Pedestrian Detection: An Evaluation of the State of the Art

Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona

Abstract—Pedestrian detection is a key problem in computer vision, with several applications that have the potential to positively impact quality of life. In recent years, the number of approaches to detecting pedestrians in monocular images has grown steadily. However, multiple datasets and widely varying evaluation protocols are used, making direct comparisons difficult. To address these shortcomings, we perform an extensive evaluation of the state of the art in a unified framework. We make three primary contributions: (1) we put together a large, well-annotated and realistic monocular pedestrian detection dataset and study the statistics of the size, position and occlusion patterns of pedestrians in urban scenes, (2) we propose a refined per-frame evaluation methodology that allows us to carry out probing and informative comparisons, including measuring performance in relation to scale and occlusion, and (3) we evaluate the performance of sixteen pre-trained state-of-the-art detectors across six datasets. Our study allows us to assess the state of the art and provides a framework for gauging future efforts. Our experiments show that despite significant progress, performance still has much room for improvement. In particular, detection is disappointing at low resolutions and for partially occluded pedestrians.

Index Terms—pedestrian detection, object detection, benchmark, evaluation, dataset, Caltech Pedestrian Dataset

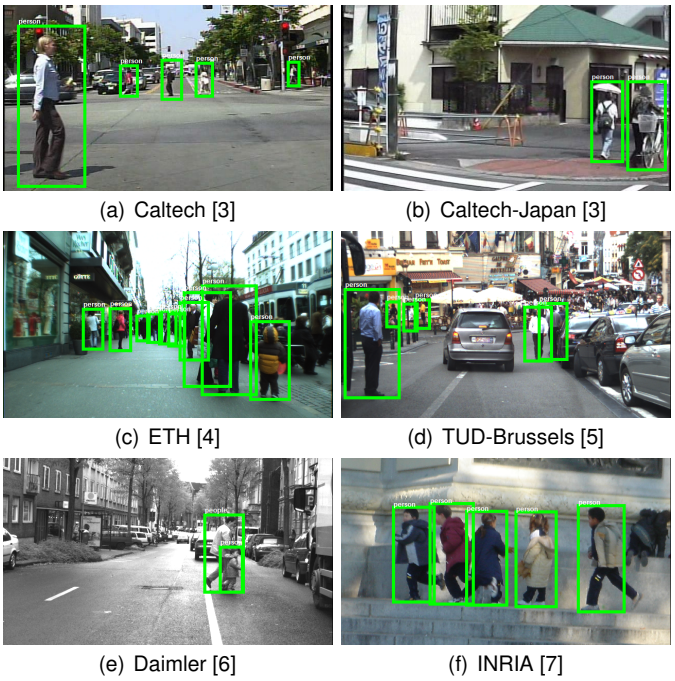
1 Introduction

People are among the most important components of a machine's environment, and endowing machines with the ability to interact with people is one of the most interesting and potentially useful challenges for modern engineering. Detecting and tracking people is thus an important area of research, and machine vision is bound to play a key role. Applications include robotics, entertainment, surveillance, care for the elderly and disabled, and content-based indexing. Just in the US, nearly 5,000 of the 35,000 annual traffic crash fatalities involve pedestrians [1], hence the considerable interest in building automated vision systems for detecting pedestrians [2].

While there is much ongoing research in machine vision approaches for detecting pedestrians, varying evaluation protocols and use of different datasets makes direct comparisons difficult. Basic questions such as "Do current detectors work well?", "What is the best approach?", "What are the main failure modes?" and "What are the most productive research directions?" are not easily answered.

Our study aims to address these questions. We focus on methods for detecting pedestrians in individual monocular images; for an overview of how detectors are incorporated into full systems we refer readers to [2]. Our approach is three-pronged: we collect, annotate and study a large dataset of pedestrian images collected from a vehicle navigating in urban traffic; we develop informative evaluation methodologies and point out pitfalls in previous experimental procedures; finally, we compare the performance of sixteen pre-trained pedestrian detectors on six publicly available datasets, including our own. Our study allows us to assess the state of the art and suggests directions for future research.

All results of this study, and the data and tools for reproducing them, are posted on the project website: *www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/*.



(a) Caltech [3] (b) Caltech-Japan [3] (c) ETH [4] (d) TUD-Brussels [5] (e) Daimler [6] (f) INRIA [7]

Fig. 1. Example images (cropped) and annotations from six pedestrian detection datasets. We perform an extensive evaluation of pedestrian detection, benchmarking *sixteen* detectors on each of these *six* datasets. By using multiple datasets and a unified evaluation framework we can draw broad conclusion about the state of the art and suggest future research directions.

- *P. Dollár and P. Perona are with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA.*
- *C. Wojek and B. Schiele are with MPI Informatics, Saarbrücken, Germany.*

1.1 Contributions

Dataset: In earlier work [3], we introduced the Caltech Pedestrian Dataset, which includes 350,000 pedestrian bounding boxes labeled in 250,000 frames and remains the largest such dataset to date. Occlusions and temporal correspondences are also annotated. Using the extensive ground truth, we analyze the statistics of pedestrian scale, occlusion, and location and help establish conditions under which detection systems must operate.

Evaluation Methodology: We aim to quantify and rank detector performance in a realistic and unbiased manner. To this effect, we explore a number of choices in the evaluation protocol and their effect on reported performance. Overall, the methodology has changed substantially since [3], resulting in a more accurate and informative benchmark.

Evaluation: We evaluate *sixteen* representative state-of-the-art pedestrian detectors (previously we evaluated seven [3]). Our goal was to choose diverse detectors that were most promising in terms of originally reported performance. We avoid retraining or modifying the detectors to ensure each method was optimized by its authors. In addition to overall performance, we explore detection rates under varying levels of scale and occlusion and on clearly visible pedestrians. Moreover, we measure localization accuracy and analyze runtime.

To increase the scope of our analysis, we also benchmark the sixteen detectors using a unified evaluation framework on *six* additional pedestrian detection datasets including the ETH [4], TUD-Brussels [5], Daimler [6] and INRIA [7] datasets and two variants of the Caltech dataset (see Figure 1). By evaluating across multiple datasets, we can rank detector performance and analyze the statistical significance of the results and, more generally, draw conclusions both about the detectors and the datasets themselves.

Two groups have recently published surveys which are complementary to our own. Geronimo et al. [2] performed a comprehensive survey of pedestrian detection for advanced driver assistance systems, with a clear focus on full systems. Enzweiler and Gavrila [6] published the Daimler detection dataset and an accompanying evaluation of three detectors, performing additional experiments integrating the detectors into full systems. We instead focus on a more thorough and detailed evaluation of state-of-the-art detectors.

This paper is organized as follows: we introduce the Caltech Pedestrian Dataset and analyze its statistics in §2; a comparison of existing datasets is given in §2.4. In §3 we discuss evaluation methodology in detail. A survey of pedestrian detectors is given in §4.1 and in §4.2 we discuss the sixteen representative state-of-the-art detectors used in our evaluation. In §5 we report the results of the performance evaluation, both under varying conditions using the Caltech dataset and on six additional datasets. We conclude with a discussion of the state of the art in pedestrian detection in §6.

(a)

total frames	∼1000k
labeled frames	∼250k
frames w peds.	∼132k
# bounding boxes	∼350k
# occluded BB	∼126k
# unique peds.	∼2300
ave ped. duration	∼5s
ave labels/frame	∼1.4
labeling time	∼400h

(b)

Fig. 2. Overview of the Caltech Pedestrian Dataset. (a) Camera setup. (b) Summary of dataset statistics ($1k = 10^3$). The dataset is large, realistic and well-annotated, allowing us to study statistics of the size, position and occlusion of pedestrians in urban scenes and also to accurately evaluate the state or the art in pedestrian detection.

2 The Caltech Pedestrian Dataset

Challenging datasets are catalysts for progress in computer vision. The Barron et al. [8] and Middlebury [9] optical flow datasets, the Berkeley Segmentation Dataset [10], the Middlebury Stereo Dataset [11], and the Caltech 101 [12], Caltech 256 [13] and PASCAL [14] object recognition datasets all improved performance evaluation, added challenge, and helped drive innovation in their respective fields. Much in the same way, our goal in introducing the Caltech Pedestrian Dataset is to provide a better benchmark and to help identify conditions under which current detectors fail and thus focus research effort on these difficult cases.

2.1 Data Collection and Ground Truthing

We collected approximately 10 hours of 30Hz video ($\sim 10^6$ frames) taken from a vehicle driving through regular traffic in an urban environment (camera setup shown in Figure 2(a)). The CCD video resolution is 640×480, and, not unexpectedly, the overall image quality is lower than that of still images of comparable resolution. There are minor variations in the camera position due to repeated mountings of the camera. The driver was independent from the authors of this study and had instructions to drive normally through neighborhoods in the greater Los Angeles metropolitan area chosen for their relatively high concentration of pedestrians including LAX, Santa Monica, Hollywood, Pasadena, and Little Tokyo. In order to remove effects of the vehicle pitching and thus simplify annotation, the video was stabilized using the inverse compositional algorithm for image alignment by Baker and Matthews [15].

After video stabilization, 250,000 frames (in 137 approximately minute long segments extracted from the 10 hours of video) were annotated for a total of 350,000 bounding boxes around 2300 unique pedestrians. To make such a large scale labeling effort feasible we created a user-friendly labeling tool, shown in Figure 3. Its most salient aspect is an interactive procedure where the annotator labels a sparse set of frames and the system automatically predicts pedestrian positions in intermediate frames. Specifically, after an annotator labels a

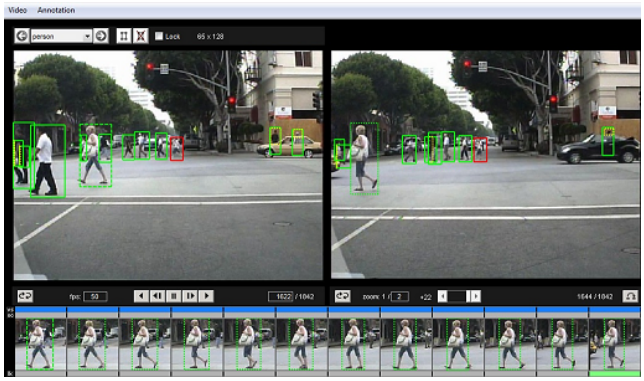


Fig. 3. The annotation tool allows annotators to efficiently navigate and annotate a video in a minimum amount of time. Its most salient aspect is an interactive procedure where the annotator labels only a sparse set of frames and the system automatically predicts pedestrian positions in intermediate frames. The annotation tool is available on the project website.

bounding box (BB) around the same pedestrian in at least two frames, BBs in intermediate frames are interpolated using cubic interpolation (applied independently to each coordinate of the BBs). Thereafter, every time an annotator alters a BB, BBs in all the unlabeled frames are re-interpolated. The annotator continues until satisfied with the result. We experimented with more sophisticated interpolation schemes, including relying on tracking; however, cubic interpolation proved best. Labeling the ∼2.3 hours of video, including verification, took ∼400 hours total (spread across multiple annotators).

For every frame in which a given pedestrian is visible, annotators mark a BB that indicates the full extent of the *entire* pedestrian (BB-full); for occluded pedestrians this involves estimating the location of hidden parts. In addition a second BB is used to delineate the visible region (BB-vis), see Figure 5(a). During an occlusion event, the estimated full BB stays relatively constant while the visible BB may change rapidly. For comparison, in the PASCAL labeling scheme [14] only the visible BB is labeled and occluded objects are marked as 'truncated'.

Each sequence of BBs belonging to a single object was assigned one of three labels. Individual pedestrians were labeled 'Person' (∼1900 instances). Large groups for which it would have been tedious or impossible to label individuals were delineated using a single BB and labeled as 'People' (∼300). In addition, the label 'Person?' was assigned when clear identification of a pedestrian was ambiguous or easily mistaken (∼110).

2.2 Dataset Statistics

A summary of the dataset is given in Figure 2(b). About 50% of the frames have no pedestrians, while 30% have two or more, and pedestrians are visible for 5s on average. Below, we analyze the distribution of pedestrian scale, occlusion and location. This serves to establish the requirements of a real world system and to help identify constraints that can be used to improve automatic pedestrian detection systems.

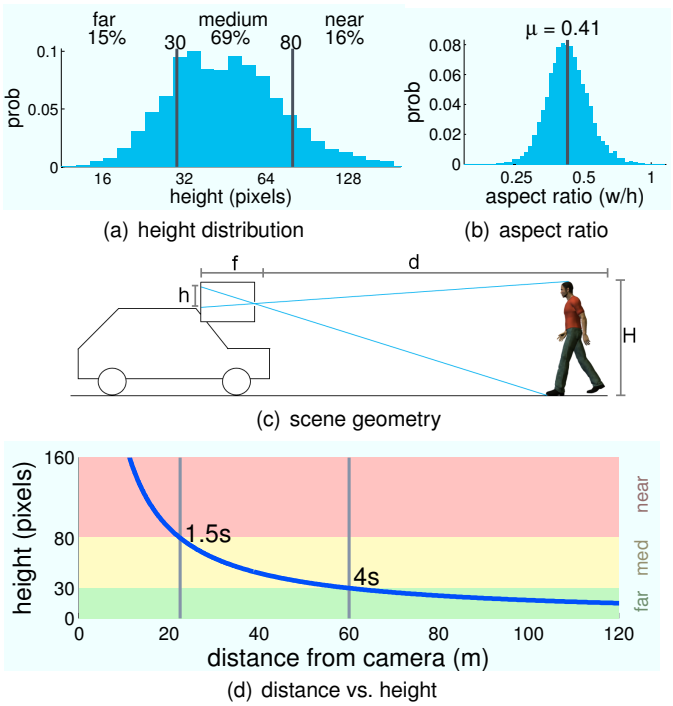


Fig. 4. (a) Distribution of pedestrian pixel heights. We define the near scale to include pedestrians over 80 pixels, the medium scale as 30-80 pixels, and the far scale as under 30 pixels. Most observed pedestrians (∼69%) are at the medium scale. (b) Distribution of BB aspect ratio; on average $w \approx .41h$. (c) Using the pinhole camera model, a pedestrian's pixel height h is inversely proportional to distance to the camera d: $h/f \approx H/d$. (d) Pixel height h as a function of distance d. Assuming an urban speed of 55 km/h, an 80 pixel person is just 1.5s away, while a 30 pixel person is 4s away. Thus, for automotive settings, detection is most important at medium scales (see §2.2.1 for details).

2.2.1 Scale Statistics

We group pedestrians by their image size (height in pixels) into three scales: near (80 or more pixels), medium (between 30-80 pixels) and far (30 pixels or less). This division into three scales is motivated by the distribution of sizes in the dataset, human performance and automotive system requirements.

In Figure 4(a), we histogram the heights of the 350,000 BBs using logarithmic sized bins. The heights are roughly lognormally distributed with a median of 48 pixels and a log-average of 50 pixels (the log-average is equivalent to the geometric mean and is more representative of typical values for lognormally distributed data than the arithmetic mean, which is 60 pixels in this case). Cutoffs for the near/far scales are marked. Note that ∼69% of the pedestrians lie in the medium scale, and that the cutoffs for the near/far scales correspond to about ± 1 standard deviation (in log space) from the log-average height of 50 pixels. Below 30 pixels, annotators have difficulty identifying pedestrians reliably.

Pedestrian width is likewise lognormally distributed, and moreover so is the joint distribution of width and height (not shown). As any linear combination of the components of a multivariate normal distribution is

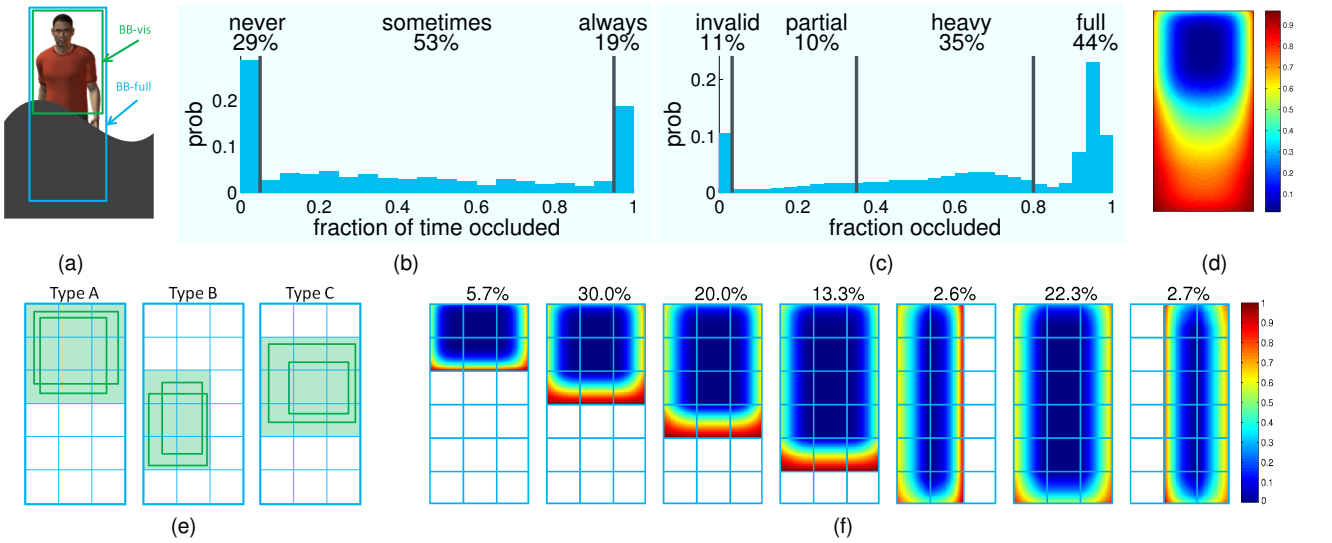


Fig. 5. Occlusion statistics. (a) For all occluded pedestrians annotators labeled both the full extent of the pedestrian (BB-full) and the visible region (BB-vis). (b) Most pedestrians (70%) are occluded in at least one frame, underscoring the importance of detecting occluded people. (c) Fraction of occlusion can vary significantly (0% occlusion indicates that a BB could not represent the extent of the visible region). (d) Occlusion is far from uniform with pedestrians typically occluded from below. (e) To observe further structure in the types of occlusions that actually occur, we quantize occlusion into a fixed number of types. (f) Over 97% of occluded pedestrians belong to just a small subset of the hundreds of possible occlusion types. Details in §2.2.2.

also normally distributed, so should the BB aspect ratio (defined as w/h) since $log(w/h) = log(w) - log(h)$. A histogram of the aspect ratios, using logarithmic bins, is shown in Figure 4(b), and indeed the distribution is lognormal. The log-average aspect ratio is $.41$, meaning that typically $w \approx .41h$. However, while BB height does not vary considerably given a constant distance to the camera, the BB width can change with the pedestrian's pose (especially arm positions and relative angle). Thus, although we could have defined the near, medium and far scales using the width, the consistency of the height makes it better suited.

Detection in the medium scale is essential for automotive applications. We chose a camera setup that mirrors expected automotive settings: 640×480 resolution, $27°$ vertical field of view, and focal length fixed at 7.5mm. The focal length in pixels is $f \approx 1000$ (obtained from $480/2/f = \tan(27°/2)$ or using the camera's pixel size of 7.5μm). Using a pinhole camera model (see Figure 4(c)), an object's observed pixel height h is inversely proportional to the distance d to the camera: $h \approx Hf/d$, where H is the true object height. Assuming $H \approx 1.8$m tall pedestrians, we obtain $d \approx 1800/h$ m. With the vehicle traveling at an urban speed of 55 km/h (~15 m/s), an 80 pixel person is just 1.5s away, while a 30 pixel person is 4s away (see Figure 4(d)). Thus detecting near scale pedestrians may leave insufficient time to alert the driver, while far scale pedestrians are less relevant.

We shall use the near, medium, and far scale definitions throughout this work. Most pedestrians are observed at the medium scale and for safety systems detection must occur in this scale as well. Human performance is also quite good in the near and medium scales but degrades noticeably at the far scale. However, most current detectors are designed for the near scale and perform poorly even at the medium scale (see §5). Thus there is an important mismatch in current research efforts and the requirements of real systems. Using higher resolution cameras would help; nevertheless, given the good human performance and lower cost, we believe that accurate detection in the medium scale is an important and reasonable goal.

2.2.2 *Occlusion Statistics*

Occluded pedestrians were annotated with two BBs that denote the visible and full pedestrian extent (see Figure 5(a)). We plot frequency of occlusion in Figure 5(b), i.e., for each pedestrian we measure the fraction of frames in which the pedestrian was at least somewhat occluded. The distribution has three distinct regions: pedestrians that are never occluded (29%), occluded in some frames (53%) and occluded in all frames (19%). Over 70% of pedestrians are occluded in at least one frame, underscoring the importance of detecting occluded people. Nevertheless, little previous work has been done to quantify occlusion or detection performance in the presence of occlusion (using real data).

For each occluded pedestrian, we can compute the fraction of occlusion as one minus the visible pedestrian area divided by total pedestrian area (calculated from the visible and full BBs). Aggregating, we obtain the histogram in Figure 5(c). Over 80% occlusion typically indicates full occlusion, while 0% is used to indicate that a BB could not represent the extent of the visible region (e.g. due to a diagonal occluder). We further subdivide the cases in between into *partial* occlusion (1-35% area occluded) and *heavy* occlusion (35-80% occluded).

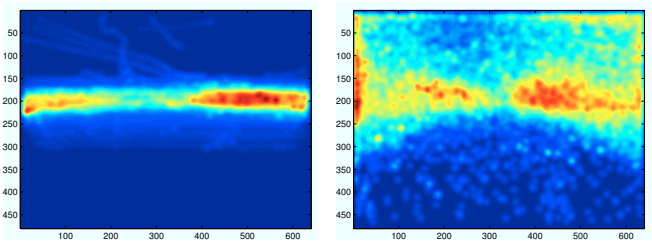
We investigated which regions of a pedestrian were most likely to be occluded. For each frame in which a pedestrian was partially to heavily occluded (1-80% fraction of occlusion), we created a binary 50×100 pixel occlusion mask using the visible and full BBs. By averaging the resulting $\sim$54k occlusion masks, we computed the probability of occlusion for each pixel (conditioned on the person being partially occluded); the resulting heat map is shown in Figure 5(d). Observe the strong bias for the lower portion of the pedestrian to be occluded, particulary the feet, and for the top portion, especially the head, to be visible. An intuitive explanation is that most occluding objects are supported from below as opposed to hanging from above (another but less likely possibility is that it is difficult for annotators to detect pedestrians if only the feet are visible). Overall, occlusion is far from uniform, and exploiting this finding could help improve the performance of pedestrian detectors.

Not only is occlusion highly non-uniform, there is significant additional structure in the types of occlusions that actually occur. Below, we show that after quantizing occlusion masks into a large number of possible types, nearly all occluded pedestrians belong to just a handful of the resulting types. To quantize the occlusions, each BB-full is registered to a common reference BB that has been partitioned into q_x by q_y regularly spaced cells; each BB-vis can then be assigned a type according to the smallest set of cells that fully encompass it. Figure 5(e) shows 3 example types for $q_x = 3, q_y = 6$ (with two BB-vis per type). There are a total of $\sum_{i=1,j=1}^{q_x,q_y} ij = q_x q_y (q_x + 1)(q_y + 1)/4$ possible types. For each, we compute the percentage of the $\sim$54k occlusions assigned to it and produce a heat map using the corresponding occlusion masks. The top 7 of 126 types for $q_x = 3, q_y = 6$ are shown in Figure 5(f). Together, these 7 types account for nearly 97% of all occlusions in the dataset. As can be seen, pedestrians are almost always occluded from either below or the side; more complex occlusions are rare. We repeated the same analysis with a finer partitioning of $q_x = 4, q_y = 8$ (not shown). Of the resulting 360 possible types the top 14 accounted for nearly 95% of occlusions. The knowledge that very few occlusion patterns are common should prove useful in detector design.

2.2.3 Position Statistics

Viewpoint and ground plane geometry (Figure 4(c)) constrain pedestrians to appear only in certain regions of the image. We compute the expected center position and plot the resulting heat map, log-normalized, in Figure 6(a). As can be seen pedestrians are typically located in a narrow band running horizontally across the center of the image (y-coordinate varies somewhat with distance/height). Note that the same constraints are not valid when photographing a scene from arbitrary viewpoints, e.g. in the INRIA dataset.

In the collected data, many objects, not just pedestrians, tend to be concentrated in this same region. In Figure 6(b) we show a heat map obtained by using BBs generated by the HOG [7] pedestrian detector with a low threshold. About half of the detections, including both true and false positives, occur in the same band as the ground truth. Thus incorporating this constraint could considerably speed up detection but it would only moderately reduce false positives.



(a) pedestrian position (b) HOG detections

Fig. 6. Expected center location of pedestrian BBs for (a) ground truth and (b) HOG detections. The heat maps are log-normalized, meaning pedestrian location is even more concentrated than immediately apparent.

2.3 Training and Testing Data

We split the dataset into training/testing sets and specify a precise evaluation methodology, allowing different research groups to compare detectors directly. We urge authors to adhere to one of four training/testing scenarios described below.

The data was captured over 11 sessions, each filmed in one of five city neighborhoods as described. We divide the data roughly in half, setting aside 6 sessions for training (S0-S5) and 5 sessions for testing (S6-S10). For detailed statistics about the amount of data see bottom row of Table 1. Images from all sessions (S0-S10) have been publicly available, as have been annotations for the training sessions (S0-S5). At this time we are also releasing annotations for the testing sessions (S6-S10).

Detectors can be trained using either the Caltech training data (S0-S5) or any 'external' data, and tested on either the Caltech training data (S0-S5) or testing data (S6-S10). This results in four evaluation scenarios:

- *Scenario ext0:* Train on any external data, test on S0-S5.
- *Scenario ext1:* Train on any external data, test on S6-S10.
- *Scenario cal0*: Perform 6-fold cross validation using S0-S5. In each phase use 5 sessions for training and the 6^{th} for testing, then merge and report results over S0-S5.
- *Scenario cal1*: Train using S0-S5, test on S6-S10.

Scenarios ext0/ext1 allow for evaluation of existing, pre-trained pedestrian detectors, while cal0/cal1 involve training using the Caltech training data (S0-S5). The results reported here use the ext0/ext1 scenarios thus allowing for a broad survey of existing pre-trained pedestrian detectors. Authors are encouraged to re-train their systems on our large training set and evaluate under scenarios cal0/cal1. Authors should use ext0/cal0 during detector development, and only after finalizing all parameters evaluate under scenarios ext1/cal1.

			Training			Testing			Height			Properties						
		imaging setup	# pedestrians	# neg. images	# pos. images	# pedestrians	# neg. images	# pos. images	10% quantile	median	90% quantile	color images	per-image eval.	no select. bias	video seqs.	temporal corr.	occlusion labels	publication
MIT	[16]	photo	924	–	–	–	–	–	128	128	128	✓						2000
USC-A	[17]	photo	–	–	–	313	–	205	70	98	133		✓					2005
USC-B	[17]	surv.	–	–	–	271	–	54	63	90	126		✓					2005
USC-C	[18]	photo	–	–	–	232	–	100	74	108	145		✓					2007
CVC	[19]	mobile	1000	6175†	–	–	–	–	46	83	164	✓		✓				2007
TUD-det	[20]	mobile	400	–	400	311	–	250	133	218	278	✓	✓					2008
Daimler-CB	[21]	mobile	2.4k	15k†	–	1.6k	10k†	–	36	36	36			✓				2006
NICTA	[22]	mobile	18.7k	5.2k	–	6.9k	50k†	–	72	72	72	✓		✓				2008
INRIA	[7]	photo	1208	1218	614	566	453	288	139	279	456	✓						2005
ETH	[4]	mobile	2388	–	499	12k	–	1804	50	90	189	✓	✓	✓	✓			2007
TUD-Brussels	[5]	mobile	1776	218	1092	1498	–	508	40	66	112	✓	✓	✓				2009
Daimler-DB	[6]	mobile	15.6k	6.7k	–	56.5k	–	21.8k	21	47	84		✓		✓	✓		2009
Caltech	[3]	mobile	192k	61k	67k	155k	56k	65k	27	48	97	✓	✓	✓	✓	✓	✓	2009

TABLE 1
Comparison of Pedestrian Detection Datasets (see §2.4 for details)

2.4 Comparison of Pedestrian Datasets

Existing datasets may be grouped into two types: (1) 'person' datasets containing people in unconstrained pose in a wide range of domains and (2) 'pedestrian' datasets containing upright, possibly moving people. The most widely used 'person' datasets include subsets of the MIT LabelMe data [23] and the PASCAL VOC datasets [14]. In this work we focus on pedestrian detection, which is more relevant to automotive safety.

Table 1 provides an overview of existing pedestrian datasets. The datasets are organized into three groups. The first includes older or more limited datasets. The second includes more comprehensive datasets including the INRIA [7], ETH [4] and TUD-Brussels [5] pedestrian datasets and the Daimler detection benchmark (Daimler-DB) [6]. The final row contains information about the Caltech Pedestrian Dataset. Details follow below.

Imaging setup: Pedestrians can be labeled in photographs [7], [16], surveillance video [17], [24], and images taken from a mobile recording setup, such as a robot or vehicle [4], [5], [6]. Datasets gathered from photographs suffer from *selection bias*, as photographs are often manually selected, while surveillance videos have restricted backgrounds and thus rarely serve as a basis for detection datasets. Datasets collected by continuously filming from a mobile recording setup, such as the Caltech Pedestrian Dataset, largely eliminate selection bias (unless some scenes are staged by actors, as in [6]) while having moderately diverse scenes.

Dataset size: The amount and type of data in each dataset is given in the next six columns. The columns are: number of pedestrian windows (not counting reflections, shifts, etc.), number of images with no pedestrians (a † indicates cropped negative windows only), and number of uncropped images containing at least one pedestrian. The Caltech Pedestrian Dataset is two orders of magnitude larger than most existing datasets.

Dataset type: Older datasets, including the MIT [16], CVC [19] and NICTA [22] pedestrian datasets and the Daimler classification benchmark (Daimler-CB) [21] tend to contain cropped pedestrian windows only. These are known as 'classification' datasets as their primary use is to train and test binary classification algorithms. In contrast, datasets that contain pedestrians in their original context are known as 'detection' datasets and allow for the design and testing of full-image detection systems. The Caltech dataset along with all the datasets in the second set (INRIA, ETH, TUD-Brussels and Daimler-DB) can serve as 'detection' datasets.

Pedestrian scale: Table 1 additionally lists the 10^{th} percentile, median and 90^{th} percentile pedestrian pixel heights for each dataset. While the INRIA dataset has fairly high resolution pedestrians, most datasets gathered from mobile platforms have median heights that range from 50-100 pixels. This emphasizes the importance of detection of low resolution pedestrians, especially for applications on mobile platforms.

Dataset properties: The final columns summarize additional dataset features including the availability of color images, video data, temporal correspondence between BBs and occlusion labels, and whether 'per-image' evaluation and unbiased selection criteria were used.

As mentioned, in our performance evaluation we additionally use the INRIA [7], ETH [4], TUD-Brussels [5] and Daimler-DB [6] datasets. The INRIA dataset helped drive recent advances in pedestrian detection and remains one of the most widely used despite its limitations. Much like the Caltech dataset, the ETH, TUD-Brussels and Daimler-DB datasets are all captured in urban settings using a camera mounted to a vehicle (or stroller in the case of ETH). While being annotated in less detail than the Caltech dataset (see Table 1), each can serve as 'detection' dataset and is thus suitable for use in our evaluation.

We conclude by summarizing the most important and novel aspects of the Caltech Pedestrian Dataset. The dataset includes $O(10^5)$ pedestrian BBs labeled in $O(10^5)$ frames and remains the largest such dataset to date. It contains color video sequences and includes pedestrians with a large range of scales and more scene variability than typical pedestrian datasets. Finally, it is the only dataset with detailed occlusion labels and one of the few to provide temporal correspondence between BBs.

3 Evaluation Methodology

Proper evaluation methodology is a crucial and surprisingly tricky topic. In general, there is no single 'correct' evaluation protocol. Instead, we have aimed to make our evaluation protocol quantify and rank detector performance in a realistic, unbiased and informative manner.

To allow for exact comparisons, we have posted the evaluation code, ground truth annotations and detection results for all detectors on all datasets on the project website. Use of the exact same evaluation code (as opposed to a re-implementation) ensures consistent and reproducible comparisons. Additionally, given all the detector outputs, practitioners can define novel performance metrics with which to re-evaluate the detectors. This flexibility is important because while we make every effort to define realistic and informative protocols, performance evaluation is ultimately task dependent.

Overall, the evaluation protocol has changed substantially since our initial version described in [3], resulting in a more accurate and informative evaluation of the state of the art. We begin with an overview of full image evaluation in §3.1. Next, we discuss evaluation using subsets of the ground truth and detections in §3.2 and §3.3, respectively. In §3.4 we propose and motivate standardizing BB aspect ratio. Finally, in §3.5, we examine the alternative per-window evaluation methodology.

3.1 Full Image Evaluation

We perform single frame evaluation using a modified version of the scheme laid out in the PASCAL object detection challenges [14]. A detection system needs to take an image and return a BB and a score or confidence for each detection. The system should perform multiscale detection and any necessary non-maximal suppression (NMS) for merging nearby detections. Evaluation is performed on the final output: the list of detected BBs.

A detected BB (BB_{dt}) and a ground truth BB (BB_{gt}) form a potential match if they overlap sufficiently. Specifically, we employ the PASCAL measure, which states that their area of overlap must exceed 50%:

$$a_o \doteq \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} > 0.5 \qquad (1)$$

The evaluation is insensitive to the exact threshold as long as it is below about .6, see Figure 7. For larger values performance degrades rapidly as improved localization accuracy is necessary; thus, to focus on detection accuracy, we use the standard threshold of .5 throughout.

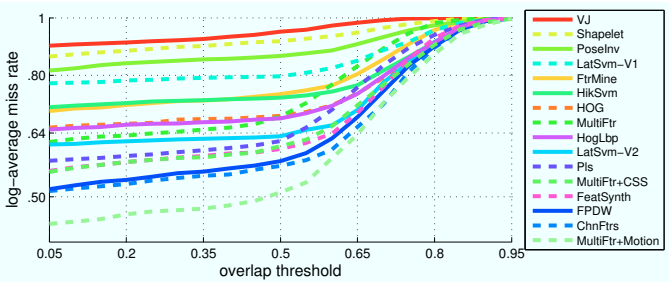


Fig. 7. Log-average miss rates for 50 pixel or taller pedestrians as a function of the threshold on overlap area (see Eqn. (1)). Decreasing the threshold below .5 has little affect on reported performance. However, increasing it over $\sim$.6 results in rapidly increasing log-average miss rates as improved localization accuracy is necessary.

Each BB_{dt} and BB_{gt} may be matched at most once. We resolve any assignment ambiguity by performing the matching greedily. Detections with highest confidence are matched first; if a detected BB matches multiple ground truth BBs, the match with highest overlap is used (ties are broken arbitrarily). In rare cases this assignment may be suboptimal, e.g. in crowded scenes [25], but in practice the effect is minor. Unmatched BB_{dt} count as false positives and unmatched BB_{gt} as false negatives.

To compare detectors we plot miss rate against false positives per image (using log-log plots) by varying the threshold on detection confidence (e.g. see Figure 11 and Figure 13). This is preferred to precision recall curves for certain tasks, e.g. automotive applications, as typically there is an upper limit on the acceptable false positives per image (FPPI) rate independent of pedestrian density.

We use the *log-average miss rate* to summarize detector performance, computed by averaging miss rate at nine FPPI rates evenly spaced in log-space in the range 10^{-2} to 10^{0} (for curves that end before reaching a given FPPI rate, the minimum miss rate achieved is used). Conceptually, the log-average miss rate is similar to the *average precision* [26] reported for the PASCAL challenge [14] in that it represents the entire curve by a single reference value. As curves are somewhat linear in this range (e.g., see Figure 13), the log-average miss rate is similar to the performance at 10^{-1} FPPI but in general gives a more stable and informative assessment of performance. A similar performance measure was used in [27].

We conclude by listing additional details. Some detectors output BBs with padding around the pedestrian (e.g. HOG outputs 128×64 BBs around 96 pixel tall people), such padding is cropped (see also §3.4). Methods usually detect pedestrians at some minimum size, to coax smaller detections we upscale the input images. For ground truth, the full BB is always used for matching, not the visible BB, even for partially occluded pedestrians. Finally, all reported results on the Caltech dataset are computed using every 30^{th} frame (starting with the 30^{th} frame) due to the high computational demands of some of the detectors evaluated (see Figure 15).

3.2 Filtering Ground Truth

Often we wish to exclude portions of a dataset during evaluation. This serves two purposes: (1) excluding ambiguous regions, e.g. crowds annotated as 'People' where the locations of individuals is unknown, and (2) evaluating performance on various subsets of a dataset, e.g. on pedestrians in a given scale range. However, we cannot simply discard a subset of ground truth labels as this would cause over-reporting of false positives.

Instead, to exclude portions of a dataset, we introduce the notion of *ignore* regions. Ground truth BBs selected to be ignored, denoted using BB_{ig}, need not be matched, however, matches are not considered mistakes either. E.g., to evaluate performance on unoccluded pedestrians, we set all occluded pedestrian BBs to ignore. Evaluation is purposely lenient: multiple detections can match a single BB_{ig}, moreover, a detection may match any *subregion* of a BB_{ig}. This is useful when the number or location of pedestrians within a single BB_{ig} is unknown as in the case of groups labeled as 'People'.

In the proposed criterion, a BB_{dt} can match any subregion of a BB_{ig}. The subregion that maximizes area of overlap (Eqn. (1)) with BB_{dt} is $BB_{dt} \cap BB_{ig}$, and the resulting maximum area of overlap is:

$$a_o \doteq \frac{area(BB_{dt} \cap BB_{ig})}{area(BB_{dt})} \tag{2}$$

Matching proceeds as before, except BB_{dt} matched to BB_{ig} do *not* count as true positives, and unmatched BB_{ig} do *not* count as false negatives. Matches to BB_{gt} are preferred, meaning a BB_{dt} can only match a BB_{ig} if it does not match any BB_{gt}, and multiple matches to a single BB_{ig} are allowed.

As discussed, setting a BB_{gt} to ignore is not the same as discarding it; in the latter case detections in the ignore regions would count as false positives. Four types of BBs are always set to ignore: any BB under 20 pixels high or truncated by image boundaries, containing a 'Person?' (ambiguous cases), or containing 'People'. Detections within these regions do not affect performance.

3.3 Filtering Detections

In order to evaluate on only a subset of the dataset, we must filter detector responses outside the considered evaluation range (in addition to filtering ground truth labels). For example, when evaluating performance in a fixed scale range, detections far outside the scale range under consideration should not influence the evaluation.

The filtering strategy used in our previous work [3] was too stringent and resulted in under-reporting of detector performance (this was also independently observed by Walk et al. [28]). Here we consider three possible filtering strategies: *strict filtering* (used in our previous work), *post filtering*, and *expanded filtering* that we believe most accurately reflects true performance. In all cases matches to BB_{gt} outside the selected evaluation range neither count as true or false positives.

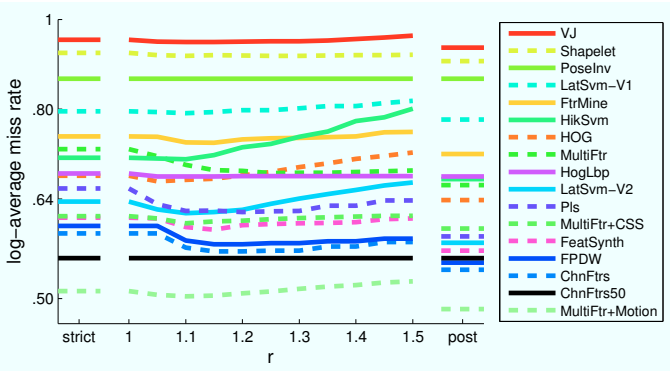


Fig. 8. Comparison of detection filtering strategies used for evaluating performance in a fixed range of scales. **Left**: Strict filtering, used in our previous work [3], undercounts true positives thus under-reporting results. **Right**: Post filtering undercounts false positives thus over-reporting results. **Middle**: Expanded filtering as a function of r. Expanded filtering with $r = 1.25$ offers a good compromise between strict and post filtering for measuring both true and false positives accurately.

Strict filtering: All detections outside the selected range are removed *prior* to matching. If a BB_{gt} inside the range was matched only by a BB_{dt} outside the range, then after strict filtering it would become a false negative. Thus, performance is under-reported.

Post filtering: Detections outside the selected evaluation range are allowed to match BB_{gt} inside the range. After matching, any unmatched BB_{dt} outside the range is removed and does *not* count as a false positive. Thus, performance is over-reported.

Expanded filtering: Similar to strict filtering, except all detections outside an *expanded* evaluation range are removed prior to evaluation. E.g., when evaluating in a scale range from S_0 to S_1 pixels, all detections outside a range S_0/r to $S_1 r$ are removed. This can result in slightly more false positives than post filtering but also fewer missed detections than strict filtering.

Figure 8 shows the log-average miss rate on 50 pixel and taller pedestrians under the three filtering strategies (see §4 for detector details) and for various choices of r (for expanded filtering). Expanded filtering offers a good compromise[1] between strict filtering (which under-reports performance) and post filtering (which over-reports performance). Moreover, detector ranking is robust to the exact value of r. Thus, throughout this work, we use expanded filtering (with $r = 1.25$).

1. Additionally, strict and post filtering are flawed as they can be easily exploited (either purposefully or inadvertently). Under post filtering, generating large numbers of detections just outside the evaluation range can increase detection rate. Under strict filtering, running a detector in the exact evaluation range ensures all detections fall within that range which can also artificially increase detection rate. To demonstrate the latter exploit, in Figure 8 we plot performance of ChnFtrs50, which is ChnFtrs [29] applied to detect pedestrians over 50 pixels. Its performance is identical under each strategy; however, its relative performance is significantly inflated under strict filtering. Expanded filtering cannot be exploited in either manner.

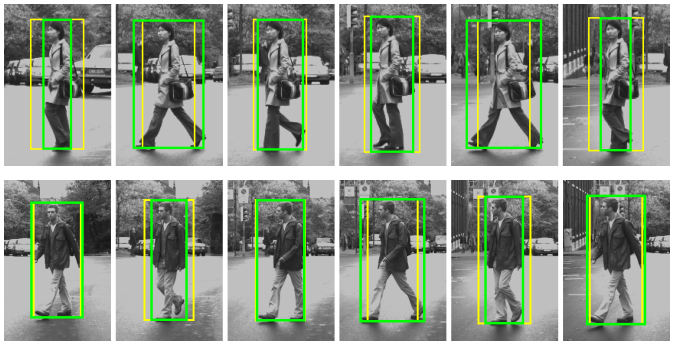


Fig. 9. Standardizing aspect ratios. Shown are profile views of two pedestrians. The original annotations are displayed in green (best viewed in color); these were used to crop fixed size windows centered on each pedestrian. Observe that while BB height changes gradually, BB width oscillates significantly as it depends on the positions of the limbs. To remove any effect *pose* may have on the evaluation of *detection*, during benchmarking width is standardized to be a fixed fraction of the height (see §3.4). The resulting BBs are shown in yellow.

3.4 Standardizing Aspect Ratios

Significant variability in both ground truth and detector BB width can have an undesirable effect on evaluation. We discuss the sources of this variability and propose to standardize aspect ratio of both the ground truth and detected BBs to a fixed value. Doing so removes an extraneous and arbitrary choice from detector design and facilitates performance comparisons.

The height of annotated pedestrians is an accurate reflection of their scale while the width also depends on *pose*. Shown in Figure 9 are consecutive, independently annotated frames from the Daimler detection benchmark [6]. Observe that while BB height changes gradually, the width oscillates substantially. BB height depends on a person's actual height and distance from the camera, but the width additionally depends on the positions of the limbs, especially in profile views. Moreover, the typical width of annotated BBs tends to vary across datasets. For example, although the log-mean aspect ratio (see §2.2.1) in the Caltech and Daimler datasets is .41 and .38, respectively, in the INRIA dataset [7] it is just .33 (possibly due to the predominance of stationary people).

Various detectors likewise return different width BBs. The aspect ratio of detections ranges from a narrow .34 for PLS to a wide .5 for MULTIFTR, while LATSVM attempts to estimate the width (see §4 for detector references). For older detectors that output uncropped BBs, we must choose the target width ourselves. In general, a detector's aspect ratio depends on the dataset used during development and is often chosen after training.

To summarize, the width of both ground truth and detected BBs is more variable and arbitrary than the height. To remove any effects this may have on performance evaluation, we propose to standardize all BBs to an aspect ratio of .41 (the log-mean aspect ratio in the Caltech dataset). We keep BB height and center fixed while adjusting the width (see Figure 9). Note that the ETH [4] and TUD-Brussels [5] evaluation protocols also suggested standardizing the aspect ratio, although to an arbitrarily chosen constant of .5. In general the exact constant has only a minor effect on reported performance; however, it is important that detector and ground truth aspect ratios match. E.g. standardizing the aspect ratios had a large positive effect on detectors that return narrow BBs (including PLS and LATSVM-V2). All results reported in this paper use the standardized aspect ratios.

3.5 Per-Window Versus Full Image Evaluation

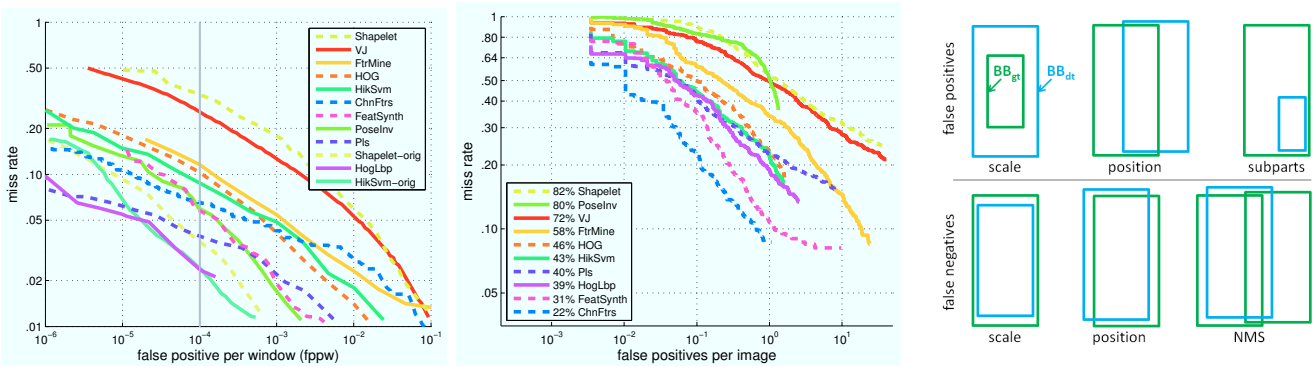
An alternative methodology for evaluating detectors based on *binary classifiers* is to measure their *per-window* (PW) performance on cropped positive and negative image windows, thus isolating classifier performance from the overall detection system. PW evaluation is commonly used to compare classifiers (as opposed to detectors) or to evaluate systems that perform automatic region of interest (ROI) generation [30]. Note that not all detectors are based on classifiers (e.g. [31], [32]), such detectors cannot be evaluated using PW metrics.

A common assumption is that better PW performance leads to better detection performance. In practice we find that PW and full image performance are only weakly correlated, see Figure 10. The PW results are reproduced from their original publications[2] (except the VJ curve, which is reproduced from [7]); the full image results were obtained by evaluating on the same pedestrians but within their original image context. While PW and full image performance are somewhat correlated, the ranking of competing methods is substantially different.

The are a number of reasons for this discrepancy. Choices made in converting a binary classifier to a detector, including choices for spatial and scale stride and non-maximal suppression (NMS), influence full image performance. Moreover, the windows tested during PW evaluation are typically not the same as the windows tested during full image detection, see Figure 10(c).

Full image metrics provide a natural measure of error of an overall detection system, and in this work we use full image metrics throughout. While the PW methodology is useful for isolating evaluation of binary classifiers (the classification task), ultimately the goal of pedestrian detection is to output the location of all pedestrians in an image (the detection task), and for this task full image metrics are appropriate. We thus advocate using full image metrics for evaluation of pedestrian detection as is standard for general object detection [14].

2. PW evaluation must be performed with care: cropped positive and negative windows obtained by different sampling procedures may contain window boundary effects that classifiers can exploit as discriminative features, leading to overfitting. We observed this for the SHAPELET [33] and HIKSVM [34] detectors, see also www.cs.sfu.ca/~mori/research/papers/sabzmeydani_shapelet_cvpr07.html and http://www.cs.berkeley.edu/~smaji/projects/ped-detector/. The original (ORIG) and corrected PW results are shown in Figure 10(a), in both cases the overfitting was discovered only after full image evaluation.



(a) per-window results (b) per-image results (c) untested cases in PW evaluation

Fig. 10. Per-window versus full image evaluation on the INRIA pedestrian dataset (see §4 for detector details and §5.2 for complete results). Legends are ordered by performance. (a) PW results reproduced from their original publications. (b) Full image results obtained by evaluating on the same pedestrians but within their original image context. While PW and full image performance are somewhat correlated, the ranking of competing methods is substantially different. (c) Illustration of six cases not tested during PW evaluation that can give rise to false positives (top) or false negatives (bottom) in full image evaluation. False positives can arise from detections on body parts or at incorrect scales or positions, while false negatives can arise from slight misalignments between the tested windows and true pedestrian locations or from NMS.

4 Detection Algorithms

We focus on computer vision algorithms for detecting pedestrians in individual monocular images, which we refer to simply as 'pedestrian detectors'. We begin with an overview of pedestrian detectors in §4.1, examining the ideas introduced in detection in the last decade. In §4.2 we enumerate and discuss in detail the 16 representative state-of-the-art detectors used in our evaluation.

4.1 Survey of the State of the Art

We review pedestrian detectors with a focus on sliding window approaches. These appear most promising for low to medium resolution settings, under which segmentation [35] or keypoint [36], [37] based methods often fail. We list abbreviations of detectors used in our evaluation in brackets [ALG]. For an overview of how detectors are incorporated into full automotive systems that utilize stereo, scene geometry, tracking, or other imaging modalities (e.g. [30], [38], [39], [40], [41]), we refer readers to [2], [42], [43]. In this work we focus on the detectors themselves.

Papageorgiou et al. [16] proposed one of the first sliding window detectors, applying support vector machines (SVM) to an over-complete dictionary of multi-scale Haar wavelets. Viola and Jones [VJ] [44] built upon these ideas, introducing integral images for fast feature computation and a cascade structure for efficient detection, and utilizing AdaBoost for automatic feature selection. These ideas continue to serve as a foundation for modern detectors.

Large gains came with the adoption of gradient-based features. Inspired by SIFT [45], Dalal and Triggs [HOG] [7] popularized histogram of oriented gradient (HOG) features for detection by showing substantial gains over intensity based features. Zhu et al. [46] sped up HOG features by using integral histograms [47]. In earlier work, Shashua et al. [48] proposed a similar representation for characterizing spatially localized parts for modeling pedestrians. Since their introduction, the number of variants of HOG features has proliferated greatly with nearly all modern detectors utilizing them in some form.

Shape features are also a frequent cue for detection. Gavrila and Philomin [49], [50] employed the Hausdorff distance transform and a template hierarchy to rapidly match image edges to a set of shape templates. Wu and Nevatia [17] utilized a large pool of short line and curve segments, called 'edgelet' features, to represent shape locally. Boosting was used to learn head, torso, leg and full body detectors; this approach was extended in [18] to handle multiple viewpoints. Similarly, 'shapelets' [33] are shape descriptors discriminatively learned from gradients in local patches; boosting was used to combine multiple shapelets into an overall detector [SHAPELET]. Liu et al. [51] proposed 'granularity-tunable' features that allow for representations with levels of uncertainty ranging from edgelet to HOG type features; an extension to the spatio-temporal domain was developed in [52].

Motion is another important cue for human perception; nevertheless, successfully incorporating motion features into detectors has proven challenging given a moving camera. Given a static camera, Viola et al. [53] proposed computing Haar-like features on difference images, resulting in large performance gains. For non-static imaging setups, however, camera motion must be factored out. Dalal et al. [54] modeled motion statistics based on an optical flow field's internal differences, thereby compensating for uniform image motion locally. While the features were successful on a per-window basis [54], for full image detection the benefit appeared minimal [55]. This was resolved by [5], who showed that certain modifications were necessary to make the motion features effective for detection.

While no single feature has been shown to outperform HOG, additional features can provide complementary

information. Wojek and Schiele [MULTIFTR] [56] showed how a combination of Haar-like features, shapelets [33], shape context [57] and HOG features outperforms any individual feature. Walk et al. [28] extended this framework by additionally combining local color self-similarity [MULTIFTR+CSS] and the motion features discussed above [MULTIFTR+MOTION]. Likewise, Wu and Nevatia [58] automatically combined HOG, edgelet and covariance features. Wang et al. [59] combined a texture descriptor based on local binary patterns (LBP) [60] with HOG [HOGLBP], additionally, a linear SVM classifier was modified to perform basic occlusion reasoning. In addition to HOG and LBP, [61] used local ternary patterns (variants of LBP). Color information and implicit segmentation were added in [62], with a performance improvement over pure HOG.

Dollár et al. [29] proposed an extension of [VJ] where Haar-like feature are computed over multiple channels of visual data [CHNFTRS], including LUV color channels, grayscale, gradient magnitude and gradient magnitude quantized by orientation (implicitly computing gradient histograms), providing a simple and uniform framework for integrating multiple feature types. In the 'Fastest Pedestrian Detector in the West' [FPDW] [63], this approach was extended to fast multiscale detection after it was demonstrated how feature computed at a single scale can be used to approximate feature at nearby scales.

Considerable effort has also been devoted to improving the learning framework. Tuzel et al. [64] utilized covariance matrices computed locally over various features as object descriptors. Since covariance matrices do not lie on a vector space, the boosting framework was modified to work on Riemannian manifolds, with improved performance. Maji et al. [34] proposed an approximation to the histogram intersection kernel for use with SVMs [HIKSVM], allowing for substantial speed-ups and thus enabling a non-linear SVM to be used in sliding-window detection. Babenko et al. [65] proposed an approach for simultaneously separating data into coherent groups and training separate classifiers for each; [5] showed that both [34] and [65] gave modest gains over linear SVMs and AdaBoost for pedestrian detection, especially when used in combination [66].

A number of groups have attempted to efficiently utilize very large feature spaces. 'Feature mining' was proposed by [67] to explore vast (possibly infinite) feature spaces using various strategies including steepest descent search prior to training a boosted classifier [FTRMINE]. These ideas were developed further by [68], who introduced a scheme for synthesizing and combining a rich family of part based features in an SVM framework [FEATSYNTH]. Schwartz et al. [69] represented pedestrians by edges, texture and color and applied partial least squares to project the features down to a lower dimensional space prior to SVM training [PLS].

To cope with articulation, the notion of parts and pose have been investigated by several authors. Mohan et al. [73] successfully extended [16] with a two stage approach: first head, arm and leg detectors were trained in a fully supervised manner, next the detectors' output was combined to fit a rough geometric model. Such fully supervised two stage approaches have been revisited over time [17], [74], [75]. Likewise, Bourdev and Malik [76] proposed to learn an exhaustive dictionary of 'poselets': parts clustered jointly in appearance and pose. Supervised pose estimation has been used in a similar manner. Lin and Davis [70] used a part-template tree to model a pedestrian's shape locally for the head, upper body and legs, and extracted HOG appearance descriptors along the shape's outline [POSEINV]. Enzweiler and Gavrila [77] labeled pedestrians as belonging to one of four canonical orientations and jointly perform classification and orientation estimation. Joint body pose estimation and person classification can also be formulated as a structured learning problem [78].

Notable early approaches for unsupervised part learning, including the constellation model [79], [80] and the sparse representation approach of [81], relied on keypoints. Leibe et al. [36] adapted the implicit shape model, also based on keypoints, for detecting pedestrians. However, as few interest points are detected at lower resolutions, unsupervised part based approaches that do not rely on keypoints have been proposed. Multiple instance learning (MIL) has been employed in order to automatically determine the position of parts without part-level supervision [82], [83]. And, in one of the most successful approaches for general object detection to date, Felzenszwalb et al. [71], [72] proposed a discriminative part based approach that models unknown part positions as latent variables in an SVM framework [LATSVM]. As part models seem to be most successful at higher resolutions, Park et al. [84] extended this to a multi-resolution model that automatically switches to parts only at sufficiently high resolutions.

4.2 Evaluated Detectors

We chose 16 representative state-of-the-art pedestrian detectors for this evaluation (see §4.1 and Table 2). Our goal was to choose a diverse set of detectors that were both representative of various lines of research and most promising in terms of originally reported performance. While we could not be exhaustive due to unavailability of many detectors and practical time and space constrains, we do believe that the selected detectors give an accurate portrait of the state of the art.

In nearly all cases we obtained *pre-trained* detectors directly from the authors as our goal was to have an unbiased evaluation of existing approaches. Any major differences from the original publications are discussed below. We thank the authors for either publishing their code online or making it available upon request.

While research in pedestrian detection is quite diverse, the approaches with the highest reported performance share many elements. These detectors typically follow a sliding window paradigm which entails feature extraction, binary classification, and dense multiscale scanning

		Features							Learning			Detection Details						Implementation			
		gradient hist.	gradients	grayscale	color	texture	self-similarity	motion	classifier	feature learn.	part based	non-maximum suppression	model height (in pixels)	scales per octave	frames per second (fps)	log-average miss rate	training data	original code	full image evaluation	publication
VJ	[44]			✓					AdaBoost			MS	96	~14	.447	95%	INRIA			'04
SHAPELET	[33]		✓	✓					AdaBoost	✓		MS	96	~14	.051	91%	INRIA			'07
POSEINV	[70]	✓							AdaBoost			MS	96	~18	.474	86%	INRIA	✓		'08
LATSVM-V1	[71]	✓							latent SVM		✓	PM	80	10	.392	80%	PASCAL	✓	✓	'08
FTRMINE	[67]	✓	✓	✓	✓				AdaBoost	✓		PM	100	4	.080	74%	INRIA	✓		'07
HIKSVM	[34]	✓							HIK SVM			MS	96	8	.185	73%	INRIA	✓		'08
HOG	[7]	✓							linear SVM			MS	96	~14	.239	68%	INRIA	✓		'05
MULTIFTR	[56]	✓		✓					AdaBoost			MS	96	~14	.072	68%	INRIA	✓	✓	'08
HOGLBP	[59]	✓				✓			linear SVM			MS	96	14	.062	68%	INRIA	✓	✓	'09
LATSVM-V2	[72]	✓							latent SVM		✓	PM	96	10	.629	63%	INRIA	✓	✓	'09
PLS	[69]	✓			✓	✓			PLS+QDA	✓		PM*	96	~10	.018	62%	INRIA	✓	✓	'09
MULTIFTR+CSS	[28]	✓					✓		linear SVM			MS	96	~14	.027	61%	TUD-MP	✓	✓	'10
FEATSYNTH	[68]	✓				✓			linear SVM	✓	✓	–	96	–	–	60%	INRIA	✓	✓	'10
FPDW	[63]	✓	✓	✓	✓				AdaBoost			PM*	100	10	6.492	57%	INRIA	✓	✓	'10
CHNFTRS	[29]	✓	✓	✓	✓				AdaBoost			PM*	100	10	1.183	56%	INRIA	✓	✓	'09
MULTIFTR+MOTION	[28]	✓					✓	✓	linear SVM			MS	96	~14	.020	51%	TUD-MP	✓	✓	'10

TABLE 2
Comparison of Evaluated Pedestrian Detectors (see §4.2 for details)

of detection windows followed by non-maximum suppression (NMS). Below we discuss each component of the evaluated detectors, including the features, learning framework, and detection details, and conclude with implementation notes; for additional details we refer readers to the original publications. Table 2, ordered by descending log-average miss rate on clearly visible pedestrians in the Caltech dataset (see §5 for details), gives an overview of each detector.

Features: The first columns in Table 2 indicate the feature types used by each detector (specified by the general category of image content extracted and not the particular instantiation). Nearly all modern detectors employ some form of gradient histograms [7]. In addition, detectors can utilize gradients directly, as well as grayscale (e.g. Haar wavelets [44]), color, texture (including LBP [60] and co-occurrence [85]), self-similarity [86] and motion [54] features. The best performing detectors tend to use a combination of cues.

Learning: The second set of columns provides details about the learning paradigm used by each detector. Support vector machines (SVMs) [16] and boosting [44] are the most popular choices due to their theoretical guarantees, extensibility, and good performance. Boosted classifiers and linear SVMs are particularly well suited due to their speed; non-linear kernels are less common, the exception being the fast histogram intersection kernel [34]. Boosting automatically performs feature selection, alternatively some detectors (indicated with a mark in the 'feature learning' column) learn a smaller or intermediate set of features prior to or jointly with classifier training. Finally a few detectors including LATSVM and FEATSYNTH are part based.

Detection Details: The next columns describe the detection scheme. Two dominant non-maximum suppression (NMS) approaches have emerged: mean shift (MS) mode estimation [55] and pairwise max (PM) suppression [71] which discards the less confident of every pair of detections that overlap sufficiently according to Eqn. (1). PM requires only a single parameter; in addition, a variant has been proposed (PM*) that allows a detection to match any subregion of another detection, resulting in improved performance (see Eqn. (2) and addendum to [29]). FEATSYNTH only tests windows returned by FTRMINE and does not require NMS. Pedestrian model height is typically around 96-100 pixels (the size of pre-cropped pedestrians in the INRIA dataset), with an additional 28-32 pixels of padding. For multiscale detection, usually around 10-14 scales per octave are scanned (with corresponding scale strides of 1.07-1.05); a fast multiscale scheme is proposed in [63]. Runtimes (for detecting over 100 pixel pedestrians in 640x480 images) and log-average miss rates (on clearly visible pedestrians) are discussed in §5.

Implementation Notes: The final columns of Table 2 list additional details. Most of the evaluated detectors were trained on the INRIA dataset [7]; two were trained on TUD motion pairs (TUD-MP) (the training set for TUD-Brussels [5]). LATSVM-V1 was trained on Pascal [14]; LATSVM-V2 used INRIA and a later version of the latent SVM framework [72]. In nearly all cases we used code obtained directly from the authors, the only exceptions being VJ and SHAPELET which were reimplemented in [56]. In a few cases the evaluated code differed from the published version: SHAPELET and HIKSVM have been corrected so they no longer overfit to boundary effects; we evaluate a variant of POSEINV based on boosting (which in our tests outperformed the much slower kernel SVM version); PLS switched to PM* NMS; and finally, the posted code for HOGLBP does not include occlusion reasoning (the improvement from occlusion reasoning was slight [59]).

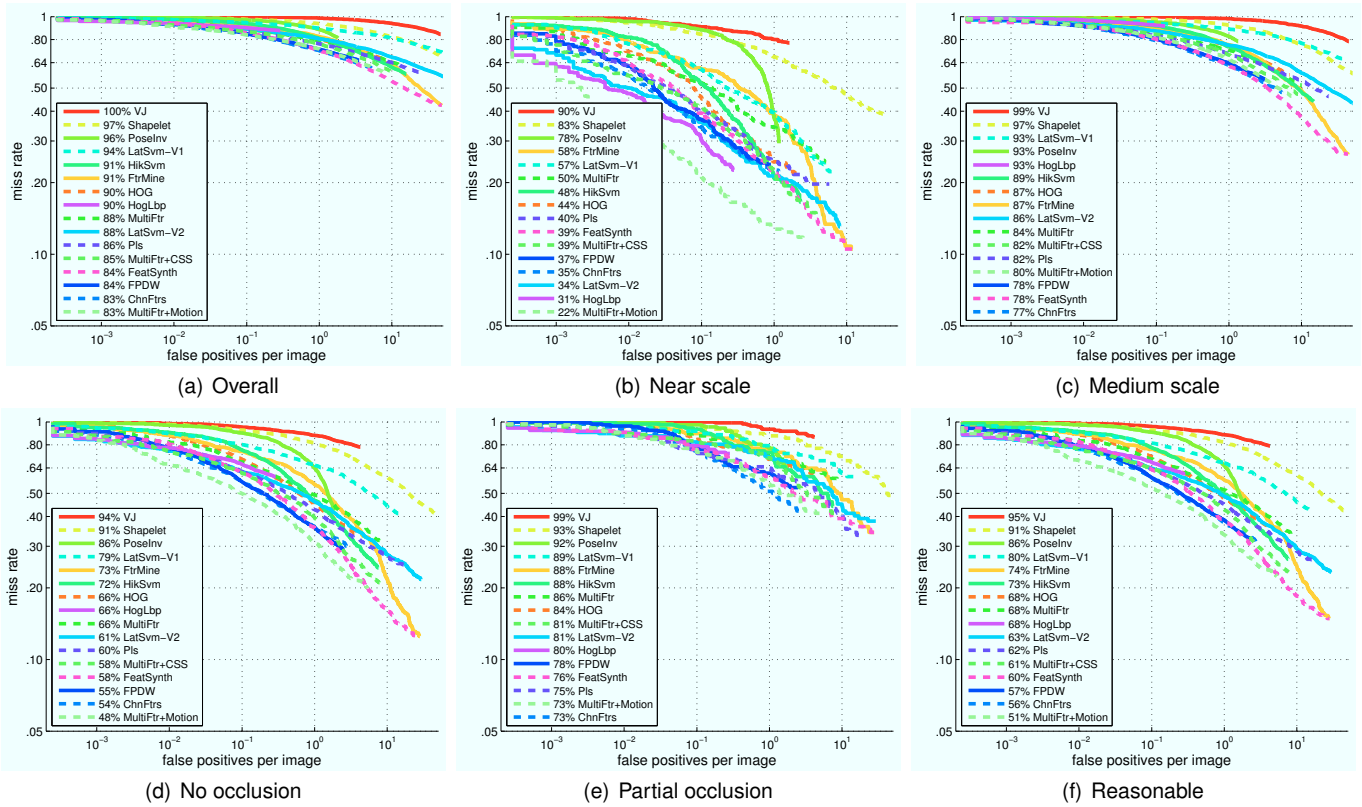


Fig. 11. Evaluation results under six different conditions on the Caltech Pedestrian Dataset. (a) Overall performance on all annotated pedestrians is unsatisfactory. (b) Performance on unoccluded pedestrians over 80 pixels tall is substantially better, (c) but degrades for 30-80 pixel pedestrians. (d) Likewise performance on unoccluded pedestrians over 50 pixels tall is better than overall performance, (e) but degrades in the presence of partial occlusion. (f) This motivates us to evaluate performance on pedestrians at least 50 pixels tall under no or partial occlusion; we refer to this as the *reasonable* evaluation setting and use it throughout.

5 Performance Evaluation

We performed an extensive evaluation of the sixteen pedestrian detectors enumerated in Table 2 under various scenarios and for multiple datasets. First, in §5.1 we evaluate performance under different conditions using the Caltech Dataset. Next we report performance on six additional datasets in §5.2 and analyze statistical significance in §5.3. Finally in §5.4 we report runtimes.

We chose to evaluate *pre-trained* detectors, obtained directly from their authors. This is an important methodological point: we assume that authors know best how to tune their algorithms, attempting to train the detectors ourselves would have opened the difficult subject of parameter tuning, making our study unwieldy. Moreover, few authors share training code, thus, insisting on re-training would have severely limited our ability to conduct a broad evaluation. Fortunately, most of the detectors were trained on the same dataset (see Table 2), making them directly comparable. Additionally, testing these pre-trained detectors on multiple other datasets allows us to study cross-dataset generalization, a topic of crucial real-world importance.

5.1 Performance on the Caltech Dataset

We first analyze performance under six conditions on the testing data in the Caltech Pedestrian Dataset. Figure 11 shows performance for the overall dataset, on near and medium scales, under no and partial occlusion, and on clearly visible pedestrians. We plot miss rate versus false positives per image (lower curves indicate better performance) and use log-average miss rate as a common reference value for summarizing performance. Legend entries display and are ordered by log-average miss rate from worst to best. We discuss the plots in detail below.

Overall: Figure 11(a) plots performance on the entire test set. MULTIFTR+MOTION slightly outperforms the other detectors, with CHNFTRS a close second. However, absolute performance is poor, with a log-average miss rate of over 80%. To understand where the detectors fail, we examine performance under various conditions.

Scale: Results for near and medium scale unoccluded pedestrians, corresponding to heights of at least 80 pixels and 30-80 pixels, respectively (see §2.2.1), are shown in Figures 11(b) and 11(c). For the near scale, MULTIFTR+MOTION performs best with a log-average miss rate of only 22%; numerous other detectors achieve still reasonable log-average miss rates around 30-40%. On the medium scale, which contains over 68% of ground truth pedestrians (see Figure 4(a)), performance degrades dramatically. CHNFTRS, FPDW and FEATSYNTH achieve the best relative performance but absolute performance is quite poor with 77-78% log-average miss rate. Moreover, the top three performing detectors on near scale

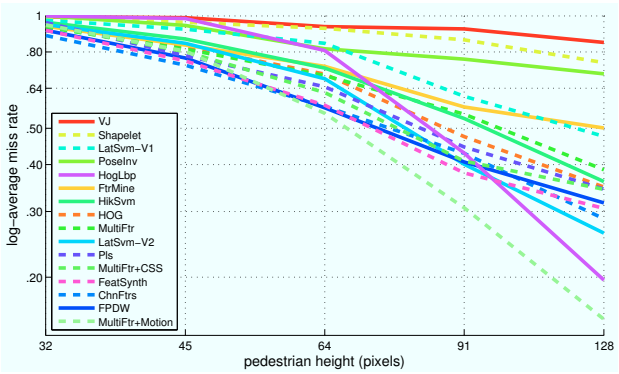


Fig. 12. Performance as a function of scale. All detectors improve rapidly with increasing scale, especially MULTIFTR+MOTION, HOGLBP and LATSVM-V2 which utilize motion, texture and parts, respectively. At small scales state-of-the-art performance has considerable room for improvement.

pedestrians degrade most. We can see this trend clearly by plotting log-average miss rate as a function of scale. Figure 12 shows performance at five scales between 32 and 128 pixels (see also §3.2 and §3.3). Performance improves for all methods with increasing scale, but most for MULTIFTR+MOTION, HOGLBP and LATSVM-V2. These utilize motion, texture and parts, respectively, for which high resolutions appear to be particularly important.

Occlusion: The impact of occlusion on detecting 50 pixel or taller pedestrians is shown in Figures 11(d) and 11(e). As discussed in §2.2.2, we classify pedestrians as unoccluded, *partially* occluded (1-35% occluded) and *heavily* occluded (35-80% occluded). Performance drops significantly even under partial occlusion, leading to a log-average miss rate of 73% for CHNFTRS and MULTIFTR+MOTION. Surprisingly, performance of part based detectors degrades as severely as for holistic detectors.

Reasonable: Performance for medium scale or partially occluded pedestrians is poor while for far scales or under heavy occlusion it is abysmal (see Figure 16). This motivates us to evaluate performance on pedestrians over 50 pixels tall under no or partial occlusion (these are clearly visible without much context). We refer to this as the *reasonable* evaluation setting. Results are shown in Figure 11(f), MULTIFTR+MOTION, CHNFTRS and FPDW perform best with log-average miss rates of 51-57%. We believe this evaluation is more representative than overall performance on all pedestrians and we use it for reporting results on all additional datasets in §5.2 and for the statistical significance analysis in §5.3.

Localization: Recall that the evaluation is insensitive to the exact overlap threshold used for matching so long as it is below ∼.6 (see §3.1 and Figure 7). This implies that nearly all detections that overlap the ground truth overlap it by at least half. However, as the threshold is increased further and higher localization accuracy is required, performance of all detectors degrades rapidly. Detector ranking is mostly maintained except MULTIFTR and PLS degrade more; this implies that all but these two detectors have roughly the same localization accuracy.

5.2 Evaluation on Multiple Datasets

To increase the scope of our analysis, we benchmarked the detectors on six additional pedestrian detection datasets including INRIA [7], TUD-Brussels [5], ETH [4] Daimler-DB [6], Caltech-Training, and Caltech-Japan. These datasets are discussed in §2.4 and Table 1; we also review their most salient aspects below. Evaluating across multiple datasets allows us to draw conclusion both about the detectors and the datasets. Here we focus on the datasets, we return to assessing detector performance using multiple datasets in §5.3. Performance results for every dataset are shown in Figure 13.

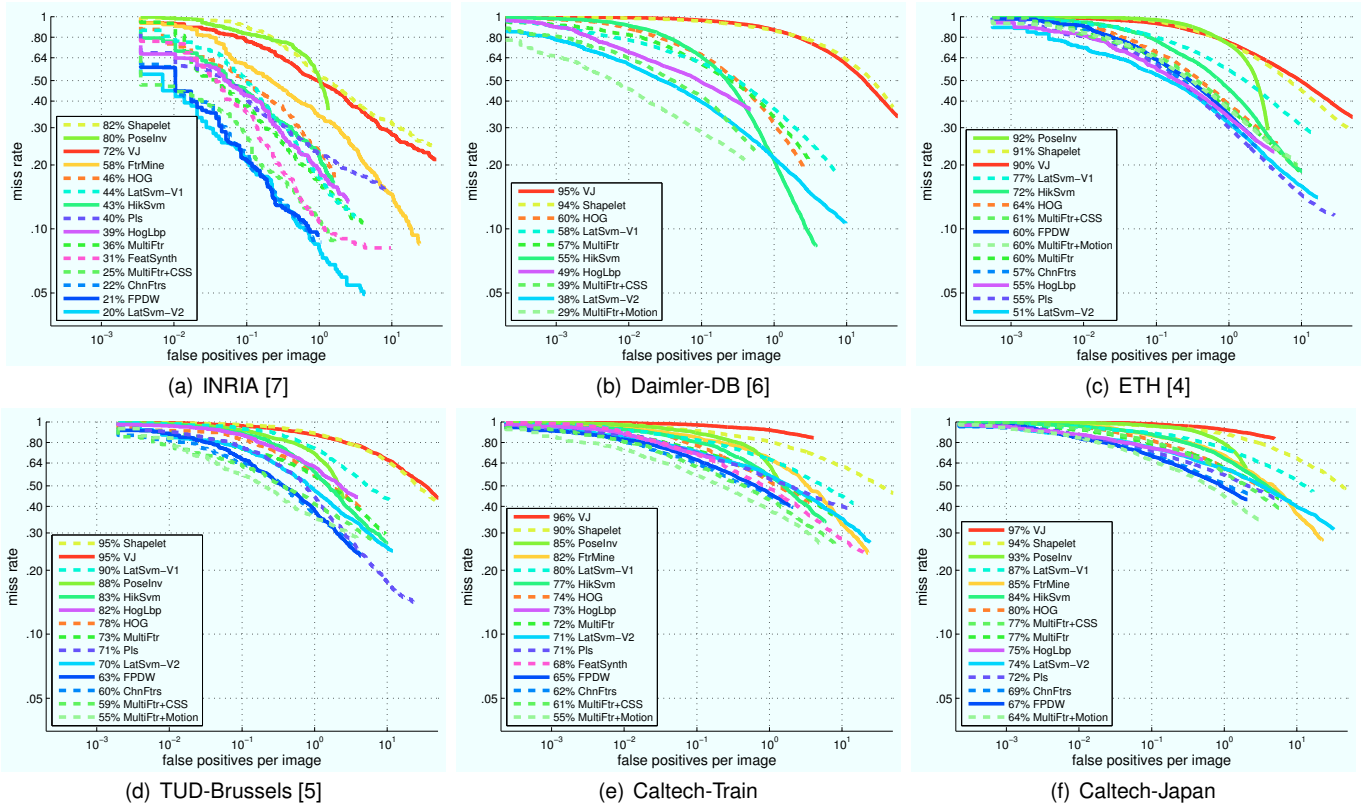
We begin with a brief review of the six datasets. INRIA contains images of high resolution pedestrians collected mostly from holiday photos (we use only the 288 test images that contain pedestrians, note that a few have incomplete labels). The remaining datasets were recorded with a moving camera in urban environments and all contain color except Daimler-DB. ETH has higher density and larger scale pedestrians than the remaining datasets (we use the refined annotations published in [5]). Caltech-Training refers to the training portion of the Caltech Pedestrian Dataset. Caltech-Japan refers to a dataset we gathered in Japan that is essentially identical in size and scope to the Caltech dataset (unfortunately it cannot be released publicly for legal reasons). Table 1 provides an overview and further details on each dataset's properties and statistics (see also Figure 1).

We benchmark performance using the reasonable evaluation setting (50 pixel or taller under partial or no occlusion), standardizing aspect ratios as described in §3.4. For Daimler-DB and INRIA, which contain only grayscale and static images, respectively, we run only detectors that do not require color and motion information. Also, FTRMINE and FTRSYNTH results are not always available, otherwise we evaluated every detector on every dataset. We make all datasets, along with annotations and detector outputs for each, available in a single standardized format on our project webpage.

Of all datasets, performance is best on INRIA, which contains high resolution pedestrians, with LATSVM-V2, CHNFTRS and FPDW achieving log-average miss rates of 20-22% (see Figure 13(a)). Performance is also fairly high on Daimler-DB (13(b)) with 29% log-average miss rate attained by MULTIFTR+MOTION, possibly due to the good image quality resulting from use of a monochrome camera. ETH (13(c)), TUD-Brussels (13(d)), Caltech-Training (13(e)), and Caltech-Testing (11(f)) are more challenging, with log-average miss rates between 51-55%, and Caltech-Japan (13(f)) is even more difficult due to lower image quality. Overall, detector ranking is reasonably consistent across datasets, suggesting that evaluation is not overly dependent on the dataset used.

5.3 Statistical Significance

We aim to rank detector performance utilizing multiple datasets and assess whether the differences between



(a) INRIA [7] (b) Daimler-DB [6] (c) ETH [4]

(d) TUD-Brussels [5] (e) Caltech-Train (f) Caltech-Japan

Fig. 13. Results on six datasets under the reasonable evaluation setting. In general, detector ranking is fairly stable across datasets. Best results are achieved on INRIA (a), which contains high-resolution pedestrians, followed by Daimler-DB (b), likely due to the good image quality. Overall performance for ETH (c), TUD-Brussels (d), Caltech-Train (e) and Caltech-Test (Figure 11(f)) are similar. Performance on Caltech-Japan (f) is slightly worse; likely due to more challenging imaging conditions. The fairly high consistency of detector ranking across multiple datasets implies that evaluation is not overly dependent on the dataset used.

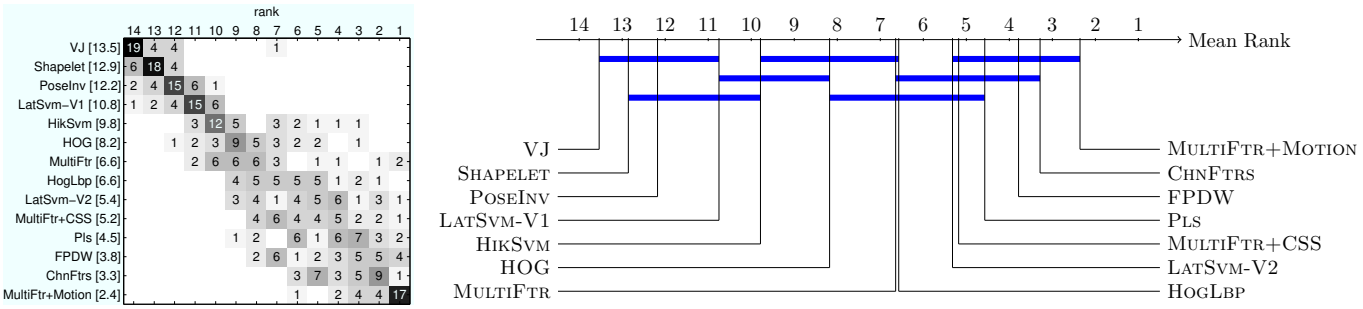
detectors are statistically significant. Two issues make such an analysis challenging: (1) dataset difficulty varies and (2) relative detector performance may change across datasets. The plots in Figure 13 clearly demonstrate both challenges. To address this, Demšar et al. [87] introduced a series of powerful statistical tests that operate on an m dataset by n algorithm performance matrix (e.g., a matrix of log-average miss rates). The key insight is to convert absolute performance on each dataset into algorithm *rank*, thus removing the effects of varying dataset difficulty. We first describe the analysis and then present results for pedestrian detectors.

We analyze statistical significance using the non-parametric Friedman test with a post-hoc analysis, this approach was also used by Everingham et al. [14] for the PASCAL VOC challenge. Contrary to ANOVA, the Friedman test does not assume a distribution on performance, but rather uses algorithm ranking. Demšar et al. [87] found this non-parametric approach to be more robust. A further in-depth study by García and Herrera [88] concludes that the Nemenyi post-hoc test which was used by [87] (and also in the PASCAL challenge [14]) is too conservative for $n \times n$ comparisons such as in a benchmark. They recommend use of more powerful post-hoc tests such as the Shaffer test that include more sophisticated logic reasoning. For our analysis we use the non-parametric Friedman test along with the Shaffer post-hoc test (code is available from [88]).

To obtain a sufficient number of performance samples we evaluate the pre-trained detectors separately on each of the 11 sets in the Caltech dataset (see §2.3), 13 sets of Caltech-Japan, 3 sequences in ETH, and 1 sequence in TUD-Brussels. We omit Daimler-DB and INRIA on which not all detectors can be tested and any detector not tested on every dataset (see §5.2). We rank detectors on each data fold based on their log-average miss rate (tested under the reasonable evaluation setting). This procedure yields a total of 28 rankings for 14 detectors.

Results are shown in Figure 14. First, we plot the number of times each detector achieved each rank in Figure 14(a). Detectors are ordered by improving mean rank (displayed in brackets). The best overall performing detector is MULTIFTR+MOTION, which ranked first on 17 of the 28 data folds and had a mean rank of 2.4. CHNFTRS and FPDW came in second and third with a mean rank of 3.3 and 3.8, respectively. VJ has the worst overall performance with a mean rank of 13.5, while HOG remains somewhat competitive with a mean rank of 8.2. Among the top performing detectors, however, variance is fairly high, with nearly every detector from MULTIFTR onward ranking first on at least one data fold.

Figure 14(b) shows the results of the significance test for a confidence level of $\alpha = 0.05$. The x-axis shows mean rank for each detector, blue bars link detectors



(a) histogram of detector ranks (b) critical difference diagram

Fig. 14. Summary of detector performance across multiple datasets and the statistical significance of the results. (a) Visualization of the number of times each detector achieved each rank. Detectors are ordered by improving mean rank (displayed in brackets); observe the high variance among the top performing detectors. (b) Critical difference diagram [87]: the x-axis shows mean rank, blue bars link detectors for which there is insufficient evidence to declare them statistically significantly different (due to the relatively low number of performance samples and fairly high variance).

for which there is insufficient evidence to declare them statistically significantly different. For example, MULTIFTR+MOTION, CHNFTRS and FPDW are all significantly better than HOG (since they are not linked). Observe, however, that the differences between the top six detector are not statistically significant, indeed, each detector tends to be linked to numerous others. This result does not change much if we relax the confidence to $\alpha = 0.1$. A similar trend was observed in Everingham et al.'s [14] analysis on the PASCAL challenge. Unfortunately, the statistical analysis requires a large number of samples, and while the 28 data folds provide a considerably more thorough analysis of pedestrian detectors than previously attempted, given their inherent variability, even more data would be necessary. We emphasize, however, that simply because we have insufficient evidence to declare the detectors statistically significantly different does *not* imply that their performance is equal.

5.4 Runtime Analysis

In many applications of pedestrian detection, including automotive safety, surveillance, robotics, and human machine interfaces, fast detection rates are of the essence. Although throughout we have focused on accuracy; we conclude by jointly considering both accuracy and speed.

We measure runtime of each detector using images from the Caltech dataset (averaging runtime over multiple frames). To compensate for detectors running on different hardware, all runtimes are normalized to the rate of a single modern machine. We emphasize that we measure the speed of binaries provided by the authors and that faster implementations are likely possible.

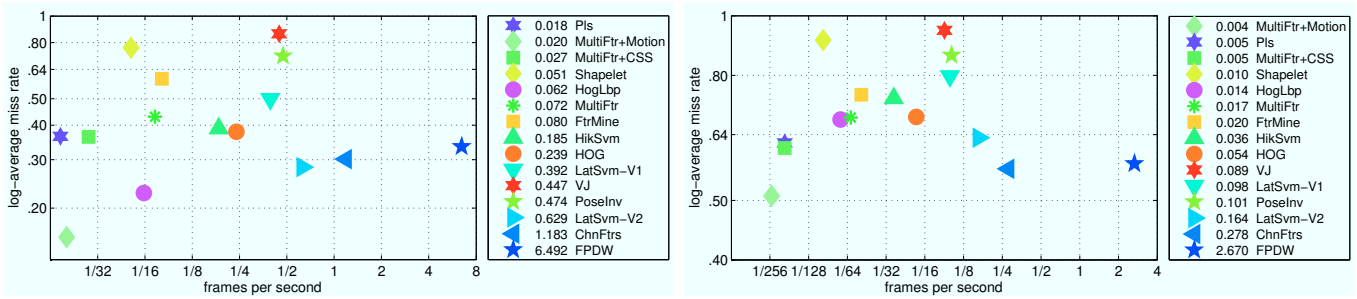
In Figure 15 we plot log-average miss rate versus runtime for each detector on 640×480 images. Legends are ordered by detection speed measured in frames per second (fps). Detection speed for pedestrians over 100 pixels ranges from $\sim$.02 fps to $\sim$6.5 fps achieved by FPDW, a sped up version of CHNFTRS. Detecting 50 pixel pedestrians typically requires image upsampling; the slowest detectors require around five *minutes* per frame. FPDW remains the fastest detector operating at $\sim$2.7 fps. Overall, there does not seem to be a strong correlation between runtime and accuracy. While the slowest detector happens to also be the most accurate (MUTLIFTR+MOTION), on pedestrians over 50 pixels the two fastest detectors, CHNFTRS and FPDW, are also the second and third most accurate, respectively.

While the frame rates may seem low, it is important to mention that all tested detectors can be employed as part of a full system (cf. [2]). Such systems may employ ground plane constraints and perform region-of-interest selection (e.g. from stereo disparity or motion), reducing runtime drastically. Moreover, numerous approaches have been proposed for speeding up detection, including speeding up the detector itself [29], [44], [46], through use of approximations [63], [89] or by using special purpose hardware such as GPUs [90] (for a review of fast detection see [63]). Nevertheless, the above runtime analysis gives a sense of the speed of current detectors.

6 DISCUSSION

This study was carried out to assess the state of the art in pedestrian detection. Automatically detecting pedestrians from moving vehicles could have considerable economic impact and the potential to substantially reduce pedestrian injuries and fatalities. We make three main contributions: a new dataset, an improved evaluation methodology and an analysis of the state of the art.

First, we put together an unprecedented object detection dataset. The dataset is large, representative and relevant. It was collected with an imaging geometry and in multiple neighborhoods that match likely conditions for urban vehicle navigation. Second, we propose an evaluation methodology that allows us to carry out probing and informative comparisons between competing approaches to pedestrian detection in a realistic and unbiased manner. Third, we compare performance of sixteen pre-trained state-of-the-art detectors across six datasets. Performance is assessed as a function of scale, degree of occlusion, localization accuracy and computational cost; moreover we gauge the statistical significance of the ranking of detectors across multiple datasets.

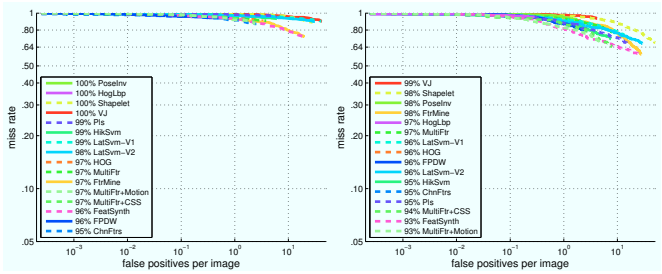


(a) accuracy versus runtime for pedestrians over 100 pixels (b) accuracy versus runtime for pedestrians over 50 pixels

Fig. 15. Log-average miss rate versus the runtime of each detector on 640×480 images from the Caltech Pedestrian Dataset. Run times of all detectors are normalized to the rate of a single modern machine, hence all times are directly comparable. (Note that the VJ implementation used did not utilize scale invariance and hence its slow speed). While the slowest detector happens to also be the most accurate (MUTLIFTR+MOTION), for pedestrians over 50 pixels the two fastest detectors, CHNFTRS and FPDW, are also the second and third most accurate, respectively.

All results of this study, and the tools for reproducing them, are posted on the project website[3]. This includes the Caltech Pedestrian Dataset, the video annotation tool (see Figure 3), and all evaluation code. We have also posted all additional datasets used in this study (INRIA, TUD-Brussels, ETH, and Daimler-DB), along with their annotations and detector outputs on each, in a standardized format. The goal is to allow all researchers to easily and accurately assess state-of-the-art performance.

Our experiments allow us to make a number of observations, and point to important directions for further research. We discuss them in the following sections.



(a) Far scale (b) Heavy occlusion

Fig. 16. Evaluation on far scales and for heavily occluded pedestrians. In both cases detection performance is abysmal and currently outside the capabilities of the state of the art.

6.1 Statistics of The Caltech Pedestrian Dataset

The Caltech Pedestrian Dataset is thoroughly annotated (including occlusion labels, temporal correspondences, and 'ignore' regions) and contains pedestrians at a wide range of scales. It thus allows us to analyze the statistics of a number of important phenomena:

Scale: Pedestrian pixel size is highly variable. Most pedestrians, however, are observed at heights of 30 to 80 pixels. This scale range also happens to be the most important for automotive settings given current sensor technology and typical imaging geometry (see Figure 4).

Occlusion: Occlusion is very frequent (see Figures 5(b) and 5(c)). Nevertheless, out of many possible occlusion patterns, few are commonly observed (see Figure 5(f)). The head is typically visible while the lower portions of a pedestrian are increasingly likely to be occluded.

Location: The distribution of pedestrian centers in images is highly concentrated along the middle band (see Figure 6(a)). However, while incorporating this constraint would speed detection it would only moderately reduce false alarms under these settings (see Figure 6(b)).

Experiments across multiple datasets show reasonably consistent detector rankings (see Figures 13 and 14), suggesting that evaluation is not overly dependent on the dataset used. Overall, however, the Caltech Pedestrians Dataset proves to be the most challenging. We thus expect that it will remain useful for a fairly long time before showing signs of saturation. Furthermore, its large size and thorough annotation enables researchers to measure performance under very low false alarm rates, small scales, and varying levels of occlusion.

6.2 Overall Performance

There is considerable room for improvement in pedestrian detection. The plots in Figure 11 show that:

(1) Performance is far from perfect even under the most favorable conditions. At the near scale, i.e. with pedestrians at least 80 pixels tall, 20-30% of all pedestrians are missed under the fairly mild goal of at most one false alarm every ten images (see Figure 11(b)).

(2) Performance degrades catastrophically for smaller pedestrians (see Figures 11(c) and 12). While pedestrians 30-80 pixels tall are most numerous and most relevant in automotive settings, around 80% are missed by the best detectors (at 1 false alarm per 10 images).

(3) Performance degrades similarly under partial occlusion (under 35% occluded), see Figure 11(e).

(4) Performance is abysmal at far scales (under 30 pixels) and under heavy occlusion (over 35% occluded), see Figure 16. Under these conditions nearly all pedestrians are missed even at high false positive rates.

The gap between current and desired performance is large and unlikely to be reached without major leaps in our understanding. One should note that single frame performance is a lower bound for the performance of

3. www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

a full system and that tracking, contextual information and the use of additional sensors can help reduce false alarms and improve detection rates (see [2]).

6.3 State of the Art Detectors

While research in pedestrian detection is quite diverse, the approaches with the best performance have many elements in common (see Table 2). These detectors typically follow a sliding window paradigm which entails feature extraction, binary classification, and dense multiscale scanning of detection windows followed by non-maximum suppression (NMS). Nearly all modern detectors employ some form of gradient histograms; in addition, the best performing detectors tend to use a combination of cues. Support vector machines and boosting are used almost exclusively although variants show promise. For multiscale detection 10-14 scales per octave are commonly tested, and for NMS, mean shift (MS) mode estimation and pairwise max (PM) suppression appear most successful. Overall, there is a high degree of convergence for the various stages of detection.

Which is the best overall detector? Figure 15 summarizes both detection accuracy and computational cost; surprisingly, there does not seem to be a hard trade-off between these two quantities. Overall, FPDW has the most appealing characteristics: it is at least one order of magnitude faster than its competitors and has amongst the best detection rates, particularly on medium scale pedestrians. If computational cost is not a consideration, then MULTIFTR+MOTION is the best choice. Note, however, that re-training on the larger Caltech Pedestrian Dataset may change the relative ranking of the detectors.

The state of the art in pedestrian detection is clearly advancing. Considerable progress has been made from earlier approaches (e.g. VJ) to the most recent ones (see Table 2). Thus, given the fast pace of technical progress in the field and the considerable room for improvement, we expect to see new detectors top the charts every year.

6.4 Research directions

Our benchmark indicates need for research in 7 areas:

(1) Small scales. Better performance is needed in the 30-80 pixel range, while most research has been focused on pedestrians over 100 pixels. Reasonable human performance at medium scales indicates that detection in this range is achievable without resorting to expensive high resolution cameras that would delay the introduction of machine vision systems to automotive settings.

(2) Occlusion. Performance degrades rapidly under even mild occlusion, including for part based detectors. The Caltech Pedestrian Dataset is the first to include occlusion labels (and a study of occlusion statistics), we hope this motivates researchers to improve this.

(3) Motion features. The detector with highest accuracy (MULTIFTR+MOTION) is the only one to utilize motion features, but the optical flow based features appear to help primarily at large scales. At low resolutions motion is very informative for human perception, thus effective motion features for this setting are needed.

(4) Temporal integration. Although full systems often utilize tracking (e.g. see [2], [6], [42]), a comparative study of approaches for integrating detector outputs over time has not been carried out. Note that full tracking may be unnecessary and methods that integrate detector outputs over a few frames may suffice [41].

(5) Context. The ground plane assumption can reduce errors somewhat, however, at low resolutions more sophisticated approaches for utilizing context are needed.

(6) Novel features. The best detectors use multiple feature types in combination with gradient histograms (see Table 2). We expect additional gains from continued research on improving feature extraction.

(7) Data. Most detectors were trained on INRIA [7]. Training using the much larger Caltech dataset should boost performance, although learning becomes more challenging due to the broad range of scale and occlusion levels. Studies are needed to see the effect of quantity and type of training data versus performance.

Acknowledgments

P. Perona and P. Dollár acknowledge the support of ONR MURI Grants #N00014-06-1-0734 and #1015-G-NA127. This project was also partially supported by the NISSAN MOTOR CO., LTD. We thank Seigo Watanabe of Nissan for collecting the video and for many useful discussions. We would also like to thank all authors that made their detection code available.

REFERENCES

[1] U. Shankar, "Pedestrian roadway fatalities," Department of Transportation, Tech. Rep., 2003. 1

[2] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey on pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, 2010. 1, 2, 10, 16, 18

[3] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: A Benchmark," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009. 1, 2, 6, 7, 8

[4] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *IEEE Intl. Conf. Computer Vision*, 2007. 1, 2, 6, 9, 14, 15

[5] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009. 1, 2, 6, 9, 10, 11, 12, 14, 15

[6] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2009. 1, 2, 6, 9, 14, 15, 18

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2005. 1, 2, 5, 6, 9, 10, 12, 14, 15, 18

[8] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. A. Burkitt, "Performance of optical flow techniques," *Intl. Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994. 2

[9] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski, "A database and eval. methodology for optical flow," in *IEEE Intl. Conf. Computer Vision*, 2007. 2

[10] D. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 530–549, 2004. 2

[11] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Intl. Journal of Computer Vision*, vol. 47, pp. 7–42, 2002. 2

[12] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006. 2

[13] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. of Technology, Tech. Rep. 7694, 2007. 2

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Intl. Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010. 2, 3, 6, 7, 9, 12, 15, 16

[15] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *Intl. Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004. 2

[16] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Intl. Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000. 6, 10, 11, 12

[17] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part det." in *IEEE Intl. Conf. Computer Vision*, 2005. 6, 10, 11

[18] ——, "Cluster boosted tree classifier for multi-view, multi-pose object det." in *IEEE Intl. Conf. Computer Vision*, 2007. 6, 10

[19] D. Gerónimo, A. Sappa, A. López, and D. Ponsa, "Adaptive image sampling and windows classification for on-board ped. det." in *Intl. Conf. Computer Vision Systems*, 2005. 6

[20] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2008. 6

[21] S. Munder and D. M. Gavrila, "An experimental study on pedestrian classification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1863–1868, 2006. 6

[22] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson, "A new pedestrian dataset for supervised learning," in *IEEE Intelligent Vehicles Symposium*, 2008. 6

[23] B. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image ann." *Intl. Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008. 6

[24] A. T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin, "ETISEO, performance evaluation for video surveillance systems," in *IEEE Intl. Conf. on Advanced Video and Signal Based Surveillance*, 2007. 6

[25] E. Seemann, M. Fritz, and B. Schiele, "Towards robust pedestrian detection in crowded image sequences," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2007. 7

[26] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986. 7

[27] M. Hussein, F. Porikli, and L. Davis, "A comp. eval. framework and a comparative study for human det." *IEEE Trans. Intelligent Transportation Systems*, vol. 10, pp. 417–427, Sept. 2009. 7

[28] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2010. 8, 11, 12

[29] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *British Machine Vision Conf.*, 2009. 8, 11, 12, 16

[30] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *Intl. Journal of Computer Vision*, vol. 73, pp. 41–59, 2007. 9, 10

[31] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *Intl. Journal of Computer Vision*, vol. 77, no. 1-3, pp. 259–289, May 2008. 9

[32] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by eff. subwindow search," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2008. 9

[33] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2007. 9, 10, 11, 12

[34] S. Maji, A. Berg, and J. Malik, "Classification using intersection kernel SVMs is efficient," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2008. 9, 11, 12

[35] C. Gu, J. J. Lim, P. Arbelaez, and J. Malik, "Recog. using regions," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009. 10

[36] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2005. 10, 11

[37] E. Seemann, B. Leibe, K. Mikolajczyk, and B. Schiele, "An evaluation of local shape-based features for pedestrian detection," in *British Machine Vision Conf.*, 2005. 10

[38] I. Alonso, D. Llorca, M. Sotelo, L. Bergasa, P. R. de Toro, J. Nuevo, M. Ocana, and M. Garrido, "Combination of feature extraction methods for SVM pedestrian detection," *IEEE Trans. Intelligent Transportation Systems*, vol. 8, no. 2, pp. 292–307, June 2007. 10

[39] M. Bajracharya, B. Moghaddam, A. Howard, S. Brennan, and L. H. Matthies, "A fast stereo-based system for detecting and tracking pedestrians from a moving vehicle," *The Intl. Journal of Robotics Research*, vol. 28, 2009. 10

[40] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "Robust multi-person tracking from a mobile platform," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, pp. 1831–1846, 2009. 10

[41] C. Wojek, S. Roth, K. Schindler, and B. Schiele, "Monocular 3d scene modeling and inference: Understanding multi-object traffic scenes," in *European Conf. Computer Vision*, 2010. 10, 18

[42] E. Dickmanns, *Dynamic Vision for Perception and Control of Motion*. Springer, 2007. 10, 18

[43] T. Gandhi and M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Trans. Intelligent Transportation Systems*, vol. 8, no. 3, pp. 413–430, Sept. 2007. 10

[44] P. A. Viola and M. J. Jones, "Robust real-time face det." *Intl. Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004. 10, 12, 16

[45] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. 10

[46] Q. Zhu, S. Avidan, M. Yeh, and K. Cheng, "Fast human detection using a cascade of histograms of oriented gradients," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2006. 10, 16

[47] F. M. Porikli, "Integral histogram: A fast way to extract histograms in cartesian spaces," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2005. 10

[48] A. Shashua, Y. Gdalyahu, and G. Hayun, "Ped. det. for driving assistance systems: Single-frame classification and system level performance," in *IEEE Intl. Conf. Intelligent Vehicles*, 2004. 10

[49] D. M. Gavrila and V. Philomin, "Real-time object det. for smart vehicles," in *IEEE Intl. Conf. Computer Vision*, 1999, pp. 87–93. 10

[50] D. M. Gavrila, "A bayesian, exemplar-based approach to hierarchical shape matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2007. 10

[51] Y. Liu, S. Shan, W. Zhang, X. Chen, and W. Gao, "Granularity-tunable gradients partition descriptors for human det." in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009. 10

[52] Y. Liu, S. Shan, X. Chen, J. Heikkila, W. Gao, and M. Pietikainen, "Spatial-temporal granularity-tunable gradients partition descriptors for human det." in *European Conf. Computer Vision*, 2010. 10

[53] P. A. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Intl. Journal of Computer Vision*, vol. 63(2), pp. 153–161, 2005. 10

[54] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conf. Computer Vision*, 2006. 10, 12

[55] N. Dalal, "Finding people in images and videos," Ph.D. dissertation, Institut Nat. Polytechnique de Grenoble, July 2006. 10, 12

[56] C. Wojek and B. Schiele, "A performance evaluation of single and multi-feature people detection," in *DAGM Symposium Pattern Recognition*, 2008. 11, 12

[57] G. Mori, S. Belongie, and J. Malik, "Efficient shape matching using shape contexts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 1832–1837, 2005. 11

[58] B. Wu and R. Nevatia, "Optimizing discrimination-efficiency tradeoff in integrating heterogeneous local features for object det." in *IEEE Conf. Computer Vision and Pattern Recognition*, 2008. 11

[59] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *IEEE Intl. Conf. Computer Vision*, 2009. 11, 12

[60] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul. 2002. 11, 12

[61] S. Hussain and B. Triggs, "Feature sets and dimensionality reduction for visual object det." in *British Machine Vision Conf.*, 2010. 11

[62] P. Ott and M. Everingham, "Implicit color segmentation features for pedestrian and object detection," in *IEEE Intl. Conf. Computer Vision*, 2009. 11

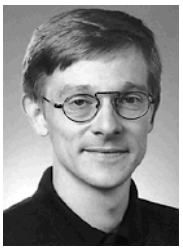
[63] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *British Machine Vision Conf.*, 2010. 11, 12, 16

[64] O. Tuzel, F. Porikli, and P. Meer, "Ped. det. via classification on riemannian manifolds," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008. 11

[65] B. Babenko, P. Dollár, Z. Tu, and S. Belongie, "Simultaneous learning and alignment: Multi-instance and multi-pose learning," in *ECCV Faces in Real-Life Images*, 2008. 11
[66] S. Walk, K. Schindler, and B. Schiele, "Disparity statistics for pedestrian detection: Combining appearance, motion and stereo," in *European Conf. Computer Vision*, 2010. 11
[67] P. Dollár, Z. Tu, H. Tao, and S. Belongie, "Feature mining for image classification," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2007. 11, 12
[68] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg, "Part-based feature synthesis for human detection," in *European Conf. Computer Vision*, 2010. 11, 12
[69] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis, "Human detection using partial least squares analysis," in *IEEE Intl. Conf. Computer Vision*, 2009. 11, 12
[70] Z. Lin and L. S. Davis, "A pose-invariant descriptor for human det. and seg." in *European Conf. Computer Vision*, 2008. 11, 12
[71] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2008. 11, 12
[72] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 99, no. PrePrints, 2009. 11, 12
[73] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object det. in images by components," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 4, pp. 349–361, Apr. 2001. 11
[74] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *European Conf. Computer Vision*, 2004. 11
[75] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue ped. classification with partial occlusion handling," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2010. 11
[76] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *IEEE Intl. Conf. Computer Vision*, 2009. 11
[77] M. Enzweiler and D. M. Gavrila, "Integrated pedestrian classification and orientation estimation," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2010. 11
[78] D. Tran and D. Forsyth, "Configuration estimates improve pedestrian finding," in *Advances in Neural Information Processing Systems*, 2008. 11
[79] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recog." in *European Conf. Computer Vision*, 2000. 11
[80] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2003. 11
[81] S. Agarwal and D. Roth, "Learning a sparse representation for object det." in *European Conf. Computer Vision*, 2002. 11
[82] P. Dollár, B. Babenko, S. Belongie, P. Perona, and Z. Tu, "Multiple component learning for object detection," in *European Conf. Computer Vision*, 2008. 11
[83] Z. Lin, G. Hua, and L. S. Davis, "Multiple instance feature for robust part-based object detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009. 11
[84] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for obj. det." in *European Conf. Computer Vision*, 2010. 11
[85] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973. 12
[86] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2007. 12
[87] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006. 15, 16
[88] S. García and F. Herrera, "An extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all pairwise comparisons," *Journal of Machine Learning Research*, vol. 9, pp. 2677–2694, 2008. 15
[89] W. Zhang, G. J. Zelinsky, and D. Samaras, "Real-time accurate object detection using multiple resolutions," in *IEEE Intl. Conf. Computer Vision*, 2007. 16
[90] C. Wojek, G. Dorkó, A. Schulz, and B. Schiele, "Sliding-windows for rapid object class localization: A parallel technique," in *DAGM Symposium Pattern Recognition*, 2008. 16

Piotr Dollár received his masters degree in computer science from Harvard University in 2002 and his PhD from the University of California, San Diego in 2007. He joined the Computational Vision lab at Caltech as a postdoctoral fellow in 2007, where he currently resides. He has worked on behavior recognition, boundary learning, manifold learning, and object and pedestrian detection, including efficient feature representations and novel learning paradigms. His general interests lie in machine learning and pattern recognition and their application to computer vision.

Christian Wojek received his masters degree in computer science from the University of Karlsruhe in 2006 and his PhD from TU Darmstadt in 2010. He was awarded a DAAD scholarship to visit McGill University from 2004 to 2005. He was with MPI Informatics Saarbrücken as a postdoctoral fellow from 2010 to 2011 and in 2011 joined Carl Zeiss Corporate Research. His research interests are object detection, scene understanding and activity recognition.



Bernt Schiele received his masters in computer science from Univ. of Karlsruhe and INP Grenoble in 1994. In 1997 he obtained his PhD from INP Grenoble in computer vision. He was a postdoctoral associate and Visiting Assistant Professor at MIT between 1997 and 2000. From 1999 until 2004 he was an Assistant Professor at ETH Zurich and from 2004 to 2010 he was a full professor of computer science at TU Darmstadt. In 2010, he was appointed scientific member of the Max Planck Society and a director at the Max Planck Institute for Informatics. Since 2010 he has also been a Professor at Saarland University. His main interests are computer vision, perceptual computing, statistical learning methods, wearable computers, and integration of multi-modal sensor data. He is particularly interested in developing methods which work under real-world conditions.

Pietro Perona graduated in Electrical Engineering from the Università di Padova in 1985 and received a PhD in EECS from the University of California at Berkeley in 1990. After a postdoctoral fellowship at MIT in 1990-91 he joined the faculty of Caltech in 1991, where is is now Allen E. Puckett Professor of Electrical Engineering and Computation and Neural Systems. His current interests are visual recognition, modeling and measuring animal behavior and Visipedia. He has worked on anisotropic diffusion, multiresolution-multi-orientation filtering, human texture perception and segmentation, dynamic vision, grouping, analysis of human motion, recognition of object categories, and modeling visual search.