An Improved Labelling for the INRIA Person Data Set for Pedestrian Detection

Matteo Taiana, Jacinto Nascimento, and Alexandre Bernardino*

Institute for Systems and Robotics, IST, Lisboa, Portugal,
mtaiana@isr.ist.utl.pt,
WWW home page: http://users.isr.ist.utl.pt/~mtaiana

Abstract. Data sets are a fundamental tool for comparing detection algorithms, fostering advances in the state of the art. The INRIA person data set is very popular in the Pedestrian Detection community, both for training detectors and reporting results. Yet, the labelling of its test set has some limitations: some of the pedestrians are not labelled, there is no specific label for the ambiguous cases and the information on the visibility ratio of each person is missing. We present a new labelling that overcomes such limitations and show that it can be used to evaluate the performance of detection algorithms in a more truthful way.

Keywords: Pedestrian Detection, INRIA person data set, labelling

1 Introduction

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 focuses 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.

* Work partially supported by the Portuguese Government – Fundação para a Ciência e a Tecnologia [PEst-OE/EEI/LA0009/2011], by the FCT project VISTA [PTDC/EIA-EIA/105062/2008] and by the project High Definition Analytics (HDA), QREN - I&D em Co-Promoção 13750. Matteo Taiana is supported by the FCT doctoral grant [SFRH/BD/43840/2008].

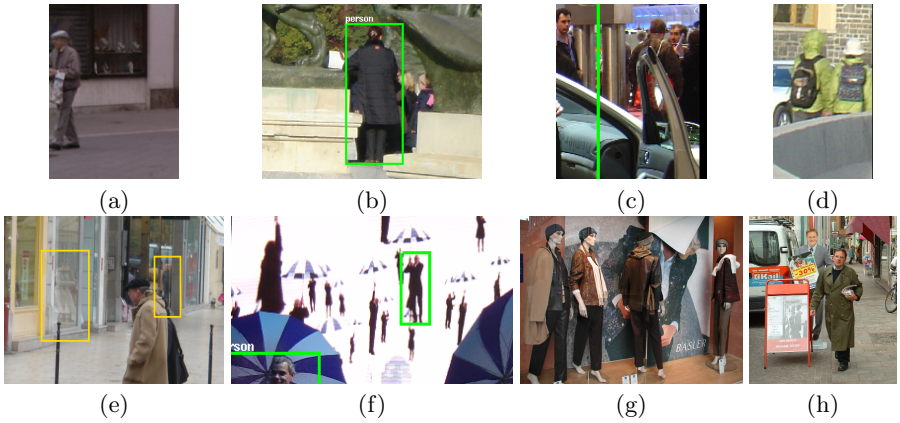


Fig. 1. Details from the INRIA test set highlighting some limitations. (a–d) Unlabelled persons. (e–h) Ambiguous cases. (e) Reflections of persons on a shop window, not labelled. (f) Some persons drawn on a wall, only one of them is labelled. (g) Some mannequins, all labelled. (h) A poster with the photo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

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 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 [1] at 0.1 False Positives Per Image (FPPI) has dropped from around 50% to around 20% since its publication (see [2]). There is still room for improvement, which explains why that data set is still popular as a benchmark [3–6], but its annotation is 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 test images are not labelled, second, an estimate of the visible part of each person is lacking and, third, 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 (see Fig. 1). In this work we propose a new labelling for the INRIA test set, elaborated following the method proposed in [2]. We argue that such labelling leads to a better evaluation of PD algorithms. The proposed annotation is available on the authors' website[1].

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

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

sets are usually compiled, while in Section 4 we describe the principles that guided the proposed labelling. In Section 5 we describe our implementation of a PD algorithm. We relate results in Section 6 and draw conclusions in Section 7.

2 Background Knowledge

Advances in PD stem mostly from research in the areas of visual feature extraction and Machine Learning, the most common classifiers being based either on AdaBoost [7] or Support Vector Machines [8]. Dense features, computed on a regular grid over the image, have been very successful. Seminal work in PD was presented in [9, 10]. One dualism in the literature contrasts part-based detectors, which explicitly model the articulation of the human body (see [11, 12]), to monolithic detectors (see [1, 3, 6]), which associate one descriptor to one detection window.

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 [9], 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 [1] 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 [13] 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 [14] (TUD) and the Caltech pedestrian data set [2] 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[2] 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.

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

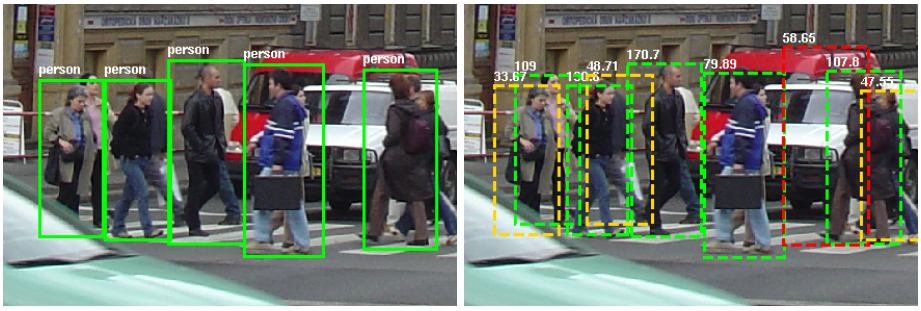
3 Labelling Strategies

The purpose of the labelling of a data set for 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 test sets are used during evaluation to determine which detections are correct, corresponding to a pedestrian. In this work we focus on the labelling of the second kind.

Most PD algorithms output a detection in the form of a rectangle on the input image. Each detection is associated to a confidence value. It is therefore natural to define the ground truth (GT) labelling in terms of a collection of rectangles as well. Such rectangles are known as detection and GT "bounding boxes", in short, "BB's".

Evaluating the performance of a PD algorithm 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 the union (Pascal VOC criterion, see [15]). 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. The labellers focused on big pedestrians: persons with a height on the image smaller than 60 pixels are not labelled. People appearing under a significant degree of occlusion were also excluded from the labelling (see Fig. 2). These choices were reasonable at the time of the publication of the data set, but current state-of-the-art algorithms can detect at least some of the partially occluded and smaller pedestrians. During evaluation, 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 occluded people. The performance of some detectors are thus under-reported (see Fig. 2 for an example of how the performance of the FPDW algorithm [16] 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 areas 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 gen-



(a) (b)

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 [16] 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.

erates 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 (see Fig. 1(e–h)), introducing noise in the evaluation process.

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 as such. The labelling was performed manually by one of the authors. As in the original annotation, we consider both cyclists and pedestrians as belonging to the "Person" class. We use rectangular BB's and the annotation scheme introduced in [2], labelling individual person as "Person", large groups of persons for which it is very difficult to label each individual as "People", and ambiguous cases as "Person?". The proposed annotation is available on the authors' website. 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 gray 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 TP's nor 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 GT BB's are centered horizontally and transformed to assume an aspect ratio of 0.41 (width/height) prior to matching (see [2] for details). Thus we design the BB's of the proposed labelling to be centered around the vertical axis of each person. The evaluation code enables researchers to perform different experiments on a single data set, using just one annotation. The minimum height and the minimum visibility ratio of the GT rectangles are specified as a parameter for the evaluation, so that all the BB's that do not match the criterion are set to "Ignore". One can for instance test the performance of a detector considering only pedestrians taller than 50 pixels and visible for at least 65%. This evaluation mode is dubbed "Reasonable" in [2]. We introduce the "Reasonable90" mode, which requires BB's to be at least 90 pixels high and keeps the minimum visible portion at 65%. We argument this choice in Section 6.

5 Our Implementation of a Pedestrian Detector

We implemented a version of the FPDW algorithm. 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 [17, 18]. We use 1 000 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 [19]), we compute the Integral Channels using publicly available code by the author[3]. We train the detector on the INRIA pedestrians training set with 4 epochs of bootstrap.

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

6 Results

6.1 Analysis of the New Labelling

The proposed annotation 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". It is common practice to compare algorithms on the original INRIA test set using the "Reasonable" evaluation mode: only people taller than 50 pixels are considered as "Person", the rest of the BB's are set to "Ignore". The condition requiring more than 65% visibility has no effect when used with the original annotation, as the GT labels do not carry information about occluded areas. It is meaningful to notice that the detections provided in the Caltech benchmark for the vast majority of the algorithms have a lower limit for the height of the detections at around 90 pixels (see Tab. 1, column 2). There are, though, some labels shorter than 90 pixels in the GT annotations. Such GT labels can never be matched by the output of most of the detectors. It is, thus, unfair to use the "Reasonable" mode on this data set, with these detections. Having access to a collection of detections spanning a limited range of heights, we decide to tune the range of the GT BB's accordingly, defining the "Reasonable90" mode.

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. 3a). The fraction of "Ignore" BB's for the new annotation is considerable, Figure 3b illustrates the amount of labels that are set to "Person" and "Ignore" for the "Reasonable90" mode, as a function of height.

6.2 Experiments

We perform two experiments in which we use the detections of many state-of-the-art algorithms together with the detections generated by our implementation of FPDW. We use the INRIA person test set and the evaluation code by Piotr Dollár. The detections, the original annotation and the evaluation code are available on the Caltech Pedestrian Benchmark website[4], the proposed annotation is available on the authors' website. In the first experiment we compare the reported performance of PD algorithms using the original labelling and selecting two different evaluation modes: "Reasonable" and "Reasonable90". We argue that "Reasonable90" is a more appropriate test mode for that data set. In the second experiment we compare the reported performance of the algorithms on the original and the proposed annotation, using the "Reasonable90" mode.

The results obtained using the original labelling and either the "Reasonable" or the "Reasonable90" mode are very similar in quality, we display the missed detection rate/FPPI curves for two representative algorithms, for the two modes (see Fig. 4). The performance of one algorithm is synthesized in the legends of

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

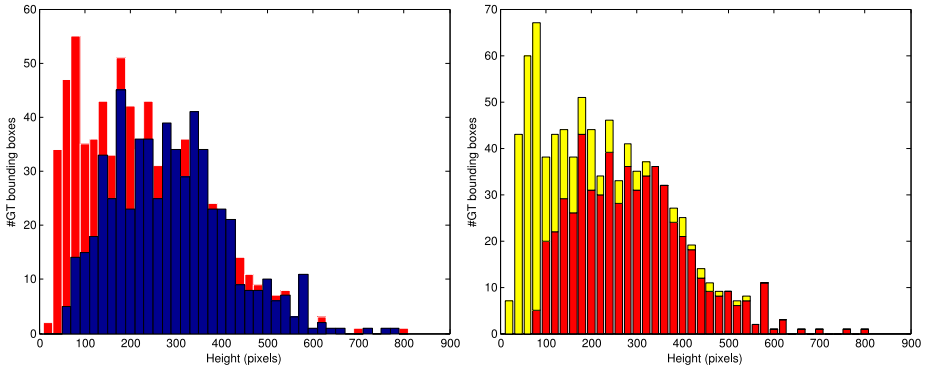
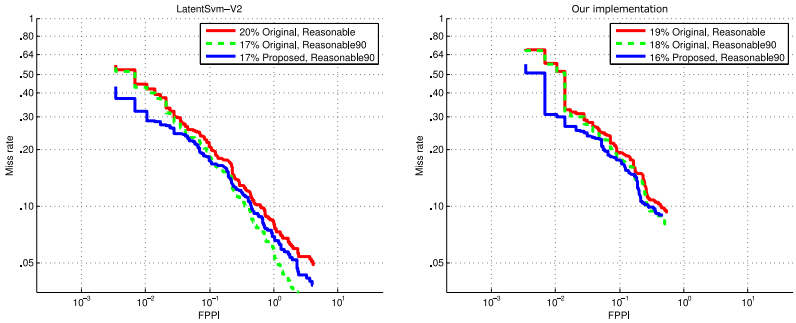


Fig. 3. Characterization of the original and the proposed labellings. (a) Histograms of the height of "Person" labels for the original (blue) and the proposed labelling (red). The proposed annotation outnumbers the original one, particularly at low heights. (b) Histogram for the proposed labelling and 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.

the plots with the log-average miss rate, the average miss rate computed between 10^-2 and 10^0 FPPI (see [2]). 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 Tab.1, columns 3–5). This result is expected, as passing from "Reasonable" to "Reasonable90" we removed from the test set some labels which were impossible for the algorithms to match.

In the second experiment we compare the performance reported using the original and the proposed annotations for the INRIA test set. We use the same set of algorithms and the "Reasonable 90" mode. We display the missed detection/FPPI plot for two representative algorithms and the two annotations, in Figure 4. 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 Tab. 1, columns 6–8). We ascribe this to the removal of the spurious 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 spurios FP's 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 Tab. 1, columns 8 and 9).



(a) (b)

Fig. 4. The reported performance of the LatSvm-V2 algorithm [11] (a) and of our implementation of FPDW [16] (b) in three conditions: original labeling and "Reasonable" mode, original labelling and "Reasonable90" mode, proposed labelling and "Reasonable90" mode. Performance is synthesized with the log-average miss rate.

			MD at 10^0FPPI			MD at 10^{-2}FPPI			
Algorithm		Minimum height	Original Reas.	Original Reas.90	Difference	Original Reas.90	Proposed Reas.90	Difference	Log av. miss rate for Prop., Reas.90
FtrMine	[20]	100.0	0.340	0.324	-0.016	0.918	0.900	-0.019	57%
LatSvm-V1	[21]	79.0	0.175	0.159	-0.015	0.806	0.835	+0.029	43%
HOG	[1]	100.0	0.231	0.215	-0.015	0.744	0.702	-0.042	42%
HikSvm	[22]	100.0	0.221	0.207	-0.014	0.766	0.681	-0.085	39%
PLS	[23]	100.0	0.226	0.212	-0.014	0.674	0.596	-0.078	38%
HogLbp	[24]	96.0	0.190	0.173	-0.017	0.665	0.629	-0.036	35%
MultiFtr+CSS	[25]	93.7	0.109	0.093	-0.016	0.469	0.425	-0.044	29%
FeatSynth	[26]	100.0	0.109	0.089	-0.019	0.754	0.738	-0.015	21%
FPDW	[16]	100.0	0.093	0.075	-0.018	0.576	0.386	-0.189	18%
ChnFtrs	[19]	100.0	0.087	0.072	-0.015	0.581	0.383	-0.198	18%
LatSvm-V2	[11]	*91.3*	*0.081*	*0.058*	*-0.024*	*0.448*	*0.319*	*-0.129*	*17%*
Our FPDW		*95.6*	*0.093*	*0.081*	*-0.013*	*0.577*	*0.307*	*-0.270*	*16%*
CrossTalk	[3]	99.2	0.098	0.079	-0.020	0.511	0.333	-0.178	15%
Mean					-0.017			-0.089	

Table 1. The performances of a set of state-of-the-art PD algorithms reported with different labellings and different evaluation modes.

7 Conclusions

In this work we discussed the importance of data sets and benchmarking procedures for the evaluation of detection algorithms. We highlighted the limitations of the labelling of the INRIA person data set and proposed a new labelling and a new evaluation mode. We showed that the proposed labelling and evaluation mode allow for a more accurate evaluation of the state-of-the-art algorithms.

References

1. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. CVPR (2005)
2. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: An Evaluation of the State of the Art. PAMI (2012)
3. Dollár, P., Appel, R., Kienzle, W.: Crosstalk Cascades for Frame-Rate Pedestrian Detection. ECCV (2012)

4. Pedersoli, M., Vedaldi, A.: A Coarse-to-fine approach for fast deformable object detection. CVPR (2011)
5. Sangineto, E., Cristani, M., Del Bue, A., Murino, V.: Learning discriminative spatial relations for detector dictionaries: An application to pedestrian detection. ECCV (2012)
6. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. CVPR (2012)
7. Freund, Y., Schapire, R.: A desicion-theoretic generalization of on-line learning and an application to boosting. Computational Learning Theory (1995)
8. Cortes, C., Vapnik, V.: Support-vector networks. ML (1995)
9. Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., Poggio, T.: Pedestrian detection using wavelet templates. CVPR (1997)
10. Gavrila, D., Philomin, V.: Real-time object detection for "smart" vehicles. ICCV (1999)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI (2010)
12. Pishchulin, L., Thorm, T., Planck, M.: Articulated People Detection and Pose Estimation: Reshaping the Future. CVPR (2012)
13. Ess, A., Leibe, B., Van Gool, L.: Depth and Appearance for Mobile Scene Analysis. ICCV (2007)
14. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. CVPR (2009)
15. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. IJCV (2010)
16. Dollár, P., Belongie, S., Perona, P.: The Fastest Pedestrian Detector in the West. BMVC (2010)
17. Bourdev, L., Brandt, J.: Robust Object Detection via Soft Cascade. CVPR (2005)
18. Zhang, C., Viola, P.: Multiple-Instance Pruning For Learning Efficient Cascade Detectors. NIPS (2007)
19. Dollár, P., Tu, Z., Perona, P.: Integral channel features. BMVC (2009)
20. Dollár, P., Tu, Z., Tao, H., Belongie, S.: Feature mining for image classification. CVPR (2007)
21. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. CVPR (2008)
22. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. CVPR (2008)
23. Schwartz, W., Kembhavi, A., Harwood, D., Davis, L.: Human detection using partial least squares analysis. ICCV (2009)
24. Wang, X., Han, T., Yan, S.: An HOG-LBP human detector with partial occlusion handling. ICCV (2009)
25. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. CVPR (2010)
26. Bar-Hillel, A., Levi, D., Krupka, E., Goldberg, C.: Part-Based Feature Synthesis for Human Detection. ECCV (2010)