

The Fastest Pedestrian Detector in the West

Piotr Dollár¹
pdollar@caltech.edu
Serge Belongie²
sjb@cs.ucsd.edu
Pietro Perona¹
perona@caltech.edu

¹ Dept. of Electrical Engineering
California Institute of Technology
Pasadena, CA, USA

² Dept. of Computer Science and Eng.
University of California, San Diego
San Diego, CA, USA

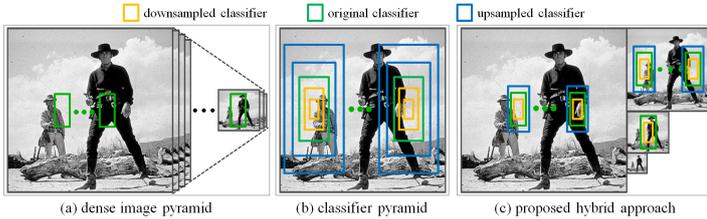


Figure 1: In many applications, detection speed is as important as accuracy. (a) A standard pipeline for performing modern multiscale detection is to create a densely sampled image pyramid, compute features at each scale, and finally perform sliding window classification (with a fixed scale model). Although effective; the creation of the feature pyramid can dominate the cost of detection, leading to slow multiscale detection. (b) Viola and Jones [6] utilized shift and scale invariant features, allowing a trained detector to be placed at any location and scale without relying on an image pyramid. Constructing such a *classifier pyramid* results in fast multiscale detection; unfortunately, most features are *not* scale invariant, including gradient histograms, significantly limiting the generality of this scheme. (c) We propose a fast method for approximating features at multiple scales using a sparsely sampled image pyramid with a step size of an entire octave and within each octave we use a classifier pyramid. The proposed approach achieves nearly the same accuracy as using densely sampled image pyramids, with nearly the same speed as using a classifier pyramid applied to an image at a single scale.

We demonstrate a multiscale pedestrian detector operating in near real time with state-of-the-art detection performance. The computational bottleneck of many modern detectors is the construction of an image pyramid, typically sampled at 8-16 scales per octave, and associated feature computations at each scale. We propose a technique to avoid constructing such a finely sampled image pyramid without sacrificing performance: our key insight is that for a broad family of features, including gradient histograms, the feature responses computed at a single scale can be used to approximate feature responses at nearby scales. This allows us to decouple the sampling of the image pyramid from the sampling of detection scales. An overview of our approach is shown if Figure 1.

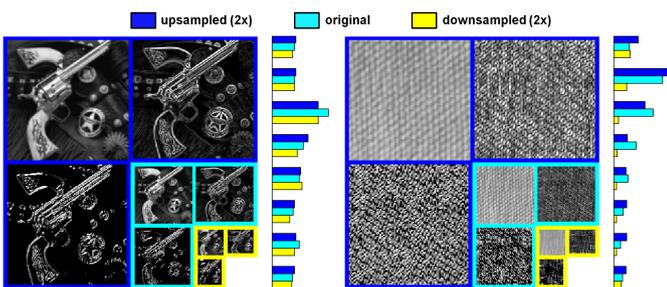


Figure 2: Approximating features in resampled images. For each image set, we take the original image (cyan border) and generate an upsampled (blue) and downsampled (yellow) version. Shown at each scale are the image (center), gradient magnitude (right), and gradient orientation (bottom). At each scale we compute a gradient histogram with 8 bins, adjusted according to the approximations developed in the paper. Assuming these approximations hold, the three normalized gradient histograms should be roughly equal. In the first case, the approximations are fairly accurate. In the second case, showing a highly structured Brodatz texture with significant high frequency content, the downsampling approximation fails.

In order to understand how information behaves in resampled images, we turn to the study of natural image statistics. Ruderman and Bialek [5] showed that various statistics of natural images are independent of the scale at which the images were captured, or in other words, that the statistics of an image are independent of the scene area corresponding to a single pixel. Using this observation, we derive an exponential law governing how feature responses vary with changes in scale, which can in turn be used to predict features in resampled images (see Figure 2). That such an approach is possible is not entirely trivial and relies on the fractal

structure of the visual world; nevertheless, the mathematical foundations we develop should be readily applicable to other problems. Full details are given in the paper.

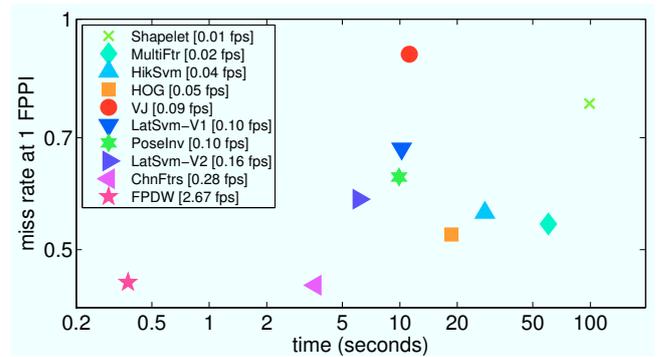


Figure 3: Time versus detection rate at 1 false positive per image on 640×480 images from the Caltech Pedestrian Dataset [4] for detecting pedestrians 50 pixels and up (times are much faster for larger pedestrians/smaller images). Run times of all algorithms are normalized to the rate of a single modern machine. *FPDW* obtains a speedup of about 10-100 compared to competing methods with a detection rate within a few percent of best reported performance.

For the detection experiments, we adopted the *ChnFtrs* detector described in [3]. No re-training was necessary for this work; instead, we rescale the pre-trained *ChnFtrs* detector [3] using the approximations developed in the paper. We refer to our method as the ‘Fastest Pedestrian Detector in the West’ (*FPDW*). Detailed timing results are reported in Figure 3. In Figure 4 we show full-image results on two datasets [2, 4]. In all cases the detection rate of *FPDW* is within 1-2% of the top performing algorithm, and always quite close to the original *ChnFtrs* classifier, all while being 1-2 orders of magnitude faster than competing methods. The proposed approach is general and should be widely applicable.

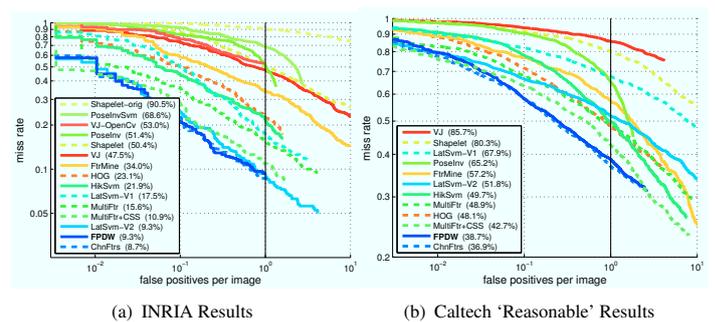


Figure 4: Detection results on the INRIA [2] and Caltech [4] pedestrian datasets (legends are ordered by miss rate at 1 false positive per image – lower is better). Results on additional datasets and under varying scenarios can be found in the paper and online at [1]. In all cases the detection rate of *FPDW* is within a few percent of *ChnFtrs* while being 1-2 orders of magnitude faster than all competing methods. Evaluation scripts, detector descriptions, and additional results are available at [1].

- [1] www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/.
- [2] N. Dalal and B. Triggs. Histogram of oriented gradient for human detection. In *CVPR*, 2005.
- [3] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009.
- [4] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009.
- [5] D. L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, 1994.
- [6] P. Viola and M. Jones. Fast multi-view face detection. In *CVPR*, 2001.