Learning Scene-Specific Pedestrian Detectors without Real Data

Hironori Hattori[1], Vishnu Naresh Boddeti[2], Kris Kitani[2], Takeo Kanade[2]
[1]Sony Corporation. [2]Carnegie Mellon University.

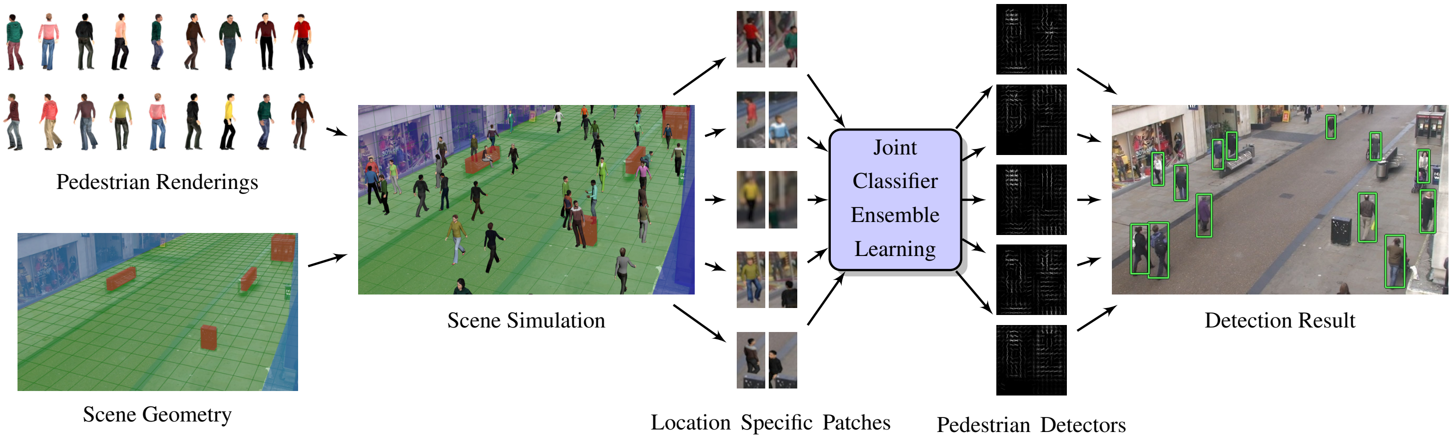

Figure 1: **Overview:** For every grid location, geometrically correct renderings of pedestrian are synthetically generated using known scene information such as camera calibration parameters, obstacles (red), walls (blue) and walkable areas (green). All location-specific pedestrian detectors are trained jointly to learn a smoothly varying appearance model. Multiple scene-and-location-specific detectors are run in parallel at every grid location.

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 (*i.e.,* 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. One can use a generic pedestrian detection algorithm trained over copious amounts of real data to 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 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. 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[2], DPM[3]).

The above 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 the form of static location of objects, ground plane and walls). In this work, we show that with this information, it is possible to generate geometrically accurate simulations of pedestrian appearance as training data (computer generated pedestrians) to act as a proxy for the real data. This allows us to learn a highly accurate scene-specific pedestrian detector. 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 that outperforms several baseline techniques.

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. The main contributions of our work are 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 discriminatively learning a large number of scene-specific *location-specific* pedestrian detectors. Our algorithmic framework makes use of highly-efficient correlation filters[1] 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.

Our experiments showed that our model outperforms several baseline approaches–classical pedestrian detection models, hybrid synthetic-real models and a baseline which leverages the scene geometry and camera calibration parameters at the inference stage–both in terms of image plane localization as well as localization in 3D for the task of scene-specific pedestrian detection. More importantly 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.

[1] Vishnu Naresh Boddeti, Takeo Kanade, and BVK Kumar. Correlation filters for object alignment. In *CVPR, 2013.*

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

[3] 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.