How Far are We from Solving Pedestrian Detection?

Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang and Bernt Schiele
Max Planck Institute for Informatics
Saarbrücken, Germany
firstname.lastname@mpi-inf.mpg.de

Abstract

Encouraged by the recent progress in pedestrian detection, we investigate the gap between current state-of-the-art methods and the "perfect single frame detector". We enable our analysis by creating a human baseline for pedestrian detection (over the Caltech dataset), and by manually clustering the recurrent errors of a top detector. Our results characterise both localisation and background-versus-foreground errors.

To address localisation errors we study the impact of training annotation noise on the detector performance, and show that we can improve even with a small portion of sanitised training data. To address background/foreground discrimination, we study convnets for pedestrian detection, and discuss which factors affect their performance.

Other than our in-depth analysis, we report top performance on the Caltech dataset, and provide a new sanitised set of training and test annotations.

1. Introduction

Object detection has received great attention during recent years. Pedestrian detection is a canonical sub-problem that remains a popular topic of research due to its diverse applications.

Despite the extensive research on pedestrian detection, recent papers still show significant improvements, suggesting that a saturation point has not yet been reached. In this paper we analyse the gap between the state of the art and a newly created human baseline (section 3.1). The results indicate that there is still a ten fold improvement to be made before reaching human performance. We aim to investigate which factors will help close this gap.

We analyse failure cases of top performing pedestrian detectors and diagnose what should be changed to further push performance. We show several different analysis, including human inspection, automated analysis of problem cases (e.g. blur, contrast), and oracle experiments (section 3.2). Our results indicate that localisation is an important source of high confidence false positives. We address this aspect by improving the training set alignment quality, both by manually sanitising the Caltech training annotations and via algorithmic means for the remaining training samples (sections 3.3 and 4.1).

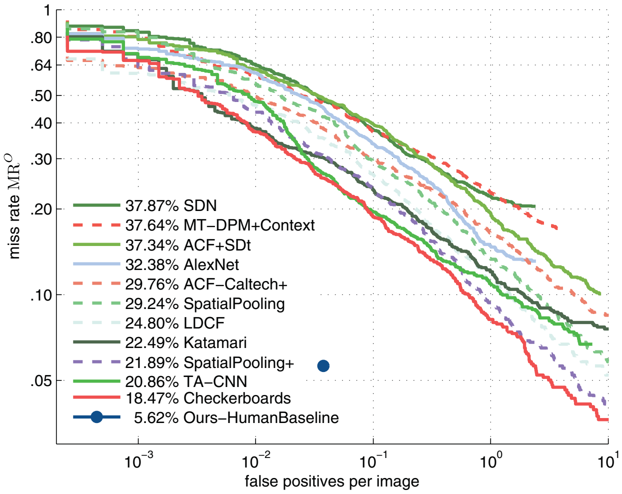


Figure 1: Overview of the top results on the Caltech-USA pedestrian benchmark (CVPR2015 snapshot). At $\sim 95\%$ recall, state-of-the-art detectors make ten times more errors than the human baseline.

To address background versus foreground discrimination, we study convnets for pedestrian detection, and discuss which factors affect their performance (section 4.2).

1.1. Related work

In the last years, diverse efforts have been made to improve the performance of pedestrian detection. Following the success of integral channel feature detector (ICF) [6, 5], many variants [21, 23, 15, 17, 22] were proposed and showed significant improvement. A recent review of pedestrian detection [3] concludes that improved features have been driving performance and are likely to continue doing so. It also shows that optical flow [18] and context information [16] are complementary to image features and can further boost detection accuracy.

By fine-tuning a model pre-trained on external data convolution neural networks (convnets) have also reached state-of-the-art performance [14, 19].

Most of the recent papers focus on introducing novelty and better results, but neglect the analysis of the resulting system. Some analysis work can be found for general object detection [1, 13]; in contrast, in the field of pedestrian detection, this kind of analysis is rarely done. In 2008, [20] provided a failure analysis on the INRIA dataset, which is relatively small. The best method considered in the 2012 Caltech dataset survey [7] had 10× more false positives at 20 % recall than the methods considered here, and no method had reached the 95 % mark.

Since pedestrian detection has improved significantly in recent years, a deeper and more comprehensive analysis based on state-of-the-art detectors is valuable to provide better understanding as to where future efforts would best be invested.

1.2. Contributions

Our key contributions are as follows:

(a) We provide a detailed analysis of a state-of-the-art pedestrian detector, providing insights into failure cases.

(b) We provide a human baseline for the Caltech Pedestrian Benchmark; as well as a sanitised version of the annotations to serve as new, high quality ground truth for the training and test sets of the benchmark. This data is public[1].

(c) We analyse the effects of training data quality. More specifically we quantify how much better alignment and fewer annotation mistakes can improve performance.

(d) Using the insights of the analysis, we explore variants of top performing methods: filtered channel feature detector [23] and R-CNN detector [12, 14], and show improvements over the baselines.

2. Preliminaries

Before delving into our analysis, let us describe the datasets in use, their metrics, and our baseline detector.

2.1. Caltech-USA pedestrian detection benchmark

Amongst existing pedestrian datasets [4, 9, 8], KITTI [10] and Caltech-USA are currently the most popular ones. In this work we focus on the Caltech-USA benchmark [7] which consists of 2.5 hours of 30Hz video recorded from a vehicle traversing the streets of Los Angeles, USA. The video annotations amount to a total of 350 000 bounding boxes covering ∼ 2 300 unique pedestrians. Detection methods are evaluated on a test set consisting of 4 024 frames. The provided evaluation toolbox generates plots for different subsets of the test set based on annotation size, occlusion level and aspect ratio. The established procedure for training is to use every 30th video frame which results in a total of 4 250 frames with ∼1 600 pedestrian cutouts. More recently, methods which can leverage more data for training have resorted to a finer sampling of the videos [15, 23], yielding up to 10× as much data for training than the standard "1×" setting.

Filter type	MR$^O_{-2}$
ACF [5]	44.2
SCF [3]	34.8
LDCF [15]	24.8
RotatedFilters	19.2
Checkerboards	18.5

Table 1: The filter type determines the ICF methods quality.

Base detector	MR$^O_{-2}$	+Context	+Flow
Orig. 2Ped [16]	48	~5pp	/
Orig. SDt [18]	45	/	8pp
SCF [3]	35	5pp	4pp
Checkerboards	19	~0	1pp

Table 2: Detection quality gain of adding context [16] and optical flow [18], as function of the base detector.

MR^O, MR^N In the standard Caltech evaluation [7] the miss rate (MR) is averaged over the low precision range of $[10^{-2}, 10^0]$ FPPI (false positives per image). This metric does not reflect well improvements in localisation errors (lowest FPPI range). Aiming for a more complete evaluation, we extend the evaluation FPPI range from traditional $[10^{-2}, 10^0]$ to $[10^{-4}, 10^0]$, we denote these MR^O_{-2} and MR^O_{-4}. O stands for "original annotations". In section 3.3 we introduce new annotations, and mark evaluations done there as MR^N_{-2} and MR^N_{-4}. We expect the MR_{-4} metric to become more important as detectors get stronger.

2.2. Filtered channel feature detectors

For the analysis in this paper we consider all methods published on the Caltech Pedestrian benchmark, up to the last major conference (CVPR2015). As shown in figure 1, the best method at the time is `Checkerboards`, and most of the top performing methods are of its same family.

The `Checkerboards` detector [23] is a generalisation of the Integral Channels Feature detector (ICF) [6], which filters the HOG+LUV feature channels before feeding them into a boosted decision forest.

We compare the performance of several detectors from the ICF family in table 1, where we can see a big improvement from 44.2% to 18.5% MR^O_{-2} by introducing filters over the feature channels and optimising the filter bank.

Current top performing convnets methods [14, 19] are sensitive to the underlying detection proposals, thus we first focus on the proposals by optimising the filtered channel feature detectors (more on convnets in section 4.2).

Rotated filters For the experiments involving training new models (in section 4.1) we use our own re-implementation of `Checkerboards` [23], based on the `LDCF` [15] codebase. To improve the training time we decrease the number of filters from 61 in the original `Checkerboards` down to 9 filters. Our so-called `RotatedFilters` are a simplified version of `LDCF`, applied at three different scales (in the same spirit as `SquaresChnFtrs` (SCF) [3]). More details on the filters are given in the supplementary material.

[1] http://www.mpi-inf.mpg.de/pedestrian_detection_cvpr16

As shown in table 1, RotatedFilters are significantly better than the original LDCF, and only 1 pp (percent point) worse than Checkerboards, yet run 6× faster at training and test time.

Additional cues The review [3] showed that context and optical flow information can help improve detections. However, as the detector quality improves (table 1) the returns obtained from these additional cues erodes (table 2). Without re-engineering such cues, gains in detection must come from the core detector.

3. Analysing the state of the art

In this section we estimate a lower bound on the remaining progress available, analyse the mistakes of current pedestrian detectors, and propose new annotations to better measure future progress.

3.1. Are we reaching saturation?

Progress on pedestrian detection has been showing no sign of slowing in recent years [23, 19, 3], despite recent impressive gains in performance. How much progress can still be expected on current benchmarks? To answer this question, we propose to use a human baseline as lower bound. We asked domain experts to manually "detect" pedestrians in the Caltech-USA test set; machine detection algorithms should be able to at least reach human performance and, eventually, superhuman performance.

Human baseline protocol To ensure a fair comparison with existing detectors, most of which operate at test time over a single image, we focus on the single frame monocular detection setting. Frames are presented to annotators in random order, and without access to surrounding frames from the source videos. Annotators have to rely on pedestrian appearance and single-frame context rather than (long-term) motion cues.

The Caltech benchmark normalises the aspect ratio of all detection boxes [7]. Thus our human annotations are done by drawing a line from the top of the head to the point between both feet. A bounding box is then automatically generated such that its centre coincides with the centre point of the manually-drawn axis, see illustration in figure 2. This procedure ensures the box is well centred on the subject (which is hard to achieve when marking a bounding box).

To check for consistency among the two annotators, we produced duplicate annotations for a subset of the test images ($\sim$ 10%), and evaluated these separately. With a Intersection over Union (IoU) $\geq$ 0.5 matching criterion, the results were identical up to a single bounding box.

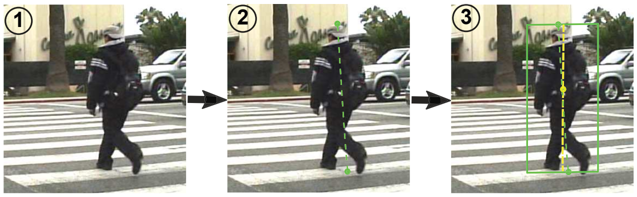


Figure 2: Illustration of bounding box generation for human baseline. The annotator only needs to draw a line from the top of the head to the central point between both feet, a tight bounding box is then automatically generated.

Conclusion In figure 3, we compare our human baseline with other top performing methods on different subsets of the test data . We find that the human baseline widely outperforms state-of-the-art detectors in all settings[2], indicating that there is still room for improvement for automatic methods.

3.2. Failure analysis

Since there is room to grow for existing detectors, one might want to know: when do they fail? In this section we analyse detection mistakes of Checkerboards, which obtains top performance on most subsets of the test set (see figure 3). Since most top methods of figure 1 are of the ICF family, we expect a similar behaviour for them too. Methods using convnets with proposals based on ICF detectors will also be affected.

3.2.1 Error sources

There are two types of errors a detector can do: false positives (detections on background or poorly localised detections) and false negatives (low-scoring or missing pedestrian detections). In this analysis, we look into false positive and false negative detections at 0.1 false positives per image (FPPI, 1 false positive every 10 images), and manually cluster them (one to one mapping) into visually distinctive groups. A total of 402 false positive and 148 false negative detections (missing recall) are categorised by error type.

False positives After inspection, we end up having all false positives clustered in eleven categories, shown in figure 4a. These categories fall into three groups: localisation, background, and annotation errors. Localisation errors are defined as false detections overlapping with ground truth bounding boxes, while background errors have zero overlap with any ground truth annotation.
Background errors are the most common ones, mainly vertical structures (e.g. figure 5b), tree leaves, and traffic lights. This indicates that the detectors need to be extended with a better *vertical context*, providing visibility over larger structures and a rough height estimate.
Localisation errors are dominated by double detections

[2] Except for IoU $\geq$ 0.8. This is due to issues with the ground truth, discussed in section 3.3.

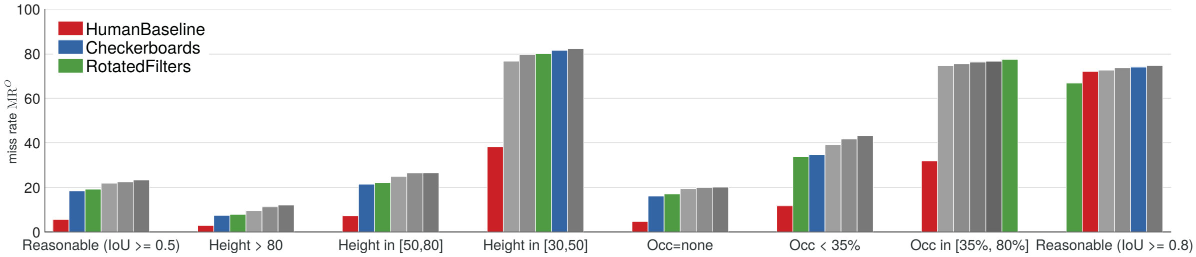
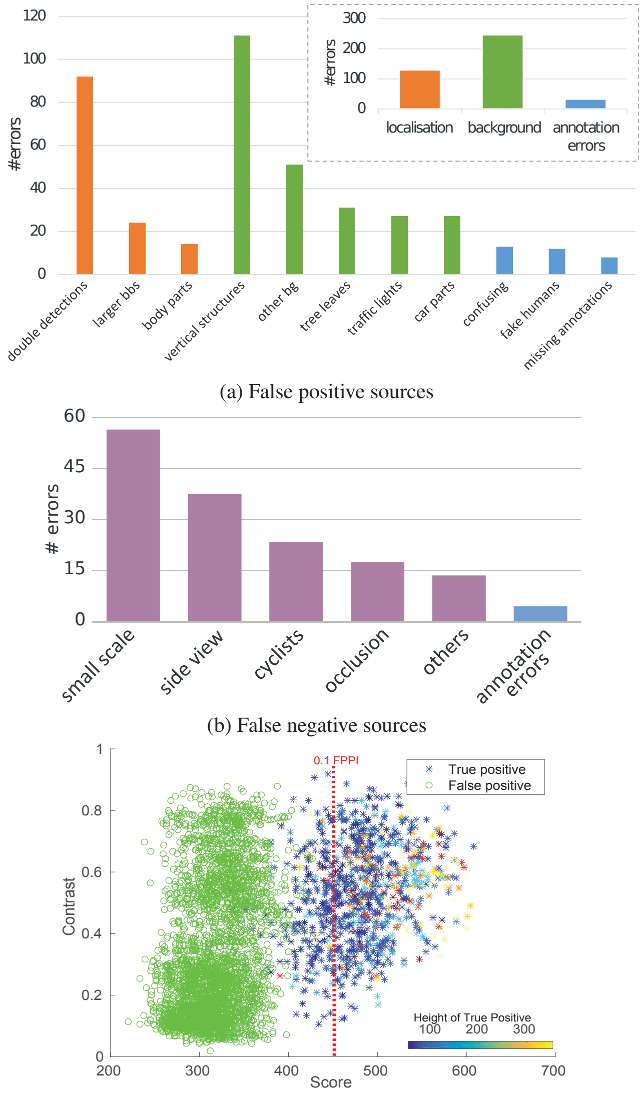


Figure 3: Detection quality (log-average miss rate) for different test set subsets. Each group shows the human baseline, the `Checkerboards` [23] and `RotatedFilters` detectors, as well as the next top three (unspecified) methods (different for each setting). The corresponding curves are provided in the supplementary material.

(high scoring detections covering the same person, e.g. figure 5a). This indicates that improved detectors need to have more localised responses (peakier score maps) and/or a different non-maxima suppression strategy. In sections 3.3 and 4.1 we explore how to improve the detector localisation.
The annotation errors are mainly missing ignore regions, and a few missing person annotations. In section 3.3 we revisit the Caltech annotations.

False negatives Our clustering results in figure 4b show the well known difficulty of detecting small and occluded objects. We hypothesise that low scoring side-view persons and cyclists may be due to a dataset bias, i.e. these cases are under-represented in the training set (most persons are non-cyclist walking on the side-walk, parallel to the car). Augmenting the training set with external images for these cases might be an effective strategy.
To understand better the issue with small pedestrians, we measure size, blur, and contrast for each (true or false) detection. We observed that small persons are commonly saturated (over or under exposed) and blurry, and thus hypothesised that this might be an underlying factor for weak detection (other than simply having fewer pixels to make the decision). Our results indicate however that this is not the case. As figure 4c illustrates, there seems to be no correlation between low detection score and low contrast. This also holds for the blur case, detailed plots are in the supplementary material. We conclude that the small number of pixels is the true source of difficulty. Improving small objects detection thus need to rely on making proper use of all pixels available, both inside the window and in the surrounding context, as well as across time.

Conclusion Our analysis shows that false positive errors have well defined sources that can be specifically targeted with the strategies suggested above. A fraction of the false negatives are also addressable, albeit the small and occluded pedestrians remain a (hard and) significant problem.



(a) False positive sources

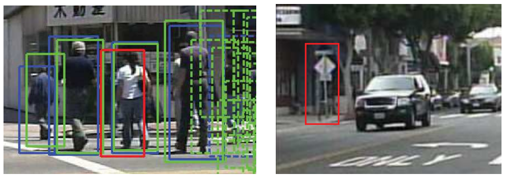
(b) False negative sources

(c) Contrast versus detection score

Figure 4: Errors analysis of `Checkerboards` [23] on the test set.

3.2.2 Oracle test cases

The analysis of section 3.2.1 focused on errors counts. For area-under-the-curve metrics, such as the ones used in Caltech, high-scoring errors matter more than low-scoring ones. In this section we directly measure the impact of localisation and background-vs-foreground errors on the detec-



(a) double detection (b) vertical structure

Figure 5: Example of analysed false positive cases (red box). Additional ones in supplementary material.

tion quality metric (log-average miss-rate) by using oracle test cases.

In the oracle case for localisation, all false positives that overlap with ground truth are ignored for evaluation. In the oracle tests for background-vs-foreground, all false positives that do not overlap with ground truth are ignored.

Figure 6a shows that fixing localisation mistakes improves performance in the low FPPI region; while fixing background mistakes improves results in the high FPPI region. Fixing both types of mistakes results zero errors, even though this is not immediately visible in the double log plot.

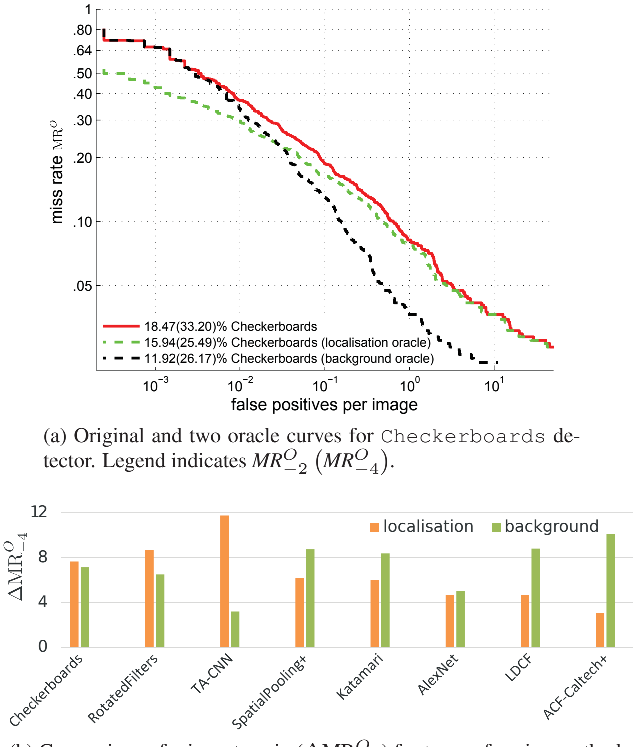
In figure 6b we show the gains to be obtained in MR^O_{-4} terms by fixing localisation or background issues. When comparing the eight top performing methods we find that most methods would boost performance significantly by fixing either problem. Note that due to the log-log nature of the numbers, the sum of localisation and background deltas do not add up to the total miss-rate.

Conclusion For most top performing methods localisation and background-vs-foreground errors have equal impact on the detection quality. They are equally important.

3.3. Improved Caltech-USA annotations

When evaluating our human baseline (and other methods) with a strict IoU ≥ 0.8 we notice in figure 3 that the performance drops. The original annotation protocol is based on interpolating sparse annotations across multiple frames [7], and these sparse annotations are not necessarily located on the evaluated frames. After close inspection we notice that this interpolation generates a systematic offset in the annotations. Humans walk with a natural up and down oscillation that is not modelled by the linear interpolation used, thus in most frames have shifted bounding box annotations. This effect is not noticeable when using the forgiving IoU ≥ 0.5, however such noise in the annotations is a hurdle when aiming to improve object localisation.

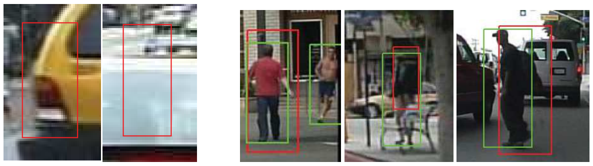
This localisation issues together with the annotation errors detected in section 3.2.1 motivated us to create a new set of improved annotations for the Caltech pedestrians dataset. Our aim is two fold; on one side we want to provide a more accurate evaluation of the state of the art, in particular an evaluation suitable to close the "last 20%" of the problem. On the other side, we want to have training annotations



(a) Original and two oracle curves for `Checkerboards` detector. Legend indicates MR^O_{-2} (MR^O_{-4}).

(b) Comparison of miss-rate gain (ΔMR^O_{-4}) for top performing methods.

Figure 6: Oracle cases evaluation over Caltech test set. Both localisation and background-versus-foreground show important room for improvement.



(a) False annotations (b) Poor alignment

Figure 7: Examples of errors in original annotations. New annotations in green, original ones in red.

and evaluate how much improved annotations lead to better detections. We evaluate this second aspect in section 4.1.

New annotation protocol Our new annotations are done both on the test and training $1\times$ set, and focus on high quality. The annotators are allowed to look at the full video to decide if a person is present or not, they are requested to mark ignore regions in areas covering crowds, human shapes that are not persons (posters, statues, etc.), and in areas that could not be decided as certainly not containing a person. Each person annotation is done by drawing a line from the top of the head to the point between both feet, the same as human baseline. The annotators must hallucinate head and feet if these are not visible. When the person is not fully visible, they must also annotate a rectangle around the largest visible region. This allows to estimate the occlusion level in a similar fashion as the original annotations.

Detector	Training data	Median IoU^O	Median IoU^N
Roerei [2]	INRIA	0.76	*0.84*
RotatedFilters	Orig. 10×	*0.80*	0.77
RotatedFilters	New 10×	0.76	*0.85*

Table 3: Median IoU of true positives for detectors trained on different data, evaluated on original and new Caltech test. Models trained on INRIA align well with our new annotations, confirming that they are more precise than previous ones. Curves for other detectors in the supplement.

The new annotations do share some bounding boxes with the human baseline (when no correction was needed), thus the human baseline cannot be used to do analysis across different IoU thresholds over the new test set.

In summary, our new annotations differ from the human baseline in the following aspects: both training and test sets are annotated, ignore regions and occlusions are also annotated, full video data is used for decision, and multiple revisions of the same image are allowed.

After creating a full independent set of annotations, we consolidated the new annotations by cross-validating with the old annotations. Any correct old annotation not accounted for in the new set, was added too.

Our new annotations correct several types of errors in the existing annotations, such as misalignments (figure 7b), missing annotations (false negatives), false annotations (false positives, figure 7a), and the inconsistent use of "ignore" regions. More examples of "original versus new annotations" provided in the supplementary material, as well as a visualisation software to inspect them frame by frame.

Better alignment In table 3 we show quantitative evidence that our new annotations are at least more precisely localised than the original ones. We summarise the alignment quality of a detector via the median IoU between true positive detections and a given set of annotations. When evaluating with the original annotations ("median IoU^O" column in table 3), only the model trained with original annotations has good localisation. However, when evaluating with the new annotations ("median IoU^N" column) *both* the model trained on INRIA data, and on the new annotations reach high localisation accuracy. This indicates that our new annotations are indeed better aligned, just as INRIA annotations are better aligned than Caltech.

Detailed IoU curves for multiple detectors are provided in the supplementary material. Section 4.1 describes the `RotatedFilters-New10×` entry.

4. Improving the state of the art

In this section we leverage the insights of the analysis, to improve localisation and background-versus-foreground discrimination of our baseline detector.

Detector	Anno. variant	MR$^O_{-2}$	MR$^N_{-2}$
ACF	Original	*36.90*	40.97
	Pruned	36.41	35.62
	New	41.29	*34.33*
RotatedFilters	Original	*28.63*	33.03
	Pruned	23.87	25.91
	New	31.65	*25.74*

Table 4: Effects of different training annotations on detection quality on validation set (1× training set). Italic numbers have matching training and test sets. Both detectors improve on the original annotations, when using the "pruned" variant (see §4.1).

4.1. Impact of training annotations

With new annotations at hand we want to understand what is the impact of annotation quality on detection quality. We will train `ACF` [5] and `RotatedFilters` models (introduced in section 2.2) using different training sets and evaluate on both original and new annotations (i.e. MR^O_{-2}, MR^O_{-4} and MR^N_{-2}, MR^N_{-4}). Note that both detectors are trained via boosting and thus inherently sensitive to annotation noise.

Pruning benefits Table 4 shows results when training with original, new and pruned annotations (using a $5/6+1/6$ training and validation split of the full training set). As expected, models trained on original/new and tested on original/new perform better than training and testing on different annotations. To understand better what the new annotations bring to the table, we build a hybrid set of annotations. Pruned annotations is a mid-point that allows to decouple the effects of removing errors and improving alignment.

Pruned annotations are generated by matching new and original annotations (IoU $\geq$ 0.5), marking as ignore region any original annotation absent in the new ones, and adding any new annotation absent in the original ones.

From original to pruned annotations the main change is removing annotation errors, from pruned to new, the main change is better alignment. From table 4 both `ACF` and `RotatedFilters` benefit from removing annotation errors, even in MR^O_{-2}. This indicates that our new training set is better sanitised than the original one.

We see in MR^N_{-2} that the stronger detector benefits more from better data, and that the largest gain in detection quality comes from removing annotation errors.

Alignment benefits The detectors from the ICF family benefit from training with increased training data [15, 23], using 10× data is better than 1× (see section 2.1). To leverage the 9× remaining data using the new 1× annotations we train a model over the new annotations and use this model

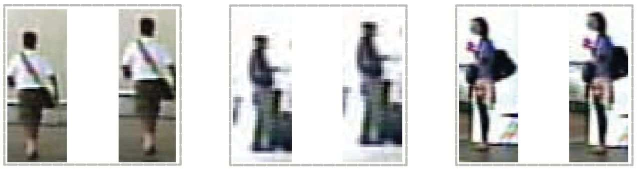


Figure 8: Examples of automatically aligned ground truth annotations. Left/right→ before/after alignment.

1× data	10× data aligned with	MR^O_{-2} (MR^O_{-4})	MR^N_{-2} (MR^N_{-4})
Orig.	Ø	19.20 (34.28)	17.22 (31.65)
Orig.	Orig. 10×	19.16 (32.28)	15.71 (28.13)
Orig.	New 1/2×	16.97 (28.01)	14.54 (25.06)
New	New 1×	16.77 (29.76)	12.96 (22.20)

Table 5: Detection quality of `RotatedFilters` on test set when using different aligned training sets. All models trained with Caltech 10×, composed with different 1 × +9× combinations.

to re-align the original annotations over the 9× portion. Because the new annotations are better aligned, we expect this model to be able to recover slight position and scale errors in the original annotations. Figure 8 shows example results of this process. See supplementary material for details.

Table 5 reports results using the automatic alignment process, and a few degraded cases: using the original 10×, self-aligning the original 10× using a model trained over original 10×, and aligning the original 10× using only a fraction of the new annotations (without replacing the 1× portion). The results indicate that using a detector model to improve overall data alignment is indeed effective, and that better aligned training data leads to better detection quality (both in MR^O and MR^N). This is in line with the analysis of section 3.2. Already using a model trained on 1/2 of the new annotations for alignment, leads to a stronger model than obtained when using original annotations.

We name the `RotatedFilters` model trained using the new annotations and the aligned 9× data, `RotatedFilters-New10×`. This model also reaches high median true positives IoU in table 3, indicating that indeed it obtains more precise detections at test time.

Conclusion Using high quality annotations for training improves the overall detection quality, thanks both to improved alignment and to reduced annotation errors.

4.2. Convnets for pedestrian detection

The results of section 3.2 indicate that there is room for improvement by focusing on the core background versus foreground discrimination task (the "classification part of object detection"). Recent work [14, 19] showed competitive performance with convolutional neural networks (convnets) for pedestrian detection. We include convnets into our analysis, and explore to what extent performance is driven by the quality of the detection proposals.

Test proposals	Proposal	+AlexNet	+VGG	+bbox reg & NMS
ACF [5]	48.0%	28.5%	22.8%	20.8%
SquaresChnFtrs [3]	31.5%	21.2%	15.9%	14.7%
LDCF [15]	23.7%	21.6%	16.0%	13.7%
Rot.Filters	17.2%	21.5%	17.8%	13.8%
Checkerboards [23]	16.1%	21.0%	15.3%	11.1%
Rot.Filters-New10×	12.9%	17.2%	11.7%	10.0%

Table 6: Detection quality of convnets with different proposals. Grey numbers indicate worse results than the input proposals. All numbers are MR^N_{-2} on the Caltech test set.

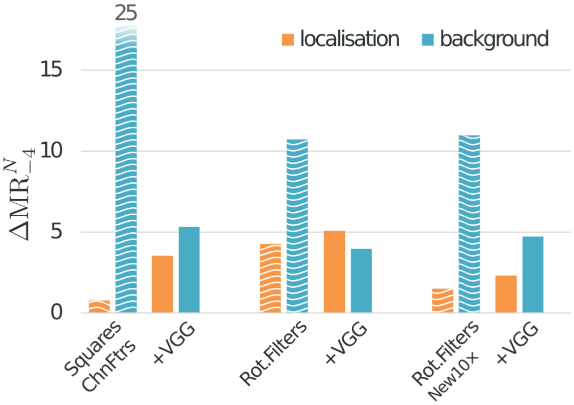


Figure 9: Oracle case analysis of proposals + convnets (after second NMS). Miss-rate gain, ΔMR^O_{-4}. The convnet significantly improves background errors, while slightly increasing localisation ones.

AlexNet and VGG We consider two convnets. 1) The AlexNet from [14], and 2) The VGG16 model from [11]. Both are pre-trained on ImageNet and fine-tuned over Caltech 10× (original annotations) using `SquaresChnFtrs` proposals. Both networks are based on open source, and both are instances of the R-CNN framework [12]. Albeit their training/test time architectures are slightly different (R-CNN versus Fast R-CNN), we expect the result differences to be dominated by their respective discriminative power (VGG16 improves 8 pp in mAP over AlexNet in the Pascal detection task [12]).

Table 6 shows that as the quality of the detection proposals improves, AlexNet fails to provide a consistent gain, eventually worsening the results of our ICF detectors (similar observation in [14]). Similarly VGG provides large gains for weaker proposals, but as the proposals improve, the gain from the convnet re-scoring eventually stalls.

After closer inspection of the resulting curves (see supplementary material), we notice that both AlexNet and VGG push background instances to lower scores, and at the

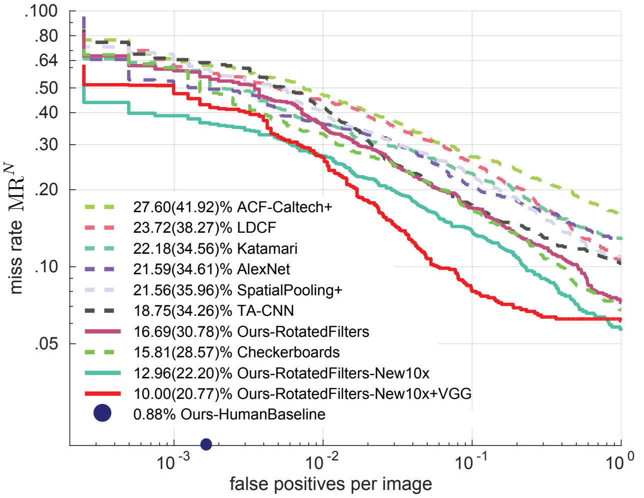


Figure 10: Detection quality on Caltech test set (reasonable subset), evaluated on the new annotations (MR^N_{-2} (MR^N_{-4})). Further results in the supplementary material.

Detector aspect	MR^O_{-2} (MR^O_{-4})	MR^N_{-2} (MR^N_{-4})
`RotatedFilters`	19.20 (34.28)	17.22 (31.65)
+ Alignment §4.1	16.97 (28.01)	14.54 (25.06)
+ New annotations §4.1	16.77 (29.76)	12.96 (22.20)
+ VGG §4.2	16.61 (34.79)	11.74 (28.37)
+ bbox reg & NMS	*14.16* (*28.39*)	*10.00* (*20.77*)
`Checkerboards`	18.47 (33.20)	15.81 (28.57)

Table 7: Step by step improvements from previous best method `Checkerboards` to `RotatedFilters-New10x+VGG`.

same time generate a large number of high scoring false positives. The ICF detectors are able to provide high recall proposals, where false positives around the objects have low scores (see [14, supp. material, fig. 9]), however convnets have difficulties giving low scores to these windows surrounding the true positives. In other words, despite their fine-tuning, the convnet score maps are "blurrier" than the proposal ones. We hypothesise this is an intrinsic limitation of the AlexNet and VGG architectures, due to their internal feature pooling. Obtaining "peakier" responses from a convnet most likely will require using rather different architectures, possibly more similar to the ones used for semantic labelling or boundaries estimation tasks which require pixel-accurate output.

Fortunately, we can compensate for the lack of spatial resolution in the convnet scoring by using bounding box regression. Adding bounding regression over VGG, and applying a second round of non-maximum suppression (first NMS on the proposals, second on the regressed boxes), has the effect of "contracting the score maps". Neighbour proposals that before generated multiple strong false positives, now collapse into a single high scoring detection. We use the usual IoU ≥ 0.5 merging criterion for the second NMS.

The last column of table 6 shows that bounding box regression + NMS is effective at providing an additional gain over the input proposals, even for our best detector `RotatedFilters-New10×`. On the original annotations `RotatedFilters-New10×+VGG` reaches 14.2% MR^O_{-2} , which improves over [14, 19]. Our best performing detector `RotatedFilters-New10×` runs on a 640 × 480 image for ~3.5 seconds, including the ICF sliding window detection and VGG rescoring. Training times are counted 1~2 days for the `RotatedFilters` detector, and 1~2 days for VGG fine-tunning.

Figure 9 repeats the oracle tests of section 3.2.2 over our convnet results. One can see that VGG significantly cuts down the background errors, while at the same time slightly increases the localisation errors.

Conclusion Although convnets have strong results in image classification and general object detection, they seem to have limitations when producing well localised detection scores around small objects. Bounding box regression (and NMS) is a key ingredient to side-step this limitation with current architectures. Even after using a strong convnet, background-versus-foreground remains the main source of errors; suggesting that there is still room for improvement on the raw classification power of the neural network.

5. Summary

In this paper, we make great efforts on analysing the failures for a top-performing detector on Caltech dataset. Via our human baseline we have quantified a lower bound on how much improvement there is to be expected. There is a 10× gap still to be closed. To better measure the next steps in detection progress, we have provided new sanitised Caltech train and test set annotations.

Our failure analysis of a top performing method has shown that most of its mistakes are well characterised. The error characteristics lead to specific suggestions on how to engineer better detectors (mentioned in section 3.2; e.g. data augmentation for person side views, or extending the detector receptive field in the vertical axis).

We have partially addressed some of the issues by measuring the impact of better annotations on localisation accuracy, and by investigating the use of convnets to improve the background to foreground discrimination. Our results indicate that significantly better alignment can be achieved with properly trained ICF detectors, and that, for pedestrian detection, convnet struggle with localisation issues, that can be partially addressed via bounding box regression. Both on original and new annotations, the described detection approach reaches top performance, see progress in table 7.

We hope the insights and data provided in this work will guide the path to close the gap between machines and humans in the pedestrian detection task.

References

[1] P. Agrawal, R. Girshick, and J. Malik. Analyzing the performance of multilayer neural networks for object recognition. In *ECCV*, 2014. 2

[2] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool. Seeking the strongest rigid detector. In *CVPR*, 2013. 6

[3] R. Benenson, M. Omran, J. Hosang, , and B. Schiele. Ten years of pedestrian detection, what have we learned? In *ECCV, CVRSUAD workshop*, 2014. 1, 2, 3, 7

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[5] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *PAMI*, 2014. 1, 2, 6, 7

[6] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009. 1, 2

[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 2012. 2, 3, 5

[8] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *PAMI*, 2009. 2

[9] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, 2008. 2

[10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2

[11] R. Girshick. Fast R-CNN. In *ICCV*, 2015. 7

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2, 7

[13] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *ECCV*, 2012. 2

[14] J. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians. In *CVPR*, 2015. 2, 7, 8

[15] W. Nam, P. Dollár, and J. H. Han. Local decorrelation for improved detection. In *NIPS*, 2014. 1, 2, 6, 7

[16] W. Ouyang and X. Wang. Single-pedestrian detection aided by multi-pedestrian detection. In *CVPR*, 2013. 1, 2

[17] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Strengthening the effectiveness of pedestrian detection with spatially pooled features. In *ECCV*, 2014. 1

[18] D. Park, C. L. Zitnick, D. Ramanan, and P. Dollár. Exploring weak stabilization for motion feature extraction. In *CVPR*, 2013. 1, 2

[19] Y. Tian, P. Luo, X. Wang, and X. Tang. Pedestrian detection aided by deep learning semantic tasks. In *CVPR*, 2015. 2, 3, 7, 8

[20] C. Wojek and B. Schiele. A performance evaluation of single and multi-feature people detection. In *DAGM*, May 2008. 2

[21] S. Zhang, C. Bauckhage, and A. B. Cremers. Informed haar-like features improve pedestrian detection. In *CVPR*, 2014. 1

[22] S. Zhang, C. Bauckhage, D. A. Klein, and A. B. Cremers. Exploring human vision driven features for pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2015. 1

[23] S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *CVPR*, 2015. 1, 2, 3, 4, 6, 7