since our approach is purely dependent on synthetic data, the algorithm requires no real-world data. To learn the set of scene-specific location-specific pedestrian detectors, we propose an efficient and scalable appearance learning framework. Our algorithmic framework makes use of highly-efficient correlation filters as our basic detection unit and globally optimizes each model by taking into the account the appearance of a pedestrian over a small spatial region. We compare our approach to several generically trained baseline models and show that our approach generates a better performing scene-specific pedestrian detector. More importantly, our experimental results over multiple data sets show that our 'data-free' approach actually outperforms models that are trained on real scene-specific pedestrian data when data is limited.

Contributions: The contribution of our work is as follows: (1) the first work to learn a scene-specific *location-specific* geometry-aware pedestrian detection model using purely synthetic data and (2) an efficient and scalable algorithm for learning a large number of scene-specific *location-specific* pedestrian detectors.

2. Related Work

3D Models for Detection. The idea of using 3D synthetic data for 2D object detection is not new. Brooks [8] used computerized 3D primitives to describe 2D images. Dhome *et al*. used computer generated models to recognize articulated objects from a single image [10]. 3D computer graphics models have been used for modeling human shape [16, 7], body-part gradient templates [1], full-body gradient templates [22] and hand appearance [27, 2, 28]. In addition to modeling people, 3D simulation has been used for multi-view car detection [25, 24, 18] and 3D indoor scene

understanding [31, 21]. Sun and Saenko [34] used virtual objects to train 2D object detector for real objects. Work by Marin *et al*. [22] used a video game rendering engine to generate synthetic training data. While they learned a single pedestrian detector applied to a mobile scenario, we learn hundreds of location sensitive models for a surveillance scenario. Synthetic data can also be used for evaluation [35]. The main benefit of computer generated data is that it does not require manual data labeling since the ground truth is known. The second benefit is that large amounts of data can be generated with little effort. We take advantage of both of these benefits in our work.

Synthesis for Domain Adaptation. One effective use of computer generated images is in the area of visual domain adaptation [23, 26, 37]. First, large repositories of synthetic 2D data can be used to bootstrap detectors. Then, the data or detectors can be adapted to real data by leveraging data from the test distribution. Pishchulin *et al*. combined synthesized real 3D human body models and a small number of labeled pedestrian bounding boxes to learn a very robust pedestrian detector [26]. Their work showed that augmenting the training set with the appropriate mix of synthetic and real data can maximize test time performance. Vazquez *et al*. [37] also showed how synthetic pedestrian data can be combined with real pedestrian data to generate robust real-world detectors. We address a different task than domain adaptation, in that we are learning the synthetic pedestrian model needed prior to the adaptation task.

Scene-Specific Adaptive Classifiers. Adaptive techniques have been proposed for surveillance scenarios [5, 32, 29, 33]. The use of scene geometry, changes in background over time and locality aware detectors can be used to greatly improve the performance of detectors for a specific scene. Our work is similar in that we use scene geometry and camera calibration parameters to generate scene-specific synthetic data. Our work is different in that we do not use real data from the scene to adapt our detector.

Scene-Specific Domain Adaptation. Adapting pre-trained models to a new domain has been an active area of research [39, 38, 40, 42, 30]. The most recent approaches can adapt detectors trained in another domain without the need for new labeled data by bootstrapping a new detector with high or low confidence detections in the new scene [43]. Our work is distinct from work on domain adaption in that we do not allow access to scene-specific real data. In domain adaptation a pre-existing pedestrian detector (or generic pedestrian data) is *augmented* with scene-specific real data to improve performance. Our work is complementary to domain adaptation work in that our proposed detector or data can be used as an input to a domain adaptation algorithm.

In this work, we have limited ourselves to a surveillance scenario where the camera is static and the scene is known. This stands in contrast to the large body of work focused primarily on pedestrian detection from a mobile platform [12, 13, 41, 11]. The mobile scenario describes a more challenging problem where the camera is undergoing ego-motion and the scene geometry is usually unknown. There is also significant work regarding the choice of features and detector for effective pedestrian detection [11]. For this work, we limit our choice to the standard HOG feature. For the classifier, we utilize a correlation filter based approach [4, 19] over the standard SVM for computational efficiency reasons as we are required to learn large numbers of templates for a scene.

3. Proposed Method

Figure 2 gives a pictorial illustration of our spatially-varying scene-specific pedestrian detection framework. We consider the surveillance setting where the following information is available: (1) intrinsic and extrinsic parameters of the static camera and (2) the geometrical layout of the scene, *i.e*., semantic labels for all the regions ("pedestrian region") in the scene where a pedestrian could possibly appear and semantic labels for obstacles in the scene where a pedestrian could either be occluded or physically cannot be present. This information is leveraged along with synthesized 3D pedestrian models to generate realistic simulations of the appearance of pedestrians for every location of the "pedestrian region". We then learn a smooth spatially-varying scene-specific discriminative appearance model for pedestrian detection. During detection, unlike the conventional approach where a single global detector is applied across the entire image, hundreds of scene-specific location-specific pedestrian detectors are applied to the scene.

3.1. Data Simulation

Most conventional pedestrian detection techniques require training images with high quality ground truth labels, which in our case would be required at every location of the "pedestrian region" of our scene. However, these labels can be very expensive to obtain and annotate for pedestrians at every location in the scene. Therefore in this paper, we adopt a simulation-based approach to artificially render the pedestrians in the scene taking into account the camera parameters and the geometrical layout of the scene *e.g*., obstacles and occlusions in the scene. We use a total of 36 different pedestrian models and at each location in the scene, we simulate pedestrians with 3 different walking configurations and 12 (every 30°) different orientations[1].

We use 3DS Max from Autodesk to recreate the geometry of the scene using the intrinsic parameters of the camera and manually labeled scene points. Object locations such as walls and obstacles are also labelled manually. Having the correct camera parameters is important as it determines

[1]More details are included in the supplementary material

the amount of perspective distortion that is applied to the synthetic pedestrians. For surveillance cameras with a wide field-of-view and small focal length, there is a significant amount of perspective distortion for people who are near the camera. This must be learned by pedestrian detectors to handle distorted (*e.g.*, tilted pedestrians with big heads) images of people.

In the extreme case of training a model for each pixel location, a typical scene considered in this paper has about 100,000 locations and at each location we simulate about 4000 pedestrians with randomly chosen appearance, walking configuration and orientation for a total of about 400 million simulated pedestrians. In practice we learn only a few hundred models with several million training images, since performance plateaus after a certain spatial resolution (more details in section 4.4).

3.2. Classifier Ensemble Learning

Since the detectors at each location in the image significantly overlap with each other it is natural to impose smoothness constraints between neighboring detectors – neighboring detectors should be similar. Therefore, we propose a joint detector learning approach while imposing smoothness constraints between neighboring detectors. A nice consequence of our framework is that the detectors are implicitly calibrated since they are jointly trained. We base our detector on the Vector Correlation Filter [4] formulation where the detector design is posed as a regression problem.

Notation: For notational ease all expressions through the rest of this paper are given for 1-D signals with K-channels. Vectors are denoted by lower-case bold ($\mathbf{x}$) and matrices in upper-case bold ($\mathbf{X}$). $\hat{\mathbf{x}} \leftarrow \mathcal{F}_K(\mathbf{x})$ and $\mathbf{x} \leftarrow \mathcal{F}_K^{-1}(\hat{\mathbf{x}})$ denotes the Fourier transform of $\mathbf{x}$ and the inverse Fourier transform of $\hat{\mathbf{x}}$, respectively, where $\hat{}$ denotes variables in the frequency domain, $\mathcal{F}_K()$ is the Fourier transform operator and $\mathcal{F}_K^{-1}()$ is the inverse Fourier transform operator with the operators acting on each of the K channels independently. Superscript † denotes the complex conjugate transpose operation.

We pose the problem of jointly learning the n detectors with m_i training samples per detector as the following optimization problem:

$$\min_{\mathbf{w_1},\ldots,\mathbf{w_n}} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\|\sum_{k=1}^{K}\mathbf{x}_i^{kj} * \mathbf{w}_i^k - \mathbf{g}_i^j\right\|_2^2 + \frac{\lambda}{2}\sum_{(i,j)\in E} c_{ij}\|\mathbf{w}_i - \mathbf{w}_j\|_2 \tag{1}$$

where $*$ denotes the correlation operation $\mathbf{x}_i^j$ is the j-th training image for the i-th detector $\mathbf{w}_i$ and $\mathbf{g}_i^j$ is the desired correlation response. The second term captures the classifier smoothness constraints for overlapping regions of the classifier and c_{ij} captures the smoothness weights and λ is the regularization parameter which trades-off the smoothness term. We adopt the Alternating Direction Method of Multipliers (ADMMs) [6] to solve the above optimization problem efficiently. The problem is now posed as:

$$\begin{aligned}\min_{\mathbf{w_1},\ldots,\mathbf{w_n}} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\|\sum_{k=1}^{K}\mathbf{x}_i^{kj} * \mathbf{w}_i^k - \mathbf{g}_i^j\right\|_2^2 \\ & + \frac{\lambda}{2}\sum_{(i,j)\in E} c_{ij}\|\mathbf{h}_i - \mathbf{h}_j\|_2^2 + \frac{\rho}{2}\|\mathbf{W} - \mathbf{H}\|_F^2 \\ s.t. \quad & \mathbf{W} = \mathbf{H}\end{aligned} \tag{2}$$

where ρ is a regularization parameter. We now form and optimize the Lagrangian for this optimization problem,

$$\begin{aligned}L(\mathbf{W},\mathbf{H},\mathbf{\Lambda}) = \; & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\|\sum_{k=1}^{K}\mathbf{x}_i^{kj} * \mathbf{w}_i^k - \mathbf{g}_i^j\right\|_2^2 \\ & + \frac{\lambda}{2}\sum_{(i,j)\in E} c_{ij}\|\mathbf{h}_i - \mathbf{h}_j\|_2^2 \\ & + \frac{\rho}{2}\|\mathbf{W} - \mathbf{H}\|_F^2 \\ & + \Lambda^T(vec(\mathbf{W}) - vec(\mathbf{H}))\end{aligned} \tag{3}$$

This problem can be solved by decomposing it into subproblems for $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{\Lambda}$, each of which can in turn be solved very efficiently.

Subproblem W:

$$\mathbf{W}^{l+1} = \arg\min_{\mathbf{W}} L(\mathbf{W}, \mathbf{H}^l, \mathbf{\Lambda}^l) \tag{4}$$

This sub-problem can be further decomposed into individual sub-problems for each of the locations in the scene in closed form in the Fourier domain i.e,

$$\begin{aligned}\mathbf{w}_i^{l+1} &= \arg\min_{\mathbf{w}} \frac{1}{2}\sum_{j=1}^{m_i}\|\sum_{k=1}^{K}\mathbf{x}_i^{kj} * \mathbf{w} - \mathbf{g}_i^j\|_2^2 \\ &\quad + \frac{\rho}{2}\|\mathbf{w} - \mathbf{h}_i^l\|_2^2 + \mathbf{\Lambda}_i^{lT}(\mathbf{w} - \mathbf{h}_i^l) \\ &= \mathcal{F}_K^{-1}\{\arg\min_{\hat{\mathbf{w}}} \frac{1}{2}\sum_{j=1}^{m_i}\|\sum_{k=1}^{K}\hat{\mathbf{X}}_i^{kj\dagger}\hat{\mathbf{w}} - \hat{\mathbf{g}}_i^j\|_2^2 \\ &\quad + \frac{\rho}{2}\|\hat{\mathbf{w}} - \hat{\mathbf{h}}_i^l\|_2^2 + \hat{\mathbf{\Lambda}}_i^{l\dagger}(\hat{\mathbf{w}} - \hat{\mathbf{h}}_i^l)\} \\ &= \mathcal{F}_K^{-1}\{(\hat{\mathbf{D}} + \rho\mathbf{I})^{-1}(\rho\hat{\mathbf{h}}_i^l + \hat{\mathbf{p}} - \hat{\mathbf{\Lambda}}_i^l)\}\end{aligned}$$

where we use the Parseval's theorem to express the objective function in the Fourier domain. $\hat{\mathbf{X}}_j^{k\dagger}\hat{\mathbf{h}}^k$ is the DFT (of size $N_{\mathcal{F}}$) of the correlation of the k-th channel of the j-th training image with the corresponding k-th channel of

the CF template where the diagonal matrix $\hat{\mathbf{X}}_j^k$ contains the vector $\hat{\mathbf{x}}_j^k$ along its diagonal and,

$$\hat{\mathbf{D}} = \frac{1}{N_{\mathcal{F}}} \begin{bmatrix} \sum_{j=1}^{m_i} \hat{\mathbf{X}}_j^1 \hat{\mathbf{X}}_j^{1\dagger} & \cdots & \sum_{j=1}^{m_i} \hat{\mathbf{X}}_j^1 \hat{\mathbf{X}}_j^{K\dagger} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{m_i} \hat{\mathbf{X}}_j^K \hat{\mathbf{X}}_j^{1\dagger} & \cdots & \sum_{j=1}^{m_i} \hat{\mathbf{X}}_j^K \hat{\mathbf{X}}_j^{K\dagger} \end{bmatrix} \tag{5}$$

$$\hat{\mathbf{p}} = \frac{1}{N_{\mathcal{F}}} \begin{bmatrix} \sum_{l=1}^{m_i} \hat{\mathbf{X}}_i^{1j} \hat{\mathbf{g}}_i^j \\ \vdots \\ \sum_{l=1}^{m_i} \hat{\mathbf{X}}_i^{Kj} \hat{\mathbf{g}}_i^j \end{bmatrix}, \hat{\mathbf{h}} = \begin{bmatrix} \hat{\mathbf{h}}^1 \\ \vdots \\ \hat{\mathbf{h}}^K \end{bmatrix}, \hat{\mathbf{w}} = \begin{bmatrix} \hat{\mathbf{w}}^1 \\ \vdots \\ \hat{\mathbf{w}}^K \end{bmatrix}$$

Subproblem H:

$$\mathbf{H}^{l+1} = \arg\min_{\mathbf{H}} L(\mathbf{W}^{l+1}, \mathbf{H}, \mathbf{\Lambda}^l) \tag{6}$$

The solution for this sub-problem results in a closed form solution which can be implemented very efficiently in the spatial domain.

$$\begin{aligned} \mathbf{H}^{l+1} &= \arg\min_{\mathbf{H}} \frac{\lambda}{2} \sum_{(i,j)\in E} c_{ij} \|\mathbf{h}_i - \mathbf{h}_j\|_2^2 \\ &\quad + \frac{\rho}{2} \|\mathbf{W}^l - \mathbf{H}\|_F^2 + \mathbf{\Lambda}^{lT}(vec(\mathbf{W}^l) - vec(\mathbf{H})) \\ &= \arg\min_{\mathbf{H}} \frac{\lambda}{2} \mathbf{H}^T \mathbf{A} \mathbf{H} + \frac{\rho}{2} \|\mathbf{W}^l - \mathbf{H}\|_F^2 \\ &\quad + \mathbf{\Lambda}^{lT}(vec(\mathbf{W}^l) - vec(\mathbf{H})) \\ &= (\lambda \mathbf{A} + \rho \mathbf{I})^{-1} (\rho vec(\mathbf{W}) + vec(\mathbf{\Lambda})) \end{aligned}$$

where $\mathbf{A}$ is a sparse adjacency matrix defining the connectivity structure (defined by the scene geometry) of the smoothness graph.

Subproblem Λ:

$$\mathbf{\Lambda}^{l+1} = \mathbf{\Lambda}^l + \rho(\mathbf{W}^{l+1} - \mathbf{H}^{l+1}) \tag{7}$$

3.3. Detection Protocol

Given a video frame, pedestrian detection is performed by running each of the spatially varying pedestrian detectors at their corresponding locations resulting in a detection response map over the entire image. To account for the height variation among pedestrians we also evaluate the detectors over a small range of scales at each location (0.95 to 1.05). Finally we apply non-maximal suppression to filter the multiple overlapping detections for each instance of the object obtained from the response map. We note that due to the spatially varying nature of the pedestrian detector, detection can no longer be performed efficiently using convolution.

4. Experimental Evaluation

4.1. Metrics

Datasets – We evaluate the efficacy of our proposed scene-specific spatially-varying pedestrian detection framework on three different datasets: an outdoor dataset, a semi-outdoor datatset and an indoor dataset.

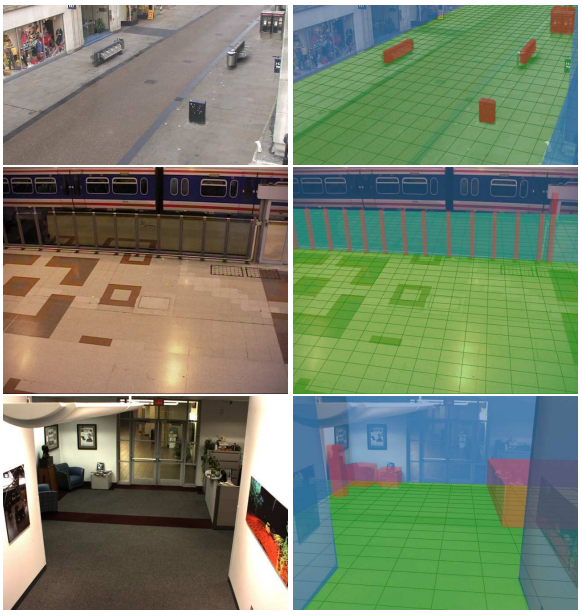


Figure 3. Three evaluation scenes with their corresponding geometric labels. Town Center [3] (top), PETS 2006 [36] (middle) and CMUSRD [17] (bottom).

Towncenter Dataset [3]: The town center dataset is a video dataset of a semi-crowded town center with a resolution of 1920 × 1080 and a frame rate of $25fps$. We down-sample the videos to a standardized resolution of 640 × 360.

PETS 2006 Dataset [36]: The PETS 2006 datatset consists of video (at a resolution of 720 × 576) of a public space including a number of pedestrians. While the dataset consists of videos captured by four different cameras we just use a single camera view for our experiments since our approach is based on a single camera. We down-sample the videos to a standardized resolution of 640 × 512.

CMUSRD [17]: The Carnegie Mellon University Surveillance Research Dataset is a new dataset for indoor surveillance. The data is collected using multiple cameras inside a building at the Carnegie Mellon University and consists of several tens of different people as subjects. While the original resolution of the data is 1280 × 960, we down-sample the videos to a standardized resolution of 640 × 480.

Baselines – We evaluate and compare against the following baseline approaches for the task of pedestrian detection.

G: A single HOG+SVM based pedestrian detector trained on INRIA pedestrian dataset [9].

G+: A single HOG+SVM based pedestrian detector trained on the INRIA pedestrian dataset augmented with negative background patches from the corresponding specific scene.

SS: A single HOG+SVM based pedestrian detector trained on real data from the corresponding specific scene.

DPM: The deformable parts based [14, 15] pedestrian de-

tector trained on the PASCAL VOC *person* class.

DPM+: We build upon the pioneering work by Hoeim et.al.,[20] to leverage the known ground truth scene geometry and camera location/viewpoint at the inference stage using the DPM pedestrian detector as our base detector.

SSV: A single HOG+SVM based pedestrian detector trained **only** on virtual pedestrians whose appearance is simulated in the specific scene under consideration.

SSV+: A single HOG+SVM based pedestrian detector trained on **both** real and virtual pedestrians whose appearance is simulated in the specific scene under consideration. This baseline is similar in spirit to the approach in [37].

SLSV(Ours): Our proposed scene-specific pedestrian detection framework with a spatially varying pedestrian appearance model. This model is learned **entirely** from virtual pedestrians whose appearance is simulated in the specific scene under consideration. In the experiments that follow we train a detector for each 16×16 image patch. The number of models learned for the Town Center, PETS 2006 and CMUSRD is 640, 879 and 348, respectively. Each model is trained using 4000 examples (2000 positive and 2000 negative). This translates to roughly 2.5 million synthetic images used to train the detectors for the Towncenter scene.

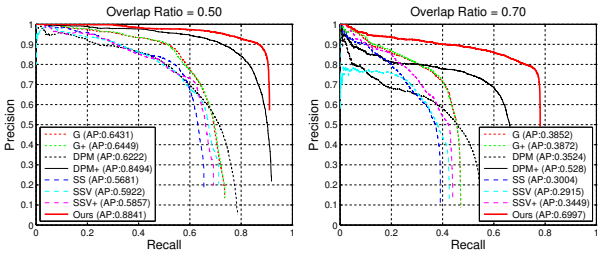
4.2. 2D Bounding Box Evaluation

We compare our proposed model to all baselines using the standard 50% overlap metric used for pedestrian detection [11]. In addition to this metric, we also include results of the 70% overlap criteria to show the 2D localization power of our approach. Results are summarized as PR curves in Fig. 5. The curves show that our approach has a significantly better recall rate due to the ability to learn accurate location specific detectors. The qualitative examples are given in Fig.4 also illustrate the ability of our method to accurately localize pedestrians. Failure cases also show that our model is not able to detect pedestrians occluded by other pedestrians since this type of occlusion was not generated during training. Table 1 shows the mean average precision

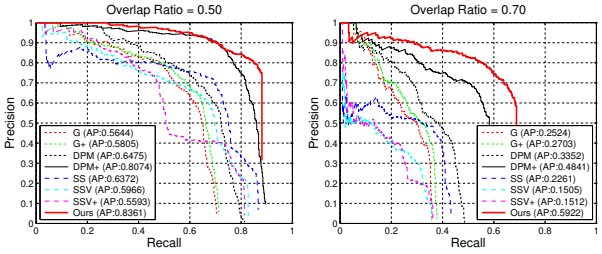
Table 1. Average precision by bounding box overlap criteria

	0.5 overlap	0.7 overlap	Change
G [9]	0.70	0.44	37%
G+	0.71	0.45	37%
DPM [14]	0.73	0.41	44%
DPM+ [20]	0.86	0.51	41%
SS	0.71	0.42	40%
SSV	0.69	0.34	50%
SSV+ [37]	0.68	0.37	46%
SLSV (Ours)	**0.90**	**0.70**	**22%**

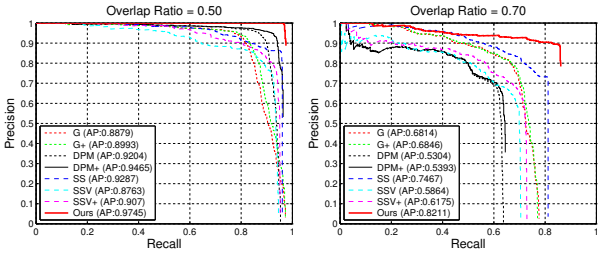
(AP) over all three datasets. Our proposed approach using purely synthetic data outperforms all baselines with an AP of 0.90. The DPM+ which uses the same geometry and camera information as our approach performs second best



(a) Town Center



(b) PETS 2006



(c) CMUSRD

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

with an AP of 0.86 followed by vanilla DPM with an AP of 0.73. All other models fall closely behind the DPM. The main difference between our approach and the other models is that specific detectors are learned for each location in the scene. Furthermore unlike DPM+ which leverages known scene geometry and camera parameters at inference our model uses the same information at the training stage.

More importantly, we observe that our approach is resilient to a more stringent criteria. Across all three datasets, the standard HOG+SVM model **G** drops by 37% ($0.70 \rightarrow 0.44$) and the DPM+ model performance drops by 41% ($0.86 \rightarrow 0.51$). In contrast, the performance of the propose method only drop by 22% ($0.90 \rightarrow 0.70$) under the tighter criteria. We will examine the localization power of our approach further in section 4.5.

4.3. Effect of the smoothness constraint

We have formulated a joint learning problem to ensure that appearance models vary smoothly over space. Figure 6 shows how overall performance is effected by changing the weight of the smoothness term λ in Equation (2). For

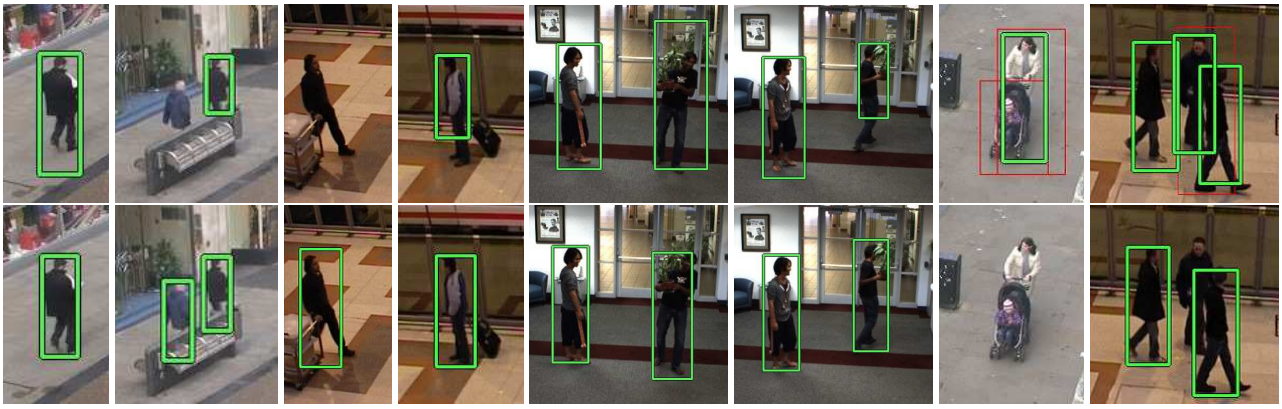


Figure 4. Sample detections of DPM (**top**) and our proposed method (**bottom**), **green** denotes true positives and **red** denotes false positives.

the CMUSRD dataset, the smoothness constrain improves performance by 8 points ($0.89 \rightarrow 0.97$). We obtain optimal performance at a value of $\lambda = 0.10$ for the CMUSRD dataset which we used for all experiments in this paper.

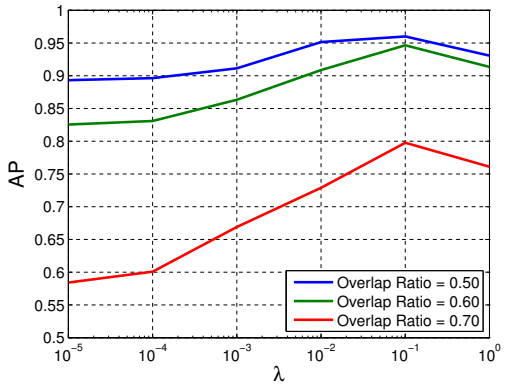


Figure 6. AP on CMUSRD for different smoothing values (λ).

4.4. Effect of grid-size resolution

While we observe from comparative experiments 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 a small portion of the Towncenter scene to understand how appearance is affected by location. Table 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.

4.5. Localization in 3D

Pedestrian detection is often used as a pre-processing step for tracking, action recognition or activity analysis. In these scenarios, it is helpful to know the precise 3D location of a person in the environment. To evaluate the performance of 3D localization we use a minimum distance metric where a detection is considered valid only if it is within 90 cm of the ground truth location. Table 3 shows mean AP scores over all three dataset. Our proposed approach performs best with a AP of 0.91. The second best is the **SS** model trained on real scene-specific data with an AP of 0.70, followed by other models trained on scene-specific data with **SSV** at an AP of 0.66 and **SSV+** at an AP of 0.65.

Table 2. Average precision by number of detectors

Patch Size	Number of Detectors	AP
8×8	371	0.802
16×16	102	0.798
32×32	30	0.764

Table 3. Average precision by 3D distance criteria

	90 cm	50 cm	Change
G [9]	0.62	0.56	10%
G+	0.62	0.56	10%
DPM [14]	0.62	0.40	35%
DPM+ [20]	0.79	0.55	30%
SS	0.70	0.63	10%
SSV	0.66	0.57	13%
SSV+ [37]	0.65	0.58	11%
SLSV (Ours)	**0.91**	**0.84**	**8%**

We also evaluate our approach with a much tighter criteria of 50 cm. Our proposed approach is most resilient to the tighter criteria with a performance drop of only 8% ($0.91 \rightarrow 0.84$). The performance of all other models, with the exception of DPM and DPM+, drops between $10\% \sim 13\%$. The performance of **DPM** and **DPM+** drops by a large 35% and 30% respectively, which indicates that the vertical localization of the DPM model is noisy.

Figure 7 compares the 3D trajectories of our proposed approach and the DPM model. Bounding box results are

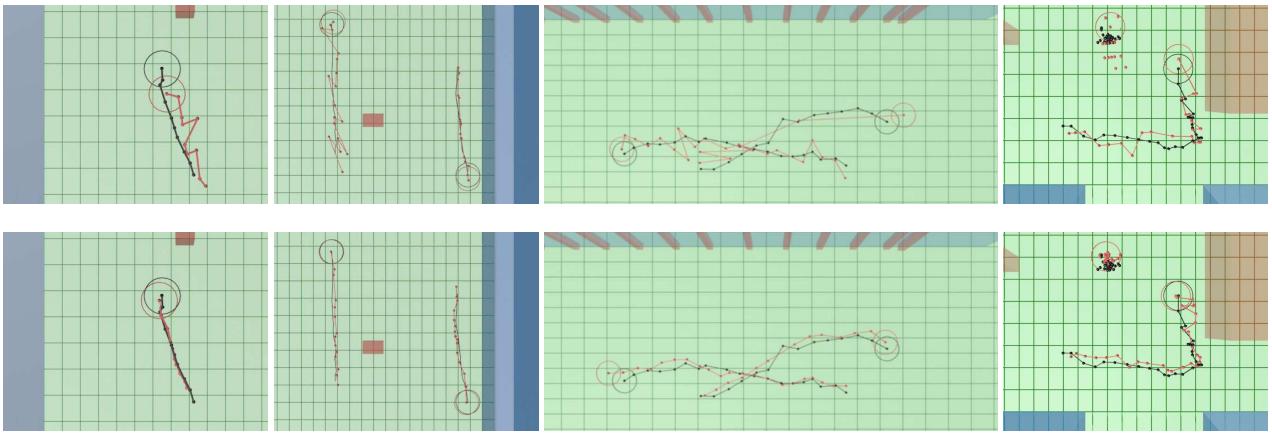


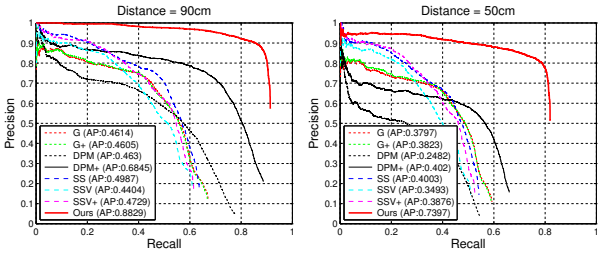
Figure 7. 3D localization trajectories of DPM (**top**) and our proposed method (**bottom**) in **red**, while the ground truth is in **black**.

projected to the ground plane using the center of the bottom of the box. Since our proposed approach is able to accurately localize pedestrians in the image plane, the projected 3D trajectories are smooth and very close to the ground truth 3D trajectories. The DPM result projected into 3D is quite jagged as the bounding box tends to move up and down during detection.

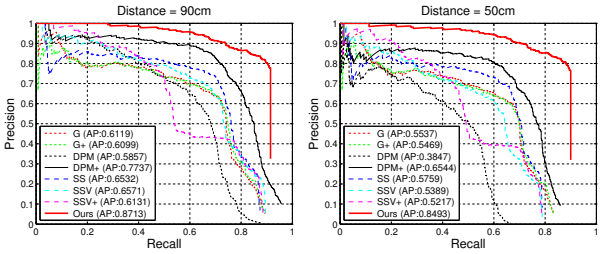
5. Conclusion

We have presented a purely synthetic approach to training scene-specific location-specific pedestrian detectors. We showed that by leveraging the parameters of the camera and known geometric layout of the scene, we are able to learn customized pedestrian models for every part of the scene. In particular, our proposed approach took into account the perspective projection of pedestrians on the image plane and also modeled pedestrian appearance under synthetic object occlusion. Our proposed algorithm jointly learns hundreds of pedestrian models using an efficient alternating algorithm, which fine tunes each pedestrian detector while also enforcing spatial smoothness between models. Our experiments showed that our model outperforms several baseline approaches in terms of image plane localization and as well as localization in 3D.

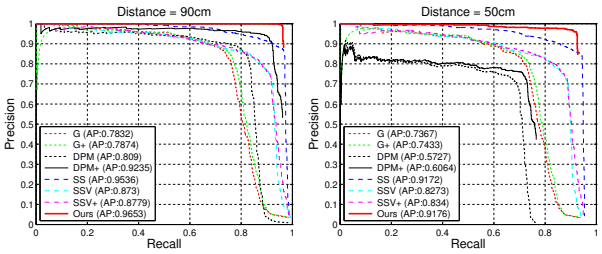
Synthesis-based training techniques are well suited for the current paradigm of data-hungry object detectors. Although, we have focused primarily on the use of scene geometry for synthesis, it is only the first step in maximizing prior scene knowledge for synthesis. We have yet to explore the more high-level semantic interpretation of the scene which can be used to generate a wider range of human poses. For example, functional attributes of the scene provide strong priors on walking direction, probable pose and likely occlusion patterns which can be used to generate a wider range of synthetic images of people. We believe that advances in functional scene understanding and improvements in human rendering techniques will enable more powerful models using our detection-from-synthesis approach.



(a) Town Center



(b) PETS 2006



(c) CMUSRD

Figure 8. Precision-Recall Curves on Town Center, PETS 2006 and CMUSRD for different amounts of distance.

References

[1] A. Agarwal and B. Triggs. A local basis representation for estimating human pose from cluttered images. In *ACCV, 2006.* 2

[2] V. Athitsos, H. Wang, and A. Stefan. A database-based framework for gesture recognition. *Personal and Ubiquitous Computing*, 14(6):511–526, 2010. 2

[3] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR, 2011.* 5

[4] V. N. Boddeti, T. Kanade, and B. Kumar. Correlation filters for object alignment. In *CVPR, 2013.* 3, 4

[5] B. Bose and E. Grimson. Improving object classification in far-field video. In *CVPR, 2004.* 3

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 4

[7] A. Broggi, A. Fascioli, P. Grisleri, T. Graf, and M. Meinecke. Model-based validation approaches and matching techniques for automotive vision based pedestrian detection. In *CVPR Workshop, 2005.* 2

[8] R. A. Brooks. Symbolic reasoning among 3-d models and 2-d images. *Artificial Intelligence*, 17(13):285 – 348, 1981. 2

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR, 2005.* 5, 6, 7

[10] M. Dhome, A. Yassine, and J.-M. Lavest. Determination of the pose of an articulated object from a single perspective view. In *BMVC, 1993.* 2

[11] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34(4):743–761, 2012. 3, 6

[12] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *PAMI*, 31(12):2179–2195, 2009. 3

[13] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *ICCV*, pages 1–8, 2007. 3

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010. 5, 6, 7

[15] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/~rbg/latent-release5/. 5

[16] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3d structure with a statistical image-based shape model. In *ICCV, 2003.* 2

[17] K. Hattori, H. Hattori, Y. Ono, K. Nishino, M. Itoh, V. N. Boddeti, and T. Kanade. Carnegie Mellon University Surveillance Research Dataset (CMUSRD). Technical report, Carnegie Mellon University, Nov. 2014. http://www.consortium.ri.cmu.edu/projSRD.php. 5

[18] M. Hejrati and D. Ramanan. Analysis by synthesis: 3d object recognition by object reconstruction. In *CVPR, 2014.* 2

[19] J. F. Henriques, J. Carreira, R. Caseiro, and J. Batista. Beyond hard negative mining: Efficient detector learning via block-circulant decomposition. In *ICCV*, 2013. 3

[20] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 80(1):3–15, 2008. 6, 7

[21] K. Lai, L. Bo, and D. Fox. Unsupervised feature learning for 3d scene labeling. In *ICRA*, 2012. 3

[22] J. Marin, D. Vázquez, D. Gerónimo, and A. M. López. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR, 2010.* 2, 3

[23] P. Matikainen, R. Sukthankar, and M. Hebert. Classifier ensemble recommendation. In *ECCV Workshop, 2012.* 3

[24] Y. Movshovitz-Attias, V. N. Boddeti, Z. Wei, and Y. Sheikh. 3d pose-by-detection of vehicles via discriminatively reduced ensembles of correlation filters. In *BMVC*, 2014. 2

[25] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In *CVPR, 2012.* 2

[26] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormahlen, and B. Schiele. Learning people detection models from few training samples. In *CVPR, 2011.* 3

[27] M. Potamias and V. Athitsos. Nearest neighbor search methods for handshape recognition. In *Proceedings of the 1st international conference on Pervasive Technologies Related to Assistive Environments*, page 30. ACM, 2008. 2

[28] J. Romero, H. Kjellstrom, and D. Kragic. Hands in action: real-time 3d reconstruction of hands in interaction with objects. In *ICRA*, 2010. 2

[29] P. M. Roth, S. Sternig, H. Grabner, and H. Bischof. Classifier grids for robust adaptive object detection. In *CVPR, 2009.* 3

[30] E. Sangineto. Statistical and spatial consensus collection for detector adaptation. In *ECCV, 2014.* 3

[31] S. Satkin, J. Lin, and M. Hebert. Data-driven scene understanding from 3d models. In *BMVC*, 2012. 3

[32] S. Stalder, H. Grabner, and L. Gool. Exploring context to learn scene specific object detectors. In *Proc. PETS*, 2009. 3

[33] S. Stalder, H. Grabner, and L. Van Gool. Cascaded confidence filtering for improved tracking-by-detection. In *ECCV, 2010.* 3

[34] B. Sun and K. Saenko. From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *BMVC, 2014.* 3

[35] G. R. Taylor, A. J. Chosak, and P. C. Brewer. Ovvv: Using virtual worlds to design and evaluate surveillance systems. In *CVPR*, pages 1–8, 2007. 3

[36] D. Thirde, L. Li, and F. Ferryman. Overview of the PETS2006 challenge. In *Proc. 9th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2006)*, pages 47–50, 2006. 5

[37] D. Vazquez, A. López, J. Marin, D. Ponsa, and D. Gerónimo. Virtual and real world adaptation for pedestrian detection. *PAMI*, 36(4):797–809, April 2014. 3, 6, 7

[38] M. Wang, W. Li, and X. Wang. Transferring a generic pedestrian detector towards specific scenes. In *CVPR, 2012.* 3

[39] M. Wang and X. Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *CVPR, 2011.* 3

[40] X. Wang, M. Wang, and W. Li. Scene-specific pedestrian detection for static video surveillance. *PAMI*, 36(2):361–374, Feb 2014. 3

[41] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *CVPR*, pages 794–801, 2009. 3

[42] J. Xu, D. Vázquez, S. Ramos, A. M. López, and D. Ponsa. Adapting a pedestrian detector by boosting lda exemplar classifiers. In *CVPR Workshop, 2013.* 3

[43] Y. Yang, G. Shu, and M. Shah. Semi-supervised learning of feature hierarchies for object detection in a video. In *CVPR, 2013.* 3