TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

A methodology for training and evaluating Pedestrian Detectors: towards practical applications

Matteo Taiana

Supervisor: Doctor Alexandre José Malheiro Bernardino
Co-Supervisor: Doctor Jacinto Carlos Marques Peixoto do Nascimento

Thesis approved in public session to obtain the PhD degree in Electrical and Computer Engineering

Jury final classification: pass with distinction

Jury

Chairperson: Chairman of the IST Scientific Board

Members of the Committee:
Doctor Vicente Javier Traver Roig
Doctor José Alberto Rosado dos Santos Victor
Doctor Alexandre José Malheiro Bernardino
Doctor Paulo José Monteiro Peixoto
Doctor Paulo Luís Serras Lobato Correia
Doctor Jacinto Carlos Marques Peixoto do Nascimento

2015

UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

A methodology for training and evaluating Pedestrian Detectors: towards practical applications

Matteo Taiana

Supervisor: Doctor Alexandre José Malheiro Bernardino
Co-Supervisor: Doctor Jacinto Carlos Marques Peixoto do Nascimento

Thesis approved in public session to obtain the PhD degree in
Electrical and Computer Engineering

Jury final classification: pass with distinction

Jury

Chairperson: Chairman of the IST Scientific Board

Members of the Committee:
Doctor Vicente Javier Traver Roig, Professor Titular, Escola Superior de Ciències Experimentals, Universitat Jaume I, Spain
Doctor José Alberto Rosado dos Santos Victor, Professor Catedrático do Instituto Superior Técnico da Universidade de Lisboa
Doctor Alexandre José Malheiro Bernardino, Professor Associado do Instituto Superior Técnico da Universidade de Lisboa
Doctor Paulo José Monteiro Peixoto, Professor Auxiliar da Faculdade de Ciências e Tecnologia da Universidade de Coimbra
Doctor Paulo Luís Serras Lobato Correia, Professor Auxiliar do Instituto Superior Técnico da Universidade de Lisboa
Doctor Jacinto Carlos Marques Peixoto do Nascimento, Professor Auxiliar do Instituto Superior Técnico da Universidade de Lisboa

Funding institution
Fundação para a Ciência e a Tecnologia

2015

Abstract

Detecting people is an important goal of many automated systems. Depending on the setting, such goal can be achieved using different kinds of sensors and recognition techniques. This thesis focusses on the problem of Pedestrian Detection, i.e., the detection of people assuming standing or walking stances, in images acquired from a moving camera. This version of the problem has clear applications in the fields of mobile robotics (to inform Human-Robot Interaction systems) and automotive (providing input to Advanced Driver Assistance Systems), among others. Detecting pedestrians is a hard problem, but years of successful research led to great advances in detection accuracy and speed.

This thesis deals with three research topics related to Pedestrian Detection. First, it concentrates on the methodology used for defining Ground Truth labels for the data sets: Pedestrian Detection systems are based on Machine Learning and, as such, they require labelled data both for training and testing. As the performance of the detectors improves, the labelling information is enriched, so that training data is better exploited and test data better highlights differences between the performances of different algorithms. Second, it studies the effect of High Definition images on detection performance: as the price of High Definition cameras drops, their use becomes more common in Video Surveillance settings. It is, therefore, important to establish a benchmark for Pedestrian Detection for such imaging conditions, in order to point out the weaknesses of the current approaches and to foster the development of detectors which exploit the high resolution images. Third, it employs a Pedestrian Detector in the design of a fully automated person Re-Identification system: Video Surveillance systems rely on human operators for the execution of many tasks. The automatization of some of such tasks is desirable, as it would allow focussing the work of the

human resources on the high-level aspects of the job. The work I performed in the three areas is outlined in the following paragraphs.

Regarding the first topic, I introduce the concept of sample "purity", identifying as "impure" the examples imaged in non-ideal conditions for detection: examples affected by partial occlusion or smaller than the detection window. I show that including slightly occluded pedestrians in the training set improves performance, even on fully visible examples, while incorporating very small pedestrians improves detection on pedestrians imaged at similar scales. Furthermore, I show that matching height ranges during experiment design and using an accurate test Ground Truth are crucial for a fair evaluation of detection performance. During this work I developed a richer and more accurate annotation for the widely used INRIA person data set.

With respect to the second topic, I collected the High Definition Analytics data set, a tool for the evaluation of Pedestrian Detectors in a Visual Surveillance scenario and for the assessment of the impact of High Definition images on the performance of Video Surveillance algorithms. I performed experiments on the data set using two detectors representative of two opposite philosophies in the state of the art. The part-based detector proved to be better on people imaged at close range, while the monolithic detector performed slightly better on fully visible people. Experiments on High Definition images show that they allow for the detection of pedestrians farther away than regular definition images do, at the price of more False Positives and longer processing times. Re-Identification experiments show that the proposed data set is very challenging and that Re-Identification algorithms based on simple features do not take advantage of High Resolution images.

Finally, to address the third challenge, I design a fully automated person Re-Identification in which a Pedestrian Detector is integrated with a standard Re-Identification module. I show that precision and recall statistics are useful to characterise the performance of the integrated system and devise two improvements with respect to the naive integration scheme. The False Positives class deals with the False Positives generated by the detector, while the Occlusion Filter uses geometrical reasoning to reject detections with a high probability of being misclassified. The two improvements afford higher Re-Identification precision at the price of a drop in recall.

I am confident that the ideas I explore in this thesis, along with the data

set and the annotation we collected, will assist the scientific community in designing the next generation of Pedestrian Detectors.

List of Acronyms

ACF Aggregate Channel Features

BB Bounding Box

CI Correct Identification

CMC Cumulative Matching Characteristic

FP False Positive

FPDW Fastest Pedestrian Detector in the West

FPPI False Positives Per Image

GT Ground Truth

HDA High Definition Analytics

II Incorrect Identification

LAMR Log-Average Miss Rate

MD Missed Detection

MHTE Minimum Height in TEst

MHTR Minimum Height in TRaining

MVTE Minimum Visibility in TEst

MVTR Minimum Visibility in TRaining

NMS Non-Maximum Suppression

OF Occlusion Filter

PD Pedestrian Detection

PD+REID fully automated Re-Identification

RE-ID Re-Identification

TP True Positive

Contents

Chapter 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 drawn between the methods that do and do not restrain the articulation of the people.

This work focusses on Pedestrian Detection (PD), i.e., the detection of people assuming poses that are common while standing or walking, in images acquired by a mobile camera. PD is important as it enables the estimation of the presence and the position of humans in the vicinity of a vision sensor. The most immediate applications of PD are in the field of automotive (smart cars and Advanced Driver Assistance Systems) and in that of mobile robotics. Advanced Driver Assistance Systems need to be aware of pedestrians in the vicinity of the vehicle they are operating on, in order to warn the driver (or override his/her commands) in case of danger. Mobile robots need to detect people in their surroundings, as the first step in complex Human–Robot Interaction systems. Other areas of application for PD are Video Surveillance, smart spaces and entertainment.

The PD 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 con-

tribute to the illumination of the scene. Background clutter also plays a role in making the detection difficult. PD has been the subject of extensive research by many research groups, with the number of scientific publications including "Pedestrian Detection" in their title steadily increasing during the last decades (see Figure 1.1), and companies such as Mobileye [Mob] bringing the technology to the market. The obvious, implicit goal of PD research is that of improving detection accuracy, but another goal has received wide attention: that of developing fast detectors. Real-time detectors are a requirement for many practical applications, in particular when vehicles or mobile robots are involved. The degree of success of PD research can be appreciated observing that the Crosstalk cascade detector [Dollár et al., 2012a] (published in 2012) is almost 150 times faster than the now classic Histogram of Oriented Gradients (HOG) detector [Dalal and Triggs, 2005] (published in 2005), while also achieving a significant improvement in detection accuracy (see [Dollár, b]).

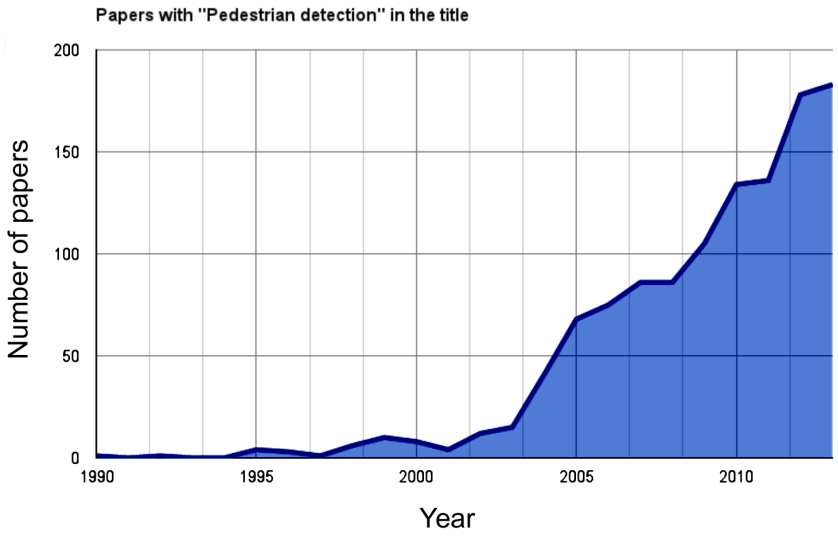


Figure 1.1: A plot of the number of published papers including "Pedestrian Detection" in their title as a function of the year of publication. Publications on Pedestrian Detection steadily increased over the last 20 years. Reproduced from [Benenson et al., 2014b].

The dominant approach to PD consists in the detection-by-classification paradigm. Such approach requires the use of an image-window classifier, designed to estimate whether a pedestrian is present in a given image win-

dow. The essence of the approach consists in running the classifier on a grid of locations on the image, collecting each positive classification as a detection. It is common practice to run the classifier on multiple scaled versions of the input image and to follow the classification step with a filtering step aimed at obtaining just one detection per imaged object. The detection-by-classification paradigm is reviewed in detail in Chapter 3.

1.1 Challenges and Contributions

In this thesis I tackle three main challenges of the PD problem. In the remainder of this section I briefly introduce each contribution, summarizing the conditions that motivated it, the contribution itself and the results it led to.

Data Set Labelling and Ground Truth – Modern PD systems rely on Machine Learning techniques: the rules of a detection system, as opposed to being designed by hand, are learnt from examples. Furthermore, the evaluation of PD systems is based on sets of labelled test images. Given the data-driven nature of PD algorithms, public data sets consisting of annotated images and standard evaluation code are needed for a fair comparison of their performances. Moreover, data sets play an important role in stimulating advances in PD performance: with the improvements in PD technology data sets become obsolete and are replaced by more challenging ones. Labelling choices, such as defining the minimum visibility and the minimum height of the training examples, have received little attention, but they do have an effect on the performance of the resulting detector. Furthermore, a fair comparison of detectors on some data sets is hindered by the way labels are used at test time.

To address these issues, I investigate the impact of sample purity on detection performance. I define as impure the examples which are imaged in non-ideal conditions for the detection-by-classification paradigm: the ones imaged under partial occlusion and the ones imaged with heights smaller than that of the detection window. I explore the effect on detection performance of the inclusion of examples with different degrees of impurity in the training set, with the goal of defining rules for selecting the examples in the training set that will lead to the best possible detector. I introduce an improved labelling for the popular INRIA person data set [Dalal and Triggs,

2005] which includes information on the visibility of the pedestrians, and is more thorough than the original one. The new labelling is exploited in the experiments.

One result of this work indicates that having accurate Ground Truth labels and carefully designing the experiments is important for a fair evaluation of PD's. Another result consists in the recommendation on which examples to include in training in order to maximize detection performance: including examples with low levels of impurity is beneficial. Incorporating slightly occluded examples (up to 10% occlusion) in training improves detection also on fully visible pedestrians. Including very small examples (two octaves smaller than the detection window) in training improves detection on pedestrians of a similar size. Finally, the experiments confirm that the degree of occlusion a pedestrian is imaged with correlates with the probability of a detection algorithm missing her/him.

High Definition Video Surveillance – The data sets commonly used for PD represent automotive scenarios [Ess et al., 2007; Dollár et al., 2009; Wojek et al., 2009; Geiger et al., 2012], with the notable exception of the INRIA data set which consists in a collection of holiday pictures [Dalal and Triggs, 2005]. However, one clear application of PD is in Video Surveillance scenarios, in which static cameras are endowed with a different perspective view from that of cameras mounted on cars, and possibly set in indoor environments. Considerations related to data set bias [Torralba and Efros, 2011; Khosla et al., 2012] suggest that detection performances measured on a data set specific for Video Surveillance would approximate the performance in real-world Video Surveillance applications more accurately than the performance measured on different (e.g., automotive) scenarios. This motivates the development of a PD data set representative of a Video Surveillance scenario. Furthermore, common PD data sets are based on low resolution images (mostly VGA: 640×480 pixels). This hinders the investigation of the effect that High Definition images have on PD and other algorithms. Establishing a high resolution benchmark for PD would also serve as a stimulus for the development of Pedestrian Detectors specific for High Definition images.

I designed the High Definition Analytics (HDA) data set for benchmarking Video Surveillance algorithms, in particular PD, person tracking and Re-Identification (RE-ID) algorithms. The data set includes image streams captured both at high and standard resolution, allowing the study of whether

high resolution affords for better Video Surveillance performances. The environment for the recordings is chosen so that it is possible to evaluate the performance of PD's in a Video Surveillance setting. The labelling and the evaluation code are designed so that PD, person RE-ID and fully automated Re-Identification (PD+REID) systems can be evaluated.

In the experiments, I compare the performance of two algorithms representative of the two main paradigms in the state of the art on the HDA data set: the Fastest Pedestrian Detector in the West [Dollár et al., 2010] and the Grammar Models detector [Girshick et al., 2011], for the monolithic and the part-based philosophies respectively. The part-based detector proves to be better at detecting pedestrians imaged at short range, possibly because of its ability of accommodating displacements of the body parts. The monolithic detector, on the other hand, proves to be slightly better at detecting fully visible pedestrians. Comparing the performance of the two algorithms on the HDA and the INRIA data set highlights that the data sets possess different peculiarities. Exploring detection on High Definition images confirmed that they allow for the detection of pedestrians standing farther from the camera than regular definition images do, but, due to the nature of the sliding window detection paradigm, also generated more False Positives and required more processing time. Experiments on RE-ID indicate that, together with CAVIAR4REID, the HDA data set is the most challenging to date. While the difficulty on CAVIAR4REID stems from low resolution images, pedestrians on HDA are hard to re-identify because of the heterogeneity in the imaging conditions (illumination changes, occlusions, etc.). Further experiments show no advantage in RE-ID when using High Definition images. This might be due, however, to the simple features we use in our implementation of a RE-ID algorithm.

Fully Automated Re-Identification – The classic set up for a RE-ID experiment requires human intervention for the selection of the test examples (i.e., image windows centred on pedestrians), leading to RE-ID systems of little practical use. The task performed by the human operators can in principle be executed by a PD module, leading to a PD+REID system. The naive integration of a PD and a RE-ID module, however, suffers because of the way the errors committed at the PD stage affect the performance of the whole system. This consideration served as motivation to design a better integration scheme.

The last contribution of this thesis consists in the design of a fully automated RE-ID system. In collaboration with Dario Figueira (see Taiana et al. [2014]; Figueira et al. [2014]), we propose two improvements to a naive integrated system: the False Positive class and the Occlusion Filter. The False Positive class models the False Positives detections generated by the PD module in a given scenario. By establishing a correct RE-ID class for the False Positives, it allows for the plotting of a sensible Cumulative Matching Characteristic (CMC) curve, which is the standard way to compare RE-ID algorithms. The Occlusion Filter exploits geometrical reasoning to filter the detections, so that only detections which have a high probability of depicting fully visible pedestrians are passed on to the RE-ID module, while ambiguous and hard to classify detections are discarded. Finally, we introduce the use of precision and recall statistics as a way to complement the information conveyed by the CMC curve.

Results indicate that the introduction of the False Positive class leads to an increase in RE-ID precision, at the price of a drop in recall. The use of the Occlusion Filter produces a similar, albeit smaller in module, change in the performance: an increase in precision and a decrease in recall. Furthermore, we show that precision and recall statistics are useful for characterising the performance of RE-ID and PD+REID systems alike.

Previous publications related to the work presented in this thesis include [Taiana et al., 2013, 2015], which explore the effect of sample purity on PD; [Nambiar et al., 2014], which introduces the HDA data set; and [Taiana et al., 2014; Figueira et al., 2014], which focus on the development of a PD+REID system.

1.2 Outline of the document

In Chapter 2 I lay the bases for the ideas presented in the rest of the document. I review the state of the art for PD, the most popular data sets used in the field, and the associated benchmarking techniques. Moreover, I list applications of PD in the fields of automotive, Video Surveillance and Human–Robot Interaction. In Chapter 3 I describe the standard architecture for PD, highlighting insight gained during my implementation of various detectors, most notably the implementation of the Fastest Pedestrian Detector in the West. In Chapter 4 I propose my ideas on the effect of impure

samples on the PD problem, at training and test time. In Chapter 5 I describe the HDA data set, characterising the data and reviewing the design choices taken in order to make it useful for benchmarking Video Surveillance algorithms in High Definition images. In Chapter 6 I motivate and describe the proposed PD+REID system, including two improvements over the naive integration architecture. In Chapter 7 I report the experiments and the results supporting the discussion. Finally, I draw conclusions and list ideas for future work in Chapter 8.

Chapter 2

Background and Related Work

2.1 Pedestrian Detection

2.1.1 Detection Paradigms in the State of the Art

Pedestrian Detectors (PD's) in the state of the art are based on a limited number of paradigms: Hough transform, branch-and-bound or detection-by-classification, with the latter being used in the overwhelming majority of detectors. Detectors based on the Hough transform [Gall et al., 2011; Lehmann et al., 2011] rely on detecting small image patches (corresponding to a head or a foot, for instance) and a voting scheme. Each detected patch votes for the presence of a person in a set of locations on the image. Votes are accumulated in the Hough space and detections are computed by finding maxima in such space. Votes in the Hough transform scheme are based on very local information, it has been shown that combining source of global information improves the detection accuracy [Leibe et al., 2005]. Efficient Subwindow Search [Lampert et al., 2009] is a detection technique based on the branch-and-bound scheme: it explores the set of all possible rectangles on one image, finding the one with the highest detection confidence. The search space is explored by hierarchically splitting it into disjoint subsets. Subsets of rectangles whose upper bound on the confidence indicates that they can not contain the maximum are discarded, leading to fast detections. The detection-by-classification paradigm is the most successful one in PD. I describe it in the following paragraph and assume its use in the rest of this

work.

2.1.2 Detection-By-Classification Paradigm

The fundamental block of a PD based on detection-by-classification 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. 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. One image pyramid is a collection of images obtained by successive scalings of one original image. 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.

2.1.3 Evolution of Pedestrian Detection Techniques

Advances in PD stem mostly from research in the areas of visual feature extraction and Machine Learning (ML), the most common classifiers being based either on AdaBoost [Freund and Schapire, 1995] or Support Vector Machines [Cortes and Vapnik, 1995]. Early work on visual PD focussed on hybrid detection/tracking systems which relied on hand-crafted models [Hogg, 1983; Rohr, 1993]. Seminal work relying on ML-based PD was presented in [Oren et al., 1997; Gavrila and Philomin, 1999]. The authors of [Viola and Jones, 2001, 2004] introduced Integral Images for faster feature computation, AdaBoost for combining many weak classifiers into a strong classifier and a Cascaded Detector for increasing the detection speed. That work focussed 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 [Dalal and Triggs, 2005] and reminiscent of SIFT [Lowe, 1999], it represents gradient information in a way that enables robust classification. A recent trend is that of combining multiple features: the Integral Channel Features [Dollár et al., 2009] exploit 10 channels of information based on colour and gradient. The authors of [Walk et al., 2010] 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). Some authors push the trend further, by combining full detectors [De Smedt and Goedemé, 2015]. One dualism in the literature contrasts monolithic detectors (see [Dalal and Triggs, 2005; Dollár et al., 2012a; Benenson et al., 2012]), which compute features at fixed locations on the detection window and deal with articulation implicitly, to part-based detectors (see [Mohan et al., 2001; Felzenszwalb et al., 2010; Pishchulin et al., 2012]), which explicitly model the articulation of the human body and allow for variable placing of the features corresponding to the human limbs.

Several of the best detectors in the state of the art are derived from the Integral Channel Features detector [Dollár et al., 2009]. Following the consideration that boosted decision trees induce decision boundaries piecewise orthogonal to the features (single-feature splits), while the features for PD can be highly correlated, Locally Decorrelated Channel Features were introduced in [Nam et al., 2014]. The Informed Haar-like Features introduced in [Zhang et al., 2014] are in practice second-order Integral Channel Features: the value of a feature consists in the difference between the integral computed on two areas on one of the image channels (in some cases a third "ignore" area is considered). The geometry defining the support of each feature is tailored to the PD problem, rather than selected randomly.

2.1.4 Partial Occlusion

Partial occlusion was identified early as a source of difficulty in the detection of pedestrians, so methods aiming at improving the detection rate on the partially occluded pedestrians were developed. Although some attempt at solving the occlusion problem were made using monolithic detectors (e.g., in [Leibe et al., 2005]), part-based detectors are the ones that have been

most frequently employed for tackling this task (see [Mohan et al., 2001; Wu and Nevatia, 2005; Wang et al., 2009; Enzweiler et al., 2010; Ouyang and Wang, 2012]). This choice is naturally motivated by fact that the part-based paradigm intrinsically affords the possibility for each part to either be visible or occluded.

Upon noticing that inter-pedestrian occlusion (one person partially occluding another) is the preponderant source of occlusion for pedestrians, algorithms were proposed that employ a "double pedestrian detector": a detector specifically trained to detect two persons at once. The detectors presented in [Pepikj et al., 2013; Tang et al., 2014] exploit multiple components and are able to detect both single persons and pairs of pedestrians, while in [Ouyang and Wang, 2013; Ouyang et al., 2015] the authors devise a probabilistic approach for fusing the detections of a "double pedestrian detector" with those obtained by a traditional single pedestrian detector.

The joint use of 16 monolithic classifiers, each of which specific for a given level and geometry of occlusion (i.e., from the bottom or from one side of the detection window), is described in [Mathias et al., 2013]. The authors design a two-step Non-Maximum Suppression for fusing the output of the multiple detectors, exploiting knowledge on the fact that the higher level of occlusion one detector is designed to work with, the worse its performance is. Furthermore, strategies are described that allow the detectors to share many features, resulting in a substantial improvement of both training and detection time.

2.1.5 Fast Pedestrian Detection

Another line of work, initiated by the aforementioned introduction of Integral Images and Cascaded Detectors, concentrates on reducing the detection time. More recently, the Fastest Pedestrian Detector in the West (FPDW) algorithm [Dollár et al., 2010] introduced the possibility of estimating features for many layers of the image pyramid instead of computing them explicitly, while the 100 FPS detector [Benenson et al., 2012] was designed to compute features at just one scale exploiting different classifier models built during training. Research on making faster detectors also focusses on Object Proposals [Zitnick and Dollár, 2014; Hosang et al., 2014], a technique based on the assumption that all objects of interest share properties which differentiate them from the background. Running an Object Proposal filter prior to

running detectors for different kind of objects quickly and greatly reduces the number of image windows to be classified by each detector.

2.1.6 Evolution of Pedestrian Detection Performance

Both detection accuracy and speed improved considerably along the years. Comparing the Viola-Jones detector [Viola and Jones, 2001] from 2001 with the Crosstalk Cascade detector from 2012, it can be seen that detection speed increased by a factor of 75, achieving 35 Frames Per Second for a VGA image, while the Log-Average Miss Rate (the average miss detection rate as computed on the logarithmic False Positives Per Image axis, see [Dollár et al., 2012b] for details) dropped from over 80% to 30% [Dollár, b], as measured on the Caltech Pedestrian data set [Dollár et al., 2012b]. The detector based on Locally Decorrelated Channel Features[Nam et al., 2014], introduced in 2014, achieves a Log-Average Miss Rate of 25%.

2.1.7 Comparison to Human Performance

Discussing the possibility of a limit for the performance of PD systems, it is interesting to evaluate the PD performance achieved by humans. The accuracy of humans has been shown to be two to three orders of magnitude better than that of current automated detectors (see [Benenson, 2015][1]). Human errors consist mostly in Missed Detections on the hardest 5% of the pedestrians, while False Positive detections are rare. Assuming the performance of automated PD systems continues improving at the current rate [Benenson et al., 2014a], superhuman accuracy will be achieved in less than five years. When the analysis is extended to include the time spent evaluating the presence of pedestrians in one image, the performance of automated detectors is shown to already surpass that of humans: automated systems can work at over 100 frames per second, while humans require a time in the order of seconds to process one image.

2.1.8 Open Challenges

The focus for PD research at the moment lies on improving the detection rates on pedestrians imaged at low resolutions or under heavy occlusion,

[1]The cited paper is currently under review, I would like to thank Rodrigo Benenson for sharing his findings before publication.

which remain unsatisfactory [Dollár, a], while further decreasing the amount of False Positive detections. On a longer time scale, I expect that the focus will shift on detecting people imaged in arbitrary poses, releasing some of the constraint of PD.

The description of interesting aspects of the detectors in the state of the art continues in Chapter 3, where the references are grounded in the discussion on the detection-by-classification architecture.

2.2 Data Sets and Benchmarking Techniques for Pedestrian Detection

Data sets for Machine Learning-based visual detectors consist of a collection of annotated images. The purpose of such data sets is twofold. First, the data of the training set is used to extract the positive and negative examples for training a detector. Second, the data of the validation and test set is used at evaluation time to determine which detections are correct and characterize the performance of the detector. However, because of finite size, every data set is bound to represent only a subset of the real world. This implies that every data sets suffers from some level of bias [Torralba and Efros, 2011; Khosla et al., 2012], making a data set acquired in one setting (automotive, industrial, and so on) particularly appropriate for estimating the performance of detectors in that specific setting. For instance, when interested in analysing the performance of PD's in the context of self-driving cars, it is advisable to train and test the detectors on a data set acquired in an automotive setting, e.g., the Caltech pedestrian data set [Dollár et al., 2012b]. The publication of data sets is an important step towards a fair comparison of the performances of PD systems, but it is not sufficient. 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. Advances in performance are furthermore stimulated by associating the publication of data sets with detection contexts, as in the case of the PASCAL Visual Object Classes challenge [Everingham et al., 2010, 2014] or the case of ImageNet [Russakovsky et al., 2104]. The lifetime of a data set

is limited: as the understanding of the problem by the scientific community grows, hurdles are conquered and data sets become obsolete.

Many data sets specific for PD have been published over the years. A first notable example is the MIT pedestrians data set [Oren et al., 1997], 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 [Dalal and Triggs, 2005] was introduced by Dalal and Triggs in 2005. It is divided in training set and test set, it provides both positive and negative examples and it provides full images in which the pedestrians are annotated. The ETH pedestrians data set [Ess et al., 2007] was introduced in 2007. It was recorded with a mobile platform moving along a sidewalk, equipped with a stereo camera. It depicts a scenario typical for a mobile robot. The TUD-MotionPairs/TUD-Brussels data set [Wojek et al., 2009] (TUD) and the Caltech pedestrian data set [Dollár et al., 2012b] 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). Another data set worth mentioning is the KITTI Vision Benchmark Suite [Geiger et al., 2012, 2013], released in 2012. The KITTI data set allows for evaluating algorithms on several visual tasks performed in the automotive context, including the detection of pedestrians and the estimation of their 3D orientation.

In spite of having been published in 2005, the INRIA data set [Dalal and Triggs, 2005] is still very commonly used both for training and for evaluating PD's. The performance of the detectors on that data set has been improving steadily: the missed detection rate at 0.1 False Positives Per Image (FPPI) has dropped from around 50% to around 20% since its publication (see [Dollár et al., 2012b]). Yet, there is still room for improvement, which explains why that data set is still widely used as a benchmark [Dollár et al., 2012a; Pedersoli and Vedaldi, 2011; Sangineto et al., 2012; Benenson et al., 2012] and for training PD's: 13 out of 16 algorithms reviewed in [Dollár et al., 2012b] are trained on it. However, its labelling is starting to show its limitations: many people appearing in the images are not labelled, there is no specific label for image areas which are ambiguous and there is no indication on the visibility ratio of each person. In Chapter 4 I discuss the concept of sample purity and describe the new labelling for the INRIA data

set that I created for performing experiments on the role of sample purity in the training and evaluation of PD's.

2.3 Applications of Pedestrian Detection

2.3.1 Automotive

The automated detection of people in urban environments has been a subject of research since as early as 1969 [Bartlett, 1969]. The main goal of PD from a moving vehicle is to avoid vehicle-pedestrian collisions, either by warning the driver of a potentially dangerous situation or by actively slowing the vehicle down. Reducing the number of accidents involving pedestrians has a big impact on society: authorities report 4743 pedestrians fatalities in the USA, during 2012 [USA, 2012]. Automotive applications arguably drive the development of visual Pedestrian Detectors, as confirmed by the number of influential PD data sets collected in urban road environments [Oren et al., 1997; Ess et al., 2007; Wojek et al., 2009; Dollár et al., 2012b]. Smart vehicles exploit rich sensors such as lidars for PD [Broggi et al., 2009], among other tasks, but PD's based exclusively on vision are advantageous in terms of price.

Requiring that a PD system be run on a car forces several constraints on the algorithm and its implementation: the system has to operate in real time and it needs to be robust in the face of swift illumination changes and varying weather conditions. Furthermore, the system has to be able to detect people when they are far enough from the vehicle, so that appropriate action can be taken to avoid a collision. This translates to the PD being able to detect people who appear small on the acquired images. In the case of an active collision avoidance system, it is important that virtually no false alarms are generated, because a false alarm which slows the vehicle down without a reason creates a dangerous situation in the flow of traffic.

2.3.2 Surveillance

The goals of Video Surveillance (VS) include detecting, counting and tracking objects of interest (people, vehicles, etc.), as well as recognizing the activities being performed by such objects. The ability to detect humans is important and can be instrumental in achieving several of the goals of

VS. The traditional pipeline of VS consists in three steps. First, motion is detected (typically through techniques of Background Subtraction [Zhao and Nevatia, 2004; Cristani et al., 2010; Brutzer et al., 2011]) and areas of the image are segmented based on it, then each segment is classified into one of the classes of interesting objects and, eventually, objects are tracked over time [Hu et al., 2004; Cristani et al., 2013]. One VS problem which has attracted significant attention in the recent past is that of person RE-ID in camera networks. Given a set of pictures of previously observed persons, a practical RE-ID system must locate and recognise such people in the stream of images flowing from the camera network.

Several data sets have been proposed for benchmarking VS algorithms, including the ones specific to PD I mentioned in Section 2.2. The CAVIAR [CAV] data set, introduced in 2004, was one of the first to provide video sequences instead of single frames. It was also the first to provide annotations of people location, identity and activity, making this data set useful for many problems. The CAVIAR 1^{st} set was acquired with a single camera at INRIA Labs Grenoble for activity recognition. CAVIAR's 2^{nd} set was acquired in a shopping mall in Lisbon using two cameras with overlapping fields of view, making it more interesting for RE-ID and multi-camera tracking. In 2011, some sequences of CAVIAR were customised for the RE-ID problem and compiled into one data set (CAVIAR4REID [Cheng et al., 2011a]), which contains a total of 72 individuals: 50 appearing in two camera views and 22 appearing in just one. The PETS data set, of which various extensions were published over time (see for instance [James and Ali, 2009]), targets the research fields of people tracking, people counting, crowd analysis and action recognition. In 2011 a collection of 90 videos from publicly available data sets was compiled into the PDds data set [García-Martín et al., 2012] . Such videos were labelled uniformly for PD and object classification tasks. Overall the collection contains 28358 frames, divided in 16 different subclasses depending on the complexity of background texture, as well as on people appearance variability and people/object interactions. This is probably one of the most complete data sets for VS but it lacks scenarios in which the same persons are imaged from different cameras. This hinders its usefulness for the RE-ID and multi-camera tracking community. Besides CAVIAR4REID, other data sets have been designed specially for the RE-ID problem. The i-LIDS data set for Re-Identification [Zheng et al., 2009] (pub-

lished in 2009) contains appearances of 119 people and was built from the i-LIDS Multiple-Camera Tracking Scenario. The presence of occlusions and quite large illumination changes makes the RE-ID task challenging on i-LIDS. The VIPeR data set [Gray and Tao, 2008], introduced in 2007, contains two views of 632 pedestrians captured from different viewpoints.

Most recent state-of-the-art algorithms in RE-ID focus on the matching problem: they require manually cropped rectangular Bounding Boxes (BB's) enclosing people, both for training (gallery) and for testing (probes) [Zhao et al., 2014; Liu et al., 2014]. However, in most RE-ID applications of interest, it is necessary to detect the location and size of people in the images in an automated way.

One choice for detecting people in a network of cameras is following the classical VS pipeline. But relying on movement detection (i.e., with Background Subtraction) is subject to some shortcomings: swift illumination changes, changes in the camera gain, shadows, movements of objects different than people (fluttering tree leaves, pets, robots, vehicles, etc.) can all be sources of false detections. Furthermore, since Background Subtraction techniques rely on the assumption of a static background, they do not adapt easily to the case of moving cameras (pan–tilt–zoom cameras, cameras mounted on robots or unstable cameras which shake due to the wind). Extensions of Background Subtraction that work with moving cameras exist [Sheikh et al., 2009], but I am not aware of any work applying them to the PD problem.

One class of algorithms which allows for the detection of people both in static and moving cameras is that of pattern recognition-based Pedestrian Detectors. Such detectors can only detect people assuming a limited range of poses, but are largely immune to the problems that affect Background Subtraction, making them a better candidate for detecting people in an automated RE-ID system.

Methods that actively integrate PD and RE-ID are still scarce in the literature. In [Corvee et al., 2012; Bak et al., 2012] pedestrian are detected and tracked, then each track is associated with a person ID. The work presented in [Mogelmose et al., 2013] relies on richer (RGB-D) sensors and, like the aforementioned approaches, employs temporal filtering in an attempt to provide clean data to the RE-ID module. None of these works studies the influence of detection errors on the performance of the integrated system.

The authors of[Li et al., 2014] compare RE-ID performance using either hand-labelled or automatically detected Bounding Boxes, but limit the analysis to the effect of misaligned Bounding Boxes, ignoring False Positive detections and Missed Detections. InChapter 6 I report my work on the integration of a PD and a RE-ID system, including two innovation which lead to a better performance of the PD+REID system.

2.3.3 Human–Robot Interaction

Human–Robot Interaction (HRI) is the field of study which focusses on the natural, safe and efficient interaction between humans and autonomous devices. Estimating the presence and the position of people in the surrounding of a robot is fundamental for such interaction to occur. The level of success in this task is typically measured by the "human awareness" metric [Steinfeld et al., 2006].

HRI applications aim at detecting people regardless of their poses (people can be standing, sitting, lying down, etc.), in such cases generic Human Detectors (HD) should be used instead of PD's, which rely on the assumption of a heavily restricted pose range. Nonetheless, Pedestrians Detectors fulfil the requirements of some HRI applications and are employed in such cases.

Human Detection has been used as part of HRI systems with diverse goals, ranging from enabling general interaction between humans and robots (see [Ruiz-del Solar et al., 2013; Jafari et al., 2014; Naseer et al., 2013]), to ensuring safety in industrial environments (see [Morzinger et al., 2011; Rybski et al., 2012]), to providing care to elderly people (see [Gross et al., 2011; Volkhardt et al., 2013]). Human Detection technology employs different sensors (and combinations of sensors), including visible-light cameras, laser rangefinders, infrared cameras and RGBD cameras, with the latter being very common in recent years. RGBD sensors provide an advantage in terms of detection performance, but visible-light cameras coupled with a PD perform better in some configurations. For instance, using one omnidirectional camera [Mekonnen et al., 2013] allows the robot to monitor a much larger portion of the surrounding area than would be possible with an RGBD sensor.

Chapter 3

Standard Architecture of Pedestrian Detectors

In this chapter I describe the detection-by-classification approach for Pedestrian Detection (PD), which is the *de facto* standard architecture for PD. The building blocks of such architecture are: feature extraction, a window classifier based on Machine Learning (ML), the sliding window scan pattern, image pyramids and Non-Maximum Suppression. An overview on how the modules are used in the system can be seen in Figure 3.1. I corroborate the description of the architecture with the details from a concrete example, my implementation of the Fastest Pedestrian Detector in the West (FPDW) [Dollár et al., 2010] and with considerations related to detection accuracy and speed. I choose to implement FPDW because its detection speed makes it suitable for robotic applications and because of its good detection accuracy. Eventually, in Section 3.4 I list the detectors I implemented and report the experiments I performed in seeking to improve their performance.

I start by describing the window classifier, because it is the central module for PD. I assume the use of generic features during its discussion (Section 3.1). I analyse feature extraction in Section 3.2 and describe the methods used to detect pedestrians imaged with arbitrary size and in arbitrary position (the sliding window approach, image pyramids and Non-Maximum Suppression) in Section 3.3.

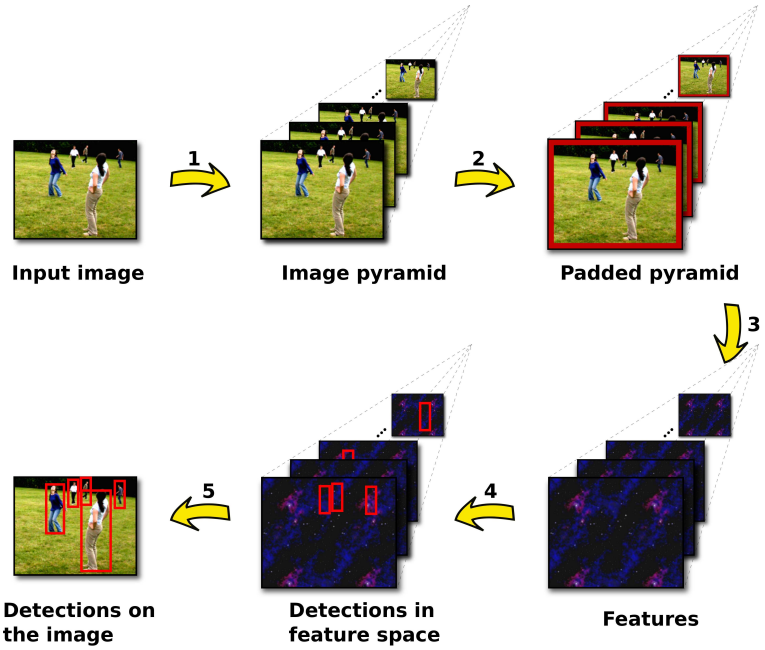


Figure 3.1: Visualization of the work flow of a standard PD system. Step 1: construction of the image pyramid. Step 2: padding of the image pyramid. Step 3: computation of the image features. Step 4: sliding-window-based detection. Step 5: Non-Maximum Suppression and mapping of the detections to the input image.

3.1 Window Classifier

The fundamental block of a detection-by-classification PD is the window classifier. It takes as input the features extracted from one image window of fixed size and computes a confidence value, also known as score. The score expresses the confidence of the classifier on the presence of a person in the window at hand. The function which maps a point in feature space to a score is learned using a set of labelled examples. A vector of features is extracted from each training window and an ML-based algorithm is used to learn the confidence function. Positive image windows for training are chosen so that each contains a centred pedestrian, while the negative ones are chosen so that they do not contain pedestrians. All the collected windows are scaled to match the size of the window classifier.

The most common ML methods used in PD are Support Vector Machines (SVM's) [Cortes and Vapnik, 1995] and AdaBoost [Freund and Schapire, 1995], I briefly describe SVM's, while focussing more on AdaBoost. This

stems from AdaBoost leading to the implementation of faster detectors, which are essential for robotics applications. SVM's are linear classifiers whose goal is to compute the optimal separating hyperplane between two classes of examples. Optimality is defined in terms of margin: in the basic case of linearly separable classes, the optimal hyperplane is defined as the one which maximizes the distance to the closest point in the training set. SVM's have been extended to work with non-separable classes and to produce non-linear classification boundaries. Non-linear classification abilities are achieved via the "kernel trick": a linear decision boundary is applied in a large, transformed version of the feature space which projects to a nonlinear boundary in the original feature space. The "kernel trick" consists in choosing the function that maps from the original to the enlarged feature space so that dot products in the transformed space can be computed quickly in terms of the variables in the original space. Variants of SVM's applied to the PD problem include Histogram Intersection Kernel SVM [Maji et al., 2008], latent SVM [Felzenszwalb et al., 2010], and multiple kernel SVM [Vedaldi et al., 2009].

3.1.1 AdaBoost

AdaBoost is an ensemble method which builds a Strong Classifier (StC) as the combination of a number of Weak Classifiers (WkC). WkC's are defined as simple classifiers which attain a classification accuracy above the level of chance. The most basic WkC is the decision stump: it computes its classification comparing the value assumed by one feature with a threshold. The most successful kind of WkC used in PD is the depth-2 decision tree (see Figure 3.2). It consists in the connection of three decision stumps, arranged as a tree. A depth-2 tree is more powerful than a stump because it bases its decisions on the value of two features, rather than just one. The superiority of depth-2 trees for PD has been confirmed empirically in [Dollár et al., 2009], [Benenson et al., 2013] and [Benenson et al., 2014a]. Using deeper trees is not common in PD, possibly due to the increased tendency of more complex models towards overfitting. In my implementation of FPDW, I use depth-2 trees as WkC's.

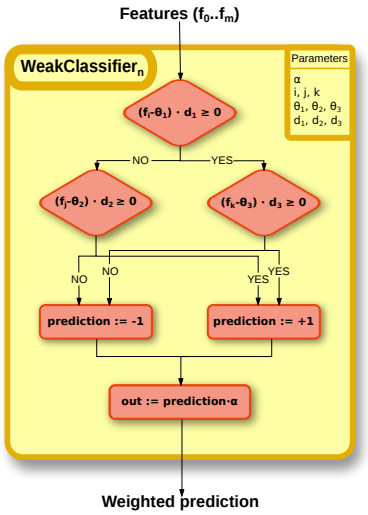


Figure 3.2: Depth-2 decision tree: a popular Weak Classifier for AdaBoost-based PD. The value assumed by a first feature is compared to a threshold. Based on the result, one of the two nodes of the second level is activated. This triggers the evaluation of a second feature against a specific threshold. This last result decides the sign of the prediction computed by the Weak Classifier. The confidence which weights the prediction (α) is learnt during the training. The features used in the three nodes of a depth-2 classifier are, in general, different from each other.

Algorithm 1 AdaBoost (adapted from [Viola and Jones, 2001])

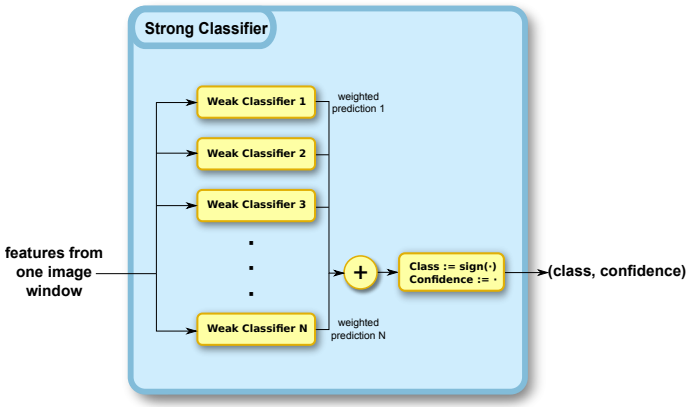
Input data consists in the features computed from each example window (x_i) and the corresponding label (y_i, which assumes the values 0 and 1 for negative and positive examples, respectively): $(x_1, y_1), \ldots (x_n, y_n)$. The number of negative and positive examples are l and m.

1: Initialize weights $w_{1,i} = \frac{1}{2l}, \frac{1}{2m}$ for negative and positive examples, respectively.
2: **for** $t = 1$ to T **do**
3: Normalize the weights: $w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$
4: Select the weak classifier (h_t) which achieves the lowest error, given the current weights. Error is computed as: $\epsilon_j = \sum_i w_i \left| h_j(x_i) - y_i \right|$.
5: Update the weights: $w_{t+1,i} = w_{t,i}\beta_t^{1-e_i}$, where $e_i = 0$ if example x_i is classified correctly, $e_i = 1$ otherwise, and $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$.
6: **end for**

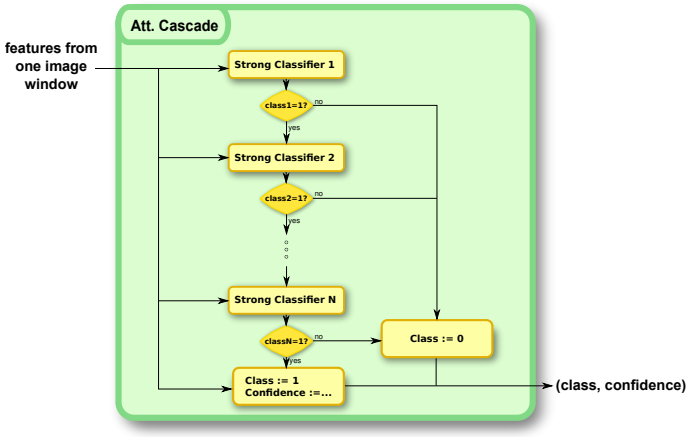
The final strong classifier is: $h(x) = \begin{cases} 1, & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0, & \text{otherwise.} \end{cases}$

(where $\alpha_t = \frac{1}{\beta_t}$)

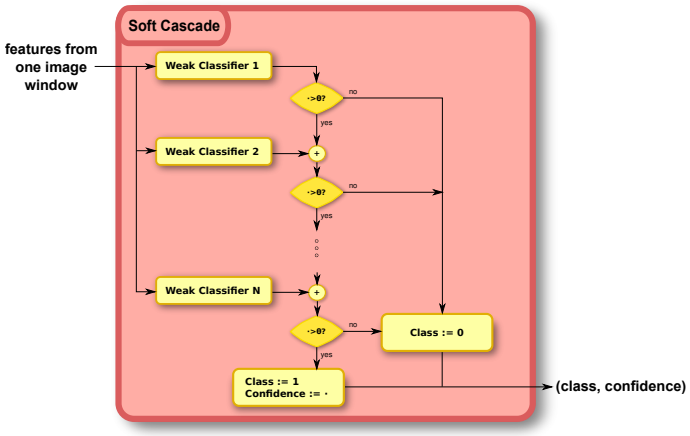
The learning algorithm of AdaBoost starts by assigning weights to the examples of the training set, then it works iteratively alternating two steps. First, it learns a WkC based on the current weights of the examples and assigns it a confidence value based on how well it performs. Second, it adjusts the weights of the examples based on the classification obtained: the weight of the examples which are misclassified by the latest WkC are increased, the others decreased. This guarantees that the subsequent WkC gives more importance to the examples which are misclassified at this stage. See Algorithm 1 for the original formulation of AdaBoost.



(a) Basic Strong Classifier



(b) Attentional Cascade



(c) Soft Cascade

Figure 3.3: Three different ways for building a Strong Classifier based on Weak Classifiers: Basic Strong Classifier (a), Attentional Cascade (b), Soft Cascade (c). In (a) all the Weak Classifiers in the structure must be evaluated in order to classify one example. In (b) the Weak Classifiers are grouped in several Strong Classifiers. The evaluation can be interrupted after the evaluation of each Strong Classifier. In (c) the evaluation can be interrupted after the evaluation of each Weak Classifier.

At classification time, each WkC produces a classification (either positive or negative) which is weighted by its confidence value. In the basic StC built with AdaBoost, the output is the sign of the sum of the weighted votes of all the WkC. In PD the signum operation is omitted and the sum of the weighted votes is interpreted as the confidence on the detection. Examples with a confidence below a user-specified threshold (typically zero) are classified as negative (see Figure 3.3(b)). I label the resulting StC as Basic Strong Classifier.

3.1.2 Attentional Cascade and Soft Cascades

In [Viola and Jones, 2001, 2004] Viola and Jones introduced the Attentional Cascade: noticing that the majority of the image windows to be evaluated are negatives (see Section 3.3 for an explanation of this phenomenon), they devised a classifier based on a series of StC's: each StC is designed to reject a large fraction of the False Positive examples, while allowing most True Positive examples to continue towards the following StC. All the examples that a StC classifies as positives are passed on to the following StC in the series (see Figure 3.3(b)). Simple StC's (StC's built using few WkC's) are located at the beginning of the cascade, while increasingly complex StC's follow. This architecture has the effect of rejecting easy negative examples with little computation, focussing the computation effort on discriminating positive examples from the hard negative ones. The resulting system is much faster than the Basic Strong Classifier, but building it is cumbersome (see [Viola and Jones, 2004]).

In [Zhang and Viola, 2007], Zhang and Viola introduced a variant of Attentional Cascade which is much easier to build, the Soft Cascade. Instead of using StC's as building blocks, Soft Cascades use WkC's: decisions on example rejection are taken after the evaluation of each WkC (see Figure 3.3(c)). In my implementation of FPDW, I use AdaBoost to build a Soft Cascade consisting of 1 000 WkC's.

In spite of the very different nature of the AdaBoost and SVM algorithms, the detection accuracy they exhibit in the PD problem is similar (see [Benenson et al., 2014a]). One key difference between the two methods resides in detection speed: the fastest PD systems in the state of the art are all based on AdaBoost and some variant of Soft Cascades (see [Dollár et al., 2012a; Benenson et al., 2013; Dollár et al., 2014]). As a note I observe that

training an AdaBoost-based classifier is slower than training an SVM-based one (in spite of the coding optimizations and approximations commonly used in practice): AdaBoost consists in an iterative process, while computing the optimal hyperplane for SVM's is a convex optimization problem.

3.1.3 Monolithic VS Part-Based Classifiers

Part-based classifiers differ from the monolithic approach I assumed so far: monolithic classifiers acknowledge the articulation of the human body only implicitly while part-based models explicitly model body parts displacements. The very successful Deformable Parts Model (DPM) detector by Felzenszwalb et al. [Felzenszwalb et al., 2010] uses a monolithic detector as a base (root detector), then refines the score of a detection based on the detection of body parts. The best location for each part inside the root detection window is determined based both on the appearance of the part and on the distance of such part from its ideal placement. The final score for the detection of a person is a combination of the score of the root detector and the score of the parts detectors, weighted by how expected their placement inside the root detector window is. In spite of their popularity, part-based detectors are outperformed by the monolithic ones in the task (see [Benenson et al., 2014a]).

3.1.4 Bootstrapping

The PD problem is intrinsically unbalanced: the goal is to learn a classifier which can differentiate between the appearance of pedestrians and the appearance of anything else. One consequence of this is that generating examples for the negative class is easy: any window on one image which does not contain a pedestrian can be used as a negative example. The abundance of negative examples poses a practical problem: training a classifier with an enormous number of examples can be unnecessarily slow or otherwise demanding in terms of computational resources. Bootstrapping allows for the use of great numbers of negative examples without the need of employing them all at the same time. In the boostrapping framework, a detector is trained multiple times, alternating two steps: the training and the mining for hard negatives. The mining for hard negatives is performed running the detector at hand on a set of negative images: images which do not contain

pedestrians. Each image corresponds to several thousands image windows. Every detection the detector generates in such cases corresponds to an error: a hard negative example that the detector was not able to classify correctly. Such negatives are collected and used to augment the negative training set for the next epoch of training. In the PD community, it is common to use 3–5 epochs of bootstrapping, after which the advantages provided by the method tend to fade.

3.2 Feature Extraction

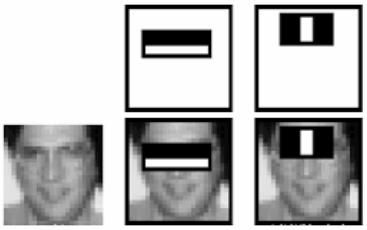
The feature extraction stage of the PD architecture has the goal of filtering the raw pixel information from the input image, mapping it to meaningful mid-level features which afford better classification at the following stage. Different kinds of features were designed over time to extract different types of information, ranging from local brightness, to colour, oriented gradients, texture, etc. Global features such as Principal Component Analysis (PCA) were employed in PD [Munder and Gavrila, 2006]. Each global feature is computed as a function of the value of all the pixels in the detection window. In contrast, local features are computed based on the values of local subset of pixels. Local features have been shown to be more effective than the global ones [Munder and Gavrila, 2006], so I focus the discussion on the former.

Oren et al. [Oren et al., 1997] introduced the use of Haar-like wavelets as features for PD. Such features define rectangular areas on the detection window and compare the average brightness of sets of such areas (see Figure 3.4(a)). The bright-and-dark pattern features computed by the Haar-like wavelets have proven extremely successful in the detection of frontal faces [Viola and Jones, 2004], but failed to perform at the same level in the case of PD. Research soon moved its focus to features describing image edges. The system described in [Gavrila, 2000] uses edge templates and the Chamfer system (a hierarchical shape-matching scheme based on distance transforms) as the first step of a Pedestrian Detector. For that system, one template corresponds to the entire silhouette of a person in a specific pose. The shape-matching step is then followed by a Radial Basis Function classifier. In [Leibe et al., 2005], edge templates and the Chamfer system are used in combination with a segmentation algorithm based on local information.

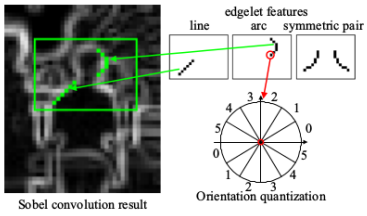
Introduced in [Wu and Nevatia, 2005, 2007], Edgelets are an example of edge-based features for which each feature encodes the characteristics of a short segment of the pedestrian's silhouette (see Figure 3.4(b)).

Arguably the single greatest improvement in PD performance came with the introduction of the Histograms of Oriented Gradients (HOG) features [Dalal and Triggs, 2005]. HOG compute descriptors over image cells (small square portions of image windows) pooling the gradient information. Different gradient orientations contribute to different bins of the resulting histogram (see Figure 3.4(c)). Image gradient information is inherently richer than edge information, as it can represent soft transitions as well as abrupt transitions in image brightness. Another advantage of the HOG features is the robustness to small image variations: the spatial pooling of the gradients used in the construction of the histograms provides invariance towards small image translations and changes in gradient patterns (i.e., one straight edge imaged in a cell will produce similar features to multiple shorter edges with a similar orientation). Robustness in the face of slight changes in gradient orientation is afforded by the angular binning: gradients within a range of orientations all contribute to the same angular bin.

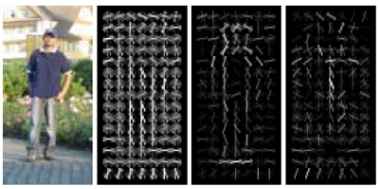
The Integral Channel Features (ICF) detector [Dollár et al., 2009] and other many other modern PD methods based on ICF (including FPDW and the Roerei detector [Benenson et al., 2013]) use features closely related to HOG, albeit designed for a cleaner integration in the detection-by-classification paradigm. The extraction of ICF consists in computing image channels (transformations of the input image) and building features in the shape of the sum over one rectangular region of one such channel. Some of the channels encode information about gradient along a specific direction, while others represent the colour and brightness information of the input image in the LUV colour space. Finally, one channel encodes the module of the image gradient (see Figure 3.4(d)). As a result, this kind of features can encode robust gradient information, but also robust brightness and colour information. Robustness is a result of spatial pooling and angular binning, much in the same way as it is for the HOG features.



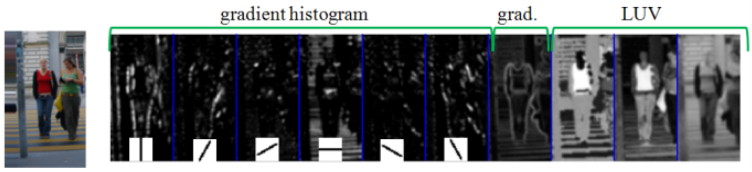
(a) Haar-like features, in a face detection application. Image reproduced from [Viola and Jones, 2001]



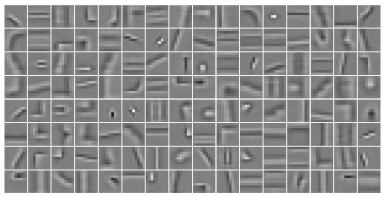
(b) Edgelets, image reproduced from [Wu and Nevatia, 2005]



(c) Histograms of Oriented Gradients, image reproduced from [Dalal and Triggs, 2005]



(d) Integral Channel Features, image reproduced from [Dollár et al., 2009]



(e) First layer of filters learned in a convolutional network, image reproduced from [Sermanet et al., 2013]

Figure 3.4: A visualization of different kinds of features: Haar-like features (based on image brightness), Edgelets (based on specific edges), HOG (based on gradient information), ICF (based on gradient, brightness and colour information) and the low level filters learned in a convolutional network model.

Other interesting features proposed over the years include the Local Binary Pattern [Wang et al., 2009] (used to encode information on texture),

Color Self Similarity[Walk et al., 2010] (a meta feature able to represent relationships such as: "the color in the area of the left and right shoulder usually matches"), and Shapelet [Sabzmeydani and Mori, 2007] (mid-level features built on parts of the detection window, aggregating gradient responses). Clearly, the quality of one input image affects that of the computed features and, eventually, the detection performance. Normalizing input images prior to computing the features has been shown to positively affect the detection performance (see [Benenson et al., 2013]). Strong of the recent success of Deep Learning approaches in a variety of CV problems: Convolutional Networks were applied to PD for the first time in [Sermanet et al., 2013], while [Luo et al., 2014] introduces Switchable Restricted Boltzmann Machines in the context of Switchable Deep Networks. The performance of such detectors is on par with that of other methods in the state of the art (see [Benenson et al., 2014a; Hosang et al., 2015]).

3.2.1 Padding

The classification windows used in PD usually do not enclose a pedestrian tightly: this stems from the observation that the information in the area around the object of interest (its context) can provide valuable information for the classification, e.g., pedestrians usually stand on a sidewalk or on the street. Thus, the detection windows for PD are designed to include a padding area around the person (see Figure 3.5). This choice has implications on the whole detection system: training examples must include padding, test images have to be padded (see the following section for details on how this affects the construction of the image pyramid), care should be taken at detection time to remove the padding space when computing the detection BB's.

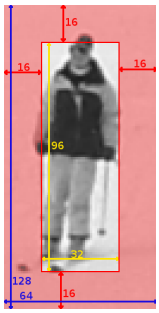


Figure 3.5: Padding for the detection window: the detection window is not designed to enclose pedestrians tightly. It instead includes a region surrounding the pedestrian, with the intent of exploiting context information.

So far I focussed on the simple case of monocular, single-frame PD. Extensions to the binocular (stereo) case and to systems which take into account video sequences (multiple frames) are straightforward: features are computed not only starting from the input image, but also from depth information (estimated via stereo) and movement information, typically estimated as a function of optic flow [Dalal et al., 2006; Walk et al., 2010].

My implementation of FPDW is based on the Integral Channel Features and uses a detection window of 96×32 pixels, plus a 16-pixel padding. I use the code by the author of FPDW to compute the image channels and my implementation for the Integral Images and the computation of the features.

3.3 Invariance to Position and Scale

In the previous sections, I examined the window classifier: a system which aims at detecting pedestrians when presented with an image window of the correct size, exhibiting a pedestrian of the correct size centred along its vertical axis. The utility of the window classifier *per se* is very limited, but when combined with the sliding window scan scheme and image pyramids it allows for the detection of all the pedestrians in one image.

3.3.1 Sliding Window

The sliding window approach makes it possible to use such a classifier to detect pedestrians of a given size anywhere in one image. It consists in running the classifier on a grid of locations on one image, obtaining a response

for each point on the grid (see Figure 3.6). One common value for the distance of any two points on the grid is 4 pixels for both the vertical and the horizontal direction. In a typical case, the majority of the points on the grid will be classified as "no person" by the detector, while the rest will be classified as "person" and will be associated with a confidence value.

Figure 3.6: The sliding window approach to detection: a window classifier is run on a grid of locations on the image. This allows for the detection of pedestrians of a fixed height all over the image. The padding area of the detection window is marked with a blue shading.

3.3.2 Image Pyramids

In order to detect pedestrians of different sizes, it is possible to run the sliding window on several scalings of the input image: the image pyramid. In PD, it is common to use pyramids with 8–16 layers per octave (one octave being the size range which goes from one image height to its half or its double). Running the same window classifier on a shrunk version of the input image corresponds to running a larger window classifier on the input image (see Figure 3.7). One detection obtained using this scheme corresponds to a detection BB of specified size and position on the input image.

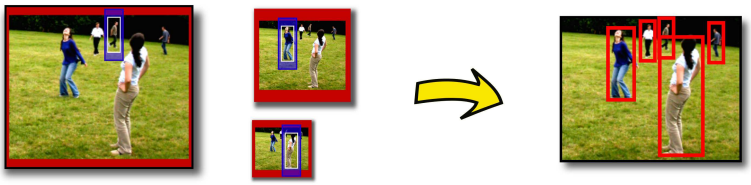


Figure 3.7: Detecting pedestrians of different sizes: the same classification window is applied to all the layers of the image pyramid (left). Applying the same window on a shrunk version of the image corresponds to applying a bigger detection window on the raw input image and results in the detection of taller pedestrians (right). Maintaining the dimensions of the detection window for the entire pyramid requires that the size of the padding on each layer be the same.

3.3.3 Border Effects

One detail is important when detecting people imaged close to the image borders: image padding needs to be done after image scaling. This ensures that the padding for every layer of the pyramid has the same size of that assumed by the detection window. In case padding is applied to the raw input image and it is then scaled with it, the resulting amount of padding is different for the different layers of the pyramid, leading to problems. Tall pedestrians close to the image border can fail to be detected due to not having enough padding, while the excess of padding can generate spurious detections when searching for short pedestrians.

3.3.4 Scale and Space Sampling

The spacing of the grid for the sliding window and the number of layers in the image pyramid concur in determining the number of image windows that have to be evaluated for detecting pedestrians on one image. The number resides in the vicinity of 100000 for a VGA image, meaning that the classification problem is very unbalanced: elements of the positive class are extremely less common than those of the negative class.

3.3.5 Non-Maximum Suppression

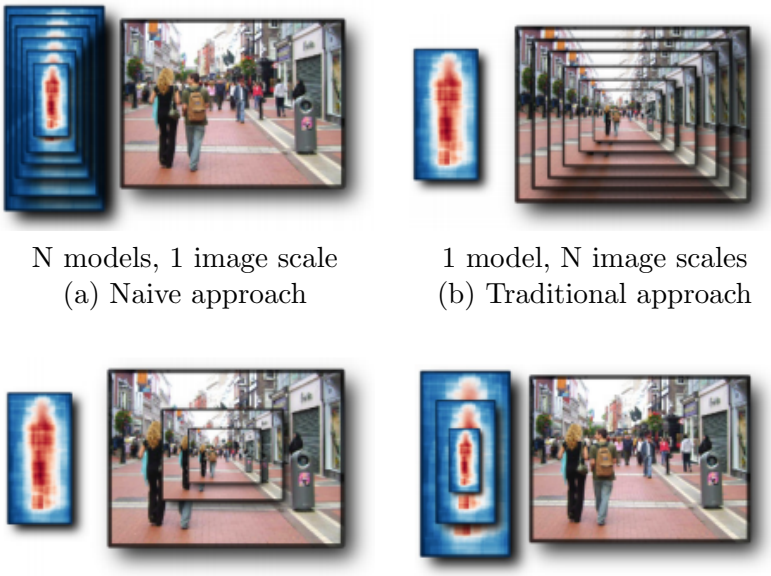
Combining image pyramids with the sliding window approach can give rise to the undesirable presence of multiple detections for a single pedestrian. Non-Maximum Suppression (NMS) techniques are used with the intent of

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. A variety of NMS methods exist. A very effective one consists in forming sets of the detections whose BB's overlap significantly (according to the PASCAL VOC criterion with a 0.6 threshold, see Section 4.1), and comparing the detections of one set pairwise, discarding the least confident detection at each step (see [Dollár et al., 2012b]).

3.3.6 Feature Interpolation and Multiple Models

One major boost in detection speed was introduced with the FPDW detector [Dollár et al., 2010]: exploiting knowledge on natural image statistics, FPDW only computes the features for one image for each octave, while the features for the other layers are estimated with a fast approximation.

Most PD approaches learn one model for the window classifier and apply it to each layer of one image pyramid. This implies disregarding the differences in image formation for objects imaged at different scales. Multi-scale approaches acknowledge such differences by learning one model for each octave of the image pyramid (see Figure 3.8). Such approaches show improved detection performance compared to the basic ones, but the improvements are minor (see [Benenson et al., 2014a]). One advantage provided by such approaches is that the features need to be computed only at the base scale (see [Benenson et al., 2012]), further improving the temporal performance respect to FPDW.



N models, 1 image scale
(a) Naive approach

1 model, N image scales
(b) Traditional approach

1 model, N/K image scales
(c) FPDW approach

N/K models, 1 image scale
(d) VeryFast approach

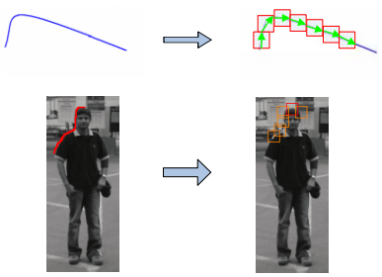
Figure 3.8: Multi scale detection approaches. The naive approach (a) disregards the similarity between pedestrians imaged at similar scales and learns a model for each scale. The traditional approach (b), in a somewhat dual fashion to (a), disregards the difference between pedestrians imaged at different scales and learns only one model. The approach introduced by FPDW (c) uses just one model like (b), but approximates the features in most layers of the image pyramid instead of computing them explicitly. The VeryFast approach presented in [Benenson et al., 2012] uses one model per image octave and computes the features only at the base scale of the image. Image adapted from [Benenson et al., 2012].

3.4 Practical Experience with Pedestrian Detectors

During the work of this thesis I implemented various PD algorithms and explored ways to improve their performance. This work did not lead to detection performance improvements in the state of the art, but the process was very valuable in terms of the experience I gained.

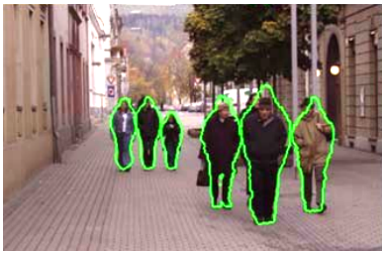
3.4.1 Adaptive Contour Features-based detector

I took the initial steps in PD implementing a simple detector based on AdaBoost and the granule features of Adaptive Contour Features [Gao et al., 2009]. The granules in these features are square patches of different sizes on the detection window. The output of a granule encodes the orientation and the magnitude of the strongest edge present on the corresponding square. One feature consist in one or more chains of such granules. The Adaptive Contour Features are attractive because they encode the contour of the person in a robust way and can be used to perform segmentation as well as detection (see Figure 3.9 for a brief description of the features). In the experiments with this detector I trained and tested on isolated image windows, rather than on full images, as was commonly done at the time. I built just one Strong Classifier combining the output of several Weak Classifiers. I obtained some preliminary results (see Figure 3.10 for a plot of classification error as a function of the number of Weak Classifiers used, also exploring the importance of bootstrap during training), but decided against implementing the full detector because building granule chains is a complex task based on heuristics. Heuristics are not inherently bad, but I decided to focus on more systematic approaches.



(a) One Adaptive Contour Feature, visualizing the granules which constitute it.

(b) Three more complex Adaptive Contour Features.



(c) Detection and segmentation results using Adaptive Contour Features.

Figure 3.9: Adaptive Contour Features. Each square represents a granule, while a set of granules constitutes one feature. Images reproduced from [Gao et al., 2009].

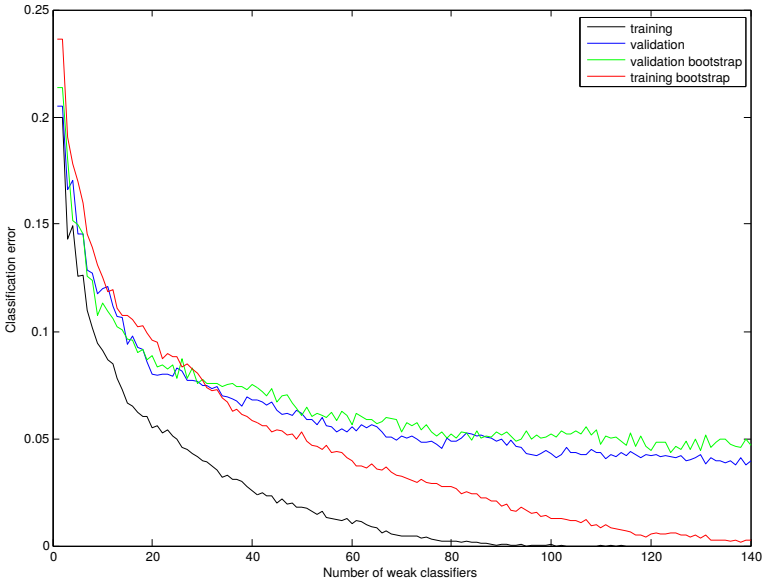


Figure 3.10: Classification error of my granule-based detector as a function of the number of weak classifiers. The training error is higher after one round of bootstrapping than after the first training. This is expected because bootstrapping augments the negative training set with hard-to-classify examples. The error on the validation set is comparable for the two cases, but, in general, a few rounds of bootstrapping are expected to reduce the validation error.

3.4.2 Edgelet-based detector

At a later stage I implemented a monolithic detector based on Edgelets [Wu and Nevatia, 2005, 2007]. Edgelets, as well as Adaptive Contour Features, work with information related to the silhouette of an object. One edgelet is a short segment of line on the detection window. The feature value associated with an edgelet encodes the affinity between the edgelet line and the edge information present on the image (see Figure 3.11). Edgelet features need to be quite dense, my implementation employed almost one million features to cover a 28×54 pixel image window. As a result both training and applying the detector was slow: one full training on the INRIA data set took around one week to be performed. The classifier was built using AdaBoost, but in this case it was an Attentional Cascade rather than a single Strong Classifier: the cascade was composed of 24 Strong Classifiers of increasing complexity. The number of Weak Classifiers used in each Strong Classifier varied between 10 and 120. This detector was implemented including the

sliding window paradigm and image pyramids, so it was able to detect people of arbitrary size and position on one image. However, the Non-Maximum Suppression step was missing, so there was no attempt to merge the multiple detections originated on one person. It was using the Edgelet detector that I developed the considerations on invariance to position and scale reported in Section 3.3. The resulting detector proved to be reasonable in terms of accuracy (see Figure 3.12), but was quite slow and very demanding in terms of working memory. Working on improving the accuracy of a detector which is slow at training or at detection time is impractical. Comparing two versions of the same detector featuring some algorithmic changes means training (and testing) each version several times, using different values for its parameters or even with just a different initialization for the randomizer. Considering that the training of my detector based on edgelets took around one week to complete, I decided to implement another detector.

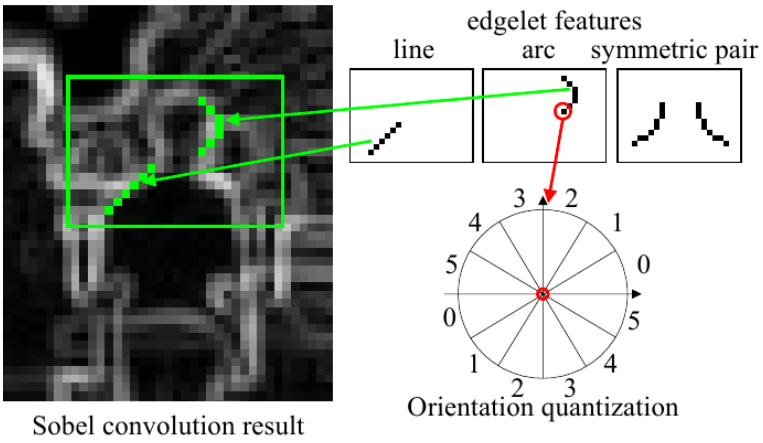


Figure 3.11: Edgelet features and their application on the Sobel response of the input image. One edgelet consists in one short segment of a line defined on the detection window. The more similar the edge information on the image to the shape of one edgelet, the higher the response of such edgelet. Image reproduced from [Wu and Nevatia, 2007]

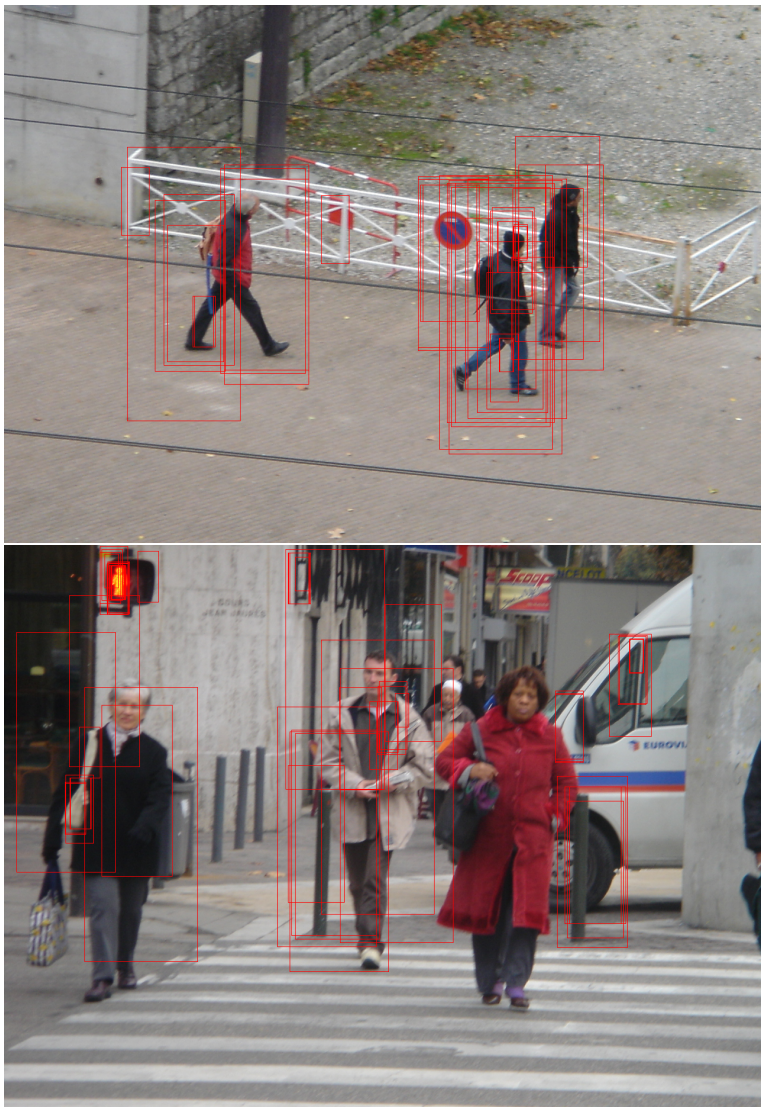


Figure 3.12: Detections generated by my Edgelet-based detector on two representative images from the INRIA data set. The quality of the detections is reasonable, with many True Positives, a few False Positives and the occasional Missed Detection. The lack of the Non-Maximum Suppression step is highlighted by the people which are associated to multiple, similarly sized detections (especially visible in the top image). The goal of Non-Maximum Suppression is to pick only the best of such detections.

3.4.3 HOG-based detector

I implemented a detector based on Histograms of Oriented Gradients (HOG) and AdaBoost. It was both faster and more accurate than the Edgelet-based

detector. I used it for some experiments, for instance the one comparing the detection performance obtained using stumps or depth-2 trees as Weak Classifiers, or the experiment augmenting the negative training set with examples of parts of pedestrians (parts like legs ans arms sometimes happen to generate False Positives). I stopped using this detector when the code for computing the channel pyramids for FPDW was made public by its authors. At that point I implemented a detector based on that.

3.4.4 Implementation of FPDW

I chose to implement the Fastest Pedestrian Detector in the West (FPDW) because of several reasons. First, its speed at detection time makes it suitable for real-time robotics applications. Second, its detection accuracy excelled at the time of its publication. Third, the low number of features it needs to use in order to achieve good accuracy (a few thousands). Such low number of features ensures that training the detector with AdaBoost is fast. Fourth, FPDW is based on the Integrated Channel Features (ICF), which encode gradient, brightness and color information (rather than contour information like Edgelets) and are easily extendable. Furthermore, ICF's have a very clean, systematic design. As already described along this chapter, FPDW is based on ICF features, but it approximates most layers of the image pyramid instead of computing them explicitly. This is the key to its detection speed. FPDW uses AdaBoost to build a Soft Cascade classifier, based on depth-2 decision trees.

3.4.5 The quest for detection accuracy

During the development of this thesis, I performed several experiments aimed at improving the detection accuracy of the detector I was using.

I tried to exploit the observation that sometimes detectors classify as people visual objects that have almost no similarity to a person (see Figure 3.13), giving origin to "weird False Positives". I believe the detector in such cases is missing the big picture: it is accumulating votes from many local features, but missing some more global cues.

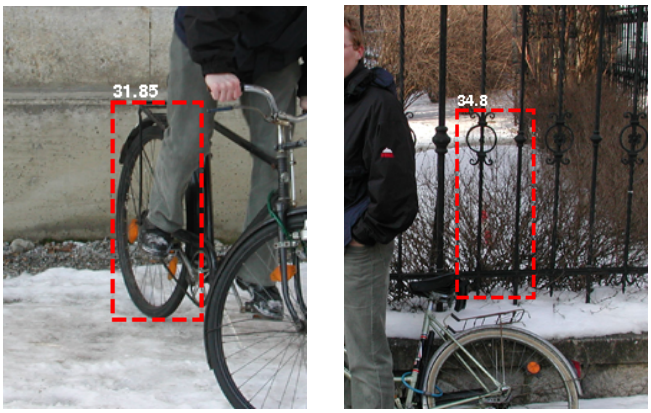


Figure 3.13: Two examples of False Positives (generated by FPDW) that have very little similarity with a person. Notice how the two detections are not centred on some human-like structure and that the detected areas are covered with thin edges.

I devised one experiment which computed features not only on the original image, but also on a segmented version of it. The idea behind this was that image edges can be generated either by object boundaries in the depicted scene (for instance the discontinuity between the leg of a person and the wall in the background), or by very thin structures or object texture. The first kind of edge should still be visible in a segmented version of an image, while the second should disappear. In principle, a detector with access to gradient information computed both on the natural and the segmented image should outperform one with access only to the former. The results of this version of the detector were not convincing, though, possibly because segmentation was performed on the input images at their base size, instead of on each layer of the image pyramid.

Another experiment stemmed by the intuition that the image content on the area occupied by a pedestrian is usually qualitatively different from the content of the area surrounding the person. I added one new feature to ICF, meant to encode the difference in image content between the two areas of a detection window (see Figure 3.14). Including this and other similar features in the pool provided to the classifier made little difference on the classification accuracy.

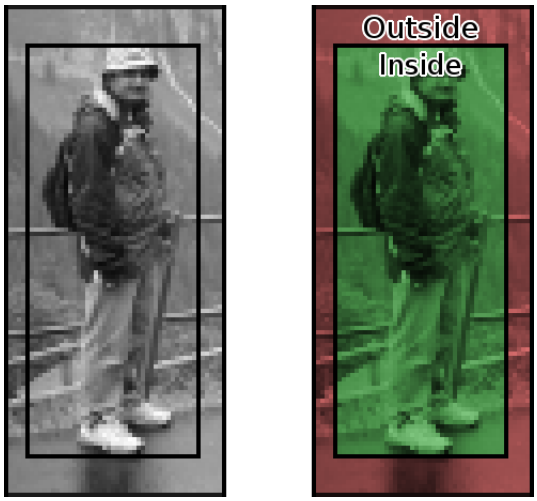


Figure 3.14: One new feature to be added to the pool of ICF: the difference between the area of the image window which is meant to contain the person and the area which is meant to contain the background (labelled as "inside" and "outside", respectively).

One more experiment aimed at eliminating the "weird False Positives" was based on feature co-occurrence. The intuition is that the voting patterns of the weak classifiers can disambiguate between real pedestrians and "weird False Positives". I clustered weak classifiers based on how they voted for the training examples and, at test time, penalized detections for which the voting patterns were not consistent with such clusters. This experiment also led to inconclusive results. In the course of this experiment I developed a tool to visualize the votes of the features of each channel. Some example visualizations are depicted in Figure 3.15.

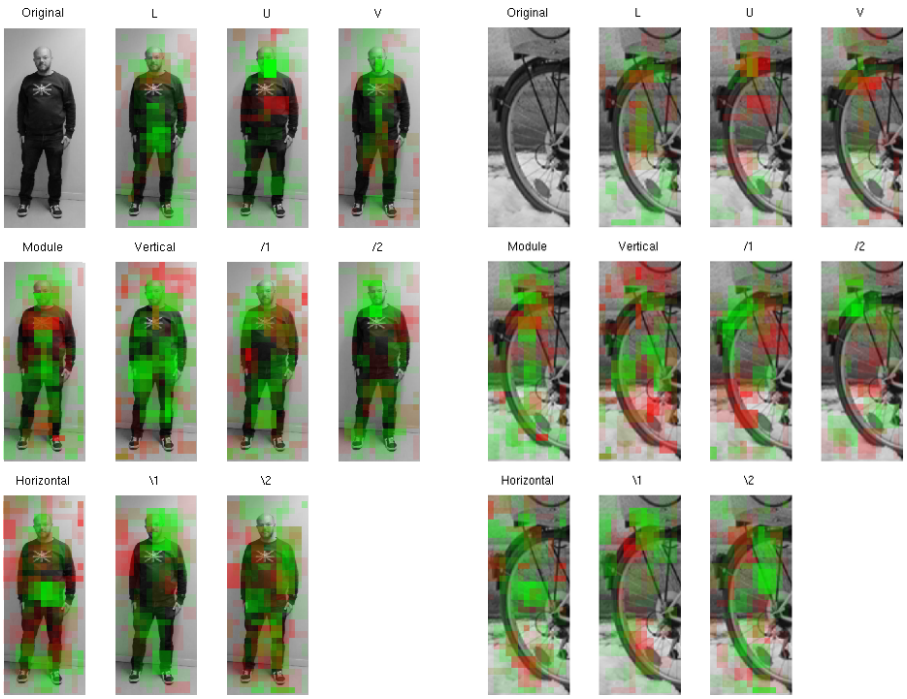


Figure 3.15: Visualization of the voting pattern on the detection of a person (left) and on the False Positive detection originated on a bicycle wheel (right). The vote of each feature is color coded based on its signum (green for positive and red for negative) and weighted according to the vote weight. Areas on the channels marked in green contribute to the detection, while the areas marked in red contribute against the detection.

I devised some other experiments, but their initial findings were inconclusive, so I did not pursue those experiments to their full extent. Considering that some False Positives are generated in correspondence with limbs of people (most notably, legs), I decided to use parts of pedestrians as negative examples during the training. As a result, a reduction of such False Positives was visible before the Non-Maximum Suppression step, but such step obtained almost the same result, while serving other purposes. One experiment regarded the use of image registration versus the use of jitter on training examples, another studied the possibility of inverting the ICF features, in an attempt to visualize the actual information encoded by the features. Yet another experiment explored the possibility of tracking pedestrians in the image pyramid space (see Figure 3.16).

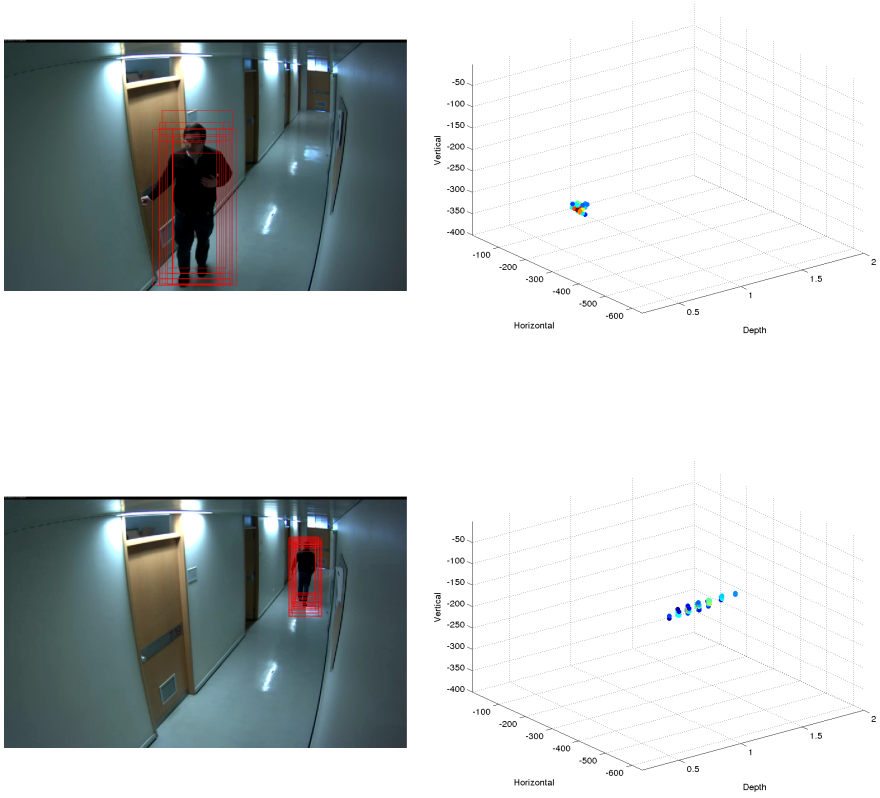


Figure 3.16: A visualization of detections (prior to applying Non-Maximum Suppression) on the image (left) and in a representation of the image pyramid space (right). Warmer colors correspond to higher detection confidences.

Chapter 4

Data Set Labelling and Ground Truth

The training and the evaluation of learning algorithms depend critically on the quality of data samples. I denote as *pure* the samples that identify clearly and without any ambiguity the class of objects of interest. For instance, in PD algorithms, I 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 I 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. I study how a sensible use of impure samples can benefit both the training and the evaluation of PD algorithms. In order to do so, I improve the labelling of one of the most widely used pedestrian data sets (INRIA) taking into account the degree of sample impurity.

I 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, I find 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 PD's and on devising fairer comparison metrics between different algorithms.

For the matters discussed in this document the only relevant person height is that measured on the image (as opposed to the real-world height). Thus, for the sake of conciseness I will write "short pedestrians", meaning "pedestrians imaged in such conditions that their projection on the image is short".

4.1 Labelling for Pedestrian Detection

The purpose of the labelling of a data set for 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. Most PD evaluation schemes 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 (BB's). Each detection BB is associated with a confidence value and is meant to tightly enclose 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 exhibiting the same aspect ratio as the positive ones. They subsequently undergo the same scaling as the positive samples do.

Labelling and Performance Evaluation: 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 Figure 4.1 and [Everingham et al., 2010]):

$$overlap = \frac{area(BB_{GT} \cap BB_{DET})}{area(BB_{GT} \cup BB_{DET})} > 0.5 \tag{4.1}$$

Figure 4.1: One example of one detection and one Ground Truth (GT) Bounding Box (BB) on one image, (a). The areas involved in the computation of a match according to the Pascal VOC criterion (b). A detection and a GT BB's are said to match when the intersection of the two rectangles is larger than half their union.

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. A True Negative occurs when a candidate image window is classified as negative and the corresponding BB does not match a GT BB. True Negatives events are not usually accounted for in the PD setting, because they do not provide insight on the performance of the detection system. Rather, they depend heavily on the parameters of the detection-by-classification approach (step of the scanning grid and number of layer of the image pyramid per octave). See Figure 4.2 for a graphical example of matching outcomes and Table 4.1 for a comparison of the terminology used in the PD and in the Pattern Recognition communities. 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 common variables for summarizing the results of a detection experiment (detection on a set of images) are the Missed Detection (MD) rate and the number of False Positives Per Image (FPPI). The MD rate is defined as the fraction of positive examples in the test set which goes undetected. FPPI is defined as the total number of FP's in the test set, divided by the number

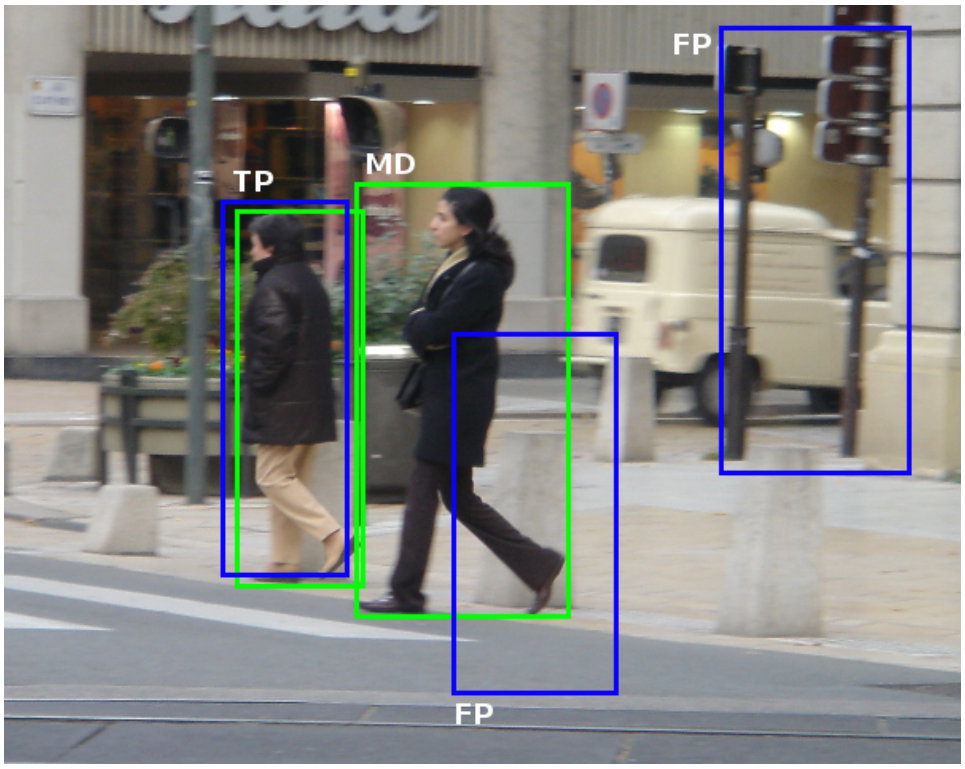


Figure 4.2: Example outcome of matching detection and Ground Truth (GT) Bounding Boxes (BB's): a True Positive (TP) is the result of a correct match between detection and GT BB's. An unmatched GT BB results in a Missed Detection (MD), while an unmatched detection results in a False Positive (FP).

Table 4.1: Possible outputs of the matching process among detection and Ground Truth (GT) Bounding Boxes (BB's). A True Positive is the outcome of a correct match between one detection and a GT BB. A detection that does not match a GT originates a False Positive (FP). A GT BB that is not matched originates a Missed Detection (MD) (often referred to as a False Negative in the Pattern Recognition community). A True Negative occurs when a candidate image window is classified as negative and the corresponding BB does not match a GT BB.

		Ground Truth	
		person	**no person**
Detection Outcome	**detection**	True Positive	False Positive
	no detection	False Negative = Missed Detection	True Negative

of images that constitute it.

$$Missed\ Detection\ rate = \frac{\#Missed\ Detections}{\#Positive\ examples} \tag{4.2}$$

$$False\ Positives\ Per\ Image = \frac{\#False\ Positives}{\#Images} \tag{4.3}$$

Both MD's and FP's are detection errors, thus, the lower the values of MD rate and FPPI, the better the performance of one algorithm. Detectors associate a confidence value (also called a score) with 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 (making them an instance of Detection Error Tradeoff plots), see Figure 4.3 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 characterized 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 [Dollár et al., 2012b] for details).

Design Choices in Labelling: Labelling for Machine Learning is not as straightforward as it seems: visual categories are not exactly defined. Pedestrians, for instance, are defined as people assuming an upright stance, but the exact point at which a posture is so extreme that a person is not regarded as a pedestrian is not defined. Design choices, some of which dic-

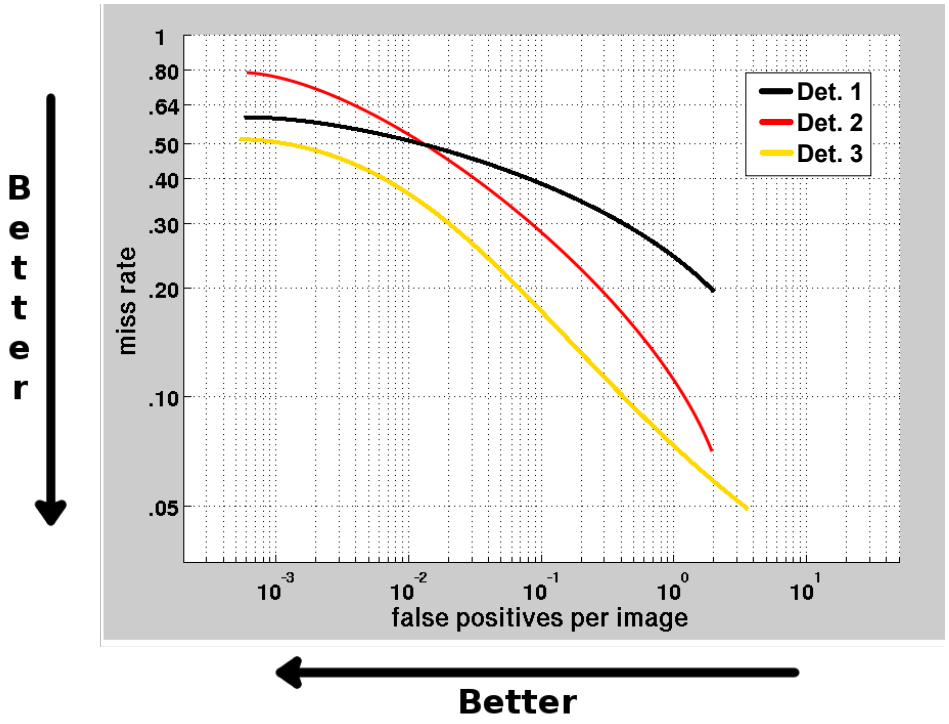


Figure 4.3: An example Missed Detection rate/False Positives Per Image plot. Different curves correspond to different algorithms. Apoint on a curve corresponds to a working point for one algorithm. In this plot the red algorithm dominates the black one for the right side of the plot, while the contrary is true for the left side. The yellow algorithm is the best of the three as it clearly dominates the others for any working point.

tated by the intended application domain for a PD, influence which examples are considered as valid. The authors of the INRIA data set [Dalal, 2005], for instance, consider cyclists as pedestrians. This choice is sensible in terms of automotive applications: both pedestrians and cyclists are vulnerable road users, an Advanced Driver Assistance System (ADAS) should detect both. Another design choice relates the desirability of a PD system to detect things which represent humans: should statues, mannequins and picture of people be detected as people? Should the performance of a PD system be penalized when it fails to detect one of such things? The answer to these questions really depends on the application domain.

The "Ignore" class for labelling was introduced in [Dollár et al., 2009] to acknowledge the fact that there is a grey area at the boundary between the "Pedestrian" and the "Non-Pedestrian" categories and with the insight that both detections and MD's 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. Matching a detection BB with an "Ignore" BB is less strict than the regular matching between a detection BB and a "Person" BB: it 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. This is so because an "Ignore" area can cover more than one object to be ignored (i.e., a group of people so tangled which is impossible to label each person individually), so a partial match is still acceptable.

Defining the Height Range for One Experiment: 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 I mentioned. Failing to do so leads to one algorithm generating detections that cannot possibly match a GT label (when the detections range is wider than the GT range) or to an algorithm not having a chance to detect some

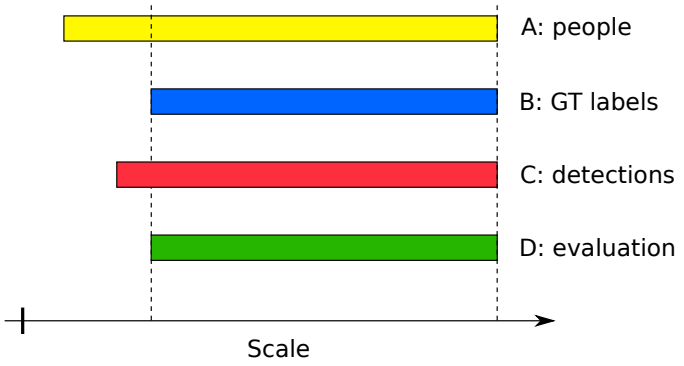


Figure 4.4: Different scale ranges to be taken into account when designing an experiment: A, the range of scales at which people appear in the test set; B, the range of scales at which people are labelled; C, the range of scales spanned by the detections produced by an algorithm and, finally, D, the range of scales taken into account during the evaluation. Correct choices for D are the subsets of the intersections of A, B and C. Other choices for D lead to inconsistencies in the evaluation, e.g., penalizing detectors for detecting unlabelled people.

of the labelled pedestrians (when the detections range is narrower than the GT range). 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 (see Figure 4.4). I apply this concept to the case of the Caltech Benchmark algorithms and the INRIA test set in Section 7.1.1, showing that changing the Minimum Height in TEsting from 50 to 90 pixels leads to a more correct evaluation of the performance of the algorithms.

4.2 Sample purity

Considering the task of visual detection and the image pyramid architecture, it is useful to define the concept of sample purity. I denote as *pure* the samples that represent the class of objects of interest as imaged in ideal conditions, i.e., when they are fully visible and imaged at a resolution larger than the detection window in size. I consider the other samples as *impure*:

the BB's of occluded pedestrians exhibit only part of the visual information a fully visible person would generate. Very small pedestrians generate peculiar image information because of the discretisation of information typical of digital image formation. Other sources of impurity are possible, such as image blur and low image contrast. I have not studied them in the present work because examples affected by high degrees of such sources of impurity are extemely rare in the popular PD data sets. In the literature the occlusion problem is referred to either in terms of the degree of visibility or in that of occlusion. In the context of this thesis the two quantities carry the same information, as highligted by the following equation:

$$occlusion = 1 - visibility \tag{4.4}$$

In the reminder of the thesis I use both quantities, each time selecting the one that makes the exposition clearer.

Purity is not commonly taken into account in PD data sets, resulting in training and evaluation based on a mixture of pure and impure samples. 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, (the *impure* ones), thus reducing its detection rate in a real world application. In testing, it leads to the unfair evaluation and comparison of different detectors since moderately impure samples, when detected, can be accounted for as FP's. Including very impure samples is also detrimental: in training it makes it hard for the learning algorithm to build a model from very difficult examples (e.g., one example of a person in which only one hand is visible or one example of a very small pedestrian). In testing, it requires algorithms to detect pedestrians from very little evidence.

In order to evaluate the effect of sample purity in training and in testing, the GT must be augmented with the visibility information: each BB needs to be labelled with the degree of occlusion the pedestrian is imaged under. The information regarding the image height of each person does not require additional labelling: it is already encoded by the size of each BB.

In the next section I describe the INRIA data set and its original labelling, and I propose a new labelling that allows the effect of sample purity to be studied.

4.2.1 A new Labelling for the INRIA Data Set

I choose to use the INRIA data set for this analysis because it is one of the data sets for PD most frequently used for training detectors. Its original labelling follows closely the general description I presented in the previous section. Each person is labelled with a rectangular BB. Only one label is possible: "UprightPerson", which includes both pedestrians and people riding a bicycle. 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, PD's 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. I discuss each of these factors in the following paragraphs.

The lack of labelling of some of the people present in the data set (see Figure 4.5(a–d)) affects both training and testing. Regarding the training, the lack of such labels prevents researchers to analyse the impact of what I 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 are thus under-reported (see Figure 4.6 for an example of how the performance of the Fastest Pedestrian Detector in the West (FPDW) algorithm [Dollár et al., 2010] 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.

There are, moreover, image patches for which it is difficult to decide

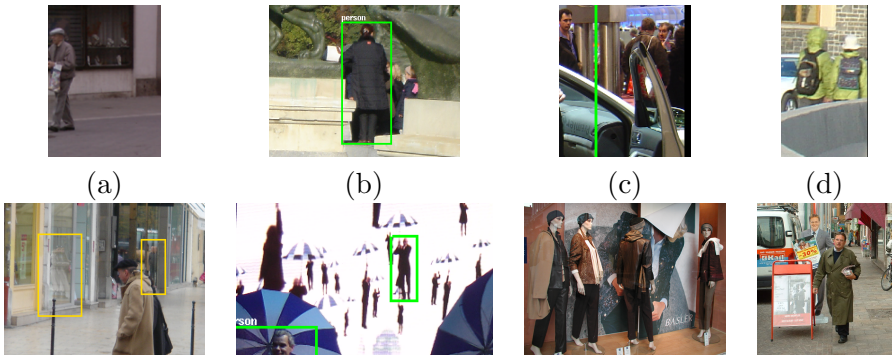
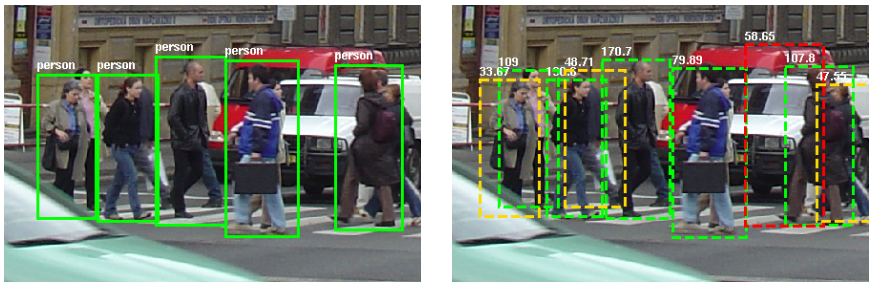


Figure 4.5: Details from the INRIA test set highlighting some limitations. (a–d) Unlabelled persons. (e–h) Ambiguous cases. (e) Reflections of persons on a shop window, not labelled. (f) Some persons drawn on a wall, only one of them is labelled. (g) Some mannequins, all labelled. (h) A poster depicting a man, not labelled.

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 Figure 4.5(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 the pure and impure fraction of the data have 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 in the algorithm introduced with the data set (96 pixels), making them "pure" for my purposes. A small fraction of the labelled pedestrians, though, is 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, I 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



(a) (b)

Figure 4.6: 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 [Dollár et al., 2010] in TP's (green), False Positives (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.

characterized by different degrees of purity.

I propose a new annotation for the data set in which I label all the pedestrians with heights greater than 25 pixels, I associate with each person the estimate of the extent of his/her visible part and mark ambiguous cases (see Figure 4.5(e–h)) as such. The labelling was performed manually. As in the original annotation, I use rectangular Bounding Boxes (BB's) and I consider both cyclists and pedestrians as belonging to the "Person" class. I base my annotations on the scheme introduced in [Dollár et al., 2012b], 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?". I label the test set according to such scheme. The proposed annotation is available on the website of the author of this thesis. For the training set, I do not label groups of people and ambiguous cases as I 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 I choose to use the two labels considering that in the future the two sets can be treated differently. In the evaluation code, the GT BB's are centred horizontally and transformed to assume an aspect ratio of 0.41 (width/height) prior to matching (see [Dollár et al., 2012b] for details).

4.2.2 Assessing the Influence of Impure Samples

In this work, I aim at determining the impact of pure and impure samples in the training and evaluation of a detection system. I 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. I 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 me 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 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".

The experiments exploring the role of sample purity are reported in Section 7.1.2. One experiment confirms that that the degree of partial occlusion of test samples negatively correlates with detection accuracy. Another shows that selecting the correct height range for the test samples used in the evaluation is important for a fair comparison of the detection performances of various algorithms. Regarding purity in training, experiments show that including examples with low levels of impurity is beneficial. I observe 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. Moreover, I observe 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 remains unchanged.

Chapter 5

The HDA data set

Data sets are fundamental for systems which rely on Machine Learning, like Pedestrian Detectors (PD's). The information they provide is used both for training and testing such systems. Public data sets, consisting of annotated images and evaluation code, lie the bases for a fair comparison of the algorithms.

One of the goals pursued during the design and compilation of a new data set is that of capturing the diversity that characterises the real world. This ensures that the performance measured on the data set is representative of the one the system will achieve in a real setting. Clearly, due to the limited size of the data sets, this goal can only be met in part: every data set is subject to some form of bias [Torralba and Efros, 2011; Khosla et al., 2012]. As a result, different data sets are better suited to estimate the performance of algorithms in different scenarios. PD algorithms are typically trained and evaluated on data sets representing automotive scenarios [Ess et al., 2007; Dollár et al., 2009; Wojek et al., 2009; Geiger et al., 2012], with the notable exception of the INRIA data set which consists in a collection of holiday pictures [Dalal and Triggs, 2005]. However, PD has clear applications in Video Surveillance scenarios, for which the conditions differ significantly from the automotive ones: Video Surveillance usually considers cameras with high mount points, with a perspective very different from that of cameras mounted on cars. Moreover, the environments in which the cameras are embedded can be as different as an indoor office scene and the lane of an urban street. In order to estimate the performance of a PD algorithm in a real-world Video Surveillance application, it is desirable to train

it and and test it on a data set representing a Video Surveillance scenario. Furthermore, despite the fact that High Definition cameras are commonly available and frequently used in surveillance tasks, most PD data sets are based on low resolution images (mostly VGA: 640×480 pixels).

Following these considerations, we decided to design the High Definition Analytics (HDA) data set with the following goals: (i) establishing a benchmark for PD algorithms specific for an office scenario, (ii) providing a benchmark featuring High Resolution images for Video Surveillance algorithms, in particular PD, person tracking and Re-Identification (RE-ID), and (iii), creating a benchmark for fully automated Re-Identification (PD+REID) systems. We think that the availability of a benchmark for PD algorithms in an office scenario will attract the attention of the Video Surveillance community on PD's. The use of cameras equipped with both standard and High Definition sensors will permit the study of the effect of High Definition on the performance of the algorithms. Moreover, the presence of Hight Resolution images will highlight the weaknesses of the Video Surveillance algorithms of the current generation for that specific case and foster the development of algorithms specific for High Definition images. To make a concrete example, we expect that the algorithms in the state of the art not to achieve real time performance on High Resolution images. Finally, we think that the creation of a benchmark for PD+REID will help the establish a community for the study of this problem, which we see as the natural evolution of classic RE-ID.

Collecting a new data set involves a huge effort, especially in terms of labelling the data and testing the evaluation code. Many members of our research group took part in the making of the HDA data set, most notably Athira Nambiar and Dario Figueira. It is to acknowledge their contributions that I use the first person plural throughout this chapter.

5.1 The HDA data set

The HDA data set was acquired recording simultaneously from 13 indoor cameras for 30 minutes. The cameras were distributed over three floors of the Institute for Systems and Robotics (part of the Instituto Superior Técnico in Lisbon, Portugal), a typical office scenario for Video Surveillance. Approximately 85 people participated in the data collection, most of them

appearing in more than one camera. The data set is heterogeneous: we used three distinct types of cameras (standard, high and very high resolution), different view types (corridors, doors, open spaces) and different frame rates. This diversity is essential for a proper assessment of the robustness of video analytics algorithms in different imaging conditions.

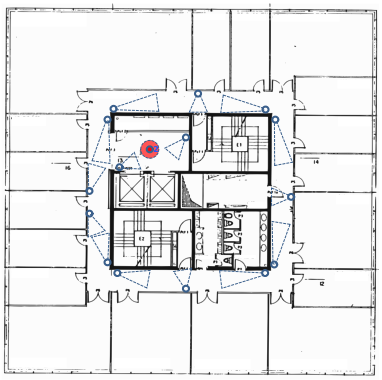
The data recordings for the HDA data set involved the use of 13 AXIS cameras, some with standard VGA resolution (AXIS 211, AXIS 212PTZ, and AXIS 215PTZ) some with 1MPixel resolution (AXIS P1344) and one of 4MPixel resolution (AXIS P1347). To save bandwidth, storage and labelling time, the sequences were not acquired at high frame rates, but at rates of 5Hz, 2Hz and 1Hz for the VGA, the 1MPixel and the 4MPixel resolution respectively. The camera poses in the three floors are depicted in Figure 5.1. Table 5.1 describes the camera network details in brief. Figure 5.2 displays one frame for each camera, highlighting differences in illumination, color balance, depth range and camera perspective.

Table 5.1: Details of the labelled camera network.

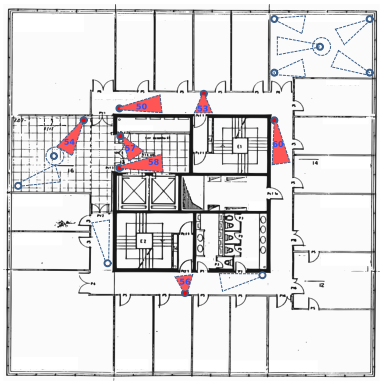
CAM	02	17	18	19	40	50	53	54	55	56	57	58	59	60
640x480	✓	✓	✓	✓	✓									
1280x800						✓	✓	✓	✓	✓	✓	✓	✓	
2560x1600														✓
fps	5	5	5	5	5	2	2	2	2	2	2	2	2	1
floor	6	8	8	8	8	7	7	7	7	7	7	7	8	7

5.1.1 Labelling for the HDA data set

The labelling for the HDA data set consists in Bounding Boxes (BB's) associated with a unique person identifier (ID) and an occlusion flag. Each person/group of people in the images is labelled by such a BB. We opted for using an occlusion flag instead of a value encoding the occlusion ratio of a person because of the much faster annotation process required by the former: given the elevated number of annotations in the data set, this choice made the labelling task more manageable. The BB's alone are used as Ground Truth (GT) in the PD task, while the information conveyed by the BB's needs to be augmented by the person ID for evaluating the RE-ID algorithms. The GT for benchmarking tracking algorithms is encoded by the ID of the BB's, together with the initial and final frame for each person appearance in a video sequence. In the process of labelling, we used the



(a) Floor 6



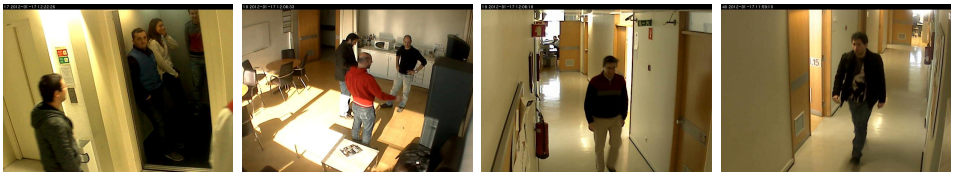
(b) Floor 7

(c) Floor 8

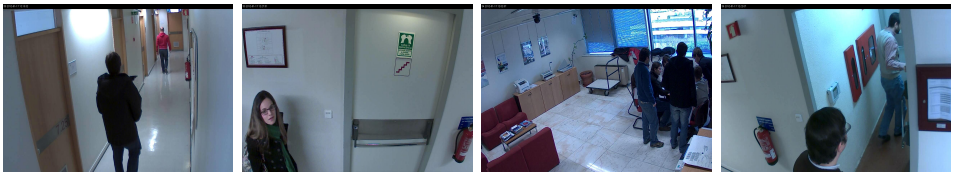
Figure 5.1: Camera poses: a visualization of the three floors of the building at which the HDA data set was acquired. The cameras marked with a red circle and an orange field of view are the ones used to record data.

following software tools: MATLAB® with the Image Processing Toolbox, Piotr Dollár's Toolbox [Dollár, d] and Detection Code [Dollár, c].

(a) Camera 02



(b) Camera 17 (c) Camera 18 (d) Camera 19 (e) Camera 40



(f) Camera 50 (g) Camera 53 (h) Camera 54 (i) Camera 56

(j) Camera 57 (k) Camera 58 (l) Camera 59 (m) Camera 60

Figure 5.2: Snapshots of the sequences acquired in the HDA data set. Notice the differences in illumination, color balance, depth range and camera perspective.

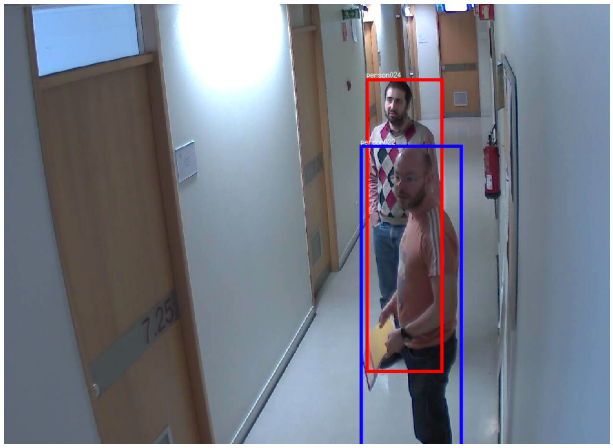
This is the list of the labelling rules:

1. Each BB is drawn so that it completely and tightly encloses the person.
2. If a person is partially occluded, the BB is drawn estimating the whole body extent.
3. Truncated people (i.e., people with projections partially outside the image boundaries) have their BB's cropped to image limits.
4. The occlusion flag is set to '0' for fully visible people, while for partially occluded and truncated people it is set to '1'.
5. A unique ID is associated with each person. In case determining the identity of a person is impossible for the labeller, the special ID 'personUnk' is used.
6. Groups of people that are impossible to label individually are labelled collectively as 'crowd'. People in front of a 'crowd' area are labelled normally.

The proposed labelling allows researchers to perform different experiments on a single test set. For instance, one could choose to test one algorithm ignoring Missed Detections on heavily occluded people, or detections on crowded regions.

We show examples of labelling in Figure 5.3. The person ID is indicated at the top of each BB. The HDA data set comprises annotations of 85 persons, of which 70 are men and 15 are women. A statistical characterization of the data is presented in Table 5.2 and Figure 5.4. One of the peculiarities of the HDA data set resides in the exceptionally wide range of peoples' BB heights: from 69 to 1075 pixels (see Figure 5.4(c)).

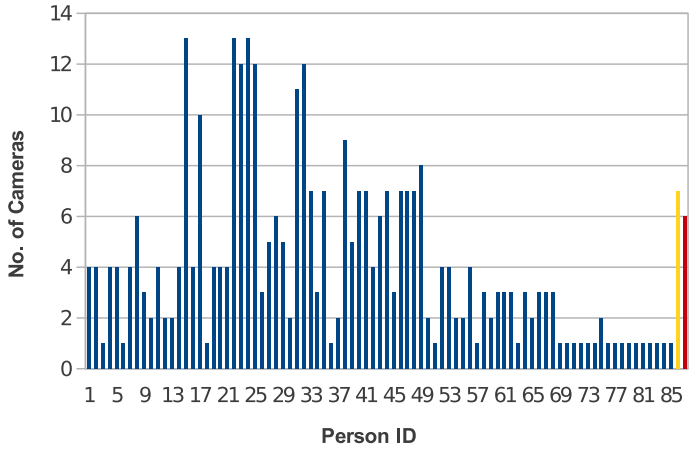
(a)



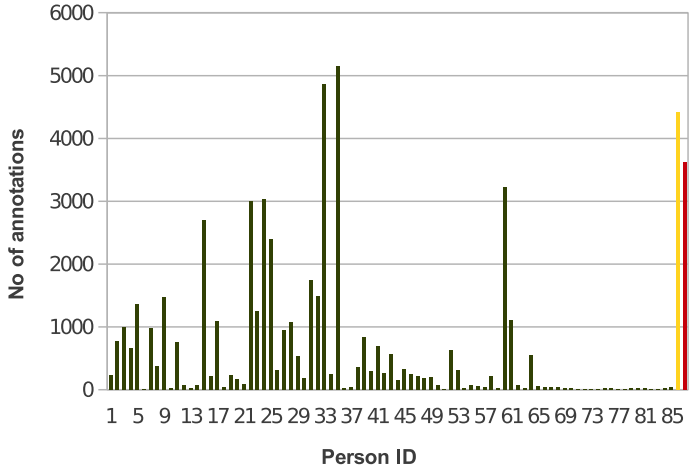
(b)

(c)

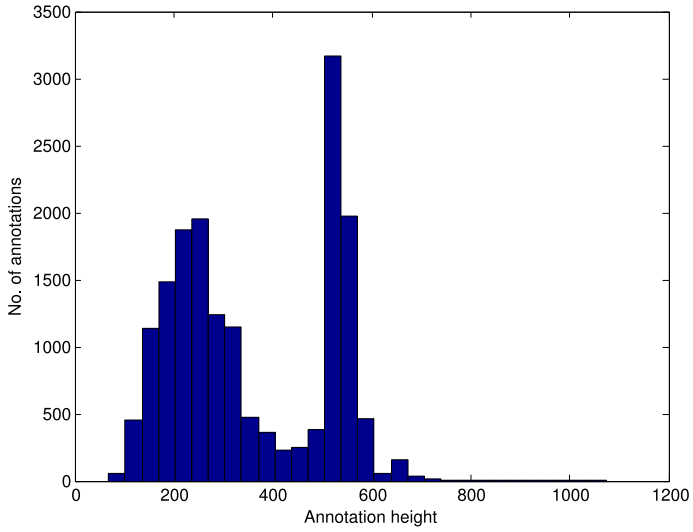
Figure 5.3: Labelling examples. (a) A fully visible (unoccluded) person. (b) Two partially occluded people. (c) A crowd with three partially occluded people in front of it. The ID of each person is indicated on top of the Bounding Boxes.



(a)



(b)



(c)

Figure 5.4: (a) Number of sequences each person appears in. Person 86 (yellow) and 87 (red) correspond to the labels 'personUnk' and 'crowd'. (b) Number of Bounding Boxes (BB's) for each person. (c) Histogram of BB height for the unoccluded people. The peaks of the VGA and the high resolution distributions are visible. The BB's span heights between 69 and 1075 pixels.

Table 5.2: Data on the number of frames, the number of annotations and the number of people for each sequence. The minimum and maximum height of unoccluded Bounding Boxes (BB's) are also reported. Camera 02 does not have person height information due to its unconventional overhead perspective.

Camera	02	17	18	19	40	50	
# frames	9819	9897	9883	9878	9861	2227	
# BB's	1832	3865	13113	18775	7855	1288	
Min. height	-	310	90	71	71	158	
Max. height	-	463	338	403	408	606	
# persons	9	26	32	34	39	20	
# frames	3521	3424	3798	3780	3721	3670	1728
# BB's	465	8703	576	3190	2291	894	1182
Min. height	69	153	619	384	395	598	212
Max. height	681	608	717	688	681	775	1075
# persons	19	12	34	43	34	34	20

We report the experiments conducted using the HDA data set in Chapter 7. Section 7.2.1 and Section 7.2.3 illustrate the performance of two state-of-the art PD algorithms on different subsets of HDA, study the influence of High Definition images on detection performance and compare the results obtained on the HDA and the INRIA data sets. The experiment show that the part-based detector outperforms the monolithic one in a scenario with strong perspective effects, while the monolithic performs slightly better in the condition of full visibility. The experimental data confirms that High Resolution images enable the detection of people imaged at greater distances, but have a tendency to generate more False Positives (FP's) and require longer processing times. Lastly, the considerable difference in detection performance on the HDA and the INRIA confirms the usefulness of creating a data set for PD which represents the Video Surveillance scenario.

Section Section 7.3.1 and Section 7.3.2 focus on the RE-ID problem, evaluating the influence of High Resolution images on the performance and comparing the RE-ID results obtained on HDA and on other data sets. The experiments show that basic RE-ID algorithms, using simple colour histograms as features, do not take advantage of High Resolution images. We speculate that more sophisticated features would better exploit high resolution information, leading to a better RE-ID performance on high resolution images.

The results indicate the HDA data set as the most challenging to date, together with CAVIAR4REID. While the difficulty in re-identifying people in CAVIAR4REID stems mostly from the low resolution of the images, in HDA the problem is hard because of the mixture of cameras with different resolutions, different perspectives and ranges, the presence of harsh illumination changes, severe occlusions, and the fact that several subjects add or remove items of clothing from one view to the next.

Chapter 6 discusses our proposals regarding a PD+REID system and Section 7.4 relates experiments performed on the HDA data set that confirm the usefulness of our proposals.

Chapter 6

Pedestrian Detection in Re-Identification

This chapter describes the work I performed on designing and implementing a fully automated Re-Identification (PD+REID) system. In Section 6.1 I describe the problem of Person Re-Identification, including the standard approach used to frame it and solve it. I introduce the problem of fully automated Re-Identification in Section 6.2, while in Section 6.3 I describe the integrated Pedestrian Detection + Re-Identification system that tackles it. In the remaining two sections (Section 6.4 and Section 6.5) I introduce two improvements to the basic integration scheme which lead to an increase in performance.

The work reported in this chapter was performed in close collaboration with Dario Figueira. The cases in which his contribution was preponderant are highlighted in the text.

6.1 Person Re-Identification

The goal of a Re-Identification (RE-ID) system is to locate and recognise known people in the stream of images flowing from a camera network. Such network is usually set to cover heterogeneous scenarios with non-overlapping (or low overlapping) fields-of-view.

RE-ID has been subject of much research in computer vision due to its usefulness in a large number of applications, e.g. Video Surveillance (VS), smart spaces, border control, crime prevention, and robotics, to quote a few.

Most real-world VS systems rely, to a large extent, on human supervision and intervention. A human operator is assigned to constantly monitor a large number of cameras that must be watched, interpreted and acted upon. This has, of course, several shortcomings: it is costly, inaccurate and subject to human errors. An effective RE-ID system would provide valuable information to a human operator working in VS, e.g., extracting from the camera network data the video clips in which a suspect is present, given one query image in which he/she appears.

Figure 6.1: A typical Re-Identification (RE-ID) algorithm is based on a gallery set: a data base that contains cropped images depicting the persons to be re-identified at evaluation time. People cropped from another set of images (probes) are matched to such data base with the intent of recognising their identities. Classically, RE-ID algorithms are evaluated with manually cropped probes. In this work I study the effect of using automatic probe detection in the full RE-ID system.

Re-identifying people in a camera network is challenging due to a multitude of factors: visual similarity among different people, occlusions, poor quality of video data and varying imaging conditions (illumination, viewing angle, distance ranges, etc.). Finally, people may change their clothing and other appearance traits over time (possibly for disguise purposes).

In the classic set up for a RE-ID experiment, training and test examples are manually selected: each region occupied by an upright, fully visible person is cropped from the raw images. The identity of the person is stored

alongside the image crop. RE-ID consists in estimating the identity of one example from the test set (probe) using some sort of matching among the training examples (gallery) and the probe (see Figure 6.1).

6.2 Fully Automated RE-ID

Fully automated RE-ID can be accomplished by using a software module to detect people and crop their image regions, and carefully designing the connection of that module to a standard RE-ID module. Detecting people can be achieved by the standard Video Surveillance framework or by pattern recognition-based Pedestrian Detectors (PD's). The standard Video Surveillance framework first step is to segment regions of an image into foreground and background based on movement, typically using a Background Subtraction algorithm. The second and final step consists in the classification of each foreground image segment into either person or not person. Relying on the detection of movement makes such detectors vulnerable in a series of conditions: rapid changes in the brightness of the image (due to illumination changes, moving shadows, changes in the camera parameters or TV screens captured by the camera) and moving objects other than people (vehicles, robots, tree leaves, etc.) can all be a source of false detections. Groups of people moving together also pose a problem, as they might be segmented as just one object and be classified either as one person or no person at all. Moreover, Background Subtraction methods assume static backgrounds and do not adapt easily to the case of moving cameras, e.g., cameras mounted on moving robots, cameras mounted on unstable supports (high poles subject to wind) or pan-tilt-zoom cameras. Although extensions of Background Subtraction that work with moving cameras do exist [Sheikh et al., 2009], they are not, to the best of my knowledge, applied to the PD problem.

Pattern recognition-based PD's, on the other hand, allow for the detection of people both in static and moving cameras. They are largely immune to the problems that affect Background Subtraction, but can only detect people imaged in a restricted range of poses. I choose to use pattern recognition-based PD's in the proposed fully automated RE-ID system because of the increasing presence of robotic platforms equipped with cameras and because the pose constraints imposed by PD's are not limiting for the RE-ID systems in the state of the art: RE-ID systems also assume to be working with standing

or walking people. I identify the combined PD and RE-ID system as PD+REID.

6.3 Integration of PD and RE-ID

The integration of PD and RE-ID poses several challenges. The False Positives (FP's) and Missed Detections (MD's) generated by the PD have an impact on the performance of the compounded system: FP's lead to cropped images that are impossible to correctly associate to the ID of a person, while MD's do not generate cropped images, making the identification of a missed person impossible. Moreover, even true positive detections can turn the standard RE-ID problem, which involves manually selected Bounding Boxes (BB's), into a more difficult one. First, the size and position of a detection BB are bound to be less precise than the corresponding BB selected by a human operator. Second, common test sets for person RE-ID consist exclusively of fully visible people, while detections can match people imaged under varying degrees of occlusion.

Another challenge that arises from working with an integrated system is that of performing a fair evaluation, taking into account all the kinds of error that can occur in the compounded system. The correct output for the compounded systems are Correct Identifications (CI's): a CI happens when a person is detected and the most likely class estimated by the RE-ID module (the rank-1 estimate) is correct. The errors the PD+REID system can generate are: FP's and MD's at the level of the PD module, and Incorrect Identifications (II's) at the level of the RE-ID module (when rank-1 estimate does not match the true ID of the person).

It is common to evaluate standard RE-ID systems using Cumulative Matching Characteristic (CMC) curves. A CMC curve shows how often, on average, the correct person ID is included in the best K matches for each test image. When evaluating an integrated system, the CMC curve penalizes detection FP's, but ignores MD's. In order to appreciate the effect of MD's, we decided to complement the CMC curves plot with precision and recall statistics (credit for this idea belongs to Dario Figueira):

- $\text{Precision} = \frac{\text{Correct Identifications}}{\text{True Positive Detections} + \text{False Positive Detections}} = \frac{\text{Correct Identifications}}{\text{Number of Detections}}$
- $\text{Recall} = \frac{\text{Correct Identifications}}{\text{True Positive Detections} + \text{Missed Detections}} = \frac{\text{Correct Identifications}}{\text{Number of Person Appearances}}$

Because of the combination of detection errors and RE-ID errors, the naive, direct connection of a PD system to a RE-ID one yields poor results. In order to limit their influence, I introduce two improvements to the naive integration scheme: the False Positive class (FP class) and the Occlusion Filter (OF). The use of a FP class stems from observations by Dario Figueira: FP detections (the ones which do not correspond to a person in the image) are impossible to be correctly classified by the RE-ID system. Defining a FP class means explicitly modelling the typical FP's of a data set, making a correct classification of FP detections possible. This in turn allows for a coherent evaluation of the performance of the integrated system. The OF is a processing block which lies between the PD and RE-ID systems, its goal being to reduce the incidence of partially occluded detections in the data fed to the RE-ID system. See Figure 6.2 for the block diagram of the fully automated PD+REID system.

In the next three sections, I describe (i) the baseline system, resulting from naive integration of PD and RE-ID, (ii) the FP class, and (iii) the OF. I list experimental results confirming the usefulness of the improvements I propose in Section 7.4. In the experiments I limit the complexity of the problem by constraining the scenario with the closed-space assumption: I require that the access to the surveilled area be granted exclusively to people listed in the training data. This avoids having to manage the case of an unknown person appearing in the test set.

6.4 False Positives Class

Integrating a PD and a RE-ID module without appropriately managing the FP's produced by the PD is suboptimal. It leads to poor performance in real-world applications and to inconsistencies in the evaluation of the performance of the integrated system. Each FP detection corresponds to an image window not centred on a person. This means that there is no correct person ID to be associated with such image window. The RE-ID module will always fail to classify it, generating II's, which will degrade the performance of the integrated system. From a practical point of view, when a Video Surveillance operator is interested in locating the appearances of a person in a video stream, the FP detections constitute a problem because some of them will appear in the results of such search (e.g., the operator looks for

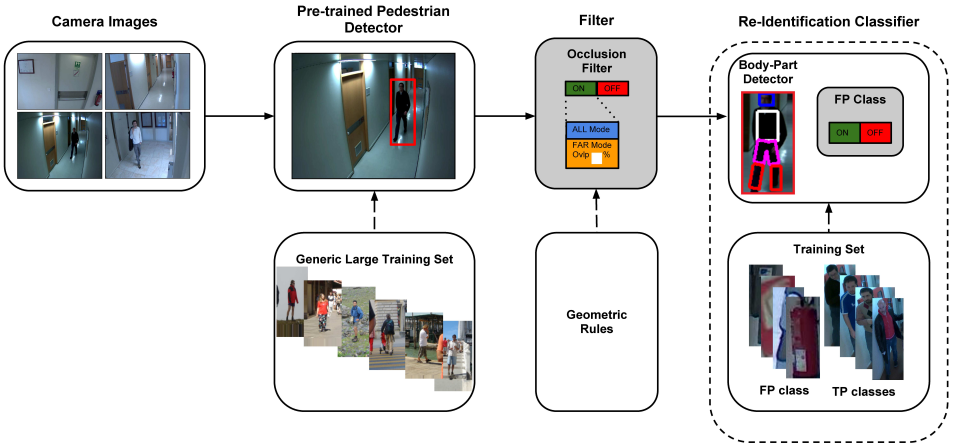


Figure 6.2: Architecture of the proposed **FP+REID!** (**FP+REID!**) system. The images acquired by a camera network are processed by a Pedestrian Detection algorithm to extract candidate Bounding Boxes (BB's). The BB's are optionally processed by the Occlusion Filter (OF) Lastly, the RE-ID module computes the features corresponding to each BB and classifies it. The classification can optionally take into account a False Positive (FP) class.

a appearances of PersonX and the system returns an image without PersonX, but with a FP detection, not centred on a person and erroneously re-identified as PersonX). From the point of view of performance evaluation, the II's resulting from FP detections give rise to CMC curves which do not reach 100% accuracy, rendering the comparison between RE-ID and PD+REID systems not straightforward. Observing that the appearance of the FP's in a given scenario is not completely aleatory, but is worth modelling (see Figure 6.3), I introduce a FP class for the RE-ID module. This means that FP detections for a given scenario are collected during training and used to build the gallery entry for the FP class. In these conditions a correct output for when a FP is presented on the RE-ID's input exists: the FP class. This change allows me to coherently evaluate the performance of the integrated system, generating well behaved CMC curves.

6.5 Occlusion Filter

The OF is a filtering block between the PD and the RE-ID modules. It exploits geometrical reasoning to reject BB's depicting partially occluded people, which can harm the performance of the RE-ID stage, because a BB's

Figure 6.3: An example of the False Positive detections which are used to train the False Positive Class.

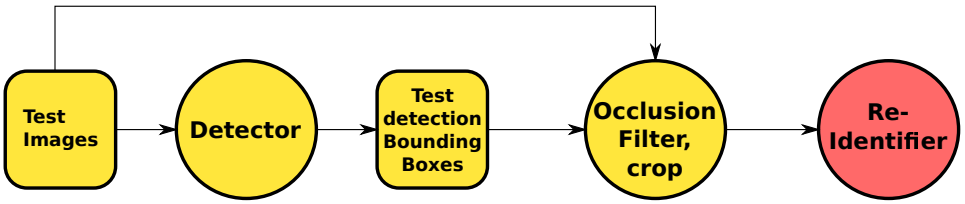
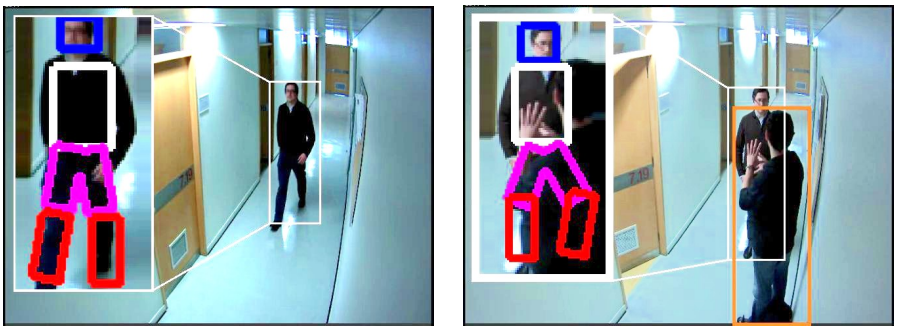


Figure 6.4: Block diagram of the Detection subsystem employed in the integrated PD+REID system. First, a detector is run on the test images. Then, the resulting BB's are filtered by the Occlusion Filter (OF) and the corresponding crops are generated. Such crops form the input data for the RE-ID system.

including a person appearing under partial occlusion and one including the same person imaged under full visibility conditions generate different features. When the partial occlusion is caused by a second person standing between the camera and the original person, the extracted features can be a mixture of those generated by the two people, making the identity classification especially hard (see illustration in Figure 6.5). For this reason, it would be advantageous for the RE-ID module to receive only BB's depicting fully visible people.

Though the visibility information is not available to the filter, it can be estimated quite accurately with a heuristic based on scene geometry: in a typical scenario the camera's perspective projection makes pedestrians closer to it extend to relatively lower regions of the image. The filter computes the overlap among all pairs of detections in one image and rejects the one in each overlapping pair for which the lower side of the BB is higher (as illustrated in Figure 6.6). Considering the mismatch between the shape of the pedestrians' bodies and that of the BB's, it is clear that an overlap between BB's does not always imply an overlap between the corresponding



(a) (b)

Figure 6.5: Example of body part detection for feature extraction in two cases: (a) a person appearing in full visibility condition and (b) under partial occlusion. The behaviour in (a) is correct, while in (b) part of the features describing the appearance of the occluded person are actually computed on image regions belonging to the occluding person. The contrast of both images was enhanced for visualization purposes.

pedestrians' projections on the image. I define an overlap threshold for the filter, considering as overlapping only detections whose overlap is above such threshold.

The experiments regarding the FP class and the OF are described in Section 7.4. The results show that introducing the FP class leads to an increase in RE-ID precision, at the price of a drop in recall. Moreover, using the FP class enables a meaningful evaluation of the combined system with a CMC curve. The introduction of the OF leads to a small improvement in the precision of the RE-ID system (thanks to the removal of ambiguous and hard to classify detections), at the same time inducing a correspondingly small drop in recall.

Figure 6.6: An example of geometrical reasoning: two detection Bounding Boxes (BB's) have a high degree of overlap (yellow area). The woman is deemed as occluded because of the comparison between the lower sides of the two BB's (see the arrows): the lower boundary of her BB is higher in the image than that of the man.

Chapter 7

Experiments and Results

The following experiments were designed for testing the ideas presented in Chapter 4, Chapter 5, and Chapter 6. In Section 7.1 I present results on the influence of purity in the training and testing of Pedestrian Detection (PD). In Section 7.2 I list results in PD on the High Definition Analytics (HDA) data set. Section 7.3 presents results on Re-Identification (RE-ID) on the same data set, while Section 7.4 reports results on the integration of a PD in a fully automated person RE-ID scheme. Throughout this chapter I refer to algorithms and data sets defined in other parts of this thesis, that information is summarized in Table 7.1. Dario Figueira played a fundamental role in all the experiments involving person RE-ID. To acknowledge this fact, in the text describing such experiments I use the first person plural.

Exp. Group	Section	Exp.	Data	PD Alg.	RE-ID Alg.
Purity	7.1	0, 1, 2, 3	INRIA	My FPDW, ACF,	–
				Caltech Algs.	–
HDA PD	7.2	4, 5, 6	HDA	My FPDW,	–
				Grammar Models	
HDA RE-ID	7.3	7, 8	HDA	–	Nearest Neighbour
HDA PD+REID	7.4	9	HDA	ACF	Nearest Neighbour

Table 7.1: A summary describing the goal of the experiments, plus the data sets and the algorithms used in each experiment.

7.1 The Influence of Sample Purity on Pedestrian Detection

For clarity, I describe the evaluation protocol and the test labels I use in the experiments in Section 7.1.1. I motivate the choices regarding the protocol and support them comparing results obtained using the standard and the proposed protocol. I describe the experiments in Section 7.1.2.

7.1.1 Experiment 0 - Evaluation Protocol and Test Labels

In the beginning of this section I define the parameters for running an experiment and the measures used to evaluate its results. Then, I discuss the height ranges relevant for the evaluation and the relationships among them. Finally, I highlight the impact of a more accurate test set labelling on the evaluation of the performance of the detectors, laying the bases for the experiments.

Parameters and Evaluation Variables for an Experiment

Performing a detection experiment in this context 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 I set are:

- Minimum Height in TRaining (MHTR)
- Minimum Height in TEst (MHTE)
- Minimum Visibility in TRaining (MVTR)
- Minimum Visibility in TEst (MVTE)

For evaluating the results of one experiment I use the *de facto* standard measures of Missed Detection (MD) rate and False Positives Per Image (FPPI). MD rate represents the fraction of positive examples which are not detected. FPPI is defined as the average of False Positive Per Image. The performance of one detector is described by a curve in the MD rate/FPPI space. Each point on the curve corresponds to an operating point for the PD algorithm. The performance of one detector is summarized by the Log-Average Miss Rate (LAMR), the average Miss Detection rate (as computed

on the logarithmic FPPI axis) between 10^{-2} and 10^{0} FPPI. Please refer to Section 4.1 for a more in-depth explanation of the evaluation protocol.

Height Ranges Matching for a Fair Evaluation

As described in Section 4.1, the height range used during evaluation should be a subset of the intersection of three other ranges: the range of scales at which people appear in the test set, the range of scales spanned by the detections produced by an algorithm and, finally, the range of scales for which people are labelled in the Ground Truth (GT). This is not the case when comparing the detections provided in the Caltech benchmark for algorithms in the state of the art on the original INRIA test set, using the "Reasonable" evaluation mode. The "Reasonable" mode corresponds to setting the MHTE to 50 pixels and the MVTE to 0.65 (see [Dollár et al., 2012b] 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 7.2, 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 and that specified by the evaluation mode 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 MD rate of zero.

I 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. I compare the reported performance of PD algorithms using the original labelling and selecting either the "Reasonable" or the "Reasonable90" evaluation mode. I 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. I evaluate the detections of a number of state-of-the-art algorithms provided with the Caltech benchmark and the detections generated by my implementation of the Fastest Pedestrian Detector in the West (FPDW), trained on the original labelling.

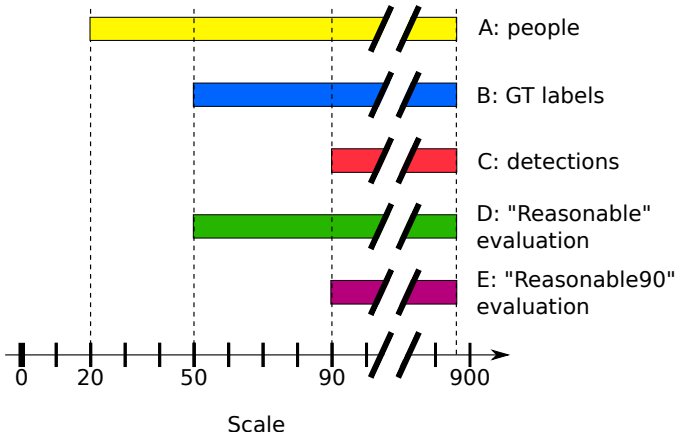


Figure 7.1: Scale ranges evaluating the performance of the detectors considered in the Caltech Benchmark, on the INRIA test set, with the "Reasonable" mode. A, the range of scales at which people appear in the test set, starting roughly at 20 pixels and ending at around 900; B, the range of scales at which people are labelled, starting at around 50 pixels; C, the range of scales spanned by the detections published with the benchmark, for various algorithms, starting at around 90 pixels; D, the range of scales taken into account by the "Reasonable" evaluation mode, starting at 50 pixels; Finally, E, the range of scales taken into account by the "Reasonable90" evaluation mode, starting at 90 pixels. The GT labels with heights between 50 and 90 pixels cannot be matched by the detections provided with the data set: using the "Reasonable" evaluation mode underreports the performance of the detectors. The "Reasonable90" evaluation mode, on the other hand, allows for a more truthful assessment of the performance.

I display the MD rate/FPPI curves for one representative algorithm, for the two modes, in Figure 7.2. Using the "Reasonable90" evaluation mode reports slightly lower MD rates, especially at relatively high (10^0) FPPI levels (see the results for all the tested algorithms in Table 7.2, 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.

			MD at 10^0FPPI		
Algorithm		Min. det. height	Reasonable (MHTE=50)	Reasonable90 (MHTE=90)	Difference
FtrMine	[Dollár et al., 2007]	100.0	0.340	0.324	-0.016
LatSvm-V1	[Felzenszwalb et al., 2008]	79.0	0.175	0.159	-0.015
HOG	[Dalal and Triggs, 2005]	100.0	0.231	0.215	-0.015
HikSvm	[Maji et al., 2008]	100.0	0.221	0.207	-0.014
PLS	[Schwartz et al., 2009]	100.0	0.226	0.212	-0.014
HogLbp	[Wang et al., 2009]	96.0	0.190	0.173	-0.017
FeatSynth	[Bar-Hillel et al., 2010]	100.0	0.109	0.089	-0.019
MultiFtr+CSS	[Walk et al., 2010]	93.7	0.109	0.093	-0.016
FPDW	[Dollár et al., 2010]	100.0	0.093	0.075	-0.018
ChnFtrs	[Dollár et al., 2009]	100.0	0.087	0.072	-0.015
LatSvm-V2	[Felzenszwalb et al., 2010]	*91.3*	*0.081*	*0.058*	*-0.024*
My FPDW		95.6	0.093	0.081	-0.013
CrossTalk	[Dollár et al., 2012a]	99.2	0.098	0.079	-0.020
Mean					-0.017

Table 7.2: 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 MD 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 Figure 7.2 for a graphical comparison of the performance of LatSvm-V2 in the two cases.

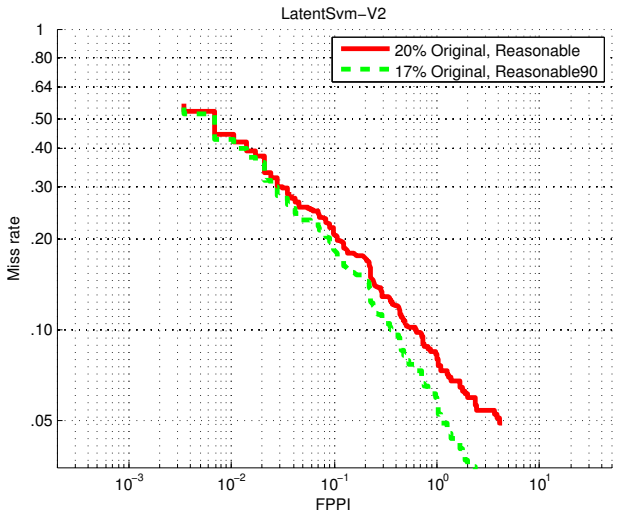


Figure 7.2: The reported performance of the LatSvm-V2 algorithm [Felzenszwalb et al., 2010] using the original labelling and the "Reasonable" or the "Reasonable90" evaluation modes, in red and green respectively. Performance is summarized in the legend with the Log-Average Miss Rate. Using the "Reasonable90" evaluation mode instead of the "Reasonable" one reports slightly lower MD 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.

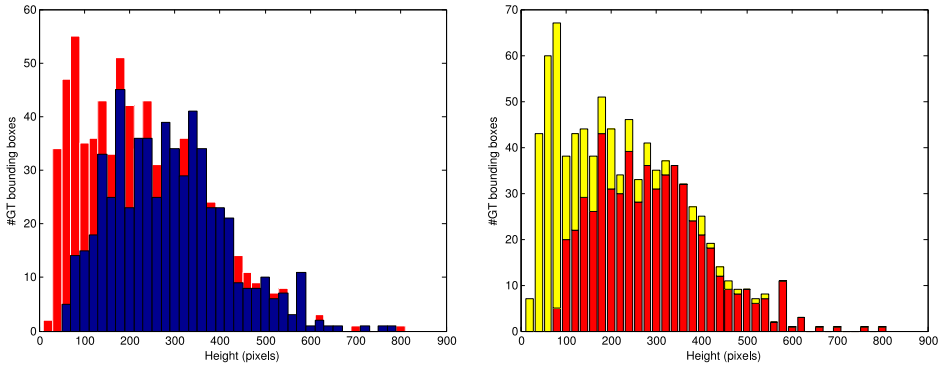
A deeper analysis of the INRIA test set reveals that only a fraction of the pedestrians is labelled: the higher the level of occlusion, the more unlikely people are to be labelled, while the smallest pedestrians are unlabelled as a whole.

The proposed test labelling and its influence on evaluation

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 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 PD's 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. I 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 Figure 7.3(a)). The fraction of "Ignore" BB's for the new annotation is considerable, Figure 7.3(b) 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 Figure 7.4.

Comparing the performance reported by testing using either the original or the proposed annotations, it can be observed that the proposed annotation reports better performances for the algorithms at low FPPI values and better differentiates the performance of the various detectors. I use the same set of algorithms mentioned in the previous subsection and the "Reasonable90" mode: MHTE = 90 pixels, MVTE = 0.65. I display the MD rate/FPPI plot for one representative algorithm and the two annotations, in Figure 7.5. Two effects can be seen: the miss rate is minimally higher at

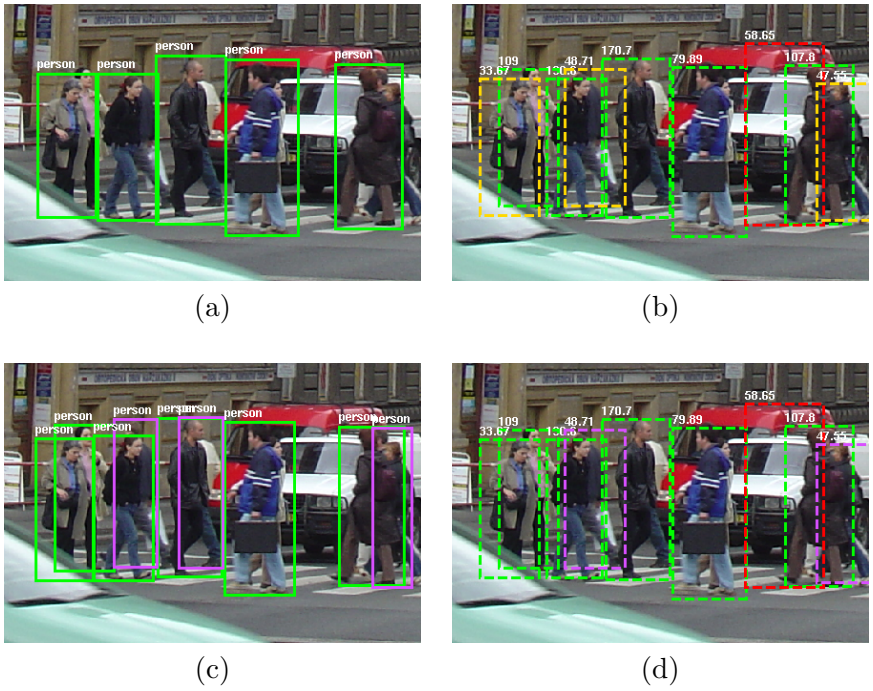


(a) (b)

Figure 7.3: Characterization of the original and the proposed labellings of the INRIA 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" is considerable and does influence the assessment of the detection performance.

high FPPI values for the proposed labelling, I ascribe this to the introduction of more occluded pedestrians in the test set, which makes the problem more difficult. The other effect, the most significant one, is the average drop of 8.9% for the MD rates at low FPPI values (10^{-2}) (see the results for all the tested algorithms in Table 7.3, columns 2–4). I 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 I am 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 better overall are the ones that benefit the most from using the proposed labelling (see Table 7.3, columns 4 and 5).

I showed that the fair evaluation of the performance of a detector in one experiment requires setting a height range for the evaluation in a principled way. Such height range should be a subset of three other ranges: the range spanned by people heights in the test set, the one spanned by Ground



(a) (b)

(c) (d)

Figure 7.4: Comparison of one evaluation performed with the original (a,b) and the proposed (c,d) INRIA test set labelling. 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, False Positives (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.

		MD at 10^{-2}FPPI			
Algorithm		Original labelling	Proposed labelling	Difference	LAMR, proposed labelling
FtrMine	[Dollár et al., 2007]	0.918	0.900	-0.019	57%
LatSvm-V1	[Felzenszwalb et al., 2008]	0.806	0.835	+0.029	43%
HOG	[Dalal and Triggs, 2005]	0.744	0.702	-0.042	42%
HikSvm	[Maji et al., 2008]	0.766	0.681	-0.085	39%
PLS	[Schwartz et al., 2009]	0.674	0.596	-0.078	38%
HogLbp	[Wang et al., 2009]	0.665	0.629	-0.036	35%
FeatSynth	[Bar-Hillel et al., 2010]	0.754	0.738	-0.015	29%
MultiFtr+CSS	[Walk et al., 2010]	0.469	0.425	-0.044	21%
FPDW	[Dollár et al., 2010]	0.576	0.386	-0.189	18%
ChnFtrs	[Dollár et al., 2009]	0.581	0.383	-0.198	18%
LatSvm-V2	[Felzenszwalb et al., 2010]	*0.448*	*0.319*	*-0.129*	*17%*
My FPDW		0.577	0.307	-0.270	16%
CrossTalk	[Dollár et al., 2012a]	0.511	0.333	-0.178	15%
Mean				-0.089	

Table 7.3: 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 TEst (MHTE) of 90 pixels and a Minimum Visibility in TEst (MVTE) of 0.65. LAMR indicates the Log-Average Miss Rate for each algorithm. Using the proposed labelling reports considerably lower MD 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 Figure 7.5 for a visualization of the performance of the LatSvm-V2 detector in the two evaluation cases.

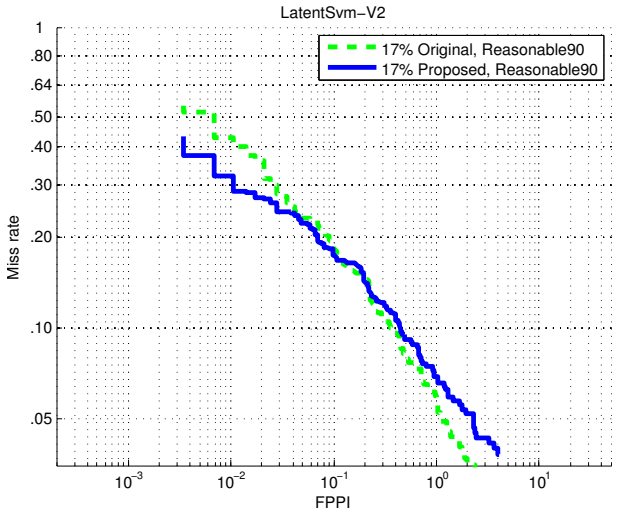


Figure 7.5: The reported performance of the LatSvm-V2 algorithm [Felzenszwalb et al., 2010] using the original and the proposed labelling, in green and blue respectively. The evaluation mode is "Reasonable90" in both cases. Performance is summarized in the legend with the Log-Average Miss Rate. Using the proposed annotation instead of the original one reports considerably lower MD rates at low FPPI values. I 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 FP's with the original annotation and as TP's with the proposed one. Using the proposed annotation also reports slightly increased MD rates at high FPPI values. I 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.

Truth labels and the one spanned by the detections generated by the detector. Specifically, when testing algorithms on the INRIA test set, using the detections from the Caltech benchmark, the "Reasonable90" mode provides for a fairer evaluation than using the default "Reasonable" mode. Additionally, I presented a new labelling for the INRIA test set and showed the influence the more accurate test annotations have on the reported detection performance of the algorithms.

7.1.2 Experiments on the Influence of Sample Purity

In this section I present the results of three experiments that measure the influence of different degrees of impurity in the testing and training of PD's. 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 7.4. For all the experiments I 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[1], the proposed annotation is available on the author's website.

Exp.	Description	Train. labelling	Test labelling	Det. Algorithm
1	Partial occl. in testing	Original	Proposed	FPDW + Caltech Algs.
2	Partial occl. in training	Proposed	Proposed	ACF
3	"Short" ex. in training	Proposed	Proposed	ACF

Table 7.4: Summary of the experiments: the variable whose influence is studied in each experiment, the algorithms and the training and test labelling used in each experiment are listed. "Caltech Algs." refers to a set of algorithms whose detections are distributed with the Caltech Pedestrian Detection Benchmark (see Table 7.2 for a list of the detectors).

Experiment 1 - Influence of partial occlusion in the test set on detection performance

In this experiment I evaluate the impact of the amount of partial occlusion of the examples in the test set on the detection performance. I tackle, thus, a single source of impurity. It has been shown in [Dollár et al., 2012b] that even a modest amount of occlusion (visibility ratio as low as 0.65) has a highly detrimental effect on the performance of PD's. Those results were obtained using the Caltech data set. I perform a similar experiment on the INRIA data set and confirm that the finding has general validity. I use the detections generated by my implementation of FPDW, trained on the original training set, as well as the detections of several algorithms distributed with the Caltech benchmark. I define 7 test modes which correspond to as many

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

test sets. I 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. I include in this experiment only pedestrians imaged with heights greater than 90 pixels.

I observed that lowering the minimum degree of visibility of the test examples negatively affects the detection performance (see Figure 7.6(a) for the MD rate/FPPI curves of FPDW tested with different degrees of visibility and Figure 7.6(c) for a visualization of the relationship between Minimum Visibility in TEst (MVTE) 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 I 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 Figure 7.6(b)).

Conclusion This experiment shows that partial occlusion correlates with the difficulty in detecting pedestrians: the higher the amount of occlusion, the harder it is for a detector to detect people. This result holds for the INRIA test set and all the tested detectors, including the part-based ones.

Proposed labelling of INRIA 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 Figure 7.7(a)). Each 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 Figure 7.7(b)).

The two following experiments are devoted to assessing the importance

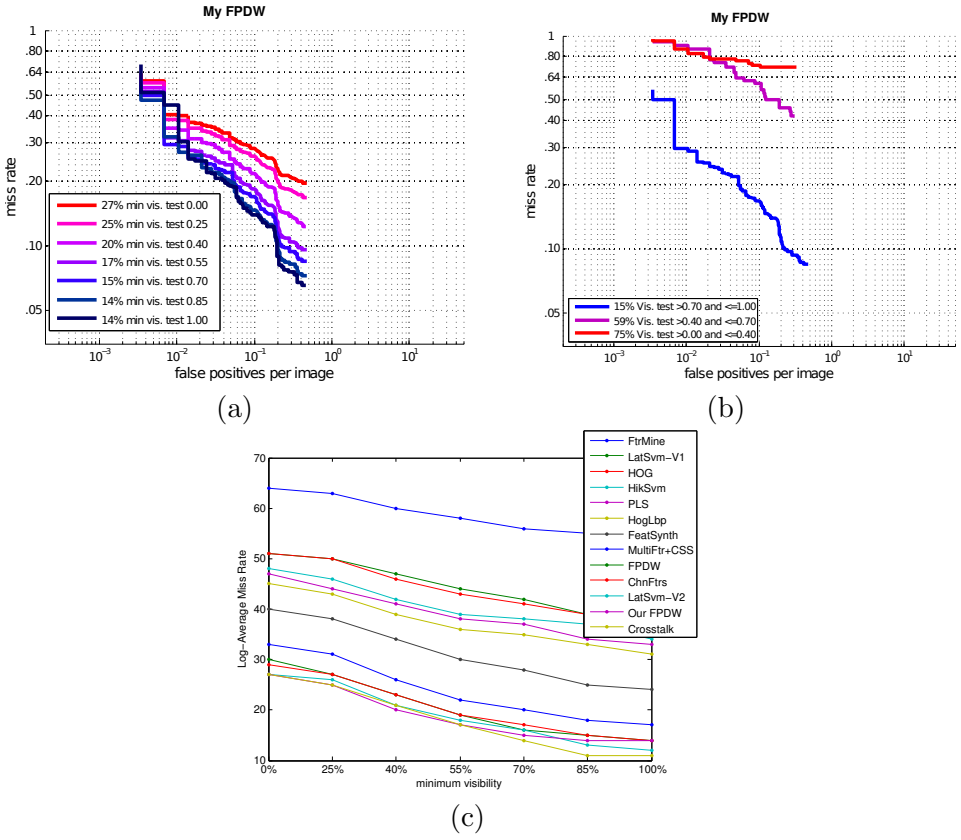


Figure 7.6: The effect of partial occlusion in testing on the accuracy of PD's. (a) Reducing the minimum visibility of the pedestrians in the test set decreases the detection performance of my implementation of FPDW. (b) The detection accuracy of FPDW varies greatly as a function of visibility. The three 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.

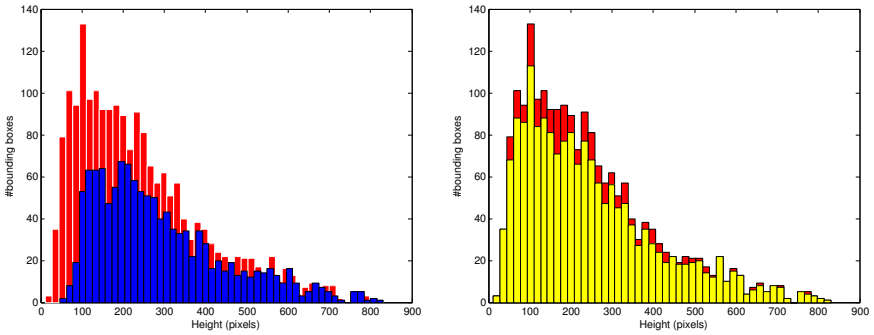


Figure 7.7: Characterization of the original and the proposed labellings of the training set. (a) Histograms of the height of "Person" labels for the original (blue) and the proposed labelling (red). The proposed annotation outnumbers the original one, particularly at lower heights. (b) Histogram for 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.

of including impure, e.g., small and partially occluded positive examples in the training set. I evaluate the effect of partial occlusion in Experiment 2, while I gauge the impact of including small pedestrians in the training set in the Experiment 3. I use the Aggregate Channel Features (ACF) detector and the proposed labelling both for training and for testing. I choose to use the ACF detector for these experiments because of the great reduction in training time it allows for, compared to my implementation of FPDW.

I filter the full training set varying two thresholds, one on the minimum height, the other on the minimum visibility:

- MHTR $\in \{25, 50, 100\}$
- MVTR $\in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$

I train the detector with the training set resulting from each of the 18 combinations and, for comparison, I train it also with the original INRIA set. To account for the stochastic parts of the ACF algorithm, I repeat each training 10 times with a different randomisation 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 I use the proposed labelling also for the test set. This

allows me to set the minimum height and minimum visibility for a person to be considered in the test set:

- MHTE $\in \{25, 50, 100\}$
- MVTE $\in \{0.65, 1.0\}$

This yields a total of 6 evaluation modes. I adjusted the depth of the image pyramid so that it is possible to detect pedestrians 25 pixels tall or taller.

Experiment 2 - Influence of partial occlusion in the training set on detection performance

In this experiment I evaluated the impact of varying the MVTR while fixing the rest of the experiment parameters: MHTR, MHTE and MVTE. I repeated the test for each of the 18 fixed parameters combinations (MHTR $\in \{25, 50, 100\}$, MHTE $\in \{25, 50, 100\}$, MVTE $\in \{0.6, 1.0\}$). In each case, MVTR spanned the 0.5–1.0 interval with 6 equally spaced values.

For the first analysis, I fixed MHTR to 25 pixels, MHTE to 100 pixels and MVTE to 0.65. I compared the performance of the ACF detector when trained including examples with different degrees of visibility. I display the performance achieved by the different training modes (as measured by the LAMR) in Figure 7.8(a). It can be seen that including partially occluded pedestrians (up to 0.8 visibility) in the training is advantageous, as it leads to lower LAMRs.

In the second part of the experiment I set the MHTR and MHTE to 100 pixels and the MVTE to 1.0, i.e., I tested on fully visible pedestrians. The results depicted in Figure 7.8(b) 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 Figure 7.8(c)).

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 Figure 7.8(d)). 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 7.5), to the increasing difficulty in labelling when dealing with smaller pedestrians or to some other phenomenon.

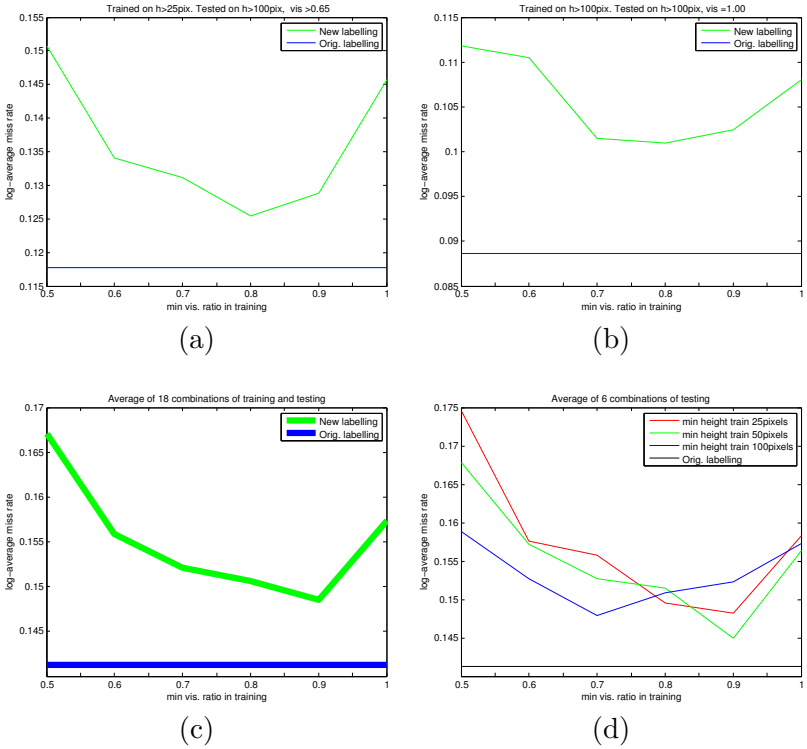


Figure 7.8: (a) The performance of ACF trained with different levels of occlusion and evaluated with a Minimum Visibility in TEst (MVTE) of 0.65. Including partially visible examples in training is advantageous: the best performance is obtained with a Minimum Visibility in TRaining (MVTR) value of 0.8. (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 originating this effect is unclear at the moment (see text for conjectures).

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

Table 7.5: The number of positive examples in the proposed labelling, in each of three height bins.

Conclusion This experiment shows that including pedestrians imaged under moderate partial occlusion in the training is advantageous, even when testing exclusively on fully visible pedestrians. The optimal minimum visibility for including pedestrians in the training set varies depending on other factors, but gravitates around the value of 0.8.

Experiment 3 - Influence of "short" examples in the training set on detection performance

In this experiment I assess the impact of the inclusion of examples smaller than the detection window in the training set. I observe that including small pedestrians in the training has a strong positive effect on the detection of pedestrians in a similar range of heights. I 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, I 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 7.5. I introduce three new testing modes based on this partition of the heights. I consider training with the "tall" class the baseline (as 100 pixels coincides with the detection window height) and I evaluate the effect of including smaller pedestrians in the training. I 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 Figure 7.9(a)). 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 Figure 7.9(b)). The advantage is not visible, though, when testing on the full set. I 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 $1/15^{th}$ of the element of the "tall" class. A comparison of the best detection performances for the "short" range, obtained with various MHTR is visible in Figure 7.10. Improving the detection performance on the "small" pedestrians is fundamental for automotive applications: pedestrians which appear small on the

image are the ones who stand far from the car, detecting such a pedestrian in a dangerous situation would give the driver of the car enough time to respond to an emergency.

When testing on the "medium" class, including elements of the same height range has a small positive effect on detection accuracy (see Figure 7.9(c)), 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 Figure 7.9(d)).

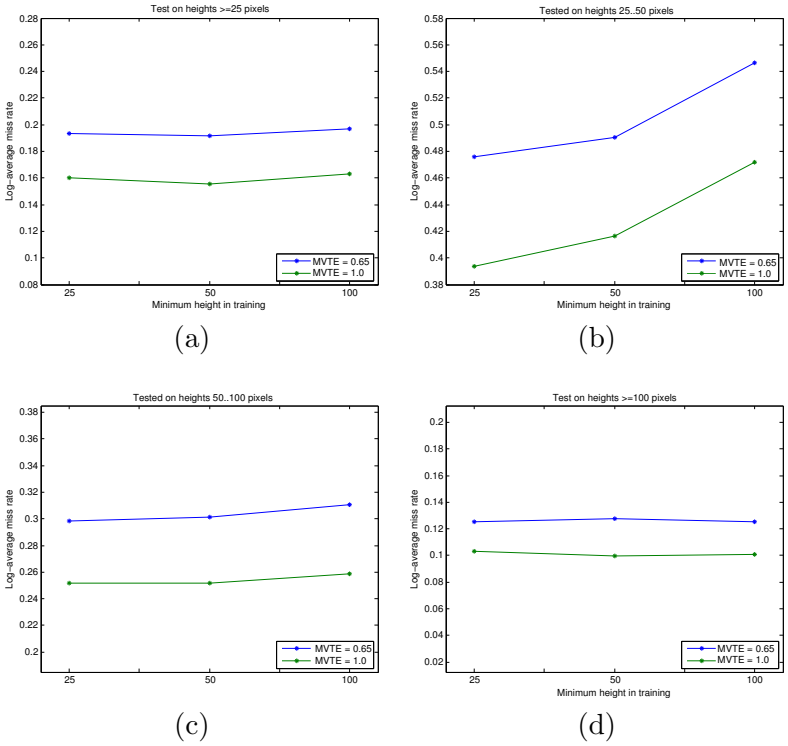


Figure 7.9: Log-Average Miss Rate obtained varying minimum training height, ACF detector. The four plots display the results obtained with two test modes: full visibility (MVTE= 1.0) in green and Minimum Visibility in TEst (MVTE) of 0.65 in blue. (a) When testing on pedestrians spanning the whole range 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 smaller pedestrians does not change the detection performance on the big pedestrians much (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).

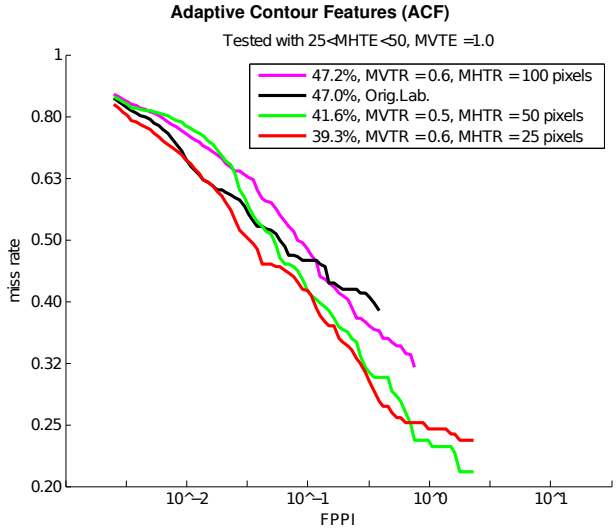


Figure 7.10: Best detection performances for the "short" range, testing with full visibility. Including pedestrians imaged at heights between 25 and 50 pixels in the training (red line) produces a detector which dominates the others for most FPPI values.

Conclusion This experiment shows that including training examples two octaves smaller than the detection window has a positive effect on the detection of pedestrians in the same range. I speculate that short pedestrians contribute to the detection system knowledge on the appearance of people when imaged at low resolutions.

7.2 Pedestrian Detection on the HDA Data Set

In this set of experiments I gauge the quality of the PD problem presented by the HDA data set and exploit the data set to compare the performance of two detectors under different imaging conditions. I compare the performance of the detectors on HDA and on the INRIA data set, concluding that each data set is better suited to evaluate performances in the scenario it depicts. Moreover, I exploit the High Definition images of HDA to perform a study on the influence of High Definition on PD performance.

I evaluate the performance of two PD systems on the HDA data set: the Fastest Pedestrian Detector in the West (FPDW) [Dollár et al., 2010] and the Grammar Models detector [Girshick et al., 2011] (also known as Discriminatively Trained Deformable Part Models, release 5). These two systems are state-of-the-art representatives of two distinct paradigms for PD: monolithic (human as a whole) and part-based (human as a composition of parts), respectively. I use the code provided by the authors [Girshick et al.] for Grammar Models, with a model trained on the Pascal VOC 2010 data set [Everingham et al., 2010], while for FPDW I use my own implementation, trained on the INRIA person data set.

For evaluation I use a customized version of the code provided by [Dollár et al., 2012b]. Detections and missed detections on a 'crowd' area of the image are not penalized. My data set annotation is consistent with the behaviour of the evaluation algorithm for most of the labels. However, because the GT BB's in HDA are designed to enclose the full extent of the projection of a person on one image, this can lead to BB's that are not horizontally centred on the targets, mainly when the pose of a person's arms or legs is very asymmetrical. In such cases the matching algorithm can report a FP or a MD instead of a TP.

In Experiment 4 (Section 7.2.1) I compare the PD performance obtained by two state-of-the-art algorithms on subsets of the HDA data set. In Experiment 5 (Section 7.2.2) I compare the PD performance obtained on the HDA and on the INRIA data sets. In Experiment 6 (Section 7.2.3) I assess the effect on PD performance of High vs. Low Resolution images.

7.2.1 Experiment 4 - PD Performance in Different Scenarios of HDA

In order to evaluate the characteristics of the HDA data set with respect to PD algorithms, I partition the video sequences based on characteristic views. I form the groups 'long range' (sequences of corridors, 19, 40, 50 and 60), 'mid range' (sequences of big rooms 18 and 54) and 'short range' (sequences of cameras pointed towards doors or inside small rooms 17, 53, 56, 57, 58 and 59). I leave sequence 02 out (see top image in Figure 5.2), as in that case the camera is pointed down from the ceiling: the projections of people onto the image plane are so different from the typical pedestrian projections that both detectors completely fail at the recognition task.

Considering that only Grammar Models handles occlusions explicitly, I evaluate the detections using two modes: in the 'base' mode I consider all the BB's composing the GT, while in the 'full visibility' mode I only consider the BB's that are completely visible. Moreover, Grammar Models estimates a BB enclosing the full person even when it observes only a part of it. This can lead to detection BB's with parts well outside the image. Such BB's would not match the GT of people only partially inside the image, as in the GT the BB's are bounded to the image limits. I thus crop each detection so that it lies inside the image.

I present the results in the form of MD rate/FPPI plots in Figure 7.11. The overall performance indicator shown is the LAMR (see Section 7.1.1 for the definition).

Conclusion The performances of FPDW and Grammar Models are quite similar for the long and the mid range sequences, with a little advantage for FPDW in the fully visible mode. For the short range setup instead, Grammar Models is the clear winner. I speculate that the advantage of Grammar Models in the short range sequences depends on its part-based structure, which can accommodate the perspective deformations between training and test data better than the monolithic structure of FPDW.

7.2.2 Experiment 5 - Comparing PD Performance on HDA and on INRIA

In this experiment I compare the performances of FPDW and Grammar Models on the proposed HDA data set and on the INRIA person data set [Dalal

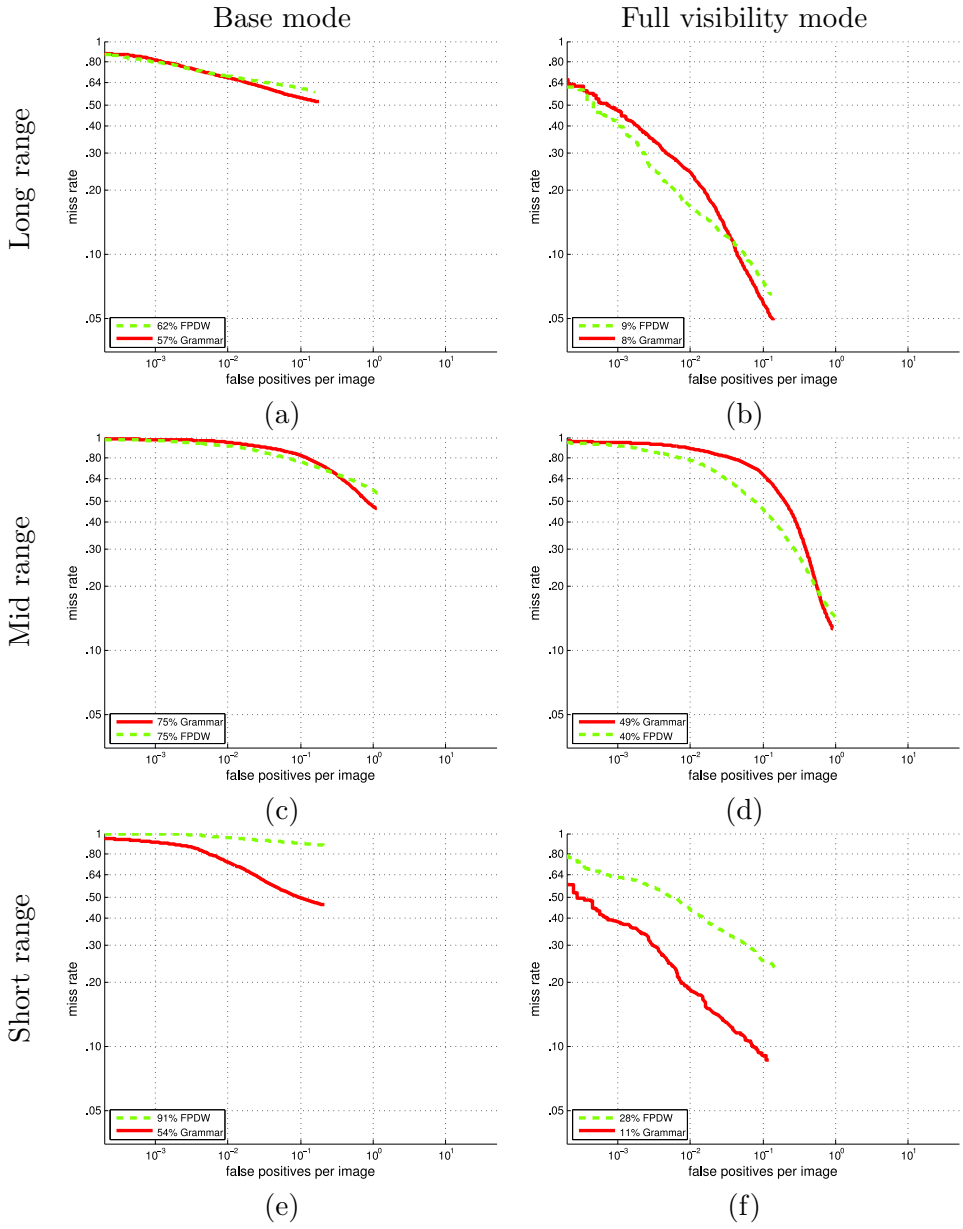


Figure 7.11: MD rate/FPPI plots for different groups of image sequences (rows) and for two different evaluation modes (columns). Lower curves and curves more on the left side of the plots indicate better performance.

and Triggs, 2005]. I use the proposed annotation for the INRIA data set, which includes information on the degree of occlusion that affects each person. The two data sets are quite different in nature: the INRIA data set consists in a collection of holiday photos, while the HDA data set portrays a typical Video Surveillance scenario.

The performance of the two algorithms varies from one data set to the other: using the Full visibility evaluation mode reports moderate differences between FPDW and Grammar Models, while using the Base mode the differences are more extreme (see Figure 7.12). This confirms that the performance of a detector depends heavily on the application scenario and that data sets specific to different applications are needed.

The large difference in the results between the Base and the Fully visible evaluation mode, on both data sets, confirms that occlusion poses a severe challenge to PD algorithms. The fact that FPDW outperforms the Grammar Models detector on the INRIA data set is to be expected: FPDW was trained on INRIA and, as shown in [Benenson et al., 2014a], detection performance degrades when testing on a test set which does not match the training set used.

Conclusion The performances of the two detectors on the HDA and the INRIA data sets are quite different. This confirms the usefulness of designing a data set representing a typical Video Surveillance scenario for PD. I expect the performance measured on the HDA data to be more representative of that achievable in a real Video Surveillance setting than the one measured on the INRIA data set.

7.2.3 Experiment 6 - PD Performance at Different Image Resolutions

In this experiment I assess the impact of image resolution on the PD task. The imaged height of a pedestrian has a strong correlation with the difficulty of detection: smaller pedestrians are more difficult to detect. Most state-of-the-art methods struggle with heights under 80 pixels and even humans start having difficulties for heights under 30 pixels (see [Dollár et al., 2012b]). This grants a clear advantage to detection performed on higher resolution images: some people with imaged height under the 30 pixels threshold with a VGA camera would exhibit a height above that threshold if imaged with a higher resolution camera.

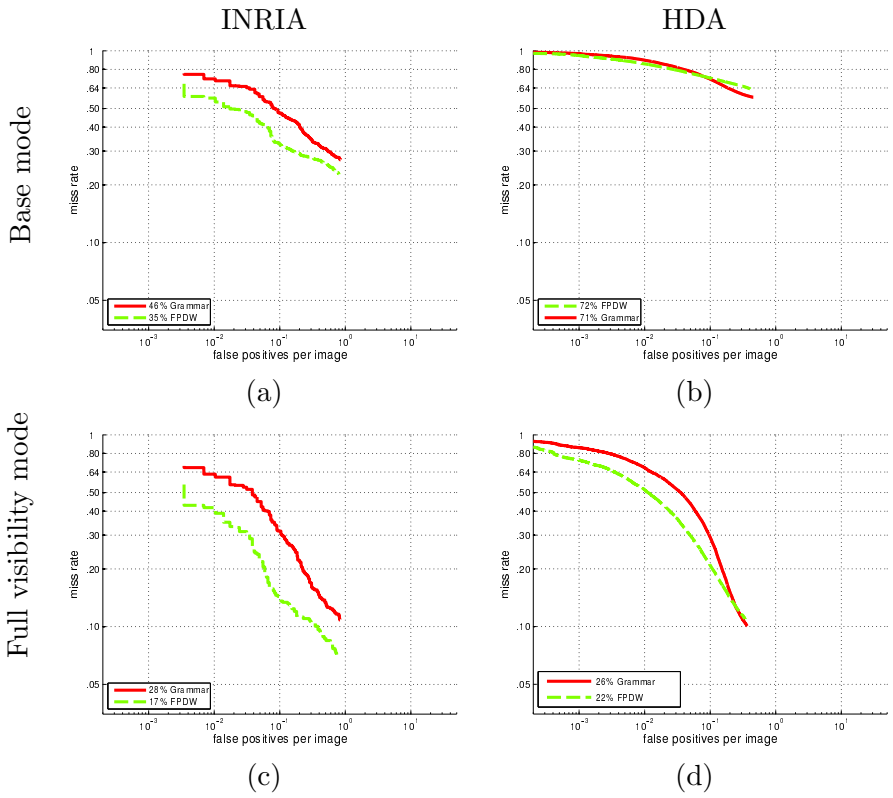


Figure 7.12: MD rate/FPPI plots comparing the performance of FPDW and Grammar Models on the INRIA and the HD data sets. There is a clear difference in the performances on the two data sets, confirming the usefulness of creating a data set for PD representing a Video Surveillance scenario. Results on other data sets for FPDW and Grammar Models are publicly available (see [Dollár et al., 2012b] and [Girshick et al., 2011], respectively).

I downsample the images of the high and medium resolution sequences to a resolution of 640×400, using bilinear interpolation, and compare the performance of the detectors on the two versions. I set the detectors up so that they are able to detect pedestrians at all of the imaged heights that appear in the data set, both at the original and at the downsampled resolution.

Two effects contribute to change the performance of a detector when it is run on the lower resolution version of an image. First, the detector incurs in more MD's due to the aforementioned phenomenon. Second, the "sliding window" paradigm of the PD algorithms implies that a detector has to evaluate a much smaller number of windows when run on the lower resolution version of an image. Statistically, this leads to the detector generating less FP's, while the number of TP's remains constant. The two contributions create a balancing effect. It must be noticed that evaluating a different number of windows translates to different execution times for the two resolution modes, detecting pedestrians in low resolution images being, of course, faster.

Figure 7.13 graphically presents the results. The performance of Grammar Models on the far range sequences degrades when run on lower resolution images, while the performance of FPDW doesn't change significantly (see Figure 7.13(a)). I observe that the farthest, and thus, smallest pedestrians are the ones where the difference is felt. I speculate that Grammar Models suffers more than FPDW from the reduction in resolution because it relies on part detectors which require finer image details to work at their best.

For the mid range sequence, the performance of both algorithms changes very little when changing the resolution (see Figure 7.13(b)). I speculate this happens because the people in this sequence are not small enough to trigger the phenomenon that can be observed in the long range sequences. For the short-range sequences a clear reduction in FP's (especially for FPDW) at high FPPI rates/low detection confidence can be observed. I ascribe this reduction to the smaller number of windows classified, as well as to the smoothing of image compression artifacts performed by the bilinear interpolation: image areas containing artifacts which are classified as persons with a low confidence in the original images, are classified as background in the LR images. Camera 59 was excluded from the short range set because of

a specific occurrence: in the original size images, the fire extinguisher sign was very often detected as a person by FPDW, while this did not happen in the subsampled images. This peculiar event had a big influence on the results, but was deemed not to be of general interest.

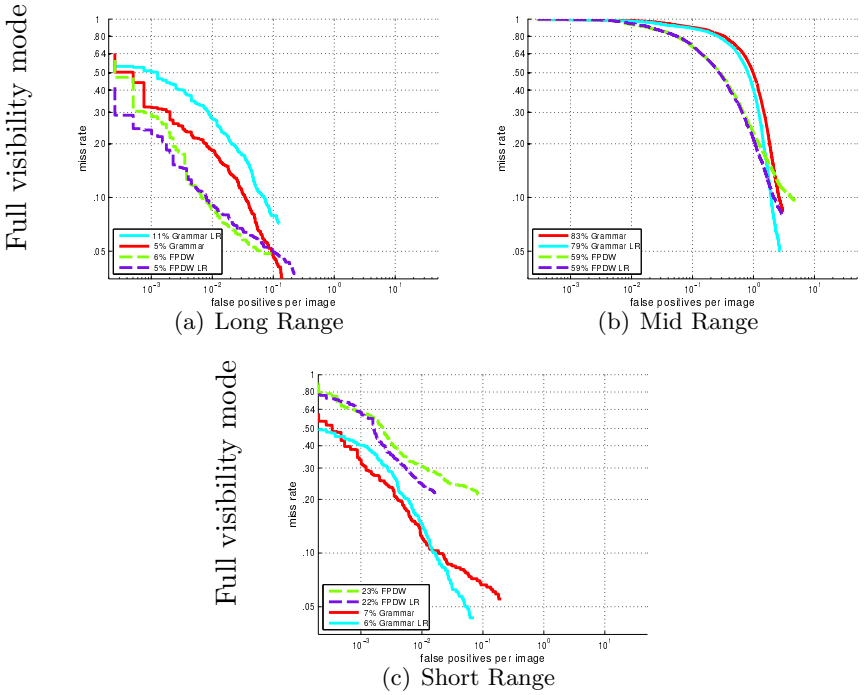


Figure 7.13: MD rate/FPPI plots comparing the performance of the detectors on the original and the Low Resolution (LR) version of the images, for the three ranges I defined. Long range comprises camera 50 and 60, mid range is just camera 54, while short range includes cameras 53, 56, 57 and 58.

Conclusion This experiment shows that HR cameras can be advantageous for detecting people, especially at greater distances, but at the same time this approach has a tendency for generating more FP's and is slower than using standard resolution cameras. When processing time is not an issue, the practitioner should set each camera to the lowest possible resolution that enables the detection of people at the farthest visible point of a given scenario. This ensures the least amount of MD's while also minimizing the number of FP's.

7.3 Pedestrian Re-Identification on the HDA Data Set

In these experiments we evaluate the HDA data set from the RE-ID point of view. Moreover, we exploit the High Definition images of HDA to assess the role of resolution in RE-ID.

We use a RE-ID architecture which is common to many algorithms in the state of the art [Cheng et al., 2011b; Figueira et al., 2013]. The input is provided in the form of cropped images, both for training (gallery) and for testing (probes). A body-part detector is run on each input image and one colour histogram is built based on each detected region. Then, the histograms are concatenated to form the feature vector for classification. Finally, we use a simple Nearest Neighbour classifier for assigning a person ID to a test example. We use two different algorithms to detect body parts: Andriluka's [Andriluka et al., 2009] (PS) and Fenzenswalb's [Girshick et al.] (DTDPM v5).

To evaluate the performance of RE-ID algorithms we use the Cumulative Matching Characteristic (CMC) curve. Such curve shows how often the correct person ID is included in the best K matches for each test image. The overall algorithm performance is measured by the nAUC, the normalized Area Under the CMC curve. The larger the area, the better the RE-ID performance.

7.3.1 Experiment 7 - Comparing RE-ID Performance on HDA and Other Data Sets

This experiment compares the RE-ID performance on the HDA data set with that obtained on other publicly available data sets (CAVIAR4REID, iLIDS4REID and VIPeR, described in Section 2.3.2). The design of the experiment is different from that of Experiment 5. In that case the performance of pre-trained detectors was evaluated on different data sets, while in this experiment one evaluation consists in training and testing one RE-ID algorithm on a given data set.

The gallery and the probes (training and test data) for the HDA data set were collected by manually selecting one detection per person, per camera. In total, we collected 250 detections from 67 pedestrians. Only instances in which people are fully visible were chosen for this experiment. For the other

data sets we used the gallery and probes they provide.

As common in RE-ID work, the evaluation was repeated 100 times randomising the data and the results are reported after averaging over the repetitions. For each subset, the training and the test set for one of the 100 repetitions were formed selecting one image per pedestrian for training and one image per pedestrian for testing. Results are shown in Figure 7.14.

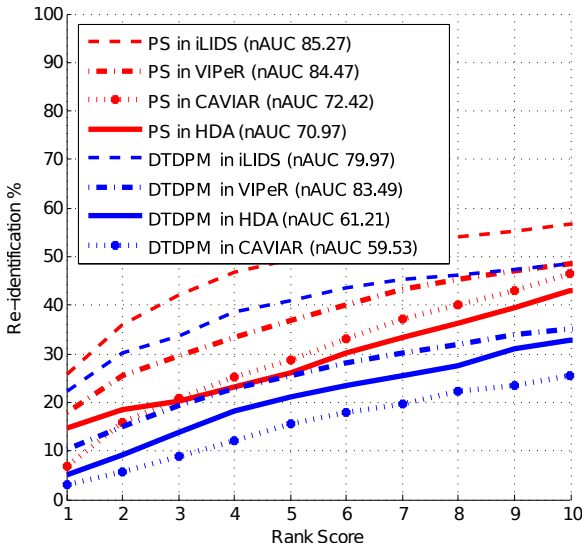


Figure 7.14: Comparing RE-ID performance on different data sets iLIDS, VIPeR, CAVIAR and HDA. PS and DTDPM indicate the algorithm used to detect the body parts used as a base for building the colour histograms. See text for details. HDA proves to be one of the most challenging data sets (low nAUC and low correct re-identifications at low ranks).

Conclusion We observe that, together with CAVIAR4REID, the HDA data set is one of the most challenging. The difficulty of performing RE-ID on CAVIAR4REID stems mostly from the low resolution of its images: it is difficult to differentiate people even for a human operator. The difficulty in re-identifying people on the HDA data set, on the other hand, stems from the mixture of cameras with different resolutions, different perspectives and ranges, the presence of harsh illumination changes, severe occlusions, and the fact that several subjects add or remove items of clothing from one view to the next (i.e., removing a jacket), making HDA one of the most challenging RE-ID data sets to date.

7.3.2 Experiment 8 - RE-ID Performance at Different Image Resolutions

This RE-ID experiment is aimed at evaluating the influence of using high vs. low resolution cameras on the RE-ID performance. In principle, higher resolution should allow for the extraction of more discriminative features, which should lead to better RE-ID performances.

High Resolution Cameras VS Low Resolution Cameras

For the first part of the experiment we use actual high and low resolution data: we partition the HDA data set based on resolution and compare the RE-ID performance on the different subsets. The two gallery and probe sets for this experiment result from a partition of the gallery and probe sets described in Experiment 7. Detections generated on images from camera 50 and above were used for the HR set, while detections generated on images from camera 40 and below for the LR set (see Table 5.2). In total, we collected 150 detections from 35 pedestrians for the HR subset and 100 detections from 32 pedestrians for the LR subset. The evaluation was repeated 100 times randomising the data and the results are reported after averaging over the repetitions, using the same scheme as in Experiment 7.

In Figure 7.15 we can observe that both PS and DTDPM v5 perform better in high resolution cameras than in regular ones, supporting the hypothesis that high resolution images carry more discriminative information than low resolution ones. It is important to highlight, though, that in this part of the experiment we compare two sets of images which differ for more factors than just the resolution, e.g., illumination and camera perspective. Such factors could be the cause for the difference in performance.

Isolating the Effect of Resolution

In the second part of this experiment we compare the RE-ID performance on the high resolution subset with that on the same subset, after its resolution is artificially reduced to VGA via bilinear interpolation. This allows us to measure the effect exclusively due to the difference in resolution. We performed the same analysis as in the first part of the experiment and, in this case, no strong difference in the RE-ID performance was observed between the high resolution and the low resolution case.

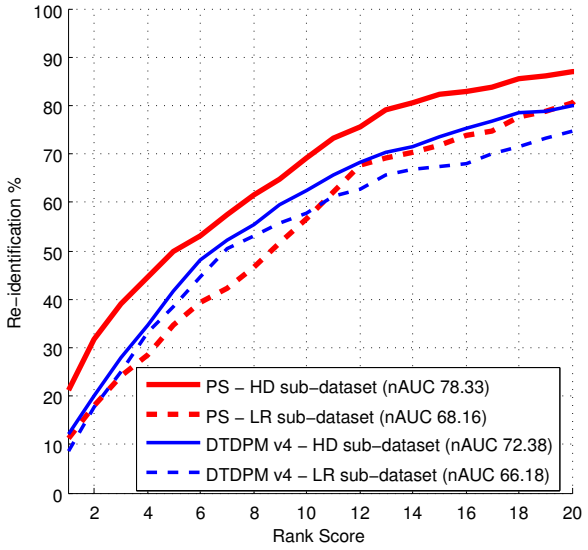


Figure 7.15: CMC curves comparing performances on High Resolution (HR) and Low Resolution (LR) subsets of the HD data set (higher curves correspond to better performance). The RE-ID performance is higher on the HR set, but this appears to be related to other factors peculiar to the two set of images, rather than resolution. See text for details.

Conclusion No clear evidence exists for recommending the use of high-resolution images for RE-ID with the algorithm we employed. The improvements obtained in the first part of this experiment may be due to better image quality or simpler environmental conditions in the high resolution subset, or to some other factor we did not control during the experiment. Nevertheless, we believe that more discriminative features than the ones used in this work (simple colour histograms) can better exploit high-resolution information, leading to a better RE-ID performance on high-resolution images.

7.4 Experiment 9 - PD for Fully Automated Re-Identification

In this last experiment we evaluate the performance of a fully automated Re-Identification (PD+REID) system, comparing it to the performance reported by the same RE-ID module in a classical RE-ID experiment. We introduce the False Positive (FP) class and the Occlusion Filter (OF), as improvements over the naive integration scheme. We show that the use of the FP class enables a meaningful comparison between the performance computed in a classical RE-ID experiment and that measured on the integrated system. We show the usefulness of the precision and recall statistics to characterise RE-ID performances and, finally, we observe that the use of the FP class and the OF have a positive impact on RE-ID performance of the integrated system.

7.4.1 Setup

We use the HDA data set (see Chapter 5), the ACF detector (see Chapter 3) and the same basic algorithm described in Section 7.3 for RE-ID (using the PS body part detector by Andriluka [Andriluka et al., 2009]). We simulate the closed-space assumption for the RE-ID problem by using the images from 7 cameras (ID's from 50 to 59) for creating the gallery (training set), while images from camera 60 are used for collecting the probes (test set).

7.4.2 Baselines: the MANUAL modes

We define two evaluation modes based on hand-cropped images of people: the "MANUAL$_{clean}$" and the "MANUAL$_{all}$" modes. The "MANUAL$_{clean}$" implements the *de facto* standard evaluation method for RE-ID algorithms: the test examples are hand-cropped images of fully visible people. The "MANUAL$_{all}$" mode is more challenging, because it uses all of the hand-labelled persons in the test set, including the partially visible ones. First, we compare the RE-ID results obtained using the two manual modes, then we compare such results with those obtained by the integrated PD+REID system.

The annotation for camera 60 comprises 1182 BB's. Because of the closed-space assumption, we build the test set for the MANUAL$_{all}$ evaluation modality using only the 1097 BB's which depict people who appear both

in camera 60 and in some of the training cameras. For the MANUAL$_{clean}$ modality, we use only the fully visible pedestrians in the test set: 467 out of 1097.

For the experiments with the PD+REID system, the ACF detector produces 2579 BB's in the test video sequence, 1167 of which are FP's. In the evaluation modes in which the OF is active, 233 detections are filtered out, leaving a total of 2309 BB's.

The results of the experiments are visualized in Figure 7.16 (CMC curves) and Figure 7.17 (P/R points). Table 7.6 lists the corresponding numerical values: the precision and recall statistics for rank-1, including the F-score (harmonic mean of precision and recall), while CMC curves are summarized by the values of the rank-1 point and those of the normalized area under the curves (nAUC).

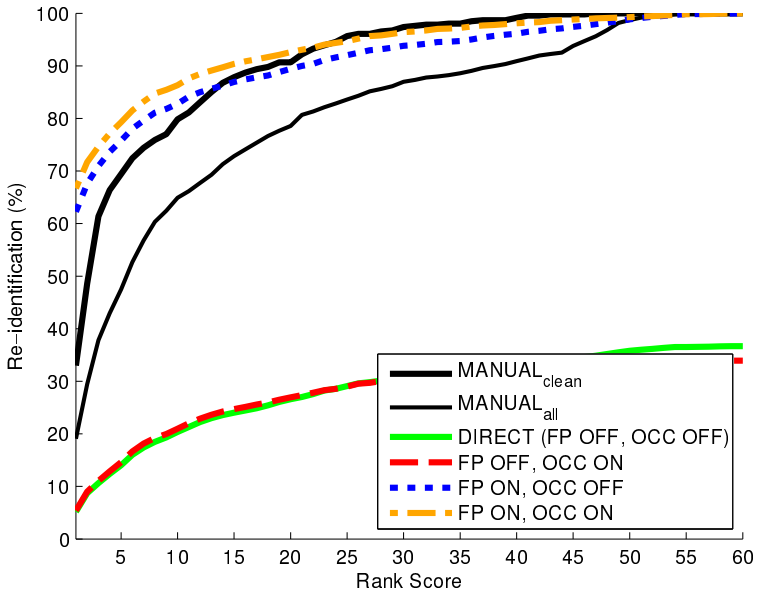


Figure 7.16: CMC curves of the six RE-ID experiments with the experimental setups described in the text.

Comparing the MANUAL$_{all}$ with the MANUAL$_{clean}$ experiment allows us to measure the difference in performance caused by the introduction of partially occluded exemplars in the test set. The MANUAL$_{clean}$ and MANUAL$_{all}$ baseline cases perform as expected. MANUAL$_{clean}$ receives the cleanest possible input (only fully visible pedestrians) and exhibits the

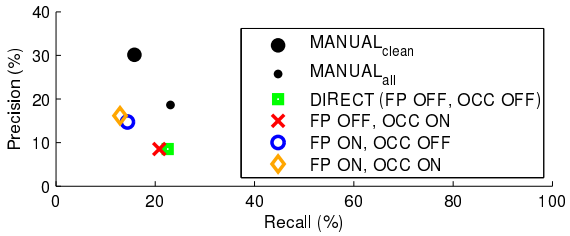


Figure 7.17: Visualization of the precision-recall values as per Table 7.6.

Exp.	# BB's	Precision (%)	Recall (%)	F-score (%)	1st Rank (%)	nAUC (%)
MANUAL$_{clean}$	462	30.1	15.8	20.7	18.7	90.6
MANUAL$_{all}$	1097	18.6	23.1	20.6	31.2	82.4
DIRECT (FP OFF, OCC OFF)	2542	8.5	22.5	12.4	5.2	29.1
FP OFF, OCC ON	2309	8.5	20.8	12.1	5.4	27.8
FP ON, OCC OFF	2542	14.7	14.4	14.6	62.3	91.3
FP ON, OCC ON	2309	16.2	12.9	14.4	66.7	93.5

Table 7.6: Statistics for the different evaluation modes presented in the text. We list the number of Bounding Boxes (**# BB's**) processed in each case, and report the results in terms of **Precision**, **Recall**, **F-score**, the CMC curves' **1st rank** and its normalized area (**nAUC**). Note that we define the precision and recall statistics so that they are not affected by the quality of RE-ID in the FP class, while CMC is affected by them. This leads to the very high **1st Rank** values for the modes with FP class turned ON. These values are of little practical interest. The precision-recall values can be visualized in Figure 7.17.

highest precision of all experiments. MANUAL$_{all}$, on the other hand, receives BB's for all the pedestrian appearances (including the ones affected by partial visibility) and reaches the highest values for recall. The CMC curve of MANUAL$_{clean}$ outperforms that of MANUAL$_{all}$ for all ranks, confirming that partial occlusion is detrimental for RE-ID.

7.4.3 Naive integration

Comparing the two manual modes with the naive integration of PD and RE-ID (the DIRECT mode) we observe a drop in performance for the PD+REID system (see Figure 7.16). This loss is mostly due to the fact that FP detections are always misclassified in the DIRECT mode. The CMC curve is thus limited in the highest accuracy value it can reach. Such loss is mitigated when

we consider the two integration modules discussed in Chapter 6: the FP class and the OF.

7.4.4 Dealing with False Positives

We highlight the improvement provided by using the FP class in Figure 7.16. Such gain is due to two factors: (i) Adding a FP class allows all detections to be correctly identified at some rank, enabling the CMC curve to reach 100% accuracy; (ii) In this experiment, most of the FP are quite easy to re-identify, as they are generated by static objects in the scene (i.e, doors, fire extinguishers). This causes the low-rank part of the curve to lay even higher than that of the manual modes. Note that we define the precision and recall statistics so that they are not affected by the quality of RE-ID in the FP class.

7.4.5 Dealing with Partial Occlusion

In this experiment the OF was set to reject BB's with an occlusion of at least 30% (See Figure 7.18). The improvement afforded by the OF when the FP class is activated can be seen in Figure 7.16: the orange dash-dot curve is always higher than the blue dotted one. The results reported in Table 7.6 comply with our expectations. Filtering out difficult to re-identify cases leads to an increase in precision. At the same time, because the OF reduces the number of detections passed on to the RE-ID module, it induces a drop in recall. When the FP class is deactivated, comparing between the "DIRECT" vs "FP OFF, OCC ON" modes, we do not see any increase in accuracy or precision, possibly because the general values for these statistics are so low that the increase afforded by the OF is not significative.

Conclusion In this experiment we defined two evaluation modes for RE-ID based on GT bounding boxes. We showed that the mode which includes only fully visible people (MANUAL_{clean}) achieves the highest RE-ID precision of all modes, while the one which includes all appearances, disregarding the degree of occlusion (MANUAL_{all}), achieves the highest recall. We showed that the naive integration of PD and RE-ID achieves a good recall, but poor precision, and cannot be compared to the manual modes on a CMC curve. Integrating a FP class leads to an increase in precision, at the price of a drop in recall. Using the FP class enables comparing the integrated system

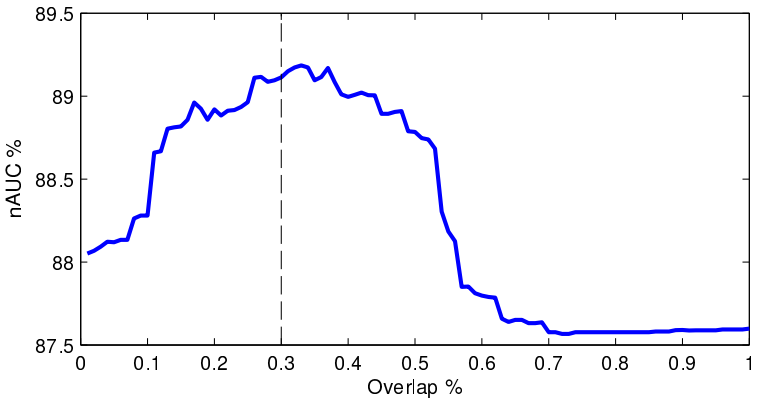


Figure 7.18: The performance of a PD+REID system as a function of the value of the minimum overlap threshold for the OF. The plot refers to an experiment that we reported in [Taiana et al., 2014]. That experiment is similar to Experiment 9, so we decided to include only this plot in this thesis. The maximum in the plot indicates that it is best for the OF to reject detections with an overlap value of at least 0.3.

to a manual one via a CMC curve, but the comparison is not completely fair, because FP examples are easier to re-identify than people. Using the OF in combination with the FP class leads to a small improvement in precision and a correspondingly small drop in recall. The proposed system with FP class and OF achieves the highest precision among the fully automated methods.

Chapter 8

Conclusions

This thesis focusses on the field of Pedestrian Detection (PD) and its applications. The work concentrated on three main areas. First, I explored the effect of data labelling on the training and on the evaluation of Pedestrian Detectors. I concluded that a careful selection of the labelled examples in the training set may have a significant impact on detection performance. I showed that a fair comparison of detection algorithms depends on the quality of the test set labelling and experiment design. I confirmed that partial occlusion in the test samples has a negative correlation with detection performance. Second, I designed a data set to measure PD performance in a Video Surveillance scenario and assess the effect of High Definition images on the performance of Video Surveillance algorithms. Third, I proposed the integration of a PD algorithm and a Re-Identification (RE-ID) module to build a fully automated Re-Identification (PD+REID) system. I introduced two schemes which improved the performance of the naive integrated system: the False Positive class and the Occlusion Filter. Furthermore, I highlighted the usefulness of precision and recall statistics to characterize the performance of the integrated system. Each contribution is reported in detail in the following paragraphs.

I proposed a new labelling for the popular INRIA person data set, which enables the user to assess the effect of impure data on PD's. The proposed labelling is richer and more accurate than the original one, making the data set better suited for benchmarking modern PD systems. I showed that selecting the correct height range for the test samples used in the evaluation is important for a fair comparison of the detection performances of various

algorithms. I confirmed that the degree of partial occlusion of test samples negatively correlates with detection accuracy, even for part-based detectors. I observed that including partially occluded examples (with a visibility of at least 90%) in the training set improves the detection performance both on fully visible and on partially visible pedestrians. Moreover, I 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 remains unchanged. This result is especially relevant for the case of automotive applications, in which detecting pedestrians far from the vehicle allows sufficient time for the automated driving system or the driver to respond. Summarizing, I showed that for achieving the best possible performance it is useful to include examples affected by a low level of impurity in the training set of a detector. The results related to the effect of sample purity on detection have been published in [Taiana et al., 2013, 2015].

Another contribution of this work was to design and create the High Definition Analytics (HDA) data set for benchmarking Video Surveillance algorithms and assess the role of High Definition images on their performance (See Chapter 5). The set up of HDA allows for the evaluation of Pedestrian Detection algorithms in a video analytics scenario, as well as the benchmarking of PD+REID algorithms. We believe that the HDA benchmark will stimulate the development of Video Surveillance algorithms specific for High Definition images. The HDA data set is heterogeneous, including footage acquired with different resolutions, from different view points, under different lighting conditions, etc. Such diversity is a key for a robust evaluation of the performance of Video Surveillance algorithms. I evaluated the performance of two PD systems which are representative of the two main paradigms in the state of the art: the Fastest Pedestrian Detector in the West (FPDW) [Dollár et al., 2010] for monolithic detectors and Grammar Models [Girshick et al., 2011] for part-based detectors on various scenarios of HDA. The Grammar Model detector proved to have an edge when detecting people imaged at a short range. I believe such advantage stems from the ability of a part-based detector to accommodate shifts of the body parts with respect to their most common position. The monolithic detector, on the other hand, performs slightly better in the condition of full visibility. Detection performance on the HDA and INRIA data sets proved to be considerably different, confirm-

ing the usefulness of creating a data set for PD which represents the Video Surveillance scenario. Experiments on the role of High Definition confirmed the intuition that higher definition images allow for the detection of farther pedestrians, but due to the image pyramid/sliding window scheme they also lead to more False Positive detections and require longer processing times. In an effort to evaluate the characteristics of the HDA data set in terms of RE-ID, we compared the performance of two versions of a simple RE-ID system on HDA and other RE-ID data sets. We observed that the HDA data set is, together with CAVIAR4REID, the most challenging to date, arguably because of the mixture of cameras with different resolutions, different perspectives and ranges, the presence of harsh illumination changes, severe occlusions, and the fact that several subjects add or remove items of clothing from one view to the next. Furthermore, we performed experiments aimed at evaluating the effect of high resolution images on RE-ID performance. The results of such experiments show that for basic RE-ID algorithms, using simple colour histograms as features, high resolution images are not advantageous. Nevertheless, we expect that more sophisticated features would better exploit high resolution information, leading to a better RE-ID performance on high resolution images. We believe this makes the HDA data set a valuable tool for the RE-ID community to explore features specific to high resolution images. The description of the HDA data set and the results of the related PD and RE-ID experiments have been published in [Nambiar et al., 2014].

The last contribution of this thesis consists in the design of a fully automated Re-Identification (PD+REID) system (see Chapter 6). The classic set up for a RE-ID experiment requires human intervention for the selection of the test examples, leading to RE-ID systems of little practical use. A fully automated Re-Identification system consisting in the integration of a RE-ID and a RE-ID module allows to overcome such limitation. We showed that precision and recall statistics are useful for characterising the performance of RE-ID and PD+REID systems alike. We proposed two improvements to a naive fully automated Re-Identification system: the False Positive (FP)False Positive class models the FP detections generated by the PD module in a given scenario. The use of the FP class leads to an increase in RE-ID precision, at the price of a drop in recall. Furthermore, it enables a meaningful evaluation of the combined system with a Cumulative Matching Characteristic (CMC). The Occlusion Filter (OF) exploits geometrical reasoning to

filter the detections, so that only detections which have a high probability of depicting fully visible pedestrians are passed on to the RE-ID module. The introduction of the OF leads to a slight improvement in the precision of the RE-ID system (thanks to the removal of ambiguous and hard to classify detections), at the same time inducing a corresponding drop in recall. The results related to the development of a fully automated Re-Identification system have been published in [Taiana et al., 2014; Figueira et al., 2014]·

8.1 Future work

In the following paragraphs I list concrete steps for possible developments of the work presented in this thesis.

Positive Training Example Value – Lapedriza et al. notice in [Lapedriza et al., 2013] that some positive examples are better than others for training detectors. They define the concept of training value for an example as the performance of a detector trained using only such positive example and a negative training set. The performance is evaluated on the rest of the training set, including positive and negative examples. The authors show that excluding the positive examples with the worst value from the training leads to an increase in detection performance, with respect to the base case (see Figure 8.1). I plan to compare the training value of examples with their level of impurity. I expect that the level of occlusion negatively correlate with training value. If this holds, it would be interesting to characterise the examples which have a low level of impurity and, at the same time, a low training value. Such examples might indicate sources of impurity I am not yet considering, e.g., image blur and unusual poses. I expect that pursuing this development would require three months' work in case the detector implemented during this thesis can be easily integrated in Lapedriza's scheme. In case the use of the Exemplar Support Vector Machine used by Lapedriza and described in [Malisiewicz et al., 2011] turns out to be necessary, I expect the achievement of meaningful results to be delayed by one month.

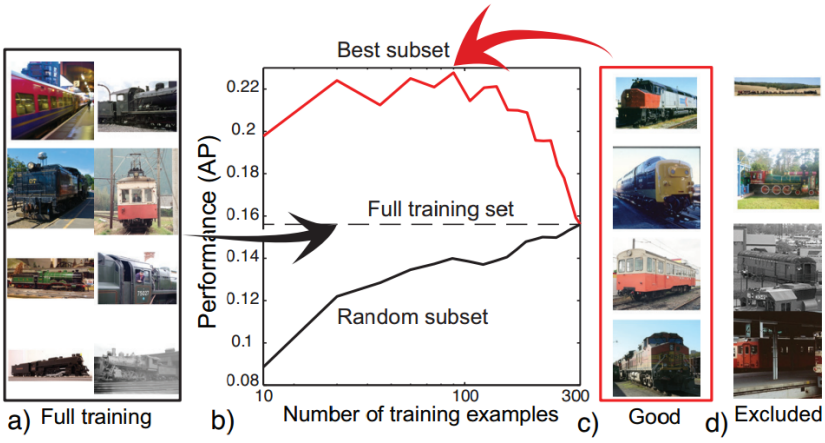


Figure 8.1: Visual explanation of the concept of "positive training value", image reproduced from [Lapedriza et al., 2013]. (a) lists positive examples from the entire training set. (c) and (d) show positive training examples with high and low training value, respectively. (b) plots the performance of two detectors as a function of the number of positive examples used in their training. The black line corresponds to a detector for which the positive training examples are added to the training set in no particular order, while for the detector corresponding to the red line, the examples are added to the training set in decreasing order of training value. The performance of the "red" detector dominates that of the "black" detector, testifying to the usefulness of computing "positive training values". The maximum of the red plot corresponds to the training set which leads to the best detection performance. Interestingly, such training set is a subset of the total training set, meaning that it is better to discard the positive examples with the worst value before training.

Detecting Waving People from a Mobile Robot – Pedestrians Detectors do not rely on the assumption of static cameras. This feature permits their deployment on mobile robots, with clear applications in the field of Human-Robot Interaction. Running a Pedestrian Detector on mobile robots like Vizzy (see Figure 8.2) would enable them to be aware of the presence of people in their surroundings. However, once a robot reaches such level of awareness, a mechanism for humans to attract its attention becomes necessary. Such mechanism can be provided by connecting a Pedestrian Detector to a gesture detector, such as the waving detector described in [Moreno and Santos-Victor, 2013]: exploiting such a system, humans would just need to wave one hand at the robot to require assistance. The integration of a Pedes-

trian Detector with a gesture detector requires additional work to associate people detections along consecutive frames and to manage PD errors: False Positives and Missed Detections. Such work can be performed by a Particle Filter-based tracker (like the one described in [Okuma et al., 2004]), which must be aware of the movements of the robot and of how such movements influence the projection of people onto the image. I expect the development of the tracker and the integration of the Pedestrian Detector, the tracker and the hand waving detector to take approximately five months.

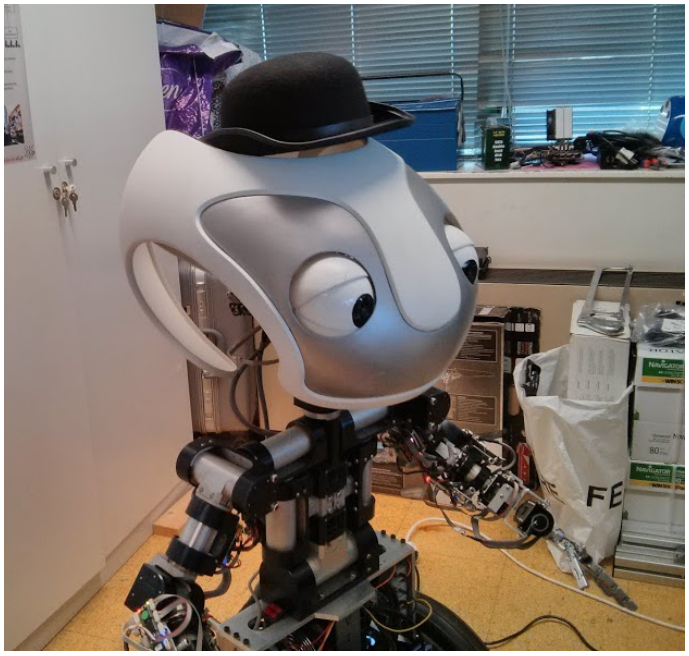
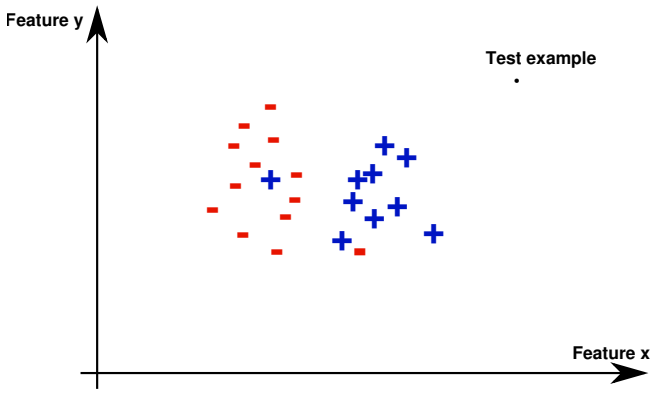


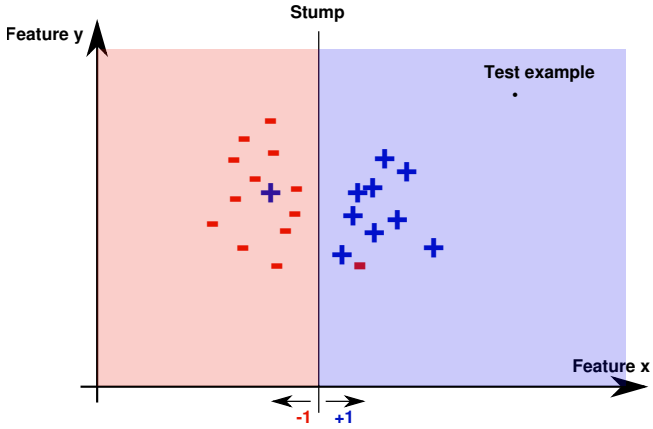
Figure 8.2: Vizzy, the mobile robot of VisLab.

Bounded Stumps as Weak Classifiers – Decision stumps are Weak Classifiers very frequently employed in Pedestrian Detectors. One decision stump partitions the range of values possible for a feature according to one threshold, into a positive and a negative interval (see Figure 8.3(b)). However, a decision stump does not differentiate between value ranges of the feature for which there is evidence in the training set (there are positive or negative examples) and the ranges for which there is no training evidence: examples in both areas are classified with the same level of confidence. This appears to be suboptimal. Bounded stumps, stumps which cast a neutral vote for the areas of the feature space with no training evidence, seem to be-

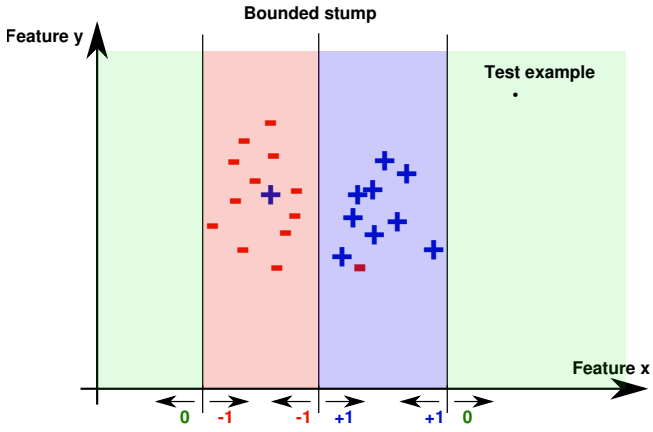
have in a more sensible way in the aforementioned case (see Figure 8.3(c)). Bounded stumps consist in three thresholds: the usual threshold separating positive and negative examples, and two additional thresholds which delimit the ranges of the feature for which no training information is available. I plan to compare the efficacy of normal decision stumps and bounded decision stumps in a Pedestrian Detector based on AdaBoost. However, some peculiarities of the current PD approaches, namely the fact that some of the features are bounded (Integral Channel Features, as well as Histograms of Oriented Gradients saturate at the value of 0.2) might limit the impact of using bounded stumps. I expect this line of work to require two months for producing results.



(a) Data for a toy classification problem



(b) Decision stump classification



(c) Bounded stump classification

Figure 8.3: Bounded stumps as weak classifiers. (a) depicts a toy classification problem with two features, positive and negative training examples (indicated by the + and - symbols, respectively) and one test example. (b) shows the classification operated by a decision stump learnt on the problem: examples in the light blue and light red areas are classified as positives and negatives, respectively. (c) depicts the classification operated by a bounded decision stump. Compared to a regular stump, this classifier benefits from the addition of two light green areas, which cast a neutral vote. Such areas correspond to the ranges of the decision feature for which there is no evidence in the training set.

Bibliography

Caviar data set website. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

Mobileye website. http://www.mobileye.com.

US Department of Transportation, traffic safety facts. http://www-nrd.nhtsa.dot.gov/Pubs/811888.pdf, 2012.

Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. *CVPR*, 2009.

Slawomir Bak, Etienne Corvée, Francois Brémond, and Monique Thonnat. Boosted human re-identification using Riemannian manifolds. *ImaVis*, 2012.

Aharon Bar-Hillel, Dan Levi, Eyal Krupka, and Chen Goldberg. Part-Based Feature Synthesis for Human Detection. *ECCV*, 2010.

Peter G Bartlett. Pedestrian detection system, 1969. US Patent 3,462,692.

Rodrigo Benenson. PAPER CURRENTLY UNDER REVIEW. 2015.

Rodrigo Benenson, Markus Mathias, Radu Timofte, and Luc Van Gool. Pedestrian detection at 100 frames per second. *CVPR*, 2012.

Rodrigo Benenson, Markus Mathias, Tinne Tuytelaars, and Luc Van Gool. Seeking the strongest rigid detector. *CVPR*, 2013.

Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? *ECCV Workshop*, 2014a.

Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? *ECCV Workshop talk slides*, 2014b.

Alberto Broggi, Pietro Cerri, Stefano Ghidoni, Paolo Grisleri, and Ho Gi Jung. A new approach to urban pedestrian detection for automatic braking. *Intelligent Transportation Systems*, 2009.

Sebastian Brutzer, Benjamin Hoferlin, and Gunther Heidemann. Evaluation of background subtraction techniques for video surveillance. *CVPR*, 2011.

Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. *BMVC*, 2011a.

Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. *BMVC*, 2011b.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *ML*, 1995.

Etienne Corvee, Slawomir Bak, and Francois Bremond. People detection and re-identification for multi surveillance cameras. *VISAPP*, 2012.

Marco Cristani, Michela Farenzena, Domenico Bloisi, and Vittorio Murino. Background subtraction for automated multisensor surveillance: a comprehensive review. *EURASIP Journal on Advances in Signal Processing*, 2010.

Marco Cristani, R Raghavendra, Alessio Del Bue, and Vittorio Murino. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, 2013.

Navneet Dalal. INRIA Person Dataset. http://pascal.inrialpes.fr/data/human/, 2005.

Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. *CVPR*, 2005.

Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. *ECCV*, 2006.

Floris De Smedt and Toon Goedemé. Open framework for combined pedestrian detection. *VISAPP*, 2015.

Piotr Dollár, Z Tu, H Tao, and S Belongie. Feature mining for image classification. *CVPR*, 2007.

Piotr Dollár, Z. Tu, and P. Perona. Integral channel features. *BMVC*, 2009.

Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. 2009.

Piotr Dollár, Serge Belongie, and Pietro Perona. The Fastest Pedestrian Detector in the West. *BMVC*, 2010.

Piotr Dollár, R Appel, and W Kienzle. Crosstalk Cascades for Frame-Rate Pedestrian Detection. *ECCV*, 2012a.

Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian Detection: An Evaluation of the State of the Art. *PAMI*, 2012b.

Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *PAMI*, 2014.

Piotr Dollár. Caltech pedestrian detection benchmark, detection accuracy under different conditions. http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/rocs/UsaTestRocs.pdf, a.

Piotr Dollár. Caltech pedestrian detection benchmark, algorithms runtime performance. http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/files/timing.pdf, b.

Piotr Dollár. Caltech pedestrian detection evaluation code. http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/DollarEvaluationCode, c.

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

Markus Enzweiler, Angela Eigenstetter, Bernt Schiele, and Dariu M Gavrila. Multi-cue pedestrian classification with partial occlusion handling. *CVPR*, 2010.

Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and Appearance for Mobile Scene Analysis. *ICCV*, 2007.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes (VOC) challenge. *IJCV*, 2010.

Mark Everingham, S M Ali Eslami, Luc Van Gool, Christopher K I Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge – a Retrospective. *IJCV*, 2014.

Pedro F Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. *CVPR*, 2008.

Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010.

Dario Figueira, Loris Bazzani, Ha Quang Minh, Marco Cristani, Alexandre Bernardino, and Vittorio Murino. Semi-supervised multi-feature learning for person re-identification. *AVSS*, 2013.

Dario Figueira, Matteo Taiana, Athira Nambiar, Jacinto Nascimento, and Alexandre Bernardino. The HDA+ data set for research on fully automated re-identification systems. *ECCV Workshop*, 2014.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Computational Learning Theory*, 1995.

Juergen Gall, Angela Yao, Nima Razavi, Luc Van Gool, and Victor Lempitsky. Hough forests for object detection, tracking, and action recognition. *PAMI*, 2011.

Wei Gao, Haizhou Ai, and Shihong Lao. Adaptive contour features in oriented granular space for human detection and segmentation. *CVPR*, 2009.

Álvaro García-Martín, José M. Martínez, and Jesús Bescós. A corpus for benchmarking of people detection algorithms. *Pattern Recognition Letters*, 2012.

Dariu M Gavrila. Pedestrian detection from a moving vehicle. *ECCV*, 2000.

Dariu M Gavrila and Vasanth Philomin. Real-time object detection for "smart" vehicles. *ICCV*, 1999.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *CVPR*, 2012.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013.

Ross B Girshick, Pedro F Felzenszwalb, and McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/~rbg/latent-release5/.

Ross B Girshick, Pedro F Felzenszwalb, and McAllester. Object detection with grammar models. *PAMI*, 2011.

Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *ECCV*, 2008.

Horst-Michael Gross, Christof Schröter, Steffen Mueller, Michael Volkhardt, Erik Einhorn, Andreas Bley, Christian Martin, Tim Langner, and Matthias Merten. Progress in developing a socially assistive mobile home robot companion for the elderly with mild cognitive impairment. *IROS*, 2011.

David Hogg. Model-based vision: a program to see a walking person. *ImaVis*, 1983.

Jan Hosang, Rodrigo Benenson, and Bernt Schiele. How good are detection proposals, really? *ARXIV*, 2014.

Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. Taking a deeper look at pedestrians. *CVPR*, 2015.

Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank. A survey on visual surveillance of object motion and behaviors. *Systems, Man, and Cybernetics*, 2004.

Omid Hosseini Jafari, Dennis Mitzel, and Bastian Leibe. Real-time rgb-d based people detection and tracking for mobile robots and head-worn cameras. *ICRA*, 2014.

Ferryman James and Shahrokni Ali. An overview of the pets2009 challenge. *PETS Workshop*, 2009.

Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei Efros, and Antonio Torralba. Undoing the Damage of Dataset Bias. *ECCV*, 2012.

Christoph H Lampert, Matthew B Blaschko, and Thomas Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *PAMI*, 2009.

Agata Lapedriza, Hamed Pirsiavash, Zoya Bylinskii, and Antonio Torralba. Are all training examples equally valuable? *ARXIV*, 2013.

Alain Lehmann, Bastian Leibe, and Luc Van Gool. Fast prism: Branch and bound hough transform for object class detection. *IJCV*, 2011.

Bastian Leibe, Edgar Seemann, and Bernt Schiele. Pedestrian detection in crowded scenes. *CVPR*, 2005.

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. *CVPR*, 2014.

Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu. Semi-supervised coupled dictionary learning for person re-identification. *CVPR*, 2014.

David G. Lowe. Object recognition from local scale-invariant features. *ICCV*, 1999.

Ping Luo, Yonglong Tian, Xiaogang Wang, and Xiaoou Tang. Switchable deep network for pedestrian detection. *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014.

Subhransu Maji, Alexander C Berg, and Jitendra Malik. Classification using intersection kernel support vector machines is efficient. *CVPR*, 2008.

Tomasz Malisiewicz, Abhinav Gupta, Alexei Efros, et al. Ensemble of exemplar-svms for object detection and beyond. *ICCV*, 2011.

Mayeul Mathias, Rodrigo Benenson, Radu Timofte, and Luc Van Gool. Handling occlusions with franken-classifiers. *ICCV*, 2013.

Alhayat Ali Mekonnen, C Briand, Frédéric Lerasle, and Ariane Herbulot. Fast hog based person detection devoted to a mobile robot with a spherical camera. *IROS*, 2013.

Andreas Mogelmose, Thomas B Moeslund, and Kamal Nasrollahi. Multimodal person re-identification using RGB-D sensors and a transient identification database. *IWBF*, 2013.

Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio. Example-based object detection in images by components. *PAMI*, 2001.

Plinio Moreno and José Santos-Victor. Waving detection using the fuzzy-boost algorithm and flow-based features. *Image Analysis and Recognition*, 2013.

Roland Morzinger, Marcus Thaler, Severin Stalder, Helmut Grabner, and Luc Van Gool. Improved person detection in industrial environments using multiple self-calibrated cameras. *AVSS*, 2011.

Stefan Munder and Dariu M Gavrila. An experimental study on pedestrian classification. *PAMI*, 2006.

Woonhyun Nam, Piotr Dollár, and Joon Hee Han. Local decorrelation for improved detection. *NIPS*, 2014.

Athira Nambiar, Matteo Taiana, Dario Figueira, Jacinto Nascimento, and Alexandre Bernardino. A multi-camera video data set for research on high-definition surveillance. *Int. Journal of Machine Intelligence and Sensory Signal Processing*, 2014.

Tayyab Naseer, Jurgen Sturm, and Daniel Cremers. Followme: Person following and gesture recognition with a quadrocopter. *IROS*, 2013.

Kenji Okuma, Ali Taleghani, Nando De Freitas, James J Little, and David G Lowe. A boosted particle filter: Multitarget detection and tracking. *ECCV*, 2004.

Michael Oren, Constantine Papageorgiou, Pawan Sinha, Edgar Osuna, and Tomaso Poggio. Pedestrian detection using wavelet templates. *CVPR*, 1997.

Wanli Ouyang and Xiaogang Wang. A discriminative deep model for pedestrian detection with occlusion handling. *CVPR*, 2012.

Wanli Ouyang and Xiaogang Wang. Single-pedestrian detection aided by multi-pedestrian detection. *CVPR*, 2013.

Wenzhuo Ouyang, Xuan Zeng, and Xiongfei Wang. Single-pedestrian detection aided by 2-pedestrian detection. *PAMI*, 2015.

Marco Pedersoli and Andrea Vedaldi. A Coarse-to-fine approach for fast deformable object detection. *CVPR*, 2011.

Bojan Pepikj, Michael Stark, Peter Gehler, and Bernt Schiele. Occlusion patterns for object class detection. *CVPR*, 2013.

Leonid Pishchulin, Thorsten Thorm, and Max Planck. Articulated People Detection and Pose Estimation: Reshaping the Future. *CVPR*, 2012.

Karl Rohr. Incremental recognition of pedestrians from image sequences. *CVPR*, 1993.

Javier Ruiz-del Solar, Mauricio Correa, Rodrigo Verschae, Fernando Bernuy, Patricio Loncomilla, Mauricio Mascaró, Romina Riquelme, and Felipe Smith. Bender – A General-Purpose Social Robot with Human-Robot Interaction Abilities. *Human-Robot Interaction*, 2013.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV-SUBMITTED*, 2104.

Paul Rybski, Peter Anderson-Sprecher, Daniel Huber, Chris Niessl, and Reid Simmons. Sensor fusion for human safety in industrial workcells. *IROS*, 2012.

Payam Sabzmeydani and Greg Mori. Detecting pedestrians by learning shapelet features. *CVPR*, 2007.

Enver Sangineto, Marco Cristani, Alessio Del Bue, and Vittorio Murino. Learning discriminative spatial relations for detector dictionaries: An application to pedestrian detection. *ECCV*, 2012.

William R Schwartz, Aniruddha Kembhavi, David Harwood, and Larry Davis. Human detection using partial least squares analysis. *ICCV*, 2009.

Pierre Sermanet, Koray Kavukcuoglu, Sandhya Chintala, and Yann LeCun. Pedestrian detection with unsupervised multi-stage feature learning. *CVPR*, 2013.

Yaser Sheikh, Omar Javed, and Takeo Kanade. Background subtraction for freely moving cameras. *ICCV*, 2009.

Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. Common metrics for human-robot interaction. *Human-Robot Interaction*, 2006.

Matteo Taiana, Jacinto Nascimento, and Alexandre Bernardino. An improved labelling for the INRIA person data set for pedestrian detection. *IbPRIA*, 2013.

Matteo Taiana, Dario Figueira, Athira Nambiar, Jacinto Nascimento, and Alexandre Bernardino. Towards fully automated person re-identification. *VISAPP*, 2014.

Matteo Taiana, Jacinto Nascimento, and Alexandre Bernardino. On the purity of training and testing data for learning: The case of pedestrian detection. *Neurocomputing*, 2015.

Siyu Tang, Mykhaylo Andriluka, and Bernt Schiele. Detection and tracking of occluded people. *IJCV*, 2014.

Antonio Torralba and Alexei Efros. Unbiased look at dataset bias. *CVPR*, 2011.

Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple kernels for object detection. *ICCV*, 2009.

Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. *CVPR*, 2001.

Paul Viola and Michael Jones. Robust real-time face detection. *IJCV*, 2004.

Michael Volkhardt, Friederike Schneemann, and Horst-Michael Gross. Fallen person detection for mobile robots using 3d depth data. *Systems, Man, and Cybernetics*, 2013.

Stefan Walk, Nikodem Majer, Konrad Schindler, and Bernt Schiele. New features and insights for pedestrian detection. *CVPR*, 2010.

Xiaoyu Wang, TX Han, and Schuicheng Yan. An HOG-LBP human detector with partial occlusion handling. *ICCV*, 2009.

Christian Wojek, Stefan Walk, and Bernt Schiele. Multi-cue onboard pedestrian detection. *CVPR*, 2009.

Bo Wu and Ram Nevatia. Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors. *IJCV*, 2007.

Bo Wu and Ramakant Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. *ICCV*, 2005.

Cha Zhang and Paul Viola. Multiple-Instance Pruning For Learning Efficient Cascade Detectors. *NIPS*, 2007.

Shaoting Zhang, Christian Bauckhage, and Armin Cremers. Informed haar-like features improve pedestrian detection. 2014.

Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. *CVPR*, 2014.

Tao Zhao and Ramakant Nevatia. Tracking multiple humans in complex situations. *PAMI*, 2004.

Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. *BMVC*, 2009.

Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. *ECCV*, 2014.