Ten Years of Pedestrian Detection, What Have We Learned?

Rodrigo Benenson Mohamed Omran Jan Hosang Bernt Schiele

Max Planck Institute for Informatics
Saarbrücken, Germany
firstname.lastname@mpi-inf.mpg.de

Abstract Paper-by-paper results make it easy to miss the forest for the trees.We analyse the remarkable progress of the last decade by discussing the main ideas explored in the 40+ detectors currently present in the Caltech pedestrian detection benchmark. We observe that there exist three families of approaches, all currently reaching similar detection quality. Based on our analysis, we study the complementarity of the most promising ideas by combining multiple published strategies. This new decision forest detector achieves the current best known performance on the challenging Caltech-USA dataset.

1 Introduction

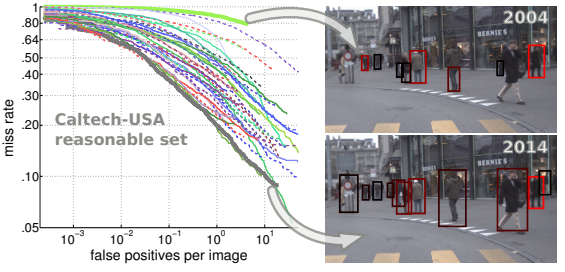
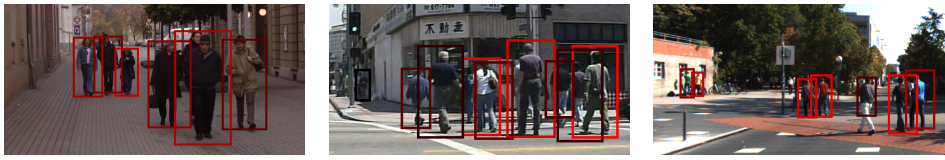


Figure 1: The last decade has shown tremendous progress on pedestrian detection. What have we learned out of the 40+ proposed methods?

Pedestrian detection is a canonical instance of object detection. Because of its direct applications in car safety, surveillance, and robotics, it has attracted much attention in the last years. Importantly, it is a well defined problem with established benchmarks and evaluation metrics. As such, it has served as a playground to explore different ideas for object detection. The main paradigms for object detection "Viola&Jones variants", HOG+SVM rigid templates, deformable part detectors (DPM), and convolutional neural networks (ConvNets) have all been explored for this task.

The aim of this paper is to review progress over the last decade of pedestrian detection (40+ methods), identify the main ideas explored, and try to quantify which ideas had the most impact on final detection quality. In the next sections we review existing datasets (section 2), provide a discussion of the different approaches (section 3), and experiments reproducing/quantifying the recent years' progress (section 4, presenting experiments over $\sim$ 20 newly trained detector models). Although we do not aim to introduce a novel technique, by putting together existing methods we report the best known detection results on the challenging Caltech-USA dataset.



(a) INRIA test set (b) Caltech-USA test set (c) KITTI test set

Figure 2: Example detections of a top performing method (`SquaresChnFtrs`).

2 Datasets

Multiple public pedestrian datasets have been collected over the years; INRIA [1], ETH [2], TUD-Brussels [3], Daimler [4] (Daimler stereo [5]), Caltech-USA [6], and KITTI [7] are the most commonly used ones. They all have different characteristics, weaknesses, and strengths.

INRIA is amongst the oldest and as such has comparatively few images. It benefits however from high quality annotations of pedestrians in diverse settings (city, beach, mountains, etc.), which is why it is commonly selected for training (see also §4.4). ETH and TUD-Brussels are mid-sized video datasets. Daimler is not considered by all methods because it lacks colour channels. Daimler stereo, ETH, and KITTI provide stereo information. All datasets but INRIA are obtained from video, and thus enable the use of optical flow as an additional cue.

Today, Caltech-USA and KITTI are the predominant benchmarks for pedestrian detection. Both are comparatively large and challenging. Caltech-USA stands out for the large number of methods that have been evaluated side-by-side. KITTI stands out because its test set is slightly more diverse, but is not yet used as frequently. For a more detailed discussion of the datasets please consult [8,7]. INRIA, ETH (monocular), TUD-Brussels, Daimler (monocular), and Caltech-USA are available under a unified evaluation toolbox; KITTI uses its own separate one with unpublished test data. Both toolboxes maintain an online ranking where published methods can be compared side by side.

In this paper we use primarily Caltech-USA for comparing methods, INRIA and KITTI secondarily. See figure 2 for example images. Caltech-USA and INRIA results are measured in log-average miss-rate (MR, lower is better), while KITTI uses area under the precision-recall curve (AUC, higher is better).

Value of benchmarks Individual papers usually only show a narrow view over the state of the art on a dataset. Having an official benchmark that collects detections from all methods greatly eases the author's effort to put their curve into context, and provides reviewers easy access to the state of the art results. The collection of results enable retrospective analyses such as the one presented in the next section.

3 Main approaches to improve pedestrian detection

Figure 3 and table 1 together provide a quantitative and qualitative overview over 40+ methods whose results are published on the Caltech pedestrian detection benchmark (July 2014). Methods marked in italic are our newly trained

Method	MR	Family	Features	Classifier	Context	Deep	Parts	M-Scales	More data	Feat. type	Training
VJ [9]	94.73%	DF	✓	✓						Haar	I
Shapelet [10]	91.37%	-	✓							Gradients	I
PoseInv [11]	86.32%	-					✓			HOG	I+
LatSvm-V1 [12]	79.78%	DPM					✓			HOG	P
ConvNet [13]	77.20%	DN				✓				Pixels	I
FtrMine [14]	74.42%	DF	✓							HOG+Color	I
HikSvm [15]	73.39%	-		✓						HOG	I
HOG [1]	68.46%	-	✓	✓						HOG	I
MultiFtr [16]	68.26%	DF	✓	✓						HOG+Haar	I
HogLbp [17]	67.77%	-	✓							HOG+LBP	I
AFS+Geo [18]	66.76%	-			✓					Custom	I
AFS [18]	65.38%	-								Custom	I
LatSvm-V2 [19]	63.26%	DPM		✓			✓			HOG	I
Pls [20]	62.10%	-	✓	✓						Custom	I
MLS [21]	61.03%	DF	✓							HOG	I
MultiFtr+CSS [22]	60.89%	DF	✓							Many	T
FeatSynth [23]	60.16%	-	✓	✓						Custom	I
pAUCBoost [24]	59.66%	DF	✓	✓						HOG+COV	I
FPDW [25]	57.40%	DF								HOG+LUV	I
ChnFtrs [26]	56.34%	DF	✓	✓						HOG+LUV	I
CrossTalk [27]	53.88%	DF			✓					HOG+LUV	I
DBN-Isol [28]	53.14%	DN					✓			HOG	I
ACF [29]	51.36%	DF	✓							HOG+LUV	I
RandForest [30]	51.17%	DF		✓						HOG+LBP	I&C
MultiFtr+Motion [22]	50.88%	DF	✓						✓	Many+Flow	T
SquaresChnFtrs [31]	50.17%	DF	✓							HOG+LUV	I
Franken [32]	48.68%	DF		✓						HOG+LUV	I
MultiResC [33]	48.45%	DPM			✓		✓	✓		HOG	C
Roerei [31]	48.35%	DF	✓					✓		HOG+LUV	I
DBN-Mut [34]	48.22%	DN			✓		✓			HOG	C
MF+Motion+2Ped [35]	46.44%	DF			✓				✓	Many+Flow	I+
MOCO [36]	45.53%	-	✓		✓					HOG+LBP	C
MultiSDP [37]	45.39%	DN	✓		✓	✓				HOG+CSS	C
ACF-Caltech [29]	44.22%	DF	✓							HOG+LUV	C
MultiResC+2Ped [35]	43.42%	DPM			✓		✓	✓		HOG	C+
WordChannels [38]	42.30%	DF	✓							Many	C
MT-DPM [39]	40.54%	DPM					✓	✓		HOG	C
JointDeep [40]	39.32%	DN			✓					Color+Gradient	C
SDN [41]	37.87%	DN				✓	✓			Pixels	C
MT-DPM+Context [39]	37.64%	DPM			✓		✓	✓		HOG	C+
ACF+SDt [42]	37.34%	DF	✓						✓	ACF+Flow	C+
SquaresChnFtrs [31]	34.81%	DF	✓							HOG+LUV	C
InformedHaar [43]	34.60%	DF	✓							HOG+LUV	C
Katamari-v1	22.49%	DF	✓		✓				✓	HOG+Flow	C+

Table 1: Listing of methods considered on Caltech-USA, sorted by log-average miss-rate (lower is better). Consult sections 3.1 to 3.9 for details of each column. See also matching figure 3. "HOG" indicates HOG-like [1]. Ticks indicate salient aspects of each method.

models (described in section 4). We refer to all methods using their Caltech benchmark shorthand. Instead of discussing the methods' individual particularities, we identify the key aspects that distinguish each method (ticks of table 1) and group them accordingly. We discuss these aspects in the next subsections.

Brief chronology In 2003, Viola and Jones applied their `VJ` detector [44] to the task of pedestrian detection. In 2005 Dalal and Triggs introduced the landmark `HOG` [1] detector, which later served in 2008 as a building block for the now classic deformable part model DPM (named `LatSvm` here) by Felzenswalb et al. [12]. In 2009 the Caltech pedestrian detection benchmark was introduced, comparing seven pedestrian detectors [6]. At this point in time, the evaluation metrics changed from per-window (FPPW) to per-image (FPPI), once the flaws of the per-window evaluation were identified [8]. Under this new evaluation metric some of the early detectors turned out to under-perform. About one third of the methods considered here were published during 2013, reflecting a renewed interest on the problem. Similarly, half of the KITTI results for pedestrian detection were submitted in 2014.

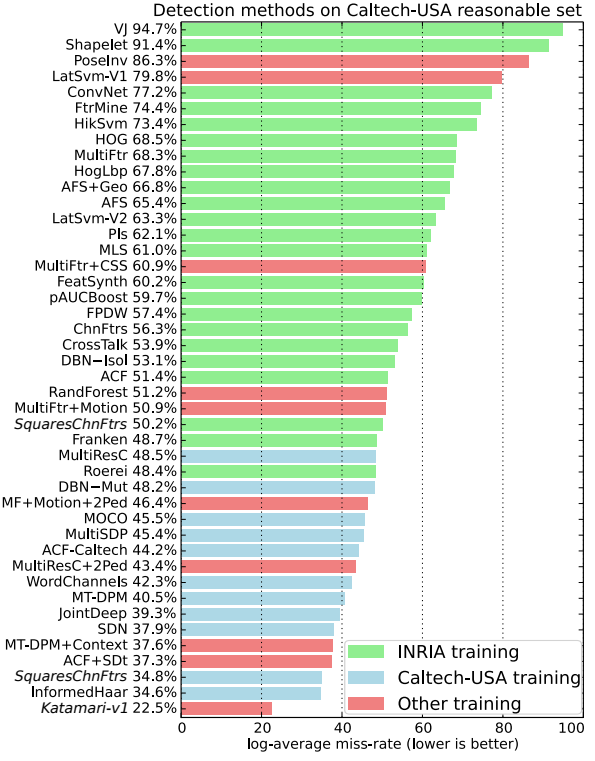


Figure 3: Caltech-USA detection results.

3.1 Training data

Figure 3 shows that differences in detection performance are, not surprisingly, dominated by the choice of training data. Methods trained on Caltech-USA systematically perform better than methods that generalise from INRIA. Table 1 gives additional details on the training data used[1]. High performing methods with "other training" use extended versions of Caltech-USA. For instance `MultiResC+2Ped` uses Caltech-USA plus an extended set of annotations over INRIA, `MT-DPM+Context` uses an external training set for cars, and `ACF+SDt`

[1] "Training" data column: I→INRIA, C→Caltech, I+/C+ →INRIA/Caltech and additional data, P→Pascal, T→TUD-Motion, I&C→both INRIA and Caltech.

employs additional frames from the original Caltech-USA videos.

3.2 Solution families

Overall we notice that out of the 40+ methods we can discern three families: 1) DPM variants (`MultiResC` [33], `MT-DPM` [39], etc.), 2) Deep networks (`JointDeep` [40], `ConvNet` [13], etc.), and 3) Decision forests (`ChnFtrs`, `Roerei`, etc.). On table 1 we identify these families as `DPM`, `DN`, and `DF` respectively.

Based on raw numbers alone boosted decision trees (`DF`) seem particularly suited for pedestrian detection, reaching top performance on both the "train on INRIA, test on Caltech", and "train on Caltech, test on Caltech" tasks. It is unclear however what gives them an edge. The deep networks explored also show interesting properties and fast progress in detection quality.

Conclusion Overall, at present, DPM variants, deep networks, and (boosted) decision forests all reach top performance in pedestrian detection (around 37 % MR on Caltech-USA, see figure 3).

3.3 Better classifiers

Since the original proposal of `HOG+SVM` [1], linear and non-linear kernels have been considered. `HikSvm` [15] considered fast approximations of non-linear kernels. This method obtains improvements when using the flawed FPPW evaluation metric (see section 3), but fails to perform well under the proper evaluation (FPPI). In the work on `MultiFtrs` [16], it was argued that, given enough features, Adaboost and linear SVM perform roughly the same for pedestrian detection.

Recently, more and more components of the detector are optimized jointly with the "decision component" (e.g. pooling regions in `ChnFtrs` [26], filters in `JointDeep` [40]). As a result the distinction between features and classifiers is not clear-cut anymore (see also sections 3.8 and 3.9).

Conclusion There is no conclusive empirical evidence indicating that whether non-linear kernels provide meaningful gains over linear kernels (for pedestrian detection, when using non-trivial features). Similarly, it is unclear whether one particular type of classifier (e.g. SVM or decision forests) is better suited for pedestrian detection than another.

3.4 Additional data

The core problem of pedestrian detection focuses on individual monocular colour image frames. Some methods explore leveraging additional information at training and test time to improve detections. They consider stereo images [45], optical flow (using previous frames, e.g. `MultiFtr+Motion` [22] and `ACF+SDt` [42]), tracking [46], or data from other sensors (such as lidar [47] or radar).

For monocular methods it is still unclear how much tracking can improve per-frame detection itself. As seen in figure 4 exploiting optical flow provides a non-trivial improvement over the baselines. Curiously, the current best results (`ACF-SDt` [42]) are obtained using coarse rather than high quality flow. In section 4.2 we inspect the complementarity of flow with other ingredients. Good success exploiting flow and stereo on the Daimler dataset has been reported [48], but similar results have yet to be seen on newer datasets such as KITTI.

Conclusion Using additional data provides meaningful improvements, albeit on modern dataset stereo and flow cues have yet to be fully exploited. As of now, methods based merely on single monocular image frames have been able to keep up with the performance improvement introduced by additional information.

3.5 Exploiting context

Sliding window detectors score potential detection windows using the content inside that window. Drawing on the context of the detection window, i.e. image content surrounding the window, can improve detection performance. Strategies for exploiting context include: ground plane constraints (`MultiResC` [33], `RandForest` [30]), variants of auto-context [49] (`MOCO` [36]), other category detectors (`MT-DPM+Context` [39]), and person-to-person patterns (`DBN-Mut` [34], `+2Ped` [35], `JointDeep` [40]).

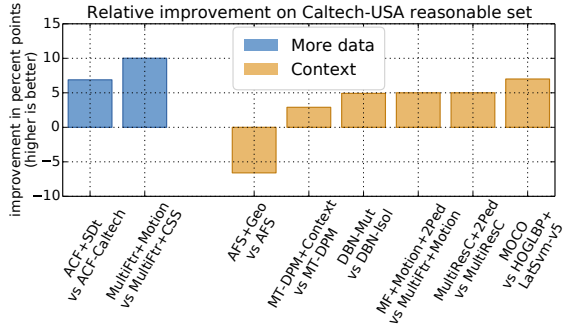


Figure 4: Caltech-USA detection improvements for different method types. Improvement relative to each method's relevant baseline ("method vs baseline").

Figure 4 shows the performance improvement for methods incorporating context. Overall, we see improvements of $3 \sim 7$ MR percent points. (The negative impact of `AFS+Geo` is due to a change in evaluation, see section 3.) Interestingly, `+2Ped` [35] obtains a consistent $2 \sim 5$ MR percent point improvement over existing methods, even top performing ones (see section 4.2).

Conclusion Context provides consistent improvements for pedestrian detection, although the scale of improvement is lower compared to additional test data (§3.4) and deep architectures (§3.8). The bulk of detection quality must come from other sources.

3.6 Deformable parts

The DPM detector [19] was originally motivated for pedestrian detection. It is an idea that has become very popular and dozens of variants have been explored.

For pedestrian detection the results are competitive, but not salient (`LatSvm` [50,12], `MultiResC` [33], `MT-DPM` [39]). More interesting results have been obtained when modelling parts and their deformations inside a deep architecture (e.g. `DBN-Mut` [34], `JointDeep` [40]).

DPM and its variants are systematically outmatched by methods using a single component and no parts (Roerei [31], SquaresChnFtrs see section 4.1), casting doubt on the need for parts. Recent work has explored ways to capture deformations entirely without parts [51,52].

Conclusion For pedestrian detection there is still no clear evidence for the necessity of components and parts, beyond the case of occlusion handling.

3.7 Multi-scale models

Typically for detection, both high and low resolution candidate windows are resampled to a common size before extracting features. It has recently been noticed that training different models for different resolutions systematically improve performance by $1 \sim 2$ MR percent points [33,31,39], since the detector has access to the full information available at each window size. This technique does not impact computational cost at detection time [53], although training time increases.

Conclusion Multi-scale models provide a simple and generic extension to existing detectors. Despite consistent improvements, their contribution to the final quality is rather minor.

3.8 Deep architectures

Large amounts of training data and increased computing power have lead to recent successes of deep architectures (typically convolutional neural networks) on diverse computer vision tasks (large scale classification and detection [54,55], semantic labelling [56]). These results have inspired the application of deep architectures to the pedestrian task.

ConvNet [13] uses a mix of unsupervised and supervised training to create a convolutional neural network trained on INRIA. This method obtains fair results on INRIA, ETH, and TUD-Brussels, however fails to generalise to the Caltech setup. This method learns to extract features directly from raw pixel values.

Another line of work focuses on using deep architectures to jointly model parts and occlusions (DBN-Isol [28], DBN-Mut [34], JointDeep [40], and SDN [41]). The performance improvement such integration varies between 1.5 to 14 MR percent points. Note that these works use edge and colour features [40,34,28], or initialise network weights to edge-sensitive filters, rather than discovering features from raw pixel values as usually done in deep architectures. No results have yet been reported using features pre-trained on ImageNet [54,57].

Conclusion Despite the common narrative there is still no clear evidence that deep networks are good at learning features for pedestrian detection (when using pedestrian detection training data). Most successful methods use such architectures to model higher level aspects of parts, occlusions, and context. The obtained results are on par with DPM and decision forest approaches, making the advantage of using such involved architectures yet unclear.

3.9 Better features

The most popular approach (about 30 % of the considered methods) for improving detection quality is to increase/diversify the features computed over the input image. By having richer and higher dimensional representations, the classification task becomes somewhat easier, enabling improved results. A large set of feature types have been explored: edge information [1,26,58,41], colour information [26,22], texture information [17], local shape information [38], covariance features [24], amongst others. More and more diverse features have been shown to systematically improve performance.

While various decision forest methods use 10 feature channels (`ChnFtrs`, `ACF`, `Roerei`, `SquaresChnFtrs`, etc.), some papers have considered up to an order of magnitude more channels [16,58,24,30,38]. Despite the improvements by adding many channels, top performance is still reached with only 10 channels (6 gradient orientations, 1 gradient magnitude, and 3 colour channels, we name these HOG+LUV); see table 1 and figure 3. In section 4.1 we study in more detail different feature combinations.

From all what we see, from `VJ` (95% MR) to `ChnFtrs` (56.34% MR, by adding HOG and LUV channels), to `SquaresChnFtrs-Inria` (50.17% MR, by exhaustive search over pooling sizes, see section 4), improved features drive progress. Switching training sets (section 3.1) enables `SquaresChnFtrs-Caltech` to reach state of the art performance on the Caltech-USA dataset; improving over significantly more sophisticated methods. `InformedHaar` [43] obtains top results by using a set of Haar-like features manually designed for the pedestrian detection task. In contrast `SquaresChnFtrs-Caltech` obtains similar results without using such hand-crafted features and being data driven instead.

Upcoming studies show that using more (and better features) yields further improvements [59,60]. It should be noted that better features for pedestrian detection have not yet been obtained via deep learning approaches (see caveat on ImageNet features in section 3.8).

Conclusion In the last decade improved features have been a constant driver for detection quality improvement, and it seems that it will remain so in the years to come. Most of this improvement has been obtained by extensive trial and error. The next scientific step will be to develop a more profound understanding of the what makes good features good, and how to design even better ones[2].

4 Experiments

Based on our analysis in the previous section, three aspects seem to be the most promising in terms of impact on detection quality: better features (§3.9), additional data (§3.4), and context information (§3.5). We thus conduct experiments on the complementarity of these aspects.

Among the three solution families discussed (section 3.2), we choose the Integral Channels Features framework [26] (a decision forest) for conducting our

[2] This question echoes with the current state of the art in deep learning, too.

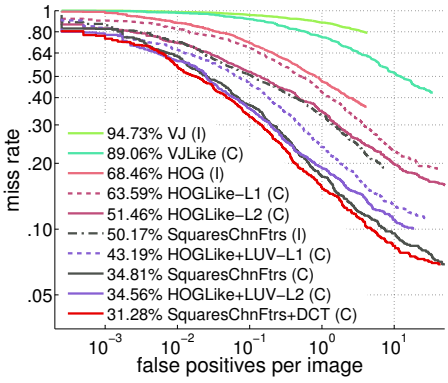


Figure 5: Effect of features on detection performance. Caltech-USA reasonable test set.

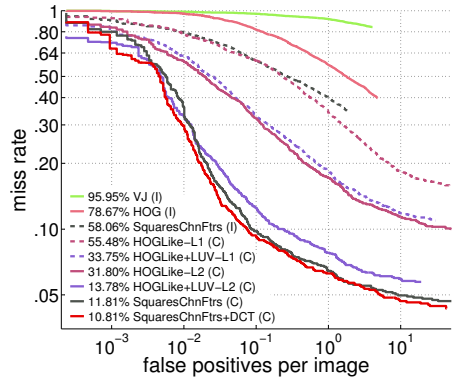


Figure 6: Caltech-USA training set performance. (I)/(C) indicates using INRIA/Caltech-USA training set.

experiments. Methods from this family have shown good performance, train in minutes∼hours, and lend themselves to the analyses we aim.

In particular, we use the (open source) `SquaresChnFtrs` baseline described in [31]: 2048 level-2 decision trees (3 threshold comparisons per tree) over `HOG+LUV` channels (10 channels), composing one 64×128 pixels template learned via vanilla AdaBoost and few bootstrapping rounds of hard negative mining.

4.1 Reviewing the effect of features

In this section, we evaluate the impact of increasing feature complexity. We tune all methods on the INRIA test set, and demonstrate results on the Caltech-USA test set (see figure 5). Results on INRIA as well as implementation details can be found in the supplementary material.

The first series of experiments aims at mimicking landmark detection techniques, such as `VJ` [44], `HOG`+linear SVM [1], and `ChnFtrs` [26]. `VJLike` uses only the luminance colour channel, emulating the Haar wavelet like features from the original [44] using level 2 decision trees. `HOGLike-L1/L2` use 8×8 pixel pooling regions, 1 gradient magnitude and 6 oriented gradient channels, as well as level 1/2 decision trees. We also report results when adding the LUV colour channels `HOGLike+LUV` (10 feature channels total). `SquaresChnFtrs` is the baseline described in the beginning of section 4, which is similar to `HOGLike+LUV` to but with square pooling regions of any size.

Inspired by [60], we also expand the 10 HOG+LUV channels into 40 channels by convolving each channel with three DCT (discrete cosine transform) basis functions (of 7×7 pixels), and storing the absolute value of the filter responses as additional feature channels. We name this variant `SquaresChnFtrs+DCT`.

Conclusion Much of the progress since `VJ` can by explained by the use of better features, based on oriented gradients and colour information. Simple tweaks to

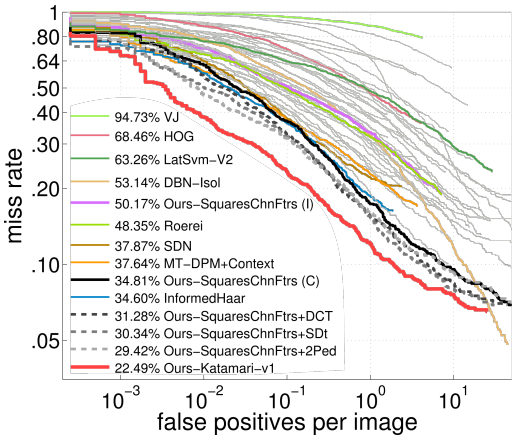


Figure 7: Some of the top quality detection methods for Caltech-USA. See section 4.2.

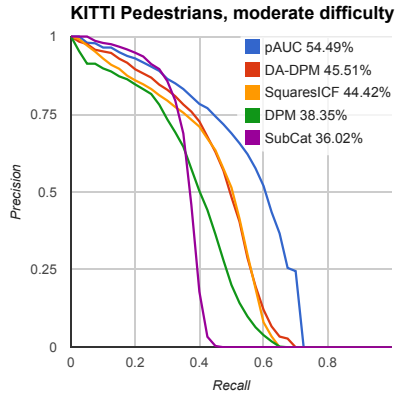


Figure 8: Pedestrian detection on the KITTI dataset.

these well known features (e.g. projection onto the DCT basis) can still yield noticeable improvements.

4.2 Complementarity of approaches

After revisiting the effect of single frame features in section 4.1 we now consider the complementary of better features (HOG+LUV+DCT), additional data (via optical flow), and context (via person-to-person interactions).

We encode the optical flow using the same SDt features from `ACF+SDt` [33] (image difference between current frame T and coarsely aligned T-4 and T-8). The context information is injected using the `+2Ped` re-weighting strategy [35] (the detection scores are combined with the scores of a "2 person" DPM detector). In all experiments both DCT and SDt features are pooled over 8×8 regions (as in [33]), instead of "all square sizes" for the HOG+LUV features.

The combination `SquaresChnFtrs+DCT+SDt+2Ped` is called `Katamari-v1`. Unsurprisingly, `Katamari-v1` reaches the best known result on the Caltech-USA dataset. In figure 7 we show it together with the best performing method for each training set and solution family (see table 1). The supplementary material contains results of all combinations between the ingredients.

Conclusion Our experiments show that adding extra features, flow, and context information are largely complementary (12 % gain, instead of $3 + 7 + 5$ %), even when starting from a strong detector.
It remains to be seen if future progress in detection quality will be obtained by further insights of the "core" algorithm (thus further diminishing the relative improvement of add-ons), or by extending the diversity of techniques employed inside a system.

4.3 How much model capacity is needed?

The main task of detection is to generalise from training to test set. Before we analyse the generalisation capability (section 4.4), we consider a necessary condition for high quality detection: is the learned model performing well on the training set?

In figure 6 we see the detection quality of the models considered in section 4.1, when evaluated over their training set. None of these methods performs perfectly on the training set. In fact, the trend is very similar to performance on the test set (see figure 5) and we do not observe yet symptoms of over-fitting.

Conclusion Our results indicate that research on increasing the discriminative power of detectors is likely to further improve detection quality. More discriminative power can originate from more and better features or more complex classifiers.

4.4 Generalisation across datasets

For real world application beyond a specific benchmark, the generalisation capability of a model is key. In that sense results of models trained on INRIA and tested on Caltech-USA are more relevant than the ones trained (and tested) on Caltech-USA.

Table 2: Effect of training set over the detection quality. Bold indicates second best training set for each test set, except for ETH where bold indicates the best training set.

Test set \ Training set	INRIA	Caltech-USA	KITTI
INRIA	*17.42* %	60.50 %	**55.83** %
Caltech-USA	**50.17** %	*34.81* %	61.19 %
KITTI	**38.61** %	28.65 %	*44.42* %
ETH	**56.27** %	76.11%	61.19 %

Table 2 shows the performance of `SquaresChnFtrs` over Caltech-USA when using different training sets (MR for INRIA/Caltech/ETH, AUC for KITTI). These experiments indicate that training on Caltech or KITTI provides little generalisation capability towards INRIA, while the converse is not true. Surprisingly, despite the visual similarity between KITTI and Caltech, INRIA is the second best training set choice for KITTI and Caltech. This shows that Caltech-USA pedestrians are of "their own kind", and that the INRIA dataset is effective due to its diversity. In other words few diverse pedestrians (INRIA) is better than many similar ones (Caltech/KITTI).

The good news is that the best methods seem to perform well both across datasets and when trained on the respective training data. Figure 8 shows methods trained and tested on KITTI, we see that `SquaresChnFtrs` (named `SquaresICF` in KITTI) is better than vanilla DPM and on par with the best known DPM variant. The currently best method on KITTI, `pAUC` [59], is a variant of `ChnFtrs` using 250 feature channels (see the KITTI website for details on the methods). These two observations are consistent with our discussions in sections 3.9 and 4.1.

Conclusion While detectors learned on one dataset may not necessarily transfer well to others, their ranking is stable across datasets, suggesting that insights can be learned from well-performing methods regardless of the benchmark.

5 Conclusion

Our experiments show that most of the progress in the last decade of pedestrian detection can be attributed to the improvement in features alone. Evidence suggests that this trend will continue. Although some of these features might be driven by learning, they are mainly hand-crafted via trial and error.

Our experiment combining the detector ingredients that our retrospective analysis found to work well (better features, optical flow, and context) shows that these ingredients are mostly complementary. Their combination produces best published detection performance on Caltech-USA.

While the three big families of pedestrian detectors (deformable part models, decision forests, deep networks) are based on different learning techniques, their state-of-the-art results are surprisingly close.

The main challenge ahead seems to develop a deeper understanding of what makes good features good, so as to enable the design of even better ones.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
2. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: CVPR, IEEE Press (June 2008)
3. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: CVPR. (2009)
4. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: Survey and experiments. PAMI (2009)
5. Keller, C., Fernandez, D., Gavrila, D.: Dense stereo-based roi generation for pedestrian detection. In: DAGM. (2009)
6. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: CVPR. (2009)
7. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and PatternRecognition (CVPR). (2012)
8. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. TPAMI (2011)
9. Viola, P., Jones, M.: Robust real-time face detection. In: IJCV. (2004)
10. Sabzmeydani, P., Mori, G.: Detecting pedestrians by learning shapelet features. In: CVPR. (2007)
11. Lin, Z., Davis, L.: A pose-invariant descriptor for human detection and segmentation. In: ECCV. (2008)
12. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR. (2008)
13. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: CVPR. (2013)
14. P. Dollár, Z. Tu, H.T., Belongie, S.: Feature mining for image classification. In: CVPR. (2007)
15. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR. (2008)

16. Wojek, C., Schiele, B.: A performance evaluation of single and multi-feature people detection. In: DAGM. (2008)
17. Wang, X., Han, X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: ICCV. (2009)
18. Levi, D., Silberstein, S., , Bar-Hillel, A.: Fast multiple-part based object detection using kd-ferns. In: CVPR. (2013)
19. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI (2010)
20. Schwartz, W., Kembhavi, A., Harwood, D., Davis, L.S.: Human detection using partial least squares analysis. In: ICCV. (2009)
21. Nam, W., Han, B., Han, J.: Improving object localization using macrofeature layout selection. In: ICCV, Visual Surveillance Workshop. (2011)
22. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. In: CVPR. (2010)
23. Bar-Hillel, A., Levi, D., Krupka, E., Goldberg, C.: Part-based feature synthesis for human detection. In: ECCV. (2010)
24. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Efficient pedestrian detection by directly optimize the partial area under the roc curve. In: ICCV. (2013)
25. Dollár, P., Belongie, S., Perona, P.: The fastest pedestrian detector in the west. In: BMVC. (2010)
26. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: BMVC. (2009)
27. Dollár, P., Appel, R., Kienzle, W.: Crosstalk cascades for frame-rate pedestrian detection. In: ECCV. (2012)
28. Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: CVPR. (2012)
29. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. PAMI (2014)
30. Marin, J., Vazquez, D., Lopez, A., Amores, J., Leibe, B.: Random forests of local experts for pedestrian detection. In: ICCV. (2013)
31. Benenson, R., Mathias, M., Tuytelaars, T., Van Gool, L.: Seeking the strongest rigid detector. In: CVPR. (2013)
32. Mathias, M., Benenson, R., Timofte, R., Van Gool, L.: Handling occlusions with franken-classifiers. In: ICCV. (2013)
33. Park, D., Ramanan, D., Fowlkes, C.: Multiresolution models for object detection. In: ECCV. (2010)
34. Ouyang, W., Zeng, X., Wang, X.: Modeling mutual visibility relationship with a deep model in pedestrian detection. In: CVPR. (2013)
35. Ouyang, W., Wang, X.: Single-pedestrian detection aided by multi-pedestrian detection. In: CVPR. (2013)
36. Chen, G., Ding, Y., Xiao, J., Han, T.X.: Detection evolution with multi-order contextual co-occurrence. In: CVPR. (2013)
37. Zeng, X., Ouyang, W., Wang, X.: Multi-stage contextual deep learning for pedestrian detection. In: ICCV. (2013)
38. Costea, A.D., Nedevschi, S.: Word channel based multiscale pedestrian detection without image resizing and using only one classifier. In: CVPR. (June 2014)
39. Yan, J., Zhang, X., Lei, Z., Liao, S., Li, S.Z.: Robust multi-resolution pedestrian detection in traffic scenes. In: CVPR. (2013)
40. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: ICCV. (2013)

41. Luo, P., Tian, Y., Wang, X., Tang, X.: Switchable deep network for pedestrian detection. In: CVPR. (2014)
42. Park, D., Zitnick, C.L., Ramanan, D., Dollár, P.: Exploring weak stabilization for motion feature extraction. In: CVPR. (2013)
43. Zhang, S., Bauckhage, C., Cremers, A.B.: Informed haar-like features improve pedestrian detection. In: CVPR. (2014)
44. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: CVPR. (2003)
45. Keller, C.G., Enzweiler, M., Rohrbach, M., Fernandez Llorca, D., Schnorr, C., Gavrila, D.M.: The benefits of dense stereo for pedestrian detection. IEEE Transactions on Intelligent Transportation Systems (2011)
46. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: Robust multi-person tracking from a mobile platform. PAMI (2009)
47. Premebida, C., Carreira, J., Batista, J., Nunes, U.: Pedestrian detection combining rgb and dense lidar data. In: IROS. (2014)
48. Enzweiler, M., Gavrila, D.: A multilevel mixture-of-experts framework for pedestrian classification. IEEE Transactions on Image Processing (2011)
49. Tu, Z., Bai, X.: Auto-context and its application to high-level vision tasks and 3d brain image segmentation. PAMI (2010)
50. Yan, J., Lei, Z., Wen, L., Li, S.Z.: The fastest deformable part model for object detection. In: CVPR. (June 2014)
51. Hariharan, B., Zitnick, C.L., Dollár, P.: Detecting objects using deformation dictionaries. In: CVPR. (2014)
52. Pedersoli, M., Tuytelaars, T., Gool, L.V.: Using a deformation field model for localizing faces and facial points under weak supervision. In: CVPR. (June 2014)
53. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. In: CVPR. (2012)
54. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: arXiv. (2014)
55. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. In: ICLR. (2014)
56. Pinheiro, P., Collobert, R.: Recurrent convolutional neural networks for scene labeling. In: JMLR. (2014)
57. Azizpour, H., Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: From generic to specific deep representations for visual recognition. CoRR (2014)
58. Lim, J., Zitnick, C.L., Dollár, P.: Sketch tokens: A learned mid-level representation for contour and object detection. In: CVPR. (2013)
59. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Strengthening the effectiveness of pedestrian detection with spatially pooled features. In: ECCV. (2014)
60. Nam, W., Dollár, P., Han, J.H.: Local decorrelation for improved detection. In: Nips. (2014)

Ten years of pedestrian detection, what have we learned? Supplementary material

6 Reviewing the effect of features

The idea behind the experiments in section 4.1 of the main paper is to demonstrate that, within a single framework, varying the features can replicate the jump in detection performance over a ten-year span ($2004 - 2014$), i.e. the jump in performance between `VJ` and the current state-of-the-art.

See figure 9 for results on INRIA and Caltech-USA of the following methods (all based on `SquaresChnFtrs`, described in section 4 of the paper):

`VJLike` uses only the luminance colour channel, emulating the original `VJ` [44]. We use 8 000 weak classifiers to compensate for the weak input feature, only square pooling regions, and level-2 trees to emulate the Haar wavelet-like features used by `VJ`.

`HOGLike-L1/L2` uses 8×8 pixel pooling regions, 6 oriented gradients, 1 gradient magnitude, and level $1/2$ decision trees ($1/3$ threshold comparisons respectively). A level-1 tree emulates the non-linearity in the original `HOG`+linear SVM features [1].

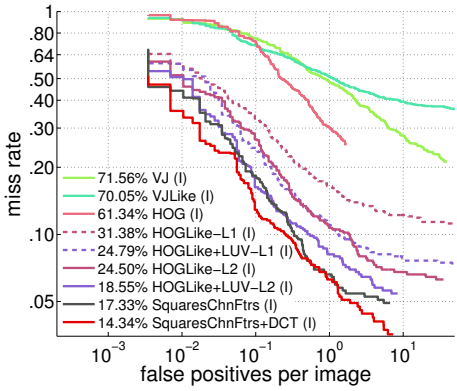
`HOGLike+LUV` is identical to `HOGLike,` but with additional LUV colour channels (10 feature channels total).

`SquaresChnFtrs` is the baseline described in the beginning of the experiments section (§4). It is similar to `HOGLike+LUV` but the size of the square pooling regions is not restricted.

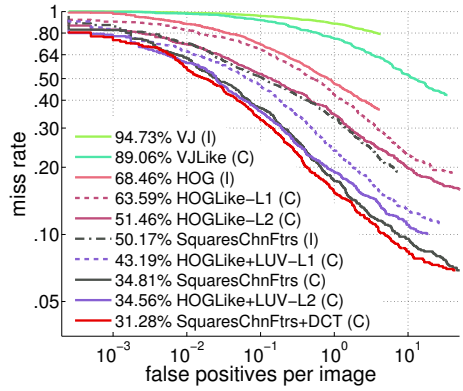
`SquaresChnFtrs+DCT` is inspired by [60]. We expand the ten HOG+LUV channels into 40 channels by convolving each of the 10 channels with three DCT (discrete cosine transform) filters (7×7 pixels), and storing the absolute value of the filter responses as additional feature channels. The three DCT basis functions we use as 2d-filters correspond to the lowest spatial frequencies. We name this variant `SquaresChnFtrs+DCT` and it serves as reference point for the performance improvement that can be obtained by increasing the number of channels.

7 Complementarity of approaches

Table 3 contains the detailed results of combining different approaches with a strong baseline, related to section 4.2 of the main paper. `Katamari-v1` combines all three listed approaches with `SquaresChnFtrs`. We train and test on the Caltech-USA dataset. It can be noticed that the obtained improvement is



(a) INRIA test set



(b) Caltech-USA reasonable test set

Figure 9: Effect of features on detection performance. (I)/(C) indicates using INRIA/Caltech-USA training set respectively.

very close to the sum of individual gains, showing that these approaches are quite complementary amongst each other.

Table 3: Complementarity between different extensions of the `SquaresChnFtrs` strong baseline. Results in MR (lower is better). Improvement in MR percent points. Expected improvement is the direct sum of individual improvements.

Method	Results	Improvement	Expected improvement
`SquaresChnFtrs`	34.81%	-	-
`+DCT`	31.28%	3.53	-
`+SDt` [33]	30.34%	4.47	-
`+2Ped` [35]	29.42%	5.39	-
`+DCT+2Ped`	27.40%	7.41	8.92
`+SDt+2Ped`	26.68%	8.13	9.86
`+DCT+SDt`	25.24%	9.57	8.00
`Katamari-v1`	*22.49%*	12.32	13.39