Neurocomputing 150 (2015) 214–226

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

On the purity of training and testing data for learning: The case of pedestrian detection

Matteo Taiana*, Jacinto Nascimento, Alexandre Bernardino

Institute for System and Robotics – Lisbon, Portugal[1]

A R T I C L E   I N F O

Article history:
Received 31 January 2014
Received in revised form
5 July 2014
Accepted 7 September 2014
Available online 14 October 2014

Keywords:
Sample purity
Pedestrian detection
Machine learning
Partial occlusion
INRIA person data set
Labelling

A B S T R A C T

The training and the evaluation of learning algorithms depend critically on the quality of data samples. We denote as *pure* the samples that identify clearly and without any ambiguity the class of objects of interest. For instance, in pedestrian detection algorithms, we consider as pure samples the ones containing persons who are fully visible and are imaged at a good resolution (larger than the detector window in size). The exclusive use of pure samples entails two kinds of problems. In training, it biases the detector to neglect slightly occluded and small sized samples (which we denote as *impure*), thus reducing its detection rate in a real world application. In testing, it leads to the unfair evaluation and comparison of different detectors since slightly impure samples, when detected, can be accounted for as false positives. In this paper we study how a sensible use of impure samples can benefit both the training and the evaluation of pedestrian detection algorithms. We improve the labelling of one of the most widely used pedestrian data sets (INRIA) taking into account the degree of sample impurity. We observe that including partially occluded pedestrians in the training improves performance, not only on partially visible examples, but also on the fully visible ones. Furthermore, we found that including pedestrians imaged at low resolutions is beneficial for detecting pedestrians in the same range of heights, leaving the performance on pure samples unchanged. However, including samples with too high a grade of impurity degrades the performance, thus a careful balance must be found. The proposed labelling will allow further studies on the role of impure samples in training pedestrian detectors and on devising fairer comparison metrics between different algorithms.

1. Introduction

Machine Learning (ML) is the field of science researching how computers can learn from data. ML has been successfully applied to many areas of knowledge, from medical [1,2] to financial [3,4], to scientific applications in general [5–7]. The results obtained by a ML system, however, depend heavily on the quality of the data used in its training [8,9]. Moreover, the evaluation and comparison of such systems depend on the quality of the data used for their testing [10]. In this paper, we focus on the application of ML to the detection of people in images.

Detecting humans in images is a challenging task that attracts the attention of the scientific community and industry alike. The problem assumes different contours depending on whether the sensor used to capture the images is fixed or mobile, whether the detection is performed on a single image or on a sequence of images, and whether the sensor is a single camera or a richer sensor providing depth information. One further distinction can be made between the methods that do and do not restrain the articulation of the persons. This work concentrates on the detection of pedestrians, i.e., people assuming poses that are common while standing or walking, in images acquired by a mobile camera. Detecting pedestrians is important as it enables the estimation of the presence and the position of humans in the vicinity of a vision sensor. The task is complex mostly because of the high variability that characterizes the pedestrians projections on the camera image plane. The appearance of a pedestrian on the image is influenced by the person's pose, his or her clothing, occlusions, and the atmospheric conditions that contribute to the illumination of the scene. Background clutter also plays a role in making the detection difficult.

The publication of data sets is an important step towards a fair comparison of the performances of Pedestrian Detection (PD) systems, but it is not enough. Standard evaluation code is also needed as different evaluation procedures can lead to discrepancies in the reported performances. Data sets are created not only with the intent of comparing the performance of algorithms, but

* Corresponding author.
E-mail address: mtaiana@isr.ist.utl.pt (M. Taiana).

[1] http://www.isr.ist.utl.pt

http://dx.doi.org/10.1016/j.neucom.2014.09.055
0925-2312/© 2014 Elsevier B.V. All rights reserved.

also with the goals of exposing the limitations of contemporary algorithms and stimulating advances in the state of the art. As such, data sets have a limited life span: as the understanding of the problem by the scientific community grows, hurdles are conquered and data sets become obsolete.

The missed detection rate for the INRIA data set [11] at 0.1 False Positives Per Image (FPPI) has dropped from around 50% to around 20% since its publication (see [12]). There is still room for improvement, which explains why that data set is still widely used as a benchmark [13–16]. The same data is also very popular for training: 13 out of 16 algorithms reviewed in [12] are trained on it. The labelling, like in many other PD data sets, consists of rectangular bounding boxes each one of which tightly enclosing a person. For the purposes of this work, we choose to enrich and extend the labelling of the INRIA person data set.

This paper is an extension of our work published in [17], in which we proposed a new labelling for the INRIA *test set* to improve the comparison between different algorithms. In this work we present a new labelling for the INRIA *training set* and show through experimental results the importance of the correct use of pure and impure samples both in the training and in the testing phase. The labelling was conducted following the method proposed in [12]. The proposed annotation is available on the authors' website.[2] We argue that the new test set labelling leads to a better evaluation of PD algorithms, while the new training set labelling enables researchers to analyse the impact of pedestrian height and visibility during training on the detection performance. In this paper we restate, for continuity of exposition, the contributions described in [17] and build on them to present new results. In the section on the evaluation protocol we show that a fair evaluation of detectors with the original labelling of the INRIA test set requires the use of a minimum resolution limit: since only pedestrians taller than 90 pixels are systematically labelled, the evaluation should ignore detections shorter than that limit. Furthermore we show that using the proposed labelling for testing produces a more truthful evaluation of the detectors. The contributions specific to this paper are: the introduction of the notion of sample purity, the elaboration of a new labelling for the training set, the results of Experiment 1, confirming that visibility plays an important role for detectability, the results of Experiment 2, showing that it is worth to include partially occluded pedestrians in the training set, even when testing on fully visible pedestrians, and the results of Experiment 3, showing that it is important to have “short” examples in the training set when the goal is to detect “short” pedestrians.

The remainder of the paper is organized as follows. In Section 2 we introduce the reader to the PD problem. In Section 3 we detail how annotations for data sets are usually compiled, while in Section 4 we describe the principles that guided the proposed labelling. In Section 5 we describe the PD used in this work and in Section 6 we define the evaluation protocol used in the experiments. We relate results in Section 7 and draw conclusions in Section 8.

2. Related work

Advances in Pedestrian Detection (PD) stem mostly from research in the areas of visual feature extraction and Machine Learning, the most common classifiers being based either on AdaBoost [18] or Support Vector Machines [19]. Seminal work in PD was presented in [20,21]. The authors of [22] introduced Integral Images for faster feature computation, AdaBoost for combining many weak classifiers into a strong classifier and a Cascaded Detector for speeding the detection up. That work focused on the recognition of frontal faces and used Haar-like features, which failed to perform as well in the person detection task. The architecture, nonetheless, became very popular for PD algorithms. Dense features, computed on a regular grid over the image, have been very successful. One example of such features, which is ubiquitously used in detection, is the Histogram of Oriented Gradients (HOG). Introduced in [11] and reminiscent of SIFT [23], it represents gradient information in a way that enables robust classification. A recent trend is that of combining multiple features: the Integral Channel Features [24] exploit 10 channels of information based on colour and gradient. The authors of [25] combine Gradient Histograms, Local Binary Patterns (to exploit texture information), Colour Self Similarity (second order statistics of colour) and Histograms of Flow (to exploit movement information). Another line of work concentrates on reducing the detection time, see for instance the Fastest Pedestrian Detector in the West (FPDW) algorithm [26]. One dualism in the literature contrasts monolithic detectors (see [11,13,16]), which compute features at fixed locations on the detection window, to part-based detectors (see [27,28]), which explicitly model the articulation of the human body and use the features where the limbs are estimated to be.

Comparing the performance of PD systems is a fairly complex matter. Many data sets have been published over the years. A first notable example is the MIT pedestrians data set [20], introduced in 1997. It includes frontal and rear views of pedestrian and only positive windows, i.e., fixed-size rectangular images designed to contain a person. The INRIA person data set [11] was introduced by Dalal and Triggs in 2005, it is divided in training set and test set and it provides both positive and negative examples. The ETH pedestrians data set [29] was introduced in 2007. It was recorded with a mobile platform moving along a sidewalk, equipped with a stereo camera. It presents a scenario typical for a mobile robot. The TUD-MotionPairs/TUD-Brussels data set [30] (TUD) and the Caltech pedestrian data set [12] were introduced in 2009 and contain sequences of images taken in automotive scenarios. The size of the data sets has grown over time, from 924 positive examples (MIT data set) to 350 000 labels over 250 000 images (Caltech data set). Each data set can be characterized in a number of ways, one important parameter being the range of sizes of the annotated pedestrians. Most PD algorithms output detections in a selected range of sizes, in order to perform a fair evaluation it is important that such ranges coincide.

The code used to evaluate the performance of a detector on a data set can considerably influence the results. Many parameters, such as the number of classes of labels used for annotating the data and the amount of padding on the candidate images, can influence the reported results. A solution for this problem is to use the same evaluation code on each algorithm. Dollár provides such a code[3] together with a collection of data sets and the detections obtained running several state-of-the-art detectors on such data sets. We adopt that evaluation code and describe its principles in Section 4.

3. Labelling strategies

The purpose of the labelling of a data set for Pedestrian Detection (PD) is twofold. First, the annotation of the training set enables the extraction of the positive and negative examples for training the detector. Second, the annotations of the validation and

[2] Proposed annotation http://users.isr.ist.utl.pt/~mtaiana/data.html

[3] Caltech Pedestrian Detection evaluation code www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/DollarEvaluationCode

test sets are used during evaluation to determine which detections are correct, corresponding to a pedestrian.

Most PD algorithms define the ground truth (GT) labelling and the detections in terms of a collection of rectangles on the images. Such rectangles are known respectively as GT and detection “Bounding Boxes”, in short, “BB's”. Each detection BB is associated to a confidence value and is meant to enclose exactly one pedestrian.

Training labels are used in the training of a PD algorithm. The positive BB's are cropped from the positive training image set and scaled to fit the detection window size. Negative samples are chosen by randomly sampling the negative training images (or the parts of the training images where no people appear) with BB's of the same aspect ratio as the positive ones. They subsequently undergo the same scaling as the positive samples do.

Test labels are used during the evaluation of the performance of a PD algorithm. Evaluating such performance on one image consists in matching detection and GT BB's and counting the occurrences of the result of the matching process. Two BB's (one detection and one GT label) are said to match if the area of intersection of the two rectangles is larger than half of the area of their union (Pascal VOC criterion, see [31]). The possible outcomes of the matching process are: True Positive (TP) when one GT BB matches one detection BB (and so one pedestrian is correctly detected), False Positive (FP) when a detection does not match any GT BB, and Missed Detection (MD) when a GT BB does not match any detection. Each GT BB can match at most one detection BB. In case there be more detections potentially matching one GT BB, the conflict can be solved by greedily assigning the detection with the highest confidence to the match, leaving the others unmatched.

The original labelling of the INRIA data set follows closely the general description. Each person is labelled with a rectangular BB. Only one label is possible: “UprightPerson”, which includes both pedestrians and people riding a bicycle (this stems from the automotive applications of PD). Sitting people are not included in the positive class. No information is present on the amount of visibility each person is imaged under.

The INRIA data set was designed in 2005 to support PD research. Since then, Pedestrian Detectors have improved dramatically and, as a result, the original labelling is now starting to show its limitations. A fair assessment of the performance of detectors on the INRIA data set is hindered by three factors: first, many persons appearing in the images are not labelled, second, there is no class label for the regions of the images that are ambiguous or difficult to be classified even by a person and thus should be ignored during the evaluation and, third, an estimate of the visible part of each person is lacking. We discuss each of these factors in the following paragraphs.

The lack of labelling of some of the people present in the data set (see Fig. 1(a–d)) affects both *training* and *testing*. Regarding the training, the lack of such labels prevents researchers to analyse the impact of what we deem pure and impure training samples on the performance of the detector. Regarding the testing, each detection on one of the unlabelled persons counts as a FP, instead of as a TP. So optimizing a detector using this labelling can lead to the undesirable effect of detecting less small and occluded people. Current state-of-the-art algorithms can detect at least some of the partially occluded and smaller pedestrians that are not marked in the original labelling. Their performance is thus under-reported (see Fig. 2 for an example of how the performance of the FPDW algorithm [26] is affected). People who have parts of their bodies outside the image boundaries are also not labelled, leading to a similar phenomenon. It is important to notice that the spurious FP's originated by the unlabelled persons tend to assume high confidence values, so they have a big impact on some regions of the performance curves of the detectors (see Section 6).

There are, moreover, image patches for which it is difficult to decide whether they should be labelled as a person or not. Such cases include the appearance on the image of a mannequin, of photographs of people, of reflections of people. It is not clear whether an algorithm that generates a detection on one of such image areas should be rewarded or penalized: this decision is very application-dependent. Only some of such occurrences are marked as “person” in the original labelling, both in the training and the test set, introducing noise in the evaluation process (see Fig. 1(e–h)).

Finally, in the original labelling there is no information on the amount of visibility each person is imaged under. Such information is not needed for a simple training or test of a PD algorithm, but it is instrumental to assess the effect of the pure and impure fraction of the data on the detection performance.

Most of the pedestrians marked in the original labelling are fully visible and larger than the size of the detection window used

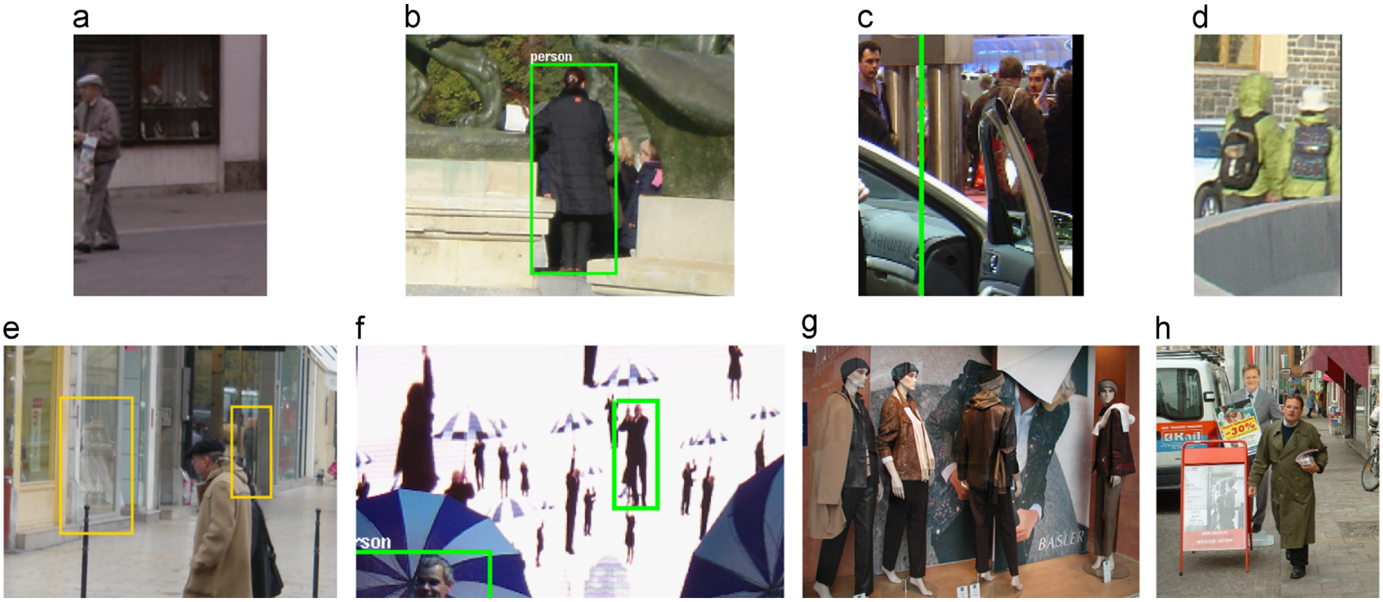

Fig. 1. Details from the INRIA test set highlighting some limitations. (a–d) Unlabelled persons. (e–h) Ambiguous cases. (e) Reflections of persons on a shop window, not labelled. (f) Some persons drawn on a wall, only one of them is labelled. (g) Some mannequins, all labelled. (h) A poster depicting a man, not labelled.

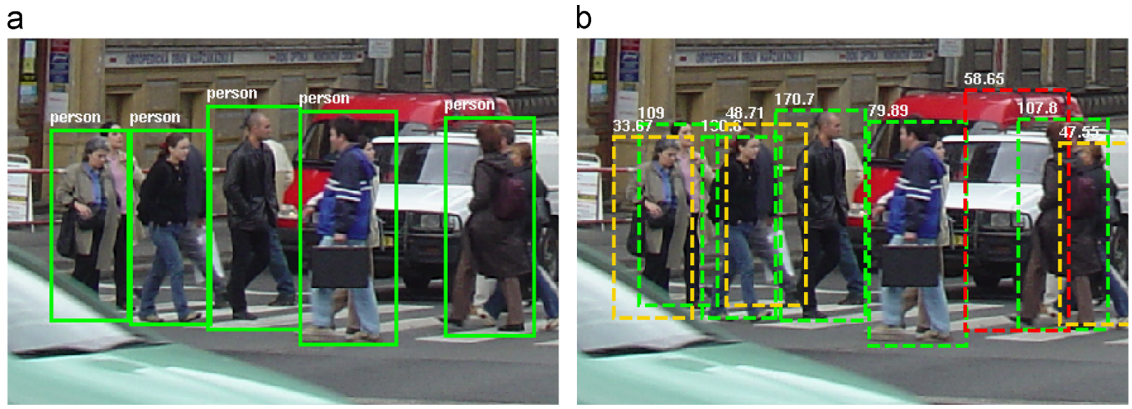

Fig. 2. The influence of labelling in the presence of mutual occlusion on the evaluation. (a) A part of image 20 of the INRIA test set showing the original labelling: only 5 persons out of 11 are marked. Some partially occluded persons are merged in the annotation with a visible one. (b) The classification of the detections produced by FPDW [26] in TP's (green), FP's (red) and FP's which significantly overlap with an unlabelled person (yellow) and thus should be considered TP's. In the whole test set, 26 out of 292 FP's ascribed to FPDW significantly overlap with an unlabelled person. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

in the algorithm introduced with the data set (96 pixels), making them "pure" for our purposes. A small fraction of the labelled pedestrians, though, are imaged under a certain degree of occlusion or are shorter than the detection window, making them "impure". The labelling, thus, results to be a mixture of pure and impure examples in unknown proportions. In this work we extend the labelling to include all, within reason, visible pedestrians and enrich it with the visibility information, allowing experiments to be run with training and test sets characterized by different degrees of purity.

4. Proposed method

We propose a new annotation for the data set in which we label all the pedestrians with heights greater than 25 pixels, we associate to each person the estimate of the extent of his/her visible part and mark ambiguous cases (see Fig. 1(e–h)) as such. The labelling was performed manually by one of the authors. As in the original annotation, we use rectangular Bounding Boxes (BB's) and we consider both cyclists and pedestrians as belonging to the "Person" class. We base our annotations on the scheme introduced in [12], which consists in labelling individual persons as "Person", large groups of persons for which it is very difficult to label each individual as "People", and ambiguous cases as "Person?". We label the test set according to such scheme. The proposed annotation is available on the authors' website. For the training set, we do not label groups of people and ambiguous cases as we believe such annotations not to be useful for training. In the Caltech evaluation code "People" and "Person?" BB's are merged in the "Ignore" class and treated as one, but we choose to use the two labels considering that in the future the two sets can be treated differently. The "Ignore" class was introduced to acknowledge the fact that there is a grey area at the boundary between the "Person" and the "Non-person" categories and with the insight that both detections and missed detections on an image area marked as "Ignore" should not be penalized. Detections that match an "Ignore" BB's are not counted as True Positives (TP's) nor False Positives (FP's) and "Ignore" BB's which are not matched by any detection are not counted as MD's. The matching between a detection BB and a "Person" BB works exactly as explained in the previous section, while matching a detection BB with an "Ignore" BB only requires that the overlap between the two is greater than half of the area of the detection. Moreover, multiple detections can match the same "Ignore" rectangle. In the evaluation code the Ground Truth (GT) BB's are centred horizontally and transformed to assume an aspect ratio of 0.41 (width/height) prior to matching (see [12] for details).

In this work we aim at determining the impact of pure and impure samples in the training and evaluation of a detection system. We consider pure the BB's enclosing pedestrians who are fully visible and imaged with a height larger than the height of the detection window in use. We deem the remaining BB's enclosing pedestrians as impure. Labelling the training set with the visibility information (the height is implicitly encoded in each BB) enables us to create various training sets with a different ratio between the pure and the impure samples. The proposed training set is filtered each time, controlling the amount of "short" and partially occluded examples used to train the detection system.

Controlling the balance between pure and impure samples during testing is allowed by the evaluation code. The minimum height and the minimum visibility ratio of the Ground Truth (GT) rectangles in the test set are specified as a parameter for the evaluation, so that all the BB's that do not match the criterion are set to "Ignore".

5. Pedestrian detectors

5.1. *Our implementation of a pedestrian detector*

We implemented a version of the Fastest Pedestrian Detector in the West (FPDW) algorithm [26]. The detector is based on a structure common to most of the pedestrian detectors in the state of the art: it combines a Machine Learning-based window classifier, the sliding window approach, image pyramids and Non-Maximum Suppression.

The fundamental block of the detector is the window classifier, which takes as input one image window of a specific size and evaluates whether it contains a person of the corresponding height. In our case the classifier is based on AdaBoost in the variant of Soft Cascades [32,33]. We use 1000 level-2 trees as weak classifiers. The output of the classifier is a real value expressing the confidence on the presence of a person in the window at hand. The sliding window approach consists in applying the window classifier on a grid of locations on one image, thus obtaining a set of confidence values. This technique allows for the detection of fixed-size pedestrians over one image, and, in order to succeed in multi-scale detection, it must be combined with image pyramids. Running the detection window on each layer of the pyramid allows for the detection of pedestrians of different sizes, but can give rise to multiple detections for a single pedestrian. Non-Maximum Suppression techniques are used with the intent of merging the positive confidence values originated by the same pedestrian, thus obtaining a detection system that returns only one detection for each pedestrian appearing in the image. As features, we use 30 000 Integral Channel Features (see [24]), we compute the Integral Channels using publicly available code by the author.[4] We train the detector on the INRIA pedestrians training set with the original labelling. We use 4 epochs of bootstrap for mining hard negative examples. We use our implementation of FPDW in experiment 1.

5.2. *The aggregated channel features detector*

The Aggregated Channel Features detector (ACF) [34] is a recent development of FPDW. In ACF the authors substitute the rectangular features of FPDW with square ones and apply additional smoothing. This removes the need for Integral Images, making the computation of a feature, and thus the whole detection process, faster. One important characteristic of the ACF code is that of using a fast implementation of AdaBoost for the training: a randomized subset of features is used for learning each node (1/16 of the total features), the values of the features are discretized to 256 bins, thus reducing the number of split points to evaluate and, finally, the number of weak classifiers is increased after each bootstrap round. Such design choices allow for a much quicker training time compared to our implementation of FPDW, but have the drawback of producing detectors with a higher variability in performance. We choose to use ACF and repeat the training 10 times for each evaluated condition, using different randomizations seeds and averaging the reported performance. We used the authors' implementation of ACF[3] in experiments 2 and 3.

5.3. *Other state-of-the-art detectors*

The Caltech Pedestrian Detection Benchmark[5] provides not only annotated data sets and evaluation code, but also detections obtained with a number of algorithms in the state of the art. This enables the users of the benchmark to evaluate the strengths and the failure modes of the different methods and to compare their performances with those of their own algorithms. The detections provided represent well the diversity of the approaches to PD. They include for instance the detections obtained with the detector based on Histogram of Oriented Gradients by Dalal and Triggs [11], the Discriminatively Trained Deformable Part Models detector by Felzenszwalb et al. [27] and the multi-feature detector by Walk et al. [25].

We use the detections provided in the Caltech benchmark to assess the impact of occlusion on the detectability of the pedestrians, as explained in Section 7.1. Having access to the detections, but not to the code of the aforementioned detectors means that we cannot train them in different conditions. Thus, we cannot use them in the experiments assessing the importance of occlusion and minimum resolution in training.

6. Evaluation protocol

In this section we motivate and describe the evaluation protocol we use in the experiments. First, we define the parameters for running an experiment and the measures used to evaluate its results. Then, we discuss the height ranges relevant for the evaluation and the relationships among them. Finally, we highlight the impact of a more accurate test set labelling on the evaluation of the performance of the detectors, laying the bases for our experiments. The results we present in this section have been published in [17], they are repeated here for completeness of the exposition.

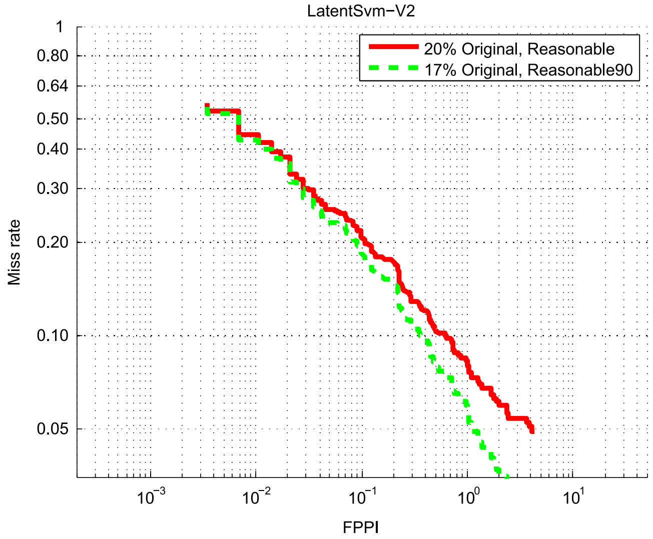

Fig. 3. The reported performance of the LatSvm-V2 algorithm [27] using the original labelling and the "Reasonable" or the "Reasonable90" evaluation modes, in red and green respectively. Performance is summarized in the legend with the Log-Average Miss Rate. Using the "Reasonable90" evaluation mode instead of the "Reasonable" one reports slightly lower missed detection rates. This is expected as some GT annotations that are impossible for the detector to match (given the detections provided in the Caltech benchmark) are accounted for in the "Reasonable" mode and ignored in the "Reasonable90" mode. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

Table 1
The performances of a set of state-of-the-art PD algorithms reported with the original INRIA labelling and the "Reasonable" or the "Reasonable90" evaluation modes. Min. det. height refers to the minimum height for the detections produced by each algorithm, in pixels. The minimum height for the detections provided with the Caltech benchmark lies between 90 and 100 pixels for most of the detectors. This makes setting the lower height limit during evaluation to 90 pixels more sensible than the default 50 pixels of the "Reasonable" mode. The Miss Detection rate at 10^0 FPPI is reported to be lower when testing with the "Reasonable90" evaluation mode than with the "Reasonable" one. This is due to the latter containing GT labels with heights under 90 pixels, which are impossible to detect correctly given the detections distributed with the Caltech benchmark. See Fig. 3 for a graphical comparison of the performance of LatSvm-V2 in the two cases.

Algorithm		Min. det. height	MD at 10^0FPPI		
			Reasonable (MHTE=50)	Reasonable90 (MHTE=90)	Difference
FtrMine	[35]	100.0	0.340	0.324	−0.016
LatSvm-V1	[36]	79.0	0.175	0.159	−0.015
HOG	[11]	100.0	0.231	0.215	−0.015
HikSvm	[37]	100.0	0.221	0.207	−0.014
PLS	[38]	100.0	0.226	0.212	−0.014
HogLbp	[39]	96.0	0.190	0.173	−0.017
FeatSynth	[40]	100.0	0.109	0.089	−0.019
MultiFtr+CSS	[25]	93.7	0.109	0.093	−0.016
FPDW	[26]	100.0	0.093	0.075	−0.018
ChnFtrs	[24]	100.0	0.087	0.072	−0.015
LatSvm-V2	[27]	*91.3*	*0.081*	*0.058*	*−0.024*
Our FPDW		95.6	0.093	0.081	−0.013
CrossTalk	[13]	99.2	0.098	0.079	−0.020
Mean					−0.017

[4] Piotr's Image and Video Matlab Toolbox (PMT) http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html

[5] Caltech Pedestrian Detection Benchmark http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

6.1. Parameters and evaluation variables for an experiment

Performing a detection experiment consists in choosing one detection algorithm, setting the visibility and height conditions for training and testing and finally running the training and testing of the detector. The variables we set are: the minimum height in training (MHTR) and in testing (MHTE), and the minimum visibility in training (MVTR) and in testing (MVTE).

For evaluating the results of one experiment we use the de facto standard measures of Missed Detection (MD) rate and False

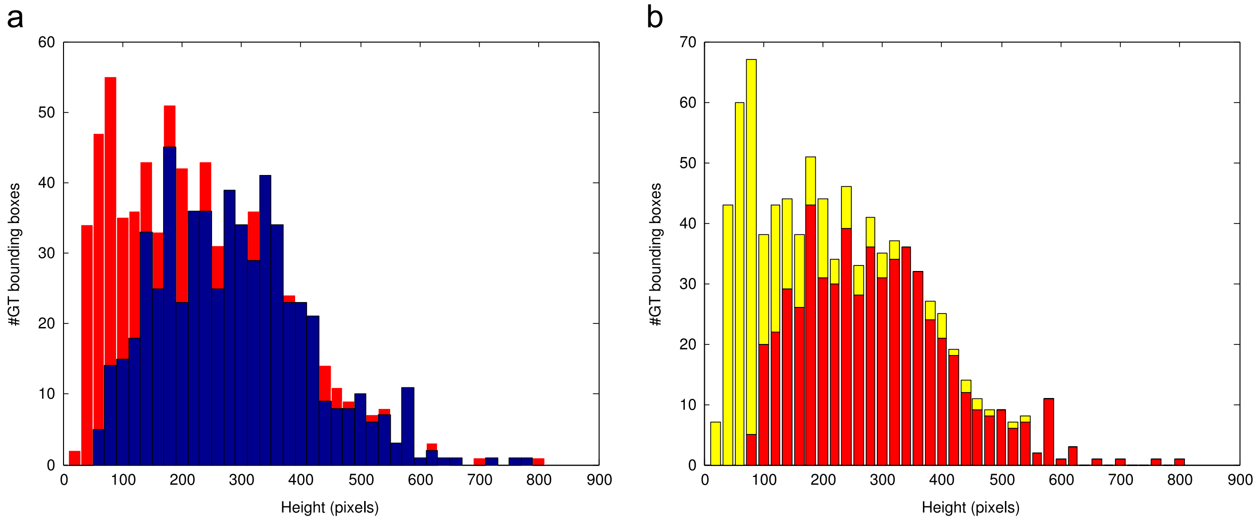

Fig. 4. Characterization of the original and the proposed labellings of the test set. (a) Histograms of the height of “Person” labels for the original (blue) and the proposed labelling (red). The proposed annotation outnumbers the original one, particularly at low heights. (b) Histogram for the proposed labelling and the “Reasonable90” mode, showing the amount of “Person” and “Ignore” BB's in red and yellow, respectively. The number of “Ignore” BB's is considerable and does influence the assessment of the detection performance. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

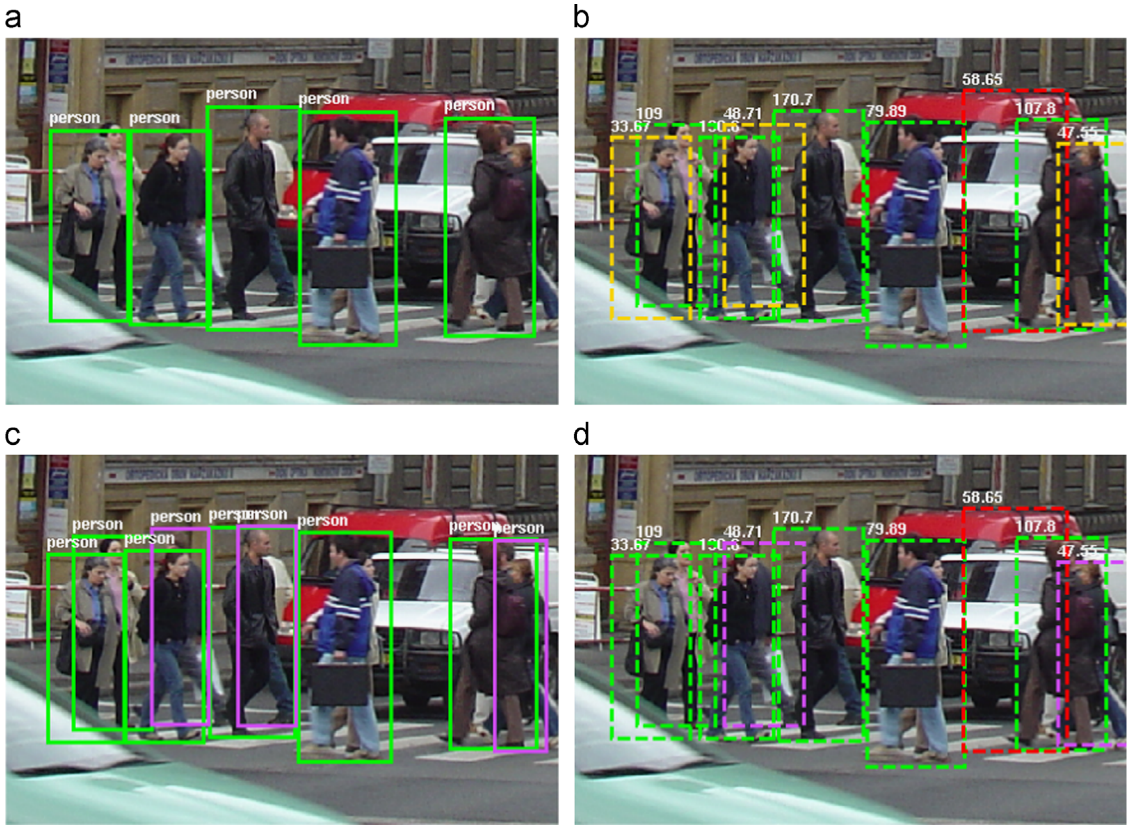

Fig. 5. Comparison of one evaluation performed with the original (a, b) and the proposed (c, d) test set labelling. The image is part of the INRIA test set, the detections were obtained with FPDW. (a) The original GT labels. (b) Evaluation with the original labels: True Positives (TP's) in green, False Positives (FP's) in red and yellow. The yellow FP's significantly overlap with unlabelled persons, hence it is unfair to consider those as errors. (c) The proposed GT labels: pink labels are the ones that the evaluation code set to “Ignore” because of excessive occlusion given the chosen evaluation mode. (d) Evaluation with the proposed labels: TP's in green, FP's in red, ignored matches in pink. Two detections match “Ignore” BB's (dashed pink lines), while one “Ignore” BB is not matched by any detection (not shown in this image). None of these events influence the evaluation of the performance of the detector. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

Positive Per Image (FPPI). MD is defined as the fraction of the positive examples in the test set which goes undetected. FPPI is defined as the total number of False Positive detections (FP's) in the test set, divided by the number of images that constitute it. Both MD's and FP's are detection errors, thus, the lower the values of MD and FPPI, the better the performance of one algorithm. Detectors associate a confidence value (also called a score) to each detection. Varying the value of the threshold on such confidence produces a curve in the MD/FPPI space. The curves are usually presented in log–log plots, see Fig. 3 for one example. Each point on the curve corresponds to an operating point for the PD algorithm. Comparing PD algorithms through curves is not always straightforward, so the performance of one detector is typically summarized by the Log-Average Miss Rate (LAMR), the average miss rate (as computed on the logarithmic FPPI axis) between 10^{-2} and 10^{0} FPPI (see [12] for details).

6.2. Height ranges matching for a fair evaluation

When testing PD algorithms, care should be taken to match several height ranges. First (A), there is the height range of the people imaged in the test set. This is implicitly defined by the images comprised in the test set. Second (B), there is the height range of the GT labels, which is decided by the authors of the data set. Ideally, it should correspond with the first range, but this is not always the case. Third (C), there is the height range of the detections generated by a particular PD system. This is typically a parameter for the algorithms and is set by the experimenters to match the second range we mentioned. Last (D), there is the range of heights taken into account by a specific evaluation mode. This should be selected by the experimenters to be a subset of the intersection of the other three ranges: it is sensible to compare the performance of algorithms for a range of heights for which there are persons in the test set, such persons are annotated and the detectors were allowed to produce detections.

It is common practice to compare algorithms on the original INRIA test set using the “Reasonable” evaluation mode and using the detections provided in the Caltech benchmark for algorithms in the state of the art. The “Reasonable” mode corresponds to setting a minimum height in testing (MHTE) of 50 pixels and a minimum visibility in testing (MVTE) of 0.65 (see [12] for details). The visibility constraint is ignored for the original labelling, since the latter provides no visibility information. In sum, people taller than 50 pixels are considered as “Person”, while the rest of the Bounding Boxes (BB's) are set to “Ignore”. For the vast majority of the detectors whose output is distributed with the Caltech benchmark, the minimum output detection height lies at around 90 pixels (see Table 1, column 2). Thus, when evaluating the performance of those detectors on the INRIA data set with the “Reasonable” mode, the height ranges of the detections (C) and that specified by the evaluation mode (D) do not match. The GT annotations of heights comprised between 50 and 90 pixels can never be matched by the output of most of the detectors. This introduces a bias in the evaluation: the affected detectors can never reach a miss rate of zero.

We define an evaluation mode that matches the range of resolution of the detections, the “Reasonable90” mode, which ignores pedestrians imaged with heights under 90 pixels or with visibilities under 0.65. We compare the reported performance of PD algorithms using the original labelling and selecting either the “Reasonable” or the “Reasonable90” evaluation mode. We argue that “Reasonable90” is a more appropriate test mode for the considered experimental setting since it matches the range of heights of the detections provided in the Caltech benchmark. We evaluate the detections of a number of state-of-the-art algorithms provided with the Caltech benchmark and the detections generated by our implementation of FPDW trained on the original labelling.

We display the Missed Detection rate/False Positive Per Image (FPPI) curves for one representative algorithm, for the two modes, in Fig. 3. Using the “Reasonable90” evaluation mode reports slightly lower missed detection rates, especially at relatively high (10^{0}) FPPI levels (see the results for all the tested algorithms in

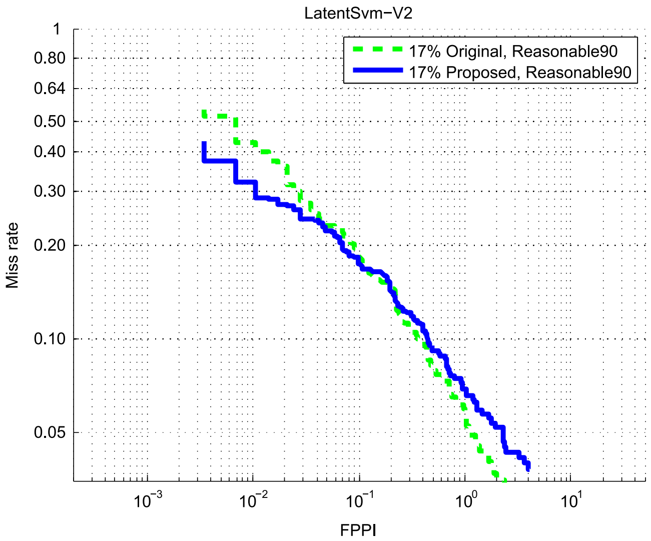

Fig. 6. The reported performance of the LatSvm-V2 algorithm [27] using the original and the proposed labelling, in green and blue respectively. The evaluation mode is “Reasonable90” in both cases. Performance is summarized in the legend with the Log-Average Miss Rate. Using the proposed annotation instead of the original one reports considerably lower Miss Detection rates at low FPPI values. We ascribe this effect to the pedestrians that are unlabelled in the original annotation and have been labelled in the proposed one: the corresponding detections are evaluated as False Positives with the original annotation and as True Positives with the proposed one. Using the proposed annotation also reports slightly increased Miss Detection rates at high FPPI values. We attribute this effect to the introduction in the test set of more occluded pedestrians, which are arguably more difficult to detect than the fully visible ones. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

Table 2
The performances of a set of state-of-the-art PD algorithms reported with the original or the proposed labelling. In both cases the evaluation mode is “Reasonable90”, which correspond to a minimum height in testing (MHTE) of 90 pixels and a minimum visibility in testing (MVTE) of 0.65. LAMR indicates the Log-Average Miss Rate for each algorithm. Using the proposed labelling reports considerably lower Miss Detection rates at 10^{-2} FPPI. This effect is likely due to the pedestrians who are unlabelled in the original annotation and who are labelled in the proposed one. The effect is stronger for the detection algorithms with a better performance (lower LAMR). See Fig. 6 for a visualization of the performance of the LatSvm-V2 detector in the two evaluation cases.

Algorithm		MD at 10^{-2}FPPI			LAMR, proposed labelling (%)
		Original labelling	Proposed labelling	Difference	
FtrMine	[35]	0.918	0.900	−0.019	57
LatSvm-V1	[36]	0.806	0.835	+0.029	43
HOG	[11]	0.744	0.702	−0.042	42
HikSvm	[37]	0.766	0.681	−0.085	39
PLS	[38]	0.674	0.596	−0.078	38
HogLbp	[39]	0.665	0.629	−0.036	35
FeatSynth	[40]	0.754	0.738	−0.015	29
MultiFtr+CSS	[25]	0.469	0.425	−0.044	21
FPDW	[26]	0.576	0.386	−0.189	18
ChnFtrs	[24]	0.581	0.383	−0.198	18
LatSvm-V2	[27]	*0.448*	*0.319*	*−0.129*	*17*
Our FPDW		0.577	0.307	−0.270	16
CrossTalk	[13]	0.511	0.333	−0.178	15
Mean				−0.089	

Table 1, columns 3–5). This result is expected, as passing from “Reasonable” to “Reasonable90” some labels which were impossible for the algorithms to match were removed from the test set. The number of such labels is low since in the original test set only a small fraction of the people with heights between 50 and 90 pixels are labelled. We conclude that the “Reasonable90” mode provides for a fairer evaluation when testing the algorithms on the INRIA test set using the detections from the Caltech benchmark.

A deeper analysis on the INRIA test set reveals that only a fraction of the pedestrians is labelled: the higher the level of occlusion, the more unlikely people are to be labelled, while the smallest pedestrians are unlabelled as a whole. We describe a new labelling for the INRIA test set and assess its impact on the evaluation in the following section.

6.3. The proposed test labelling and its influence on evaluation

With the passing of time the flaws of the labellings become evident and new labellings are needed. Evaluating detection algorithms is typically done by means of annotated test sets. Ideally, the Ground Truth annotation should be perfect. In practice though, labelling a test set is an error-prone process which reflects

Table 3
Summary of the experiments, The goal of each experiment, the algorithms and the training and test labelling used in each experiment are listed. “Others” in “FPDW + others” refers to a set of algorithms whose detections are distributed with the Caltech Pedestrian Detection Benchmark (see Section 5.3 for details and Table 1 for a list of the detectors).

Exp. #	Description	Train. labelling	Test labelling	Det. algorithm
1	Influence of partial occlusion in testing	Original	Proposed	FPDW + others
2	Influence of partial occlusion in training	Proposed	Proposed	ACF
3	Influence of “short” examples in training	Proposed	Proposed	ACF

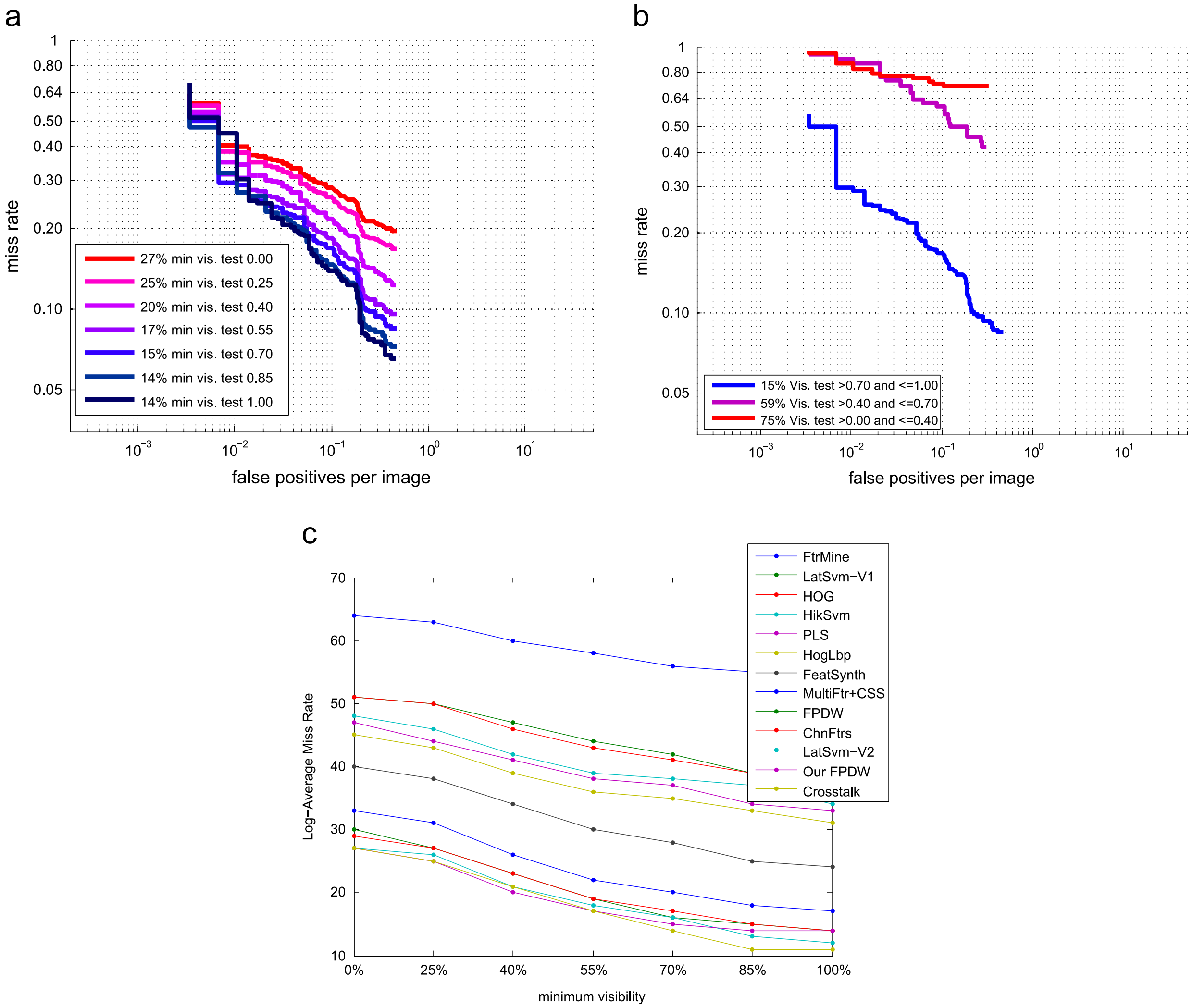

Fig. 7. The effect of partial occlusion in testing on the accuracy of Pedestrian Detectors. (a) Reducing the minimum visibility of the pedestrians in the test set decreases the detection performance of our implementation of FPDW. (b) The detection accuracy of FPDW varies greatly as a function of visibility. The three lines represent results obtained using the good visibility, average visibility and scarce visibility partitions of the test set. In the legends of (a) and (b) the performance of the test combinations is summarized with the Log-Average Miss Rate (LAMR, see text for details). (c) The effect of partial occlusion on the performance of several algorithms in the state of the art. The performance of the detectors is again summarized with the LAMR. A clear relationship holds for all the algorithms: the better the visibility, the lower the LAMR, i.e., the better the performance.

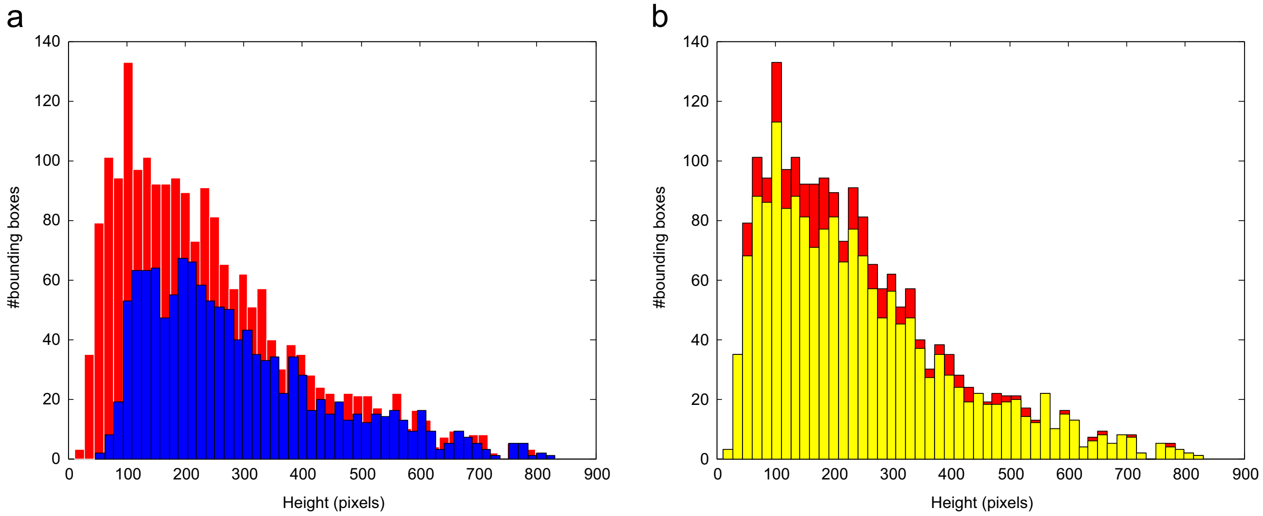

Fig. 8. Characterization of the original and the proposed labellings of the training set. (a) Histograms of the height of "Person" labels for the original (blue) and the proposed labelling (red). The proposed annotation outnumbers the original one, particularly at lower heights. (b) Histogram for the full proposed labelling and for the subset whose BB's are marked with a visibility ratio of at least 0.65, in red and yellow respectively. Most of the new BB's correspond to pedestrians imaged with a good visibility. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

the goal of the labeller. At the time of compilation of the INRIA person data set, the focus was on the detection of high-resolution, fully visible pedestrians. Meanwhile, the performance of Pedestrian Detectors has improved and the focus has shifted to partially occluded and low-resolution pedestrians. The original labelling of the INRIA test set cannot provide a good evaluation for the detections in such conditions. We propose a new annotation that enables the evaluation of the performance of algorithms on pedestrians imaged at low resolutions, and improves the accuracy of the results reported for taller pedestrians.

The proposed annotation for the test set contains a total of 879 labels, 806 of which for "Person" and 73 for "Person?" or "People". In comparison, the original annotation has 589 labels equivalent to "Person". The proposed annotation contains more labels than the original one, especially at low heights, but also at medium heights (see the comparison between the two annotations in Fig. 4a). The fraction of "Ignore" BB's for the new annotation is considerable, Fig. 4b illustrates the amount of labels that are set to "Person" and "Ignore" for the "Reasonable90" evaluation mode, as a function of height. One example of the proposed annotation, together with the effect it has on the evaluation of the detections produced by the FPDW algorithm, can be seen in Fig. 5.

Comparing the performance reported by testing using either the original or the proposed annotations, we observe that the proposed annotation reports better performances for the algorithms at low FPPI values and better differentiates the performance of the various detectors. We use the same set of algorithms mentioned in the previous subsection and the "Reasonable90" mode: $MHTE = 90$ pixels, $MVTE = 0.65$. We display the missed detection/FPPI plot for one representative algorithm and the two annotations, in Fig. 6. Two effects can be seen: the miss rate is minimally higher at high FPPI values for the proposed labelling, we ascribe this to the introduction in the test set of more occluded pedestrians, who make the problem more difficult. The other effect, the most significant one, is the average drop of 8.9% for the missed rates at low FPPI values (10^{-2}) (see the results for all the tested algorithms in Table 2, columns 2–4). We ascribe this to the removal of the spurious False Positives (FP's) generated on top of unlabelled pedestrians. Such FP's tend (correctly) to be associated with high values of confidence, ruining the reported performance especially when the number of FP's is low. A working point on the curve at (10^{-2}) FPPI for this data set means that there we are dealing with just three FP's. Adding even only one spurious FP in such conditions will damage the performance in a noticeable way. The algorithms that perform best overall are the ones that benefit the most from using the proposed labelling (see Table 2, columns 4 and 5). We conclude that for a correct evaluation of the performance of a detector it is important to label the impure examples in the test set and to treat them in a principled way.

7. Results

In this section we present the results of three experiments that measure the influence of different degrees of impurity in the testing and training of Pedestrian Detectors and describe the proposed labelling for the INRIA training set. The first experiment measures the impact of partial occlusion in the test set on the detection performance, while the last two experiments assess the influence on performance of using partially occluded and low resolution examples during the training of the detector. Information on the goal of each experiment and the labelling and the detectors used in each experiment is summarized in Table 3. For all the experiments we use the evaluation code by Piotr Dollár. The original annotation of the INRIA data set and the evaluation code are available on the Caltech Pedestrian Benchmark website,[6] the proposed annotation is available on the authors' website.

7.1. Experiment 1 – influence of partial occlusion in the test set on detection performance

In this experiment we evaluate the impact of the amount of partial occlusion of the examples in the test set on the detection performance. We tackle, thus, a single source of impurity. It has been shown in [12] that even a modest amount of occlusion (visibility ratio as low as 0.65) has a highly detrimental effect on the performance of Pedestrian Detectors. Those results were obtained using the Caltech data set. We perform a similar experiment on the INRIA data set and confirm that the finding has general validity. We use the detections generated by our

[6] Caltech Pedestrian Benchmark website http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

implementation of FPDW, trained on the original training set, as well as the detections of several algorithms distributed with the Caltech benchmark. We define 7 test modes which correspond to as many test sets. We filter the proposed test set with different constraints on the visibility to create the 7 test sets. The first test set contains only fully visible pedestrians, while the successive sets include pedestrians imaged under an increasing degree of occlusion. We include in this experiment only pedestrians imaged with heights greater than 90 pixels.

We observed that lowering the minimum degree of visibility of the test examples negatively affects the detection performance (see Fig. 7a for the MD/FPPI curves of FPDW tested with different degrees of visibility and Fig. 7c for a visualization of the relationship between minimum visibility in testing and the Log-Average Miss Rate for various detectors). This confirms the generality of the observation obtained on the Caltech data set.

In order to better gauge the impact of partial occlusion on detection we partitioned the INRIA test set into three visibility classes and evaluated the performance of FPDW: detecting on pedestrians with a good visibility (at least 0.7 visibility) leads to a Log-Average Miss Rate of 15%, while detecting on pedestrians with average and scarce visibilities (between 0.4 and 0.7, or under 0.4 visibility) leads to Log-Average Miss Rate of 59% and 75%, respectively (see Fig. 7b).

7.2. Proposed labelling for learning

The original labelling for the training set consists of 1237 “Person” BB's, while the proposed annotation contains a total of 1997 such BB's, a 60% increment. The largest increase in labelled pedestrians resides in the low height fraction of the data, but still remains significant for heights of up to 300 pixels (see Fig. 8a). Each

Table 4
The number of positive examples in the proposed labelling, in each of three height bins.

Height	Training	Test
$>25, <50$	59	47
$>50, <100$	294	150
>100	1644	682

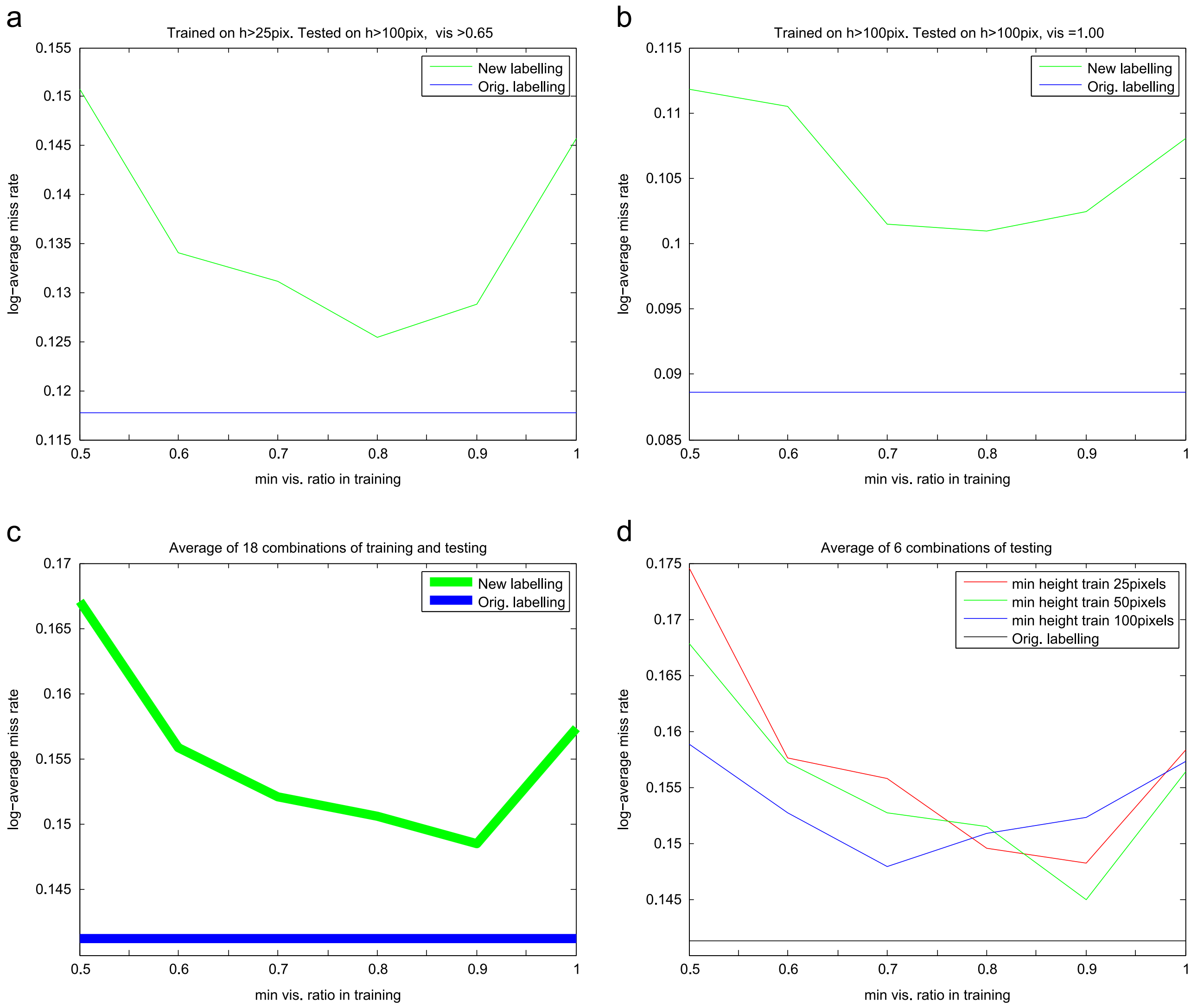

Fig. 9. (a) The performance of ACF trained with different levels of occlusion and evaluated with a minimum visibility in testing (MVTE) of 0.65. Including partially visible examples in training is advantageous: the best performance is obtained with a minimum visibility in training (MVTR) value of 0.8. (b) Including partially visible examples in training is useful even when testing exclusively on fully visible pedestrians (MVTE of 1.0). (c) The result is general: averaging over the 18 combinations of fixed parameters (see text for details) still indicates that including partially occluded examples in the training set is useful. (d) Restricting the minimum height in training (MHTR) leads to different optimal values for MVTR. The reason generating this effect unclear at the moment (see text for conjectures).

label in the proposed training set is associated with a visibility ratio, enabling different training sets to be created by setting a threshold on such quantity. Most of the newly labelled pedestrians are imaged under good visibility conditions (see Fig. 8b).

The two following experiments are devoted to assessing the importance of including impure, e.g., small and partially occluded positive examples in the training set. We evaluate the effect of partial occlusion in Experiment 2, while we gauge the impact of including small pedestrians in the training set in Experiment 3. We use the Aggregated Channel Features (ACF) detector and the proposed labelling both for training and for testing. We choose to use the ACF detector for these experiments because of the great reduction in training time it allows for, compared to our implementation of FPDW. We filter the full training set varying two thresholds: we set 3 minimum image heights for a person to 25, 50 or 100 pixels and we set 6 minimum visibility ratios to 0.5, 0.6, 0.7, 0.8, 0.9 or 1.0 (full visibility). We train the detector with the training set resulting from each of the 18 combinations and, for comparison, we train it also with the original INRIA set. To account for the stochastic parts of the ACF algorithm, we repeat each training 10 times with a different randomization seed (leading to 180 full trainings) and compute the average of the resulting performances. For a complete analysis of the influence of positive example height and visibility on the performance of PD, varying the training set is not enough. In these experiments we use the proposed labelling also for the test set. This allows us to set the minimum height for a person to be considered in the test set to 25, 50 or 100 pixels. In a similar way, we constrain the minimum visibility ratio during testing to 0.65 or 1.0 (full visibility), yielding a total of 6 evaluation modes. We adjusted the depth of the image pyramid so that it is possible to detect pedestrians 25 pixels tall or taller.

7.3. Experiment 2 – influence of partial occlusion in the training set on detection performance

In this experiment we evaluated the impact of varying the minimum visibility in training (MVTR) while fixing the rest of the

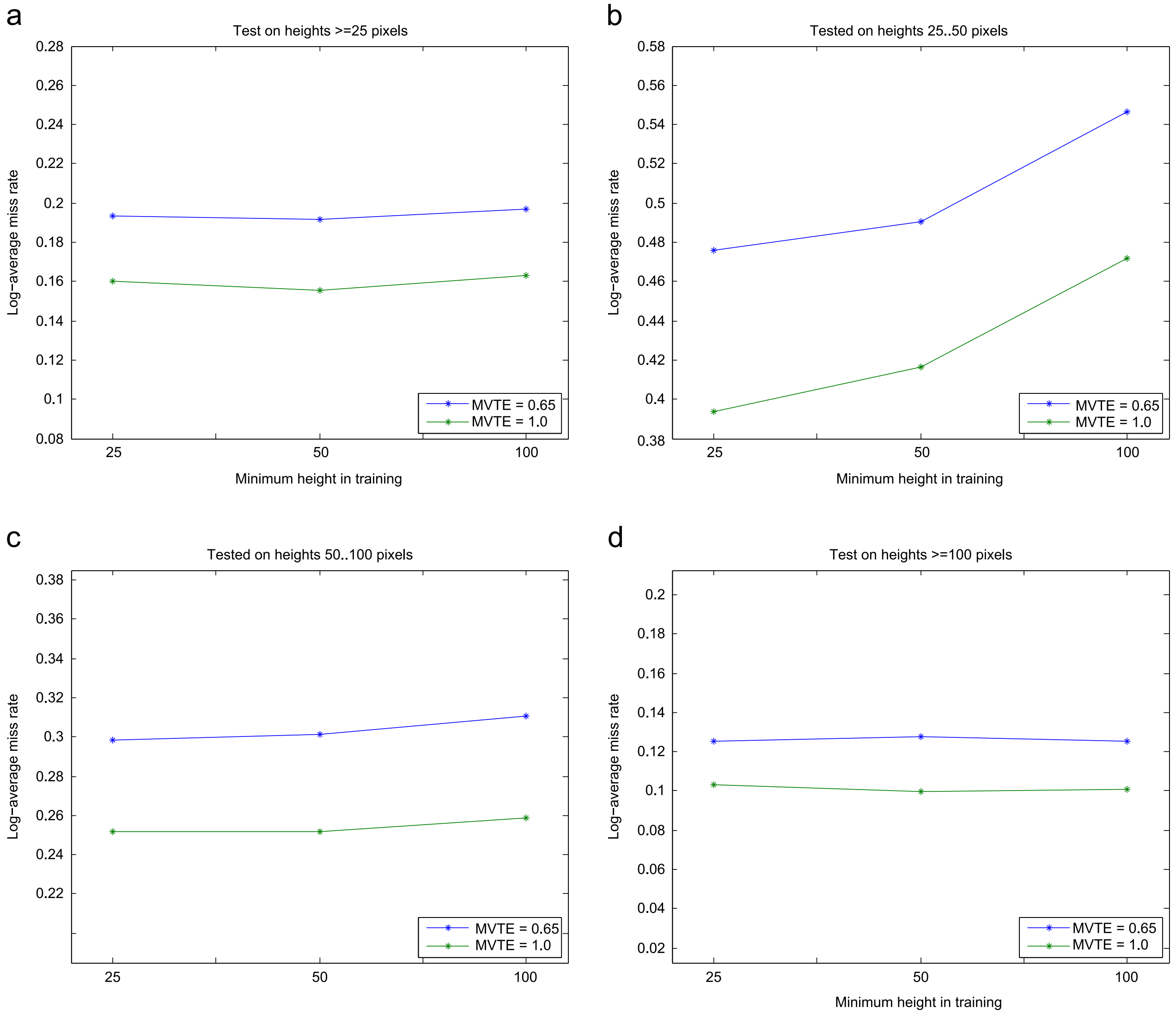

Fig. 10. Log-average miss rate obtained varying minimum training height, ACF detector. The four plots display the results obtained with two test modes: full visibility (MVTE = 1.0) in green and minimum visibility in testing (MVTE) of 0.65 in blue. (a) When testing on pedestrians spanning the whole range of heights, the different training conditions fail to produce different performances. (b) When testing on short pedestrians, it is important to include examples of similar size in the training. Including examples of medium height is also advantageous. (c) When detecting medium height pedestrians it works to include "middle sized" pedestrians in the training set, but including the small ones does not help. Including the smaller pedestrians does not change much the detection performance on the big pedestrians (b). Overall, since in the INRIA test the pedestrians with heights under 50 pixels are just 47 out of 879, the improvement of detection accuracy on this range has a small impact on the accuracy measured on the full test set (c). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

experiment parameters: the minimum height in training (MHTR), the minimum height in testing (MHTE) and the minimum visibility in testing (MVTE). We repeated the test for each of the 18 fixed parameters combinations.

For the first analysis, we fixed MHTR to 25 pixels, MHTE to 100 pixels and MVTE to 0.65. We compared the performance of the ACF detector when trained including examples with different degrees of visibility. We display the performance achieved by the different training modes (as measured by the Log-Average Miss Rate, LAMR) in Fig. 9a. It can be seen that including partially occluded pedestrians (up to 0.8 visibility) in the training is advantageous, as it leads to lower LAMR's.

In the second part of the experiment we set the MHTR and MHTE to 100 pixels and the MVTE to 1.0, i.e., we tested on fully visible pedestrians. The results depicted in Fig. 9b show that including partially occluded pedestrians in the training is advantageous even when testing exclusively on fully visible pedestrians. Lowering the required training visibility past the optimal amount, however, adversely affects performance. This behaviour is common to all the 18 combinations of training and testing constraints (see the average behaviour in Fig. 9c).

Training restricting the minimum height of the pedestrians to 25, 50 or 100 pixels leads to different optimal visibility thresholds for the training set (see Fig. 9d). It is not clear at this point if this effect is due to peculiarities of the different training subsets, to their different numerosities (see Table 4), to the increasing difficulty in labelling when dealing with smaller pedestrians or to some other phenomenon.

7.4. *Experiment 3 – influence of "short" examples in the training set on detection performance*

In this experiment we assess the impact of the inclusion of examples smaller than the detection window in the training set. We observe that including small pedestrians in the training has a strong positive effect on the detection of pedestrians in a similar range of heights. We consider the same training modes as in the previous experiment. In order to test for the influence of training pedestrians of different heights on the detection performance, we partition the pedestrians in three classes. Pedestrians between 25 and 50 pixels tall form the "short" class, those between 50 and 100 pixels tall form the "medium" class and those over 100 pixels tall form the "tall" class. The numerosity for each class in both training and testing with the proposed labelling is reported in Table 4. We introduce three new testing modes based on this partition of the heights. We consider training with the "tall" class the baseline (as 100 pixels coincides with the detection window height) and we evaluate the effect of including smaller pedestrians in the training. We report results averaged over the 6 training visibility ratios.

Testing on the full test set (pedestrians taller than 25 pixels) indicates little change when training with different subsets of the proposed data set (see Fig. 10a). Testing on the "short", "medium" and "tall" classes separately gives a better insight: when testing on the "short" class, including elements of the same height range is very beneficial, while including elements of the "medium" class is beneficial, but to a lesser extent (see Fig. 10b). The advantage is not visible, though, when testing on the full set. We ascribe the lack of impact on the full set to the small numerosity of the "short" class: in the test set its elements account for less than 1/15 of the element of the "tall" class. A comparison of the best detection performances for the "short" range, obtained with various minimum heights in training, is visible in Fig. 11. When testing on the "medium" class, including elements of the same height range has a small positive effect on detection accuracy (see Fig. 10c), while including elements from the "short" class produces little change. When testing on the "tall" class, the inclusion of shorter examples in training produces negligible variations (see Fig. 10d). Summarizing, including examples two octave smaller than the detection window has a positive effect on the detection of pedestrians in the same range. We speculate that short pedestrians contribute to the detection system knowledge on the appearance of people when imaged at low resolutions.

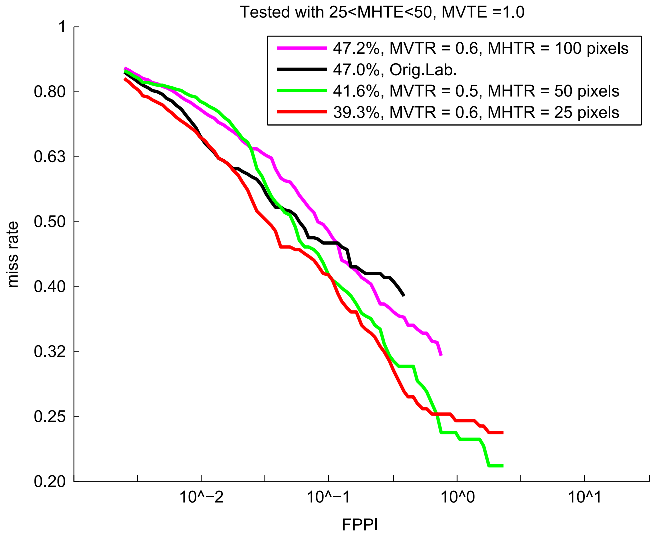

Fig. 11. Best detection performances for the "short" range, testing with full visibility. Including pedestrians imaged at heights between 25 and 50 pixels in the training (red line) produces a detector which dominates the others for most False Positive Per Image (FPPI) values. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)

8. Conclusions

In this work we discussed the importance of the sensible use of impure data in the training and evaluation of detection systems. We partitioned the positive samples in pure and impure data, the pure samples being the ones imaged in ideal conditions (in the Pedestrian Detection case, the ones imaged under full visibility and with high resolution). We proposed a new labelling for the INRIA person data set which allows for measuring the effect of impure data on Pedestrian Detectors. We showed that handling impure examples correctly is important during the evaluation phase. We observed that including partially occluded examples (up to a certain degree of occlusion) in the training set improves the detection performance both on fully visible and on partially visible pedestrians. Furthermore, we observed that the inclusion of examples imaged with heights lower than that of the detection window positively affects the detection of pedestrians in the same height range, while the performance on taller examples is unchanged. We are confident that the proposed labelling will allow for further studies on the effect of data purity on detection systems, fostering improvement in the detectors.

Acknowledgements

This work was partially supported by projects SEAGULL (QREN SI IDT 34063), FCT Pest-OE/EEI/LA0009/2013 and POETICON++ [STREP Project ICT-288382].

References

[1] Y. Wang, Y. Fan, P. Bhatt, C. Davatzikos, High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables, Neuroimage 50 (4) (2010) 1519–1535.

[2] D. Wulsin, B. Litt, E.B. Fox, Parsing epileptic events using a Markov switching process model for correlated time series, in: ICML, 2013.
[3] G. Ganeshapillai, J. Guttag, A. Lo, Learning connections in financial time series, in: ICML, 2013.
[4] N. Maknickiene, A. Maknickas, Financial market prediction system with evolino neural network and delphi method, Journal of Business Economics and Management 14 (2) (2013) 403–413.
[5] P. Smaragdis, B. Raj, The markov selection model for concurrent speech recognition, Neurocomputing 80 (2012) 64–72.
[6] R. Sznitman, A. Lucchi, P. Frazier, B. Jedynak, P. Fua, An optimal policy for target localization with application to electron microscopy, in: ICML, 2013.
[7] Y. Pang, W. Li, Y. Yuan, J. Pan, Fully affine invariant surf for image matching, Neurocomputing 85 (2012) 6–10.
[8] E. Kalapanidas, N. Avouris, M. Craciun, D. Neagu, Machine learning algorithms: A study on noise sensitivity, in: Balcan Conference in Informatics, 2003.
[9] X. Zhu, X. Wu, Class noise vs. attribute noise: a quantitative study, Artif. Intell. Rev. (2004).
[10] C.P. Lam, D.G. Stork, Evaluating classifiers by means of test data with noisy labels, in: IJCAI, 2003.
[11] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, 2005.
[12] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, Pattern Analysis and Machine Intelligence, IEEE Transactions on 34 (4) (2012) 743–761.
[13] P. Dollár, R. Appel, W. Kienzle, Crosstalk cascades for frame-rate pedestrian detection, in: ECCV, 2012.
[14] M. Pedersoli, A. Vedaldi, A Coarse-to-fine approach for fast deformable object detection, in: CVPR, 2011.
[15] E. Sangineto, M. Cristani, A. Del Bue, V. Murino, Learning discriminative spatial relations for detector dictionaries: an application to pedestrian detection, in: ECCV, 2012.
[16] R. Benenson, M. Mathias, R. Timofte, L. Van Gool, Pedestrian detection at 100 frames per second, in: CVPR, 2012.
[17] M. Taiana, J.C. Nascimento, A. Bernardino, An improved labelling for the inria person data set for pedestrian detection, in: IbPRIA, 2013.
[18] Y. Freund, R. E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: Computational learning theory, Springer, 1995, pp. 23–37.
[19] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.
[20] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, T. Poggio, Pedestrian detection using wavelet templates, in: CVPR, 1997.
[21] D. Gavrila, V. Philomin, Real-time object detection for “smart” vehicles, in: ICCV, 1999.
[22] P. Viola, M. Jones, Robust real-time object detection, in: IJCV, 2002.
[23] D. Lowe, Object recognition from local scale-invariant features, in: ICCV, 1999.
[24] P. Dollár, Z. Tu, P. Perona, Integral channel features, in: BMVC, 2009.
[25] S. Walk, N. Majer, K. Schindler, B. Schiele, New features and insights for pedestrian detection, in: CVPR, 2010.
[26] P. Dollár, S. Belongie, P. Perona, The fastest pedestrian detector in the west, in: BMVC, 2010.
[27] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, Pattern Analysis and Machine Intelligence, IEEE Transactions on 32 (9) (2010) 1627–1645.
[28] L. Pishchulin, T. Thorm, M. Planck, Articulated people detection and pose estimation: reshaping the future, in: CVPR, 2012.
[29] A. Ess, B. Leibe, L. Van Gool, Depth and appearance for mobile scene analysis, in: ICCV, 2007.
[30] C. Wojek, S. Walk, B. Schiele, Multi-cue onboard pedestrian detection, in: CVPR, 2009.
[31] M. Everingham, L. Van Gool, C. Williams, J. Winn, A. Zisserman, The Pascal visual object classes (VOC) challenge, in: IJCV, 2010.
[32] L. Bourdev, J. Brandt, Robust object detection via soft cascade, in: CVPR, 2005.
[33] C. Zhang, P. Viola, Multiple-instance pruning for learning efficient cascade detectors, in: NIPS, 2007.
[34] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection.
[35] P. Dollár, Z. Tu, H. Tao, S. Belongie, Feature mining for image classification, in: CVPR, 2007.
[36] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: CVPR, 2008.
[37] S. Maji, A. Berg, J. Malik, Classification using intersection kernel support vector machines is efficient, in: CVPR, 2008.
[38] W. Schwartz, A. Kembhavi, D. Harwood, L. Davis, Human detection using partial least squares analysis, in: ICCV, 2009.
[39] X. Wang, T. Han, S. Yan, An HOG-LBP human detector with partial occlusion handling, in: ICCV, 2009.
[40] A. Bar-Hillel, D. Levi, E. Krupka, C. Goldberg, Part-based feature synthesis for human detection, in: ECCV, 2010.

Matteo Taiana received his M.Sc. degree in Computer Engineering from Politecnico di Milano – Italy, in 2007. He is currently a Ph.D. student at the Computer and Robot Vision Laboratory of the Institute for Systems and Robotics of IST-Lisbon. His research focuses on Pedestrian Detection and Computer and Robot Vision.

Jacinto C. Nascimento (S'00 – M'06) received the E.E. degree from Instituto Superior de Engenharia de Lisboa, in 1995, the M.Sc. and Ph.D. degrees from Instituto Superior Técnico (IST), Technical University of Lisbon, in 1998 and 2003, respectively. Currently, he is an Assistant Professor with the Informatics and Computer Engineering Department, Instituto Superior Técnico, Lisbon, and a Researcher at the Institute for Systems and Robotics. Dr. Nascimento has published over 100 publications in international journals and conference proceedings, has served on program committees of many international conferences, and has been a reviewer for several international journals. His research interests include statistical image processing, pattern recognition, machine learning, medical imaging analysis, video surveillance, general visual object classification.

Alexandre Bernardino is an Assistant Professor at the Department of Electrical and Computer Engineering of IST-Lisboa and Senior Researcher at the Computer and Robot Vision Laboratory of the Institute for Systems and Robotics of IST-Lisboa. He has participated in several national and international research projects as principal investigator and technical manager. He published more than one hundred of research papers on top journals and peer-reviewed conference proceedings in the field of robotics, vision and cognitive systems. He has been associate editor in large robotics conferences and reviewer for multiple journals and conferences. His main research interests focus on the application of computer vision, machine learning, cognitive science and control theory to advanced robotics and automation systems.