Domain Adaptation of Virtual and Real Worlds for Pedestrian Detection

A dissertation submitted by **David Vázquez Bermúdez** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor of Philosophy**.

Bellaterra, June 17, 2013

Director	**Dr. Antonio López Peña** Dept. Ciències de la computació & Centre de Visió per Computador Universitat Autònoma de Barcelona
Co-Director	**Dr. Daniel Ponsa Musarra** Dept. Ciències de la computació & Centre de Visió per Computador Universitat Autònoma de Barcelona

This document was typeset by the author using LaTeX 2_ε.

The research described in this book was carried out at the Centre de Visió per Computador, Universitat Autònoma de Barcelona.

ISBN:

Printed by Ediciones Gráficas Rey, S.L.

A mis padres y abuela...

The difficulty lies, not in the new ideas,
but in escaping from the old ones
John Maynard Keynes (1883 - 1946)

Acknowledgements

I would like to dedicate some lines to the people and institutions that have supported me during these years.

First, to the Computer Vision Center, the Universitat Autònoma de Barcelona, the Personal Investigador en Formació grant, the Spanish Government (projects TRA2007-62526/AUT, TRA2010-21371-C03-01, TRA2011-29454-C03-01, TIN2011-29494-C03-02 and Consolider Ingenio 2010: MIPRCV (CSD200700018)) and the Catalan Generalitat (project CTP-2008ITT00001).

I would like to thank the members of the tribunal, Dr. M. Enzweiler, Dr. J. Poal and Dr. J.A. Rodriguez, and the European mention evaluators Prof. Dr. Theo Gevers and Dr. A. D. Bagdanov. I also want to thank the anonymous journal and conference reviewers for their enriching comments during my thesis.

To the members of Davantis that introduced me into the research of Computer Vision always with a practical point of view in mind: J. Lluís, X. Miralles, M. Balcells, J. Pagès, M. Blasco, R. Esquius, P. Cunill and J. Capdevila.

My sincere thanks to Dr. D. Gavrila that gave me the opportunity of spending my internship at Daimler AG where I have learnt so much and the great people I met there: Dr. U. Franke, Dr. U. Kressel, Dr. M. Enzweiler, Dr. K. Keller, N. Schneider, F. Flohr and F. Erbs.

During my Ph.D. I was fortunate to have several collaborations. I would like to thank J. Marín, J. Xu, S. Ramos, M. Rao, Y. Socarrás, and D. Gerónimo for their good collaborations in research for enriching the quality of the work. Specially to J. Marín who shared with me this long trip since the first day, for his continued support at good and bad times, the long nights at the CVC and H3, the absurd jokes and friendship over these years. Also to the people involved in the ADAS car project: F. Lumbreras, G. Ros, S. Ramos, J. Xu and H. Piulachs.

To all the Computer Vision Center fellows but specially to Prof. Dr. J. Villanueva, Dr. J. Lladós and Dr. M. Vanrell. To the administrative staff that had the patience to support my extensive requirements, especially to C. Pérez. To the technical staff specially to M. Paz and J. Masoliver. To all my friends from the CVC, especially, to the ones that started this Ph.D. with me: J. Marín, J. C. Rubio, J.M. Gonfaus, J. Bernal, with whom I have shared so many interesting talks, coffees, parties and now I consider part of my best friends. To my friend G. Ros that since the first moment started with me to involve the ADAS members in the reading group, Eco-Driver project, etc. To S. Ramos a brand new member of the ADAS group that I already consider one of my best friends because he is always laughing, disposed to

collaborate, help others, create new projects and he thinks so big. Finally, to the remain members of the ADAS group: Dr. A. Sappa, Dr. J. Serrat, Dr. J.M. Álvarez and A. González.

I am truly indebted and thankful to my advisor, Dr A. M. López. Tanks for the invested time, the extensive writing, the support, the advices and even convincing me to quite smoking. I consider him one of the best persons I have ever met and I admire him. Thanks for the support and guidance he showed me throughout my dissertation writing. I am sure it would have not been possible without his help. I would like also to thank to my co-advisor Dr. D. Ponsa.

I ask for excuses to all my Barceloneta friends, "La Familia", because I didn't care to much about them these last years. Of course, to J. Sobrino, my closest friend since I came to Barcelona. With her I have grown as a person and with her I have shared all the good and bad moments. She makes me laugh as nobody, with her there are no secrets and her friendship is priceless. I don't forget my dear English teacher and friend Elena who always encouraged me to go further. I will be always in debt with her because she convinced me to start this Ph.D. Finally I want to acknowledge to my family in my own language:

Tener una familia como vosotros significa mucho para mí. Sois vosotros los que habéis formado el David que hoy existe. Nunca podré agradecer el apoyo que me habéis dado, ahora y siempre, estando a mi lado sin esperar nada a cambio. Gracias por cariño que me brindáis y la amplitud de miras que me habéis inculcado sin la cual, seguramente, no habría llegado a estar donde estoy.

Finalmente, quiero también agreadecer por su comprensión y confianza a la persona que se ha ganado mi cariño, admiración y respeto. A una persona, de la que no diré su nombre, pero que sabrá quién es.

Miro con ilusión al futuro y a pesar de las dudas que se me plantean, cualquiera que sea el camino que escoja pasará por vosotros.

Abstract

Pedestrian detection is of paramount interest for many applications, *e.g.* Advanced Driver Assistance Systems, Surveillance and Media. Most promising pedestrian detectors rely on appearance-based classifiers trained with annotated samples. However, the required annotation step represents an intensive and subjective task when it has to be done by persons. Therefore, it is worth to minimize the human intervention in such a task by using computational tools like realistic virtual worlds, where precise and rich annotations of visual information can be automatically generated. Nevertheless, the use of this kind of data generates the following question: *can a pedestrian appearance model learnt with virtual-world data work successfully for pedestrian detection in real-world scenarios?*. To answer this question, we conducted different experiments that suggest that classifiers based on virtual-world data can perform well in real-world environments. However, it was also found that in some cases these classifiers can suffer the so called dataset shift problem as real-world based classifiers does. Accordingly, we have designed a domain adaptation framework, V-AYLA, in which we have explored different techniques to collect a few pedestrian samples from the target domain (real world) and combine them with many samples of the source domain (virtual world) in order to train a domain adapted pedestrian classifier. V-AYLA reports the same detection performance as the one obtained by training with human-provided pedestrian annotations and testing with real-world images from the same domain. Ideally, we would like to adapt our system without any human intervention. Therefore, as a first proof of concept we proposed the use of an unsupervised domain adaptation technique that avoids human intervention during the adaptation process. To the best of our knowledge, this is the first work that demonstrates adaptation of virtual and real worlds for developing an object detector. We also assess a different strategy to avoid the dataset shift that consists in collecting real-world samples and retrain with them, but in such a way that no bounding boxes of real-world pedestrians have to be provided. We show that the generated classifier is competitive with respect to the counterpart trained with samples collected by manually annotating pedestrian bounding boxes. The results presented on this Thesis not only end with a proposal for adapting a virtual-world pedestrian detector to the real world, but also it goes further by pointing out a new methodology that would allow the system to adapt to different situations, which we hope will provide the foundations for future research in this unexplored area.

Resumen

La detección de peatones es clave para muchas aplicaciones como asistencia al conductor, video vigilancia o multimedia. Los mejores detectores se basan en clasificadores basados en modelos de apariencia entrenados con ejemplos anotados. Sin embargo, el proceso de anotación es una tarea intensiva y subjetiva cuando es llevada a cabo por personas. Por ello, vale la pena minimizar la intervención humana en dicha tarea mediante el uso de herramientas computacionales como los mundos virtuales porque con ellos podemos obtener anotaciones variadas y precisas de forma rápida. Sin embargo, el uso de este tipo de datos genera la siguiente pregunta: *¿Es posible que un modelo de apariencia entrenado en un mundo virtual pueda funcionar de manera satisfactoria en el mundo real?* Para responder esta pregunta, hemos realizado diferentes experimentos que sugieren que los clasificadores entrenados en el mundo virtual pueden ofrecer buenos resultados al aplicarse en ambientes del mundo real. Sin embargo, también se encontró que en algunos casos estos clasificadores se pueden ver afectados por el problema conocido como el cambio en la naturaleza de los datos, igual que ocurre con los clasificadores entrenados en el mundo real. En consecuencia, hemos diseñado un sistema de adaptación de dominio, V-AYLA, en el que hemos probado diferentes técnicas para recoger unos pocos ejemplos del mundo real y combinarlos con una gran cantidad de ejemplos del mundo virtual para entrenar un detector de peatones adaptado. V-AYLA ofrece la misma precisión de detección que un detector entrenado con anotaciones manuales y probado con imágenes reales del mismo dominio. Idealmente, nos gustaría que nuestro sistema se adaptase automáticamente sin necesidad de intervención humana. Por ello, a modo de demostración, proponemos utilizar técnicas de adaptación no supervisadas que permitan eliminar completamente la intervención humana del proceso de adaptación. Hasta donde sabemos, este es el primer trabajo que muestra que es posible desarrollar un detector de objetos en el mundo virtual y adaptarlo al mundo real. Finalmente, proponemos una estrategia diferente para evitar el problema del cambio en la naturaleza de los datos que consiste en recoger ejemplos en el mundo real y reentrenar solamente con ellos pero haciéndolo de tal modo que no se tengan que anotar peatones en el mundo real. El resultado de este clasificador es equivalente a otro entrenado con anotaciones obtenidas de forma manual. Los resultados presentados en esta tesis no se limitan a adaptar un detector de peatones virtuales al mundo real, sino que va más allá, mostrando una nueva metodología que permitiría a un sistema adaptarse a cualquier nueva situación y que sienta las bases para la investigación futura en este campo todavía sin explorar.

Resum

La detecció de vianants es clau per moltes aplicacions com els sistemes d'assistència al conductor, la videovigilància o la multimèdia. Els millors detectors es basen en classificadors basats en models d'aparença entrenats amb exemples anotats. No obstant, el procés d'anotació és un procés feixuc i subjectiu quan es realitzat per persones. Per això, val la pena minimitzar la intervenció humana en aquesta tasca mitjançant l'ús d'eines computacionals com són els mons virtuals ja que amb ells podem obtenir anotacions variades i precises de forma ràpida. No obstant, la utilització d'aquestes dades, ens planteja la següent qüestió: Es possible que un model d'aparença entrenat en un món virtual pugui funcionar de manera satisfactòria en el món real? Per respondre aquesta pregunta, hem realitzat diferents experiments que suggereixen que els classificadors entrenats en el món virtual poden oferir bons resultats al aplicarse en ambients del món real. Tot i això, també s'han trobat alguns casos en que els classificadors es poden veure afectats per un problema conegut com el canvi en la naturalesa de les dades, al igual que passa amb els classificadors entrenats en el món real. Com a conseqüència de tot plegat, hem dissenyat un sistema d'adaptació de domini, V-AYLA. En aquest sistema hem provat diferents tècniques per recollir exemples del món real i combinar-los amb una gran quantitat d'exemples del món virtual, per entrenar un detector de vianants adaptat. V-AYLA presenta la mateixa precisió que un detector entrenat amb anotacions manuals i comprovat amb imatges reals del mateix domini. Idealment, ens agradaria que el nostre sistema s'adaptés automàticament sense necessitat d'intervenció humana. Per tot això, com exemple, proposem utilitzar tècniques d'adaptació no supervisada que ens permetin eliminar completament la intervenció humana en el procés d'adaptació. Actualment, aquest representa el primer treball que demostra que es possible desenvolupar un detector d'objectes en el món virtual i adaptar-lo al món real. Finalment, proposem una estratègia diferent per evitar el problema del canvi de naturalesa de dades que consisteix en agafar exemples del món real i re entrenar tan sols amb ells però fent-ho de manera que no calgui anotar els vianants en el món real. El resultat d'aquest classificador es equivalent a un altre entrenat amb milers d'anotacions manuals. Els resultats presentats en aquesta tesis no es limiten a adaptar un detector de vianants virtuals al món real, sinó que va més enllà, mostrant una nova metodologia que permetria a un sistema adaptar-se a qualsevol situació y que assenta les bases per la investigació futura en aquest camp encara sense explorar.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Advanced driver assistance systems (ADAS) aim to improve traffic safety by providing warnings and performing counteractive measures in dangerous situations. Pedestrian protection systems (PPS) are specialized in avoiding vehicle-to-pedestrian collisions. In the PPS, the key component is a forward facing image acquisition and processing system able to detect pedestrians in real-time, as well as fulfilling a demanding tradeoff between misdetections and false alarms. The challenge lies in the fact that pedestrians are very difficult to detect: they are articulated moving objects of different size that wear different clothes; moreover, the PPS image them in a continuum of distances from a vehicle moving in cluttered outdoor scenarios. In short, both pedestrians and background are highly variable. Accordingly, as the comprehensive state-of-the-art reviews found in [27, 30, 39, 45, 52, 112] reveal, research on image-based pedestrian detection for PPS has been a very relevant topic in the Computer Vision community during the last decade.

The task of an image-based pedestrian detector consists in locating the relevant pedestrians that a given image contains (*e.g.* by framing each one with a bounding box). For PPS the most widespread detection framework consists of several stages [45]: (1) a *selection of candidates* (image windows) to be classified as containing a pedestrian or not; (2) the *classification* of such windows; and (3) a *non-maximum suppression* process to remove multiple detections. As we work with videos a (4) *tracking* stage is also used to remove spurious detections, improving the candidate selection process and deriving information like the motion direction of each pedestrian. All these stages are quite relevant and can contribute on their own to achieve a reliable pedestrian detector in terms of processing time and detection performance. However, since the number of candidates (windows) per image runs from thousands to hundred of thousands, the classification stage is specially critical in such pipeline processing. Accordingly, most of the work done on image-based pedestrian detection has been focused on the classification stage, *i.e.* given a candidate window decide if it contains a pedestrian or not.

The most promising pedestrian detectors rely on appearance-based pedestrian

classifiers learnt discriminatively, *i.e.* from annotated windows; where the pedestrians and the background annotated windows act as the *examples* and the *counterexamples*, resp., to feed the chosen learning machine. In this framework, having sufficient variability in the sets of examples and counterexamples is decisive to train classifiers able to generalize properly [19]. Unfortunately, obtaining the desired variability in such sets is not easy for pedestrian detection in the PPS context, since we cannot control the real world while recording video sequences from a car. We can hypothesize that larger training sets are likely to have higher variability, which seems to be confirmed by the fact that classification performance tends to increase with the size of the training sets for some classifiers [3], including pedestrian classifiers [75]. However, while increasing the number of counterexamples is automatic and effective (*e.g.* since manual annotation of images free of examples is cheap, *bootstrapping* or *cascade* methods can be applied to gather hard false positives and retrain), having a large number of examples is expensive in the sense that many video sequences must be recorded on-board and a large amount of manual intervention is required in the annotation process. Moreover, just subjectively adding more examples does not guarantee higher variability, *i.e.* it can happen that the human annotator is just adding pedestrians too similar to the ones previously annotated. In fact, the same can happen for the images from which counterexamples must be collected.

Hence, obtaining good annotated samples (pedestrians and background) is a relevant open issue for learning reliable pedestrian classifiers within the discriminative paradigm. The ideal situation is to be able to engineer the variability of the samples at low human annotation cost. Some attempts in this line can be found in the literature of pedestrian detection. However, either the results were not good [13] or costly manual silhouette delineation was required [32]. In fact, having good annotated examples is an issue for object detection in general, as well as for category recognition, image classification and any other visual task involving discriminative machine learning. Thus, in the last years different web-based tools, as LabelMe [5] or Amazon Mechanical Turk (MTurk) [85], have been proposed for manually collecting annotated information from images. However, web-based annotation has different problems inherent to human workers whose solution still requires more research [104].

In this thesis we propose a new idea for collecting training annotations. We want to explore the synergies between modern Computer Animation and Computer Vision in order to *close the circle*: the Computer Animation community is modelling the real world by building increasingly realistic virtual worlds, thus, *can we now learn our models of interest in such controlable virtual worlds and use them successfully back in real world?*. Given that, as we have mentioned before, pedestrian classifiers heavily rely on visual appearance, in this thesis we start by focusing on a more specific instance of previous question: *Can a pedestrian appearance model learnt with virtual-world data work successfully for pedestrian detection in real-world scenarios?*. In order to address this question in the PPS context, we need to capture images at virtual urban scenarios using a virtual camera installed in a virtual car forward facing the virtual road ahead. In order to acquire such virtual-world images, we use a realistic videogame through which we obtain pixel-wise annotations of the imaged virtual pedestrians. Figure 3.1 illustrates the overall idea.

There are different possibilities to implement the proposed paradigm. We can learn a holistic (full-body) pedestrian classifier using dense descriptors [27, 30, 112], or the pedestrian silhouette [23]. Analogously, we can learn a part-based pedestrian classifier with dense descriptors [34, 74] (here we would not need to search for parts location during training, since we can know such locations thanks to the pixel-wise annotations), or using the pedestrian silhouette instead [68]. In all cases, different learning machines can be tested as well. Thus, given such a large amount of possibilities, in [73] we just followed popular wisdoms that suggests *starting from the beginning*. In particular, using virtual pedestrians and background, we trained a holistic pedestrian classifier based on histograms of oriented gradients (HOG) and linear support vector machines (Lin-SVM) [21, 22]. We tested such classifier in a dataset, made publicly available by Daimler AG [30], for pedestrian detection benchmarking in the PPS context. The obtained results were evaluated in a per-image basis and compared with a pedestrian classifier trained analogously but using real-world images. The comparison revealed that virtual and real-world based training give rise to similar classifiers. In this thesis we present a more in depth analysis than in [73] by introducing new descriptors (LBP [115], ExtHaar [30, 74, 75, 80, 110, 111], EOH [64]), new learning machines (Real-AdaBoost [92]) and a new datasets (Daimler, INRIA, Caltech-Testing [27], ETH-0,1,2 [117], TUD-Brussels [117] and CVC02 [44]); all them used before in the context of pedestrian detection, so that we can better appreciate the results of employing virtual-world samples for training.

The results of our experiments will show that we obtain the same accuracy by training with a real-world based set than by using a virtual-world based one (without doing any special selection of such virtual data, *i.e.* just driving randomly for acquiring virtual-world images). This is a very encouraging result from the viewpoint of object detection in general. Nevertheless, not only good behavior is shared between virtual- and real-world based training, but some undesired effects too. For instance, let us assume that, with the purpose of learning a pedestrian classifier, we annotated thousands of pedestrians in images acquired with a given camera. Using such camera and classifier we solve our application (PPS, video-surveillance, etc.). Later we must use a different camera or we have to apply the classifier in another similar application/context but not equal. This variation can decrease the accuracy of our classifier because the probability distribution of the training data can be now much different than before with respect to the new testing data. This situation is usually referred to as the *dataset shift* problem [88] and it is receiving increasing attention in the field of Machine Learning [7, 8, 72, 86–88] due to its applications in areas as natural language processing, speech processing, and brain-computer interfaces, to mention a few.

Dataset shift has been largely disregarded in Computer Vision, however, recently some authors have started to pay attention to this problem in the context of object recognition [10, 96]. Coming back to the example, in fact, the best we can do is to annotate the images from the new camera and learn a new classifier [7]. However, doing such new annotations is a never ending expensive procedure. In order to reduce to the minimum new annotations, the scientific challenge consists in performing some sort of adaptation between the training and testing/application domains. Virtual-world images, although photo-realistic, come from a different *eye* than those acquired

with a real camera. Thus, we inherit the dataset shift problem. Accordingly, our proposal of using virtual-world images for learning pedestrian classifiers is cast in a *domain adaptation* framework. In [108], we gave a step forward into this direction, however, only using INRIA dataset and HOG descriptors. Moreover, in this thesis we propose more alternatives than in [108] for combining descriptors coming from real- and virtual-world samples. We term our new learning framework as Virtual-AYLA[1], or just V-AYLA, which stands for *virtual*-world *a*nnotations *y*et *l*earning *a*daptively. We will see that V-AYLA will combine our virtual-world based samples with a relatively low number of real-world based ones to reach the desired performance.

Ideally, we would like to deploy our vision system in the scenario where it must operate without human intervention. Then, the system should self-learn how to distinguish the objects of interest. We are interested in exploring the self-training of a pedestrian detector for driver assistance systems. Our first approach to avoid manual labelling consisted in the use of samples coming from realistic computer graphics, so that their labels are automatically available. This would make possible the desired self-training of our pedestrian detector. In order to overcome the dataset shift, we also explore the use of unsupervised domain adaptation techniques that avoid human intervention during the adaptation process. In particular, we explore the use of the transductive SVM (T-SVM) learning algorithm in order to adapt virtual and real worlds for pedestrian detection. We term this system as V-AYLA-U (Fig. 5.1) as it is our unsupervised version of V-AYLA. V-AYLA-U will combine our virtual-world based samples with some real-world based detections to reach the desired performance.

1.1 Objectives

In summary, in this PhD dissertation we progressively address the following questions:

- *Can a pedestrian appearance model learnt with virtual-world data work successfully for pedestrian detection in real-world scenarios?*
- *Can we adapt the models learnt in the virtual scenarios to the particularities of the real ones?*
- *Can the learnt models automatically adapt to changing situations without human intervention?*

[1] *AYLA* name wants to evoke the main character (a Cro-Magnon women) of the popular saga *Earth's Children* by *Jean M. Auel*. *Ayla* is an icon of robustness and adaptability. During her childhood she is educated by Neanderthals (*the clan*), the physical appearance of them corresponds to *normal humans* for her. However, somehow, she recognizes Cro-Magnons as humans too first time she met them during her youth. *Ayla* adapts from Neanderthals custom to Cro-Magnons one, keeping the best of both worlds. She is a real survivor in such demanding primitive Earth conditions. Interestingly, *Ayla* is the Hebrew name for *oak tree*. It turns out also that there is a popular videogame that incorporates *Ayla* as character.

Therefore, bringing light to those questions are the objectives of this PhD. In the long term, our goal is to build a pedestrian detection system learned without human intervention that automatically adapts itself to the environment changes.

1.2 Contributions

Accordingly, in this thesis we present a pedestrian detector for real-world images novel from several points of view:

- We use photo-realistic virtual worlds to learn the appearance model, *i.e.* using automatic annotations combined with different state-of-the-art descriptors and learning machines.
- In order to address dataset shift, we perform domain adaptation using together virtual- and real-world samples during training. These techniques can be applied to object detection in general. In fact, to the best of our knowledge, our preliminar work [108] is the first paper considering the dataset shift problem for developing an object detector.
- We propose an unsupervised domain adaptation method, *i.e.* human intervention is not required.
- We propose an automatic method for annotating pedestrian bounding boxes with weak supervision.

1.3 Outline

The rest of the thesis is organized as follows. In Chapt. 2 we review the literature related to different aspects of our proposals. Chapt. 3 presents and discusses the results obtained by using the different pedestrian detectors we develop following the traditional (passive) learning methodology, *i.e.* assuming that training and testing domains are equal. This demonstrates that a pedestrian appearance model learnt at virtual scenarios can be successfully applied to real images. In Chapt. 4 we present the results obtained using V-AYLA methodology. In Chapt. 5 we present V-AYLA-U and an alternative to the domain adaptation strategy to create new datasets with low annotation effort. Finally, Chapt. 6 draws the main conclusions of the presented work. In addition, we include an Appendix A for quick acces to the notation used in the thesis.

Chapter 2

State of the Art

To the best of our knowledge there is neither previous proposals for pedestrian detection in particular, nor for object detection in general, where annotations coming from a photo-realistic virtual world are used to learn an appearance classifier that must operate in the real-world detection task (except for our preliminary works in [73,108]). Specially if the virtual world is domain adapted to the real one. Thus, in this chapter we review some works that indeed are related to our proposal, thought only in partial aspects: pedestrian detection, collecting annotations, engineering examples, and performing domain adaptation.

2.1 Pedestrian detection

Pedestrian detection produced a vast interest over the last years in the computer vision community. Thus, many techniques, models, features and general architectures have been proposed. Performing an extensive review of the related literature is not the aim of this work because there exist two recent good surveys in the literature [27,30, 45]. On the contrary, we focus on the object classification stage of the architecture proposed in [45] (Fig. 2.1). This module receives a set of Regions Of Interest (ROIs) to be classified as pedestrians or non-pedestrians. Among the different methods proposed in this stage we focus on the appearance based ones, which define a space of image features and then run a learning machine on examples to obtain a classifier.

Several learning machines have been used in the literature, but we can condensate the most important ones into three groups. Neural Networks (NN) [18] is a bio-inspired architecture based on layers of neurons that leads to a non-linear classifier. Support Vector Machines (SVM) [58] is a statistical method that finds a decision boundary by maximizing the margin between the different classes. Ensemble methods [61] builds a strong classifier by a combination of weak classifiers. The SVM [58] and the AdaBoost variants [36] are clearly the most widely used in the state of the art literature but recently the random forest [38, 63, 103] are attracting attention.

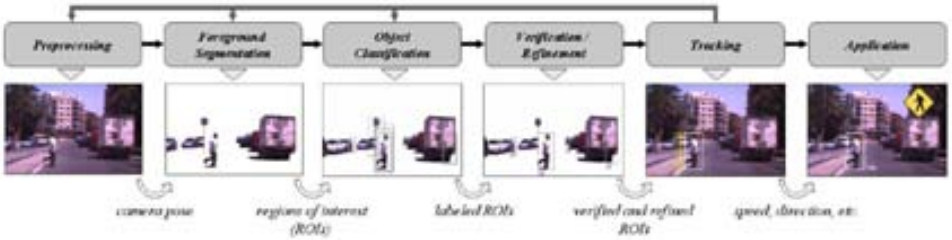

Figure 2.1: The general module-based architecture in [45] covers the structure of most of the systems. It is composed by six modules: preprocessing, foreground segmentation, object classification, verification, tracking and application.

Several feature spaces or descriptors have been proposed in the literature as well. The simplest features were proposed by Gavrila *et al.* [41] which used grey scale image pixels with a NN-LRF as a learning machine, then Zao et al. [120] used image gradient magnitudes combined with a NN. Papageorgiou *et al.* [77] introduced the Haar wavelets features that compute the pixel difference between two rectangular areas in different configurations and can be seen as large scale derivatives; they used a SVM for the classification. Viola and Jones [111] proposed an extended set of Haar wavelets features combined with an AdaBoost cascade. Gerónimo *et al.* [42] combined the Edge Orientation Histograms (EOH) with Haar wavelets in an Real-AdaBoost, resulting a robust and fast pedestrian detector.

Dalal *et al.* [22] presented the Histogram of Oriented Gradients (HOG), a SIFT [37] inspired feature that combined with a SVM is the reference feature on the state-of-the-art of pedestrian detection and have been extended by several authors. For instance, Zhu *et al.* [121] proposed to speed up the computation by changing the SVM by an AdaBoost cascade as classifier and feature selector. Wang *et al.* [115] added to [22] a texture descriptor, the local binary pattern (LBP) and an occlusion handling approach. Maji *et al.* [71] proposed a simplified version of HOG features, the multi-level oriented edge energy features, that combined with a SVM with a fast approximation of the histogram intersection kernel (HIK-SVM). Walk *et al.* [112] extended the use of this HIK-SVM with a combination of HOG, color-self similarity histograms (CSS) and histograms of flow (HOF) features. Enzweilert *et al.* [31] present a multilevel Mixture-of-Experts approach to combine information from multiple features (*i.e.* HOG and LBP) and cues (*i.e.* shape, intensity, depth and flow) with MLP and linear-SVM as expert classifiers. Tuzel et al. [106] propose a novel algorithm based on the covariance of several measures as features and LogitBoost and Riemannian manifolds to classify them.

The aforementioned references consider the pedestrians as a whole so they are called *holistic model*. Felzenszwalb et al. [35] present an approach based on Dalal's HOG detector that consists of a representation of the whole pedestrian and several representations of pedestrian parts(*e.g.* arms, torso, head, etc). This kind of models are called *part-based*. The training is done using latent SVM. It is currently one of

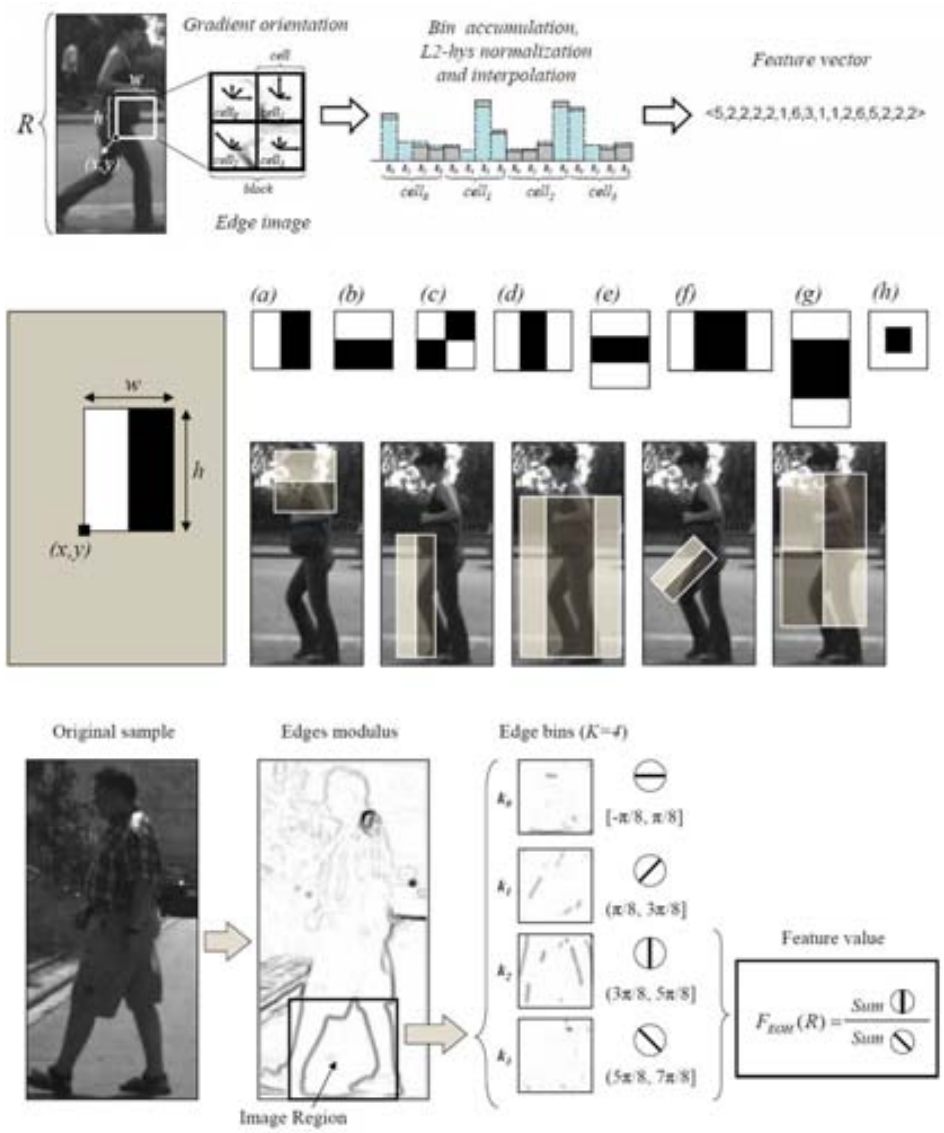

Figure 2.2: *Top: HOG features. Middle: Haar filters.* Example of a filter with parameters (x, y, w, h) with basic forms of the Extended Haar set and examples of filters that give high response in regions containing pedestrians. *Bottom: EOH features.* The feature is defined as the relation between two orientations of a region. In this case, vertical orientations are dominant with respect to the diagonal orientations (k_3), so the feature will have a high value.

the best methods for object detection and has been extended by several authors. For instance, Pedersoli *et al.* [82] proposed a coarse-to-fine approach to accelerate the detector, while Park *et al.* [81] extended it with a multi-scale approach to explicitly model pedestrians seen at different distances to the camera.

An extension of [110], the channel features, done by Dollar *et al.* [26] is recently attracting much attention because it can be fast computed thanks to the integral images and the Haar-like filters. They use cues based on orientation, colour and grayscale. They extended the work by boosting the speed of the previous system [24] by computing the feature responses at a given scale and approximating the feature responses at nearby scales. Then, they introduced the crosstalk cascade [27] where nearby detectors share information to improve the computational efficiency. Finally, Benenson *et al.* [9] proposed a system that works at more than 100 fps based on the ideas of these previous works. In particular, he proposed a combination of depth information coming from stixels [48] with a fast GPU feature computation and a cascade approach. Instead of using a multi-scale image pyramid they use a multi-scale classifier for a fixed number of scales. They follow the procedure of [24] to approximate nearby classifier scales in between the fixed ones.

In conclusion, the most promising pedestrian detectors relay on robust local descriptors (*i.e.* HOG, LBP) in combination with an SVM (*i.e.* linear-SVM, HIK-SVM) or on integral features easy to compute (*i.e.* Haar, EOH, Color) in combination with an ensemble method (*i.e.* Ral-AdaBoost, Random Forest). On top of this, part-based models, multi-resolution models, occlusion handling are build.

2.2 Collecting annotations

Since having good annotated examples is being recognized as a core issue for many Computer Vision tasks requiring learning, in the last years different web-based tools have been proposed for collecting them. A well known example is LabelMe [5] which allows human *volunteers* to localize image objects of a established class category by framing them with polygons (see Fig. 2.3). Nowadays, however, Amazon's Mechanical Turk (MTurk) [85] probably centralizes the most powerful web-based annotation force (see Fig. 2.4). MTurk allows researchers to define *human intelligence tasks* (HITs: *what* and *how*) of different difficulty (*e.g.* visual annotation tasks involve from marking points of interest to drawing polygons) to be taken by human online workers (*turkers*) which are paid for their work. Thousands of annotations have been already collected by using LabelMe and MTurk. Cost free in the former case and at a relative low cost in the latter. Unfortunately, as it is argued in [104], where these web-based tools and others are analyzed, *it is a fallacy to believe that, because good datasets are big, then they are good*.

A key reason behind such fallacy is the human factor involved in the annotation task, which poses difficulties for achieving some desirable properties of training datasets, such as large variety, precision, suitability and representativeness [104]. For instance, in general, humans performing web-based annotation tasks are not vision

Figure 2.3: LabelMe annotation tool.

experts and do not have a scientific motivation. The lack of expertise implies that such human annotators do not know what types of mistakes can be especially problematic for the posterior machine learning process, they do not know what to do in special situations that were not included in the annotation instructions, or they can just misunderstand such instructions [29]. The lack of scientific motivation makes necessary to offer some economic reguard (*e.g.* only for less than the 15% of U.S. and Indian turkers money is irrelevant [95]), but setting the appropriate one becomes a difficult issue [104]: for underpriced work, workers participate for entertainment or curiosity, while for overpriced work we can attract not very skilled workers; in both cases the quality of the annotations can suffer in terms of variety, suitability and representativeness (*e.g.* in order to increase the number of performed HITs per time unit, the turkers could focus on cases easy to annotate, thus, introducing an artificial bias) and precision (*e.g.* to increase HITs/time, the annotations could be less accurate since higher accuracy needs more time). In fact, to detect the presence of low quality annotations due to such reasons or even due to malicious workers, it is necessary to collect multiple annotations from the same image/object and assessing annotation quality from them [104]. Moreover, in the case of MTurk, we have a web-based tool that is not only focused on annotations for the Computer Vision scientific community, but for any type of HIT and *employer*. This introduces a competition among employers regarding the difficulty of the HIT and the cost, for us being a drawback the fact that annotations for Computer Vision tasks many times are quite time consuming. Altogether highlights the question of *how to collect data on the internet* as non-trivial,

Figure 2.4: MTurk annotation plataform. https://www.mturk.com/

opening a new research area [104]; and, since human work is involved in it with the respective economic reguard, *ethical questions* arise too [95].

In fact, since human beings do better a tasks if they are enjoying, some authors have pose the annotation task as a web-based game [69, 70]. However, as argued in [104], designing an entertaining game with the main purpose of collecting good annotations is difficult in itself. Note that what we propose in this thesis is to use photo-realistic videogames that are designed for entertainment, not for doing annotations (though they could be easily adapted for that), *i.e.* we would rely on one of the most powerful industries of the world[1] that will be generating and improving such virtual worlds anyway. It is true that while in [69, 70] annotations are collected directly in real-world images, we propose to use photo-realistic virtual worlds which, a priori, can pose more challenges. However, in addition to the human factor of any of such web-based tools, learning from annotations on real-world images does not avoid dataset shift.

Note that it is not only that some modern videogames, life simulators and animation films, are gaining photo-realism, but also the whole simulation pyramid for creating an artificial life is being considered: visual appearance (shape and photo-realism), kinematics, perception, behavior and cognition. For instance, see [100] for the case of animating autonomous pedestrians in crowded scenarios. This means that from such virtual worlds we could collect an enormous amount of automatically annotated information in a totally controlled manner. In other words, we could obtain training sets of large scale, variety, precision, suitability and representativeness.

[1]Videogame industry has not noticed current worldwide crisis: its incomes equal those of film and music industries together.

Thinking in the needs of the Computer Vision community, given an image taken in a virtual world form a virtual camera, intuitively it seems that we could ask for different types of automatically generated annotations, for instance:

- Scene visual annotations (*e.g.* for image classification and category recognition): what type of objects are there inside the image, what type of image is it (*e.g.* indoor, countryside, urban).
- Object visual annotations (*e.g.* for object detection and image segmentation): bounding box, silhouette (pixel-level annotation, *i.e.* object segmentation), pose, parts, identity along time of the object (tracks) and its imaged pixels (dense optical flow), even distance from the virtual camera to the object points corresponding to such imaged pixels (dense depth; *e.g.* in [51, 66] 3D points from synthetic data are used to build 3D models of real objects), as well as inter-object visual relationships (*e.g.* occlusions).
- Non-visual annotations (*e.g.* for gaining robustness using non-visual cues and working in overall scene interpretation): sounds, what are we seeing in terms of natural language, what is the functionality of each object, how objects relate each other to fulfill a task, etc.

Note that the possibility of exploring and recording such virtual worlds does not only would make possible to obtain the corresponding annotations, in fact, we could design a priori the *storyboard* to collect specifically desired annotations: viewpoint, illumination, appearance of the objects of interest (pose, texture, color, size), backgrounds, etc. It is true that some types of annotations would be quite precise (*e.g.* bounding boxes and silhouettes), while others may have some error inherent to the generative model behind the virtual world (*e.g.* pixel-wise annotations for dense optical flow). However, such errors could be estimated and taken into account, something difficult to do for human annotations due to subjectivity (in fact, it is not realistic to think in pixel-wise manual annotations for dense optical flow). Moreover, an additional benefit is that, given a virtual image, we could have all such types of annotations simultaneously, something difficult to achieve by relying on human annotations (*e.g.* the HITs in MTurk are far more constrained in terms of required annotations).

Certainly, the considerations exposed so far point out web-based annotation tools as an exciting new field. However, exploring the synergies between modern Computer Animation and Computer Vision can be also another one. This thesis aims to motivate such approach. Indeed, so far we have just foreseen the advantages of using virtual worlds for obtaining good annotations. Obviously, the annotated information is then *acquired* in a virtual world, which poses some doubts about its usefulness to learn models that must operate in the real world. For instance, let us imaging that we are developing a basic tracker. This tracker uses an appearance model for describing the objects to track and a motion model for describing how they can move. A priori, the trajectories of virtual objects projected into a virtual camera, can be expected to be useful for learning the parameters of the motion model. In fact, the equations that govern the motion of the virtual objects and the equations of the motion model to

learn may be the same. However, regarding the appearance model of the objects the situation is not that clear a priori, after all human beings can distinguish between an image acquired in a virtual world from another acquired in the real one, at least in general. On the other hand, most people assisting to movies or playing videogames set in modern animated worlds would say that the scenes and lifelike characters are quite photo-realistic. In fact, is not only a matter of the overall appearance of objects and environment (silhouettes, pose, etc.), but also in terms of low-level features as texture, since involved Computer Graphics algorithms try to approach the power spectrum of real images [84, 93, 94].

As a matter of fact, the use of virtual scenarios and virtual reality for training and evaluating human capabilities is common nowadays (pilots, surgeons, etc.). Nevertheless, beyond human observers, the question is if to the *eyes* of the appearance-based descriptors that Computer Vision algorithms use for their tasks, virtual-world appearance is sufficiently close to real-world one. Thus, since visual appearance is the primary source of information for Computer Vision algorithms, a central question is: *can we learn appearance models based on realistic virtual-world appearance and use them successfully in real-world images?*. As first proof of concept, in this thesis we address the yet more specific instance of such question: *can a pedestrian appearance model learnt in realistic virtual scenarios work successfully for pedestrian detection in real images?* (Fig. 3.1).

2.3 Engineering Examples

In the web-based annotation approach humans annotate the content of the images, but can not manipulate such context. If the images have not sufficient variability, the annotations can not add it. Thus, there have been proposed different techniques which are related to ours one in the sense of using an automatic procedure to generate the desired pedestrian examples.

In [13] synthesized examples for pedestrian detection in far infrared images (*i.e.* images capturing relative temperature) are used. In particular, a rough 3D pedestrian model encoding the morphology of a person is captured from different poses and viewpoints. The background is just roughly modelled since it is mainly dark in the used images. Each combination of pose and viewpoint constitutes a kind of grayscale template of *human relative temperature*. Then, instead of following a learning-by-examples approach to obtain a single model (classifier), a set of templates is used by a posterior pedestrian detection process based on template matching. However, the authors admit poor results, since it is difficult to handle variability due to different clothes, person size, more complex background and, in addition, computational time increases with the number of templates to be considered.

In [32] the set of examples is enlarged by transforming the shape of pedestrians (annotated in real images) as well as the texture of pedestrians and background (see Fig. 2.5). The pedestrian classifier is learnt by using a discriminative approach (NNs with LRFs, and SVM with Haar). Since these transformations encode a generative

Figure 2.5: Example of virtual pedestrian synthesis from [32] original pedestrian examples, b) shape variation, c) foreground texture variation, d) - e) joint variation of shape, foreground and back-ground texture.

model, the overall approach is seen as a generative-discriminative learning paradigm. The generative-discriminative cycle is iterated several times in a way that new synthesized examples are added in each iteration by following a *probabilistic selective sampling* to avoid redundancy in the training set. The reported results show that this procedure provides classifiers of the same performance than when increasing the number of training examples with new manually annotated ones. However, much of the improvement comes from enlarging the training set by applying jittering to the pedestrian examples as well as by introducing more counterexamples. Notice that jittering does not involve synthesizing pedestrians since it only requires shifting them inside their framing window, *i.e.* it is introduced to gain certain degree of shift invariance in the learnt classifiers. Besides, for applying the different proposed transformations the overall pedestrian silhouette must be traced, which requires a manual annotation much more labor intensive than standard bounding box framing of pedestrians. In fact, obtaining manipulated good-looking images of people by performing holistic human body transformations is in itself an area of research, specially when video is involved and thus temporal coherence is required [53].

In [2, 99] it is used a human renderer software called POSER, from Curious Labs, to randomly generate synthetic human poses for training an appearance-based human pose recovery system. In this case, these are close human views, usually from the knees up, and it must be assumed either that human detection has been performed before pose recovery, or that the camera is framing a human.

In [47] a pedestrian tracker for video surveillance (static camera) are designed and validated in controlled conditions set in realistic virtual scenarios. In [90] it is also

present a virtual simulation environment based on the Pennsylvania Station in New York for testing a distributed video surveillance camera network. However, the photo-realism is quite primitive compared to the videogame used in [47]. In these works the focus is on demonstrating the usefulness of such virtual scenarios for performance evaluation and engineering situations of interest. For instance, in [47] the tracking relies on a standard color-based mean-shift operating in the virtual scenarios during testing. In this thesis we want also to encourage the use of such virtual scenarios. In fact, as we will see, we use the same videogame than [47] to collect our virtual data. However, we go a step beyond because we want that our appearance-based detectors, learnt in virtual scenarios, can operate in real-world images.

2.4 Domain Adaptation

We will formally define *domain adaptation* in Chapt. 4, but by now let us work with the intuition already introduced in Chapt. 1: if we learn an appearance model using images coming from a *training* camera and typical environment, but after we change to another *testing* camera or environment, then we can suffer the dataset shift problem, *i.e.* the typical object poses can change, and the typical backgrounds, as well as the behavior of the descriptors in which the appearance model relies. In short, the probability distribution of the training domain is different from the one of the testing domain regarding the descriptor space. Thus, some sort of domain adaptation is required.

Domain adaptation is a fundamental problem in machine learning but it only started receiving attention recently [8, 11, 12, 16, 57, 86, 87, 97], specially in computer vision applications [10, 28, 46, 55, 60, 96, 105]. Indeed, pedestrian detection is a field where adaptation methods have been already proposed [79,113,114]. However, related fields, such as class imbalance [54], covariate shift [101], and sample selection bias [50, 118] has a longer history. There are also some related problems as multi-task learning [14], active learning [1, 102] and semi-supervised learning [15, 89] that have been studied more extensively. But, performing an extensive review of the related literature is not the aim of this work because there exist several recent good surveys in the literature: one more general [56], other focused on computer vision applications [6] and other on transfer learning [78].

As first proof of concept about how to perform domain adaptation for object detectors, in this thesis (Chapt. 4) we will use so-called *active learning* [20], *augmented descriptor space* [86], a combination of both, and a first tempt of unsupervised domain adaptation based on transductive-SVM [59]. Thus, we briefly summarize some works related to active learning and the use of the augmented descriptor space.

Aiming to minimize human annotation burden, in [1] an active learning system called SEVILLE (SEmi-atomatic VIsual LEarning) is used for developing a pedestrian detector. Starting by 215 randomly human-annotated pedestrians and sufficient background samples, it is constructed a pedestrian classifier using an AdaBoost cascade, where the weak rules are decision stumps based on one-dimensional descriptors

referred as YEF (yet even faster). This classifier is applied to unseen videos and detections are presented to a *human oracle* that must report if they correspond to actual pedestrians or to background (false positives). In fact, not all detections are presented to the oracle. First, there are examined only image windows that intersect a predefined horizon line. This reduces the application of the current classifier to around 170,000 windows. Then, from these windows, just those classified with a score falling into the *ambiguity region* of the current classifier are passed to the oracle. Once a full video is processed, the new annotated samples together with the previous ones are used to retrain a new classifier, *i.e.* the active learning follows a *batch* scheme. The process is iterated with new videos until a desired performance is achieved.

In [102] it is also used a similar active learning system, called ALVeRT (Active-Learning-based VEhicle Recognition and Tracking), to develop a vehicle detector based on Haar descriptors and a cascade-based AdaBoost learning machine. In this case, 7,500 examples are randomly human-annotated to obtain a first version of the vehicle detector (passive phase). Then, 10,000 more annotations are collected by the human oracle during the active phase.

In fact, training and testing images are coming from the same camera both in [1] and [102]. Thus, active learning is not used as a domain adaptation solution but as a method for collecting most meaningful samples (*selective sampling*) while reducing the annotation burden of pedestrians and vehicles, respectively. Note that from this point of view, active learning is a kind of bootstrapping counterpart since only difficult examples (*i.e.* those in the ambiguity region of the classifiers) must be annotated by the human oracle. Following the same point of view, training and testing a pedestrian classifier in virtual worlds, we could automatically annotate difficult virtual pedestrians analogously to a bootstrapping process. In fact, such an approach is used in [32] for increasingly generating new synthetic samples, though, as we commented before, it turned out not to be too effective. However, in this thesis, we are interested in assessing active learning as a method for domain adaptation. Thus, we will ask a human oracle to annotate some difficult pedestrians on real images, though the initial classifier is based in a large number of virtual samples. After, retraining is based on a descriptor space where some samples come from virtual world (most of them) and others from real world (some of them). We term as *cool world*[2] such joint descriptor space.

Finally, just to mention that we got the idea of using the augmented descriptor space technique from [86], where it is applied to many different problems though no one of the Computer Vision field. However, recently this idea has also been applied to object recognition as in [10] where web-annotated images are used for training and Calthech256 dataset for testing; as well as in [96] to account for different lighting and resolution conditions between real cameras. In this thesis, we challenge augmented descriptor space for a detection task. Moreover, in [10,96] the domains to be adapted are both based on real-world images, while here one of the domains correspond to virtual-world images (the adaptation being performed in the cool world). Addition-

[2] *Cool world* term is a tribute to the movie with that title. In it, there is a real world and a cool world, in the latter, real humans and cartoons live together.

ally, the descriptors involved in our detection task are different than in [10, 96].

Chapter 3

Learning Appearance in Virtual Scenarios

Detecting pedestrians in images is a key functionality to avoid vehicle-to-pedestrian collisions. The most promising detectors rely on appearance-based pedestrian classifiers trained with labelled samples. This chapter addresses the following question: *Can a pedestrian appearance model learnt with virtual-world data work successfully for pedestrian detection in real-world scenarios?* (Fig. 3.1). Our experiments suggest a positive answer, which is a new and relevant conclusion for research in pedestrian detection. More specifically, we record training sequences in virtual scenarios and then appearance-based pedestrian classifiers are learnt using HOG and LBP with linear SVM, as well as Haar and EOH with Real AdaBoost. We test such classifiers in several publicly available datasets: INRIA, Daimler, Caltech, CVC02, TUD and ETH-0,1,2. These datasets contain real-world images acquired at different conditions like personal photo albums or a moving vehicle. The obtained results are compared with the ones given by the counterpart classifiers learnt using samples coming from real images. The comparison reveals that, although virtual samples were not specially selected, both virtual and real world based training give rise to classifiers of similar accuracy in some cases while there is a gap for others.

3.1 Introduction

The most promising pedestrian detection methods rely on appearance-based pedestrian classifiers learnt from labelled samples, *i.e. examples* (pedestrians) and *counterexamples* (background). Having sufficient variability in the sets of examples and counterexamples is decisive to train classifiers able to generalize properly [19]. Unfortunately, obtaining the desired variability in such sets is not easy for pedestrian detection since we cannot control the real world while recording video sequences. We can hypothesize that larger training sets are likely to have higher variability, which

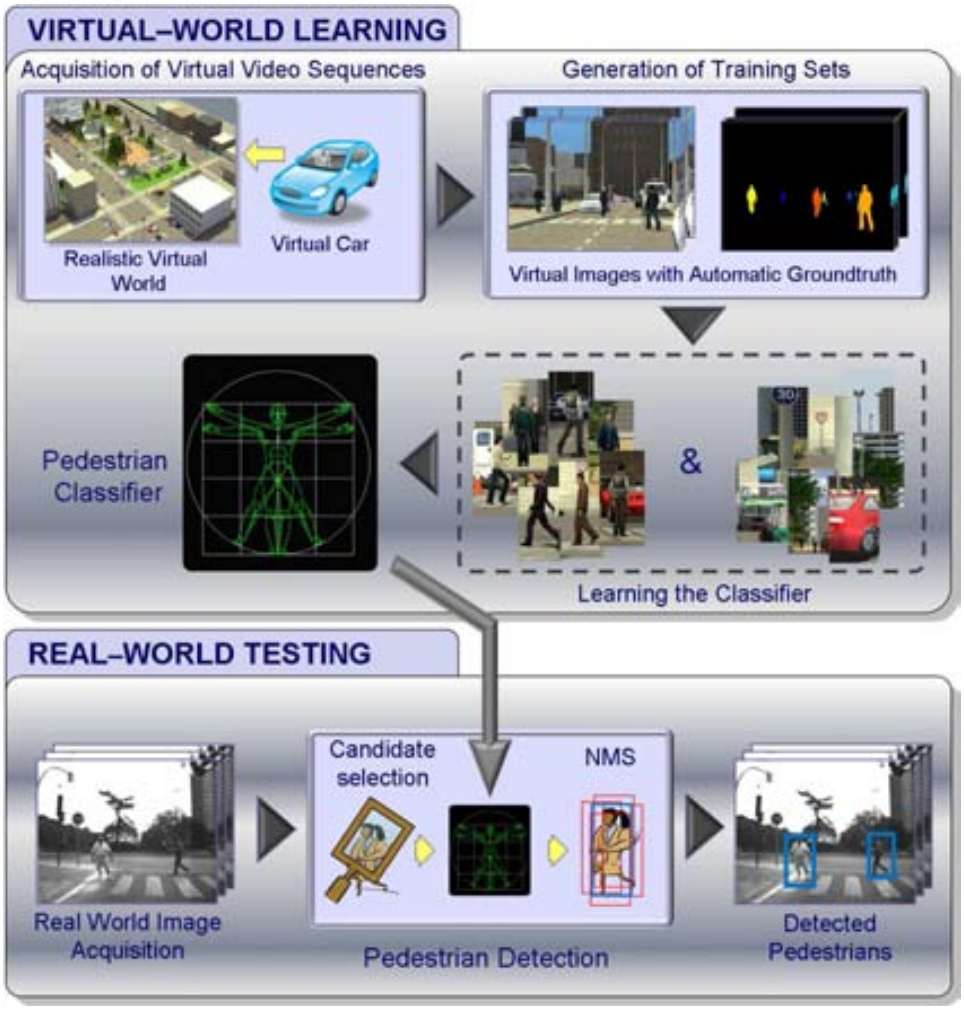

Figure 3.1: Can a pedestrian appearance model learnt at virtual scenarios be successfully applied to real images?

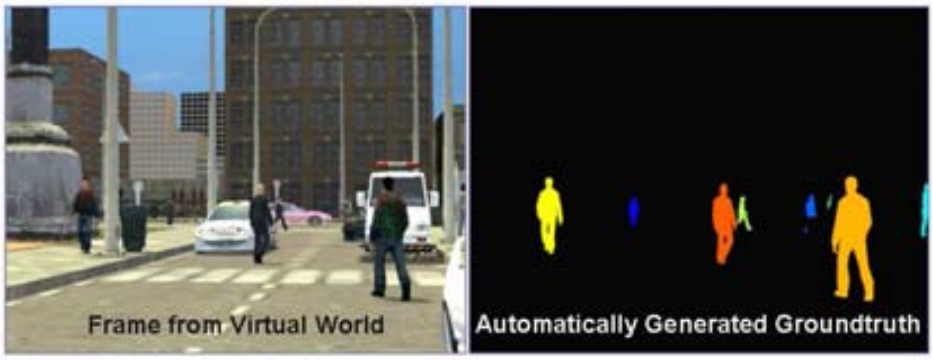

Figure 3.2: Virtual image with corresponding automatically generated pixel-wise groundtruth for pedestrians.

seems to be confirmed by the fact that classification performance tends to increase with the size of the training sets in general [3] and for some pedestrian classifiers [75] in particular. However, while increasing the number of counterexamples is automatic and effective (*e.g.* *bootstrapping* or *cascade* methods can be applied to gather false positives and retrain), having a large number of labelled examples is expensive in the sense that many video sequences must be recorded on-board and a large amount of manual intervention is required. Moreover, just subjectively adding more examples does not guarantee higher variability, *i.e.* it can happen that we are just adding pedestrians too similar to the ones we already had.

The reviewed proposals in Chapt. 2 are appealing in the sense that if we are able to use a set of automatically generated examples for learning, then we will have an easier control of its variability and cardinality and avoid human labelling for the learning phase. The central idea we propose in this chapter is the following. Rather than using rough morphological models or synthesized real examples, we propose to explore the synergies between modern Computer Animation and Computer Vision in order to *close the circle*: the Computer Animation community has modelled real world by building increasingly realistic virtual worlds, especially in the field of video games, thus, can we now learn our models of interest in such virtual worlds and use them successfully back in real world? In this chapter we focus the challenge in the *appearance of pedestrians* captured by a camera working at the visible spectrum (Fig. 3.1).

In this chapter we present an in depth analysis by testing several descriptors, learning machines and datasets used in the context of pedestrian detection, so that we can better appreciate the effect of employing virtual worlds for learning. Training with Lin-SVM we assess the behaviour of HOG, and of cell-structured local binary patterns (LBP) [115]. Since HOG is more related to overall shape and LBP to texture, following [115] we combine HOG and LBP too. We evaluate HOG and LBP separately instead of only considering the combination HOG+LBP, because we aim

to assess the behaviour of such single descriptors when transferred from virtual-world images to real-world ones; moreover they are used separately as *experts* by some mixture-of-experts pedestrian classifiers [31]. In fact, HOG alone is the key descriptor of state-of-the-art part-based object detection [34], as well as in combination with LBP [119]. Additionally, we will study a different set of descriptors that are also popular for pedestrian detection and remain competitive, they are the (extended) Haar wavelets (ExtHaar) [30, 67, 75, 111] and the edge orientation histograms (EOH) introduced in [64], as well as the combination of both [17, 43, 44]. Due to their high dimensionality some AdaBoost variant is usually employed as learning machine with these descriptors. We choose Real-AdaBoost [92] since it gave us very good results in other object detection tasks of the driver assistance context [4, 83].

The experiments we conduct here suggest a positive answer to the previous question, which we think is a new and relevant result for research in pedestrian detection. The obtained results are evaluated in a per-image basis and compared with the classifier obtained when using real samples for training. The comparison reveals that virtual-based training and real-based one give rise to similar classifiers in some cases while there is a gap for others. Furthermore, given that at this time we do not fine tune virtual training sets, the obtained outcome opens the possibility of a more custom design of these sets to obtain better classifiers, *e.g.* following active learning approaches as proposed in [32, 35] or developing our own virtual world.

The remainder of the chapter is organized as follows. Section 3.2 details the datasets, pedestrian detector stages, and evaluation methodology, that will be consistent for the remaining of the thesis. Section 3.3 details the conducted experiments while Sect. 3.4 presents the results and corresponding analysis. Moreover, Sect. 3.5 present additional experiments in terms of extra datasets and features. Finally, Sect. 3.6 summarizes the conclusions and future work.

3.2 Experimental settings

3.2.1 Pedestrian Datasets

The lack of publicly available large datasets for pedestrian detection in the ADAS context has been a recurrent problem for years [26, 30, 45]. For instance, INRIA dataset [22] has been the most widely used for pedestrian detection, however, it contains photographic pictures in which people is mainly close to the camera and on focus. Moreover, there are backgrounds that do not correspond to urban scenarios, which are the most interesting and difficult ones for detecting pedestrians from a vehicle. Fortunately, two more adapted datasets for the ADAS context have been made publicly available recently. One of them is presented by Caltech [26] and the other one by Daimler AG [30]. Also there exist other datasets that are also commonly used in the literature like, ETH-0,1,2 [117], TUD [117] and CVC02 [44].

In order to illustrate our proposal, we focus on two real-world datasets and our virtual-world one. As generic real-world dataset we have chosen the INRIA ($\mathcal{I}$)

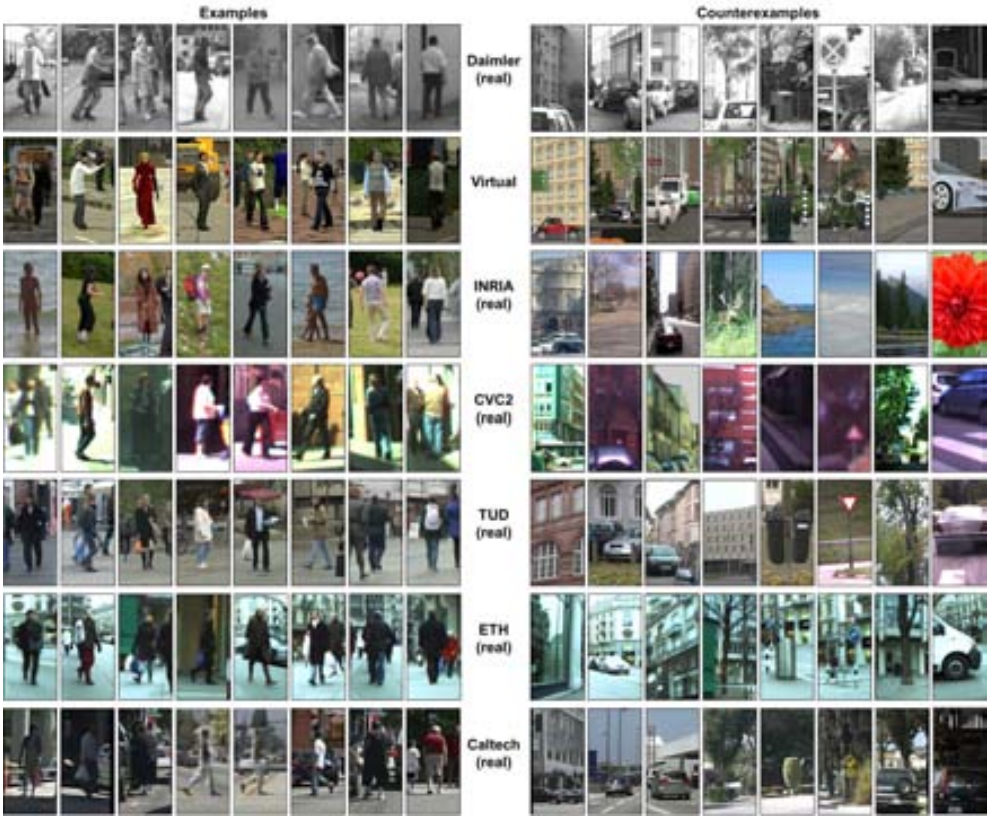

Figure 3.3: Some of the samples used to train/test from real-world images (Daimler, INRIA, CVC02, TUD, ETH and Caltech) and virtual-world ones.

one [21] since it is very well-known and still used as reference [17, 27, 112, 115]. It contains color images of different resolution (320×240 pix, 1280×960 pix, etc.) with persons photographed in different scenarios (urban, nature, indoor). As real-world dataset for driving assistance we use the one of the automotive company Daimler ($\mathcal{D}$) [30], which contains urban scenes imaged by a 640×480 pix monochrome on-board camera at different day times. Both INRIA and Daimler datasets are found divided into training and testing sets. The virtual-world dataset ($\mathcal{V}$) is generated with Half Life 2 videogame by city driving as detailed in [73]. However, for this work we have generated new virtual-world color images containing higher quality textures with anisotropic interpolation, more sequences to extract pedestrians, anti-aliased pedestrian-free images, and much more variability in urban furniture, asphalts, pavement, buildings, trees, pedestrians, etc. Emulating Daimler, virtual-world images are of 640×480 pix resolution. We use this virtual data only for training.

INRIA data includes a set of training images, $\Im_{\mathcal{I}}^{tr+}$, with the BB annotation

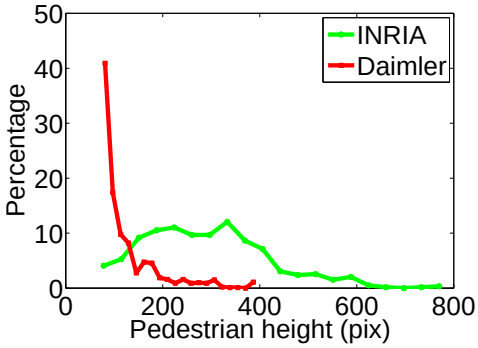

Figure 3.4: Pedestrian height distributions of $\mathcal{T}_{\mathcal{I}}^{tt}$ and $\mathcal{T}_{\mathcal{D}}^{tt}$.

of 1,208 pedestrians. Daimler training set contains 15,660 cropped pedestrians. The images containing them are not available (*i.e.* there is not a $\Im_{\mathcal{D}}^{tr+}$). These pedestrians were generated from 3,915 original annotations by jittering and mirroring. At virtual world we can acquire a set of images, $\Im_{\mathcal{V}}^{tr+}$, of any desired cardinality, with annotated pedestrians.

Training with more pedestrians could lead to better classifiers a priori. For avoiding such a potential effect, the cardinality of the smallest pedestrian training set (*i.e.* the 1,208 of INRIA) is used in our experiments. In the case of Daimler, firstly we grouped jittered and mirrored versions of the same annotation, obtaining 3,915 groups out of the 15,660 provided pedestrians. Secondly, we selected 1,208 cropped pedestrians by randomly taking either zero or one per group. In the case of the virtual-world pedestrians, we selected 1,208 randomly. In all cases, we generate a copy of each pedestrian by vertical mirroring. Thus, the number of available pedestrians for training with each dataset is 2,416. Hereinafter, we term as $\mathcal{T}_{\mathcal{I}}^{tr+}$, $\mathcal{T}_{\mathcal{D}}^{tr+}$ and $\mathcal{T}_{\mathcal{V}}^{tr+}$ these sets (of the same cardinality) from INRIA, Daimler, and virtual world, respectively. Figure 3.3 shows some of such examples.

Additionally, each dataset includes pedestrian-free images ($\Im_{\mathcal{V}}^{tr-}, \Im_{\mathcal{I}}^{tr-}, \Im_{\mathcal{D}}^{tr-}$) from which gathering counterexamples for training. INRIA provides 1,218 of such images and Daimler 6,744. As with pedestrians, we limit the number of pedestrian-free images to 1,218 per dataset. Thus, we use all the INRIA ones, for Daimler we randomly choose 1,218 out of the 6,744 available. For the virtual-world case, we drove through *uninhabited* virtual cities, collecting 6,831 pedestrian-free images, from which we randomly selected 1,218. The final number of used counterexamples from each dataset depends on *bootstrapping* (Sect. 3.2.3). Hereinafter, we note such sets of counterexamples from INRIA, Daimler and virtual world as $\mathcal{T}_{\mathcal{I}}^{tr-}$, $\mathcal{T}_{\mathcal{D}}^{tr-}$ and $\mathcal{T}_{\mathcal{V}}^{tr-}$, resp. Figure 3.3 shows some counterexamples. Accordingly, we define the training settings $\mathcal{T}_{\mathcal{X}}^{tr} = \{\mathcal{T}_{\mathcal{X}}^{tr+}, \mathcal{T}_{\mathcal{X}}^{tr-}\}$, $\mathcal{X} \in \{\mathcal{D}, \mathcal{I}, \mathcal{V}\}$.

We use the complete INRIA testing dataset ($\mathcal{T}_{\mathcal{I}}^{tt}$) consisting of 563 pedestrians in 288 frames and 453 pedestrian-free images. As Daimler testing dataset ($\mathcal{T}_{\mathcal{D}}^{tt}$) we

use 976 *mandatory* frames, *i.e.* frames containing at least one mandatory pedestrian. Daimler defines non-mandatory pedestrians as those either occluded, not upright, or smaller than 72 pix high, the rest are considered mandatory and correspond to pedestrians in the range [1.5m, 25m] away from the vehicle. There are 1,193 mandatory pedestrians in $\mathcal{T}_{\mathcal{D}}^{tt}$. Sets $\mathcal{T}_{\mathcal{I}}^{tt}$ and $\mathcal{T}_{\mathcal{D}}^{tt}$ are complementary in several aspects. $\mathcal{T}_{\mathcal{I}}^{tt}$ images are hand-shotted color photos, while $\mathcal{T}_{\mathcal{D}}^{tt}$ contains on-board monochrome video frames. This turns out in complementary resolutions of the pedestrians to be detected (Fig. 3.4). Moreover, $\mathcal{T}_{\mathcal{D}}^{tt}$ only contains urban scenes, while in $\mathcal{T}_{\mathcal{I}}^{tt}$ we found scenarios like city (916 pedestrians), beach (50), countryside (138), indoor (87) and snow (17).

3.2.2 Pedestrian Detector

In order to detect pedestrians, we *scan* a given image for obtaining windows to be classified as containing a pedestrian (*positives*) or not (*negatives*) by a learnt classifier. Since multiple positives can be due to a single pedestrian, we must *select* the best one, *i.e.* the window *detecting* the pedestrian. Figure 3.1 illustrates the idea for a pedestrian classifier learnt with virtual-world data. We will describe the learning of such classifiers in Sect. 3.2.3. In the following we briefly review the employed scanning and selection procedures.

The scanning approach is based on pyramidal sliding window [21]. It consists in constructing a pyramid of scaled images, for the range of scales in which we want to detect the pedestrians. The bottom of the pyramid (higher resolution) is the original image, while the top is limited by the size of the so-called *canonical window* (CW, Sect. 3.2.3). At the pyramid level $i \in \{0, 1, \ldots\}$, the image size is $\lceil d_x/s_p^i \rceil \times \lceil d_y/s_p^i \rceil$, being $d_x \times d_y$ the dimension of the original image ($i = 0$), and s_p a provided parameter. Opposite to [21], for building levels of lower resolution we perform down-sampling by using standard bilinear interpolation with anti-aliasing, as in [35]. Then, a fixed window of the CW size scans each pyramid level according to strides s_x and s_y, in x and y axes, resp. We experimentally found that $< s_x, s_y, s_p > := < 8, 8, 1.2 >$ is a good tradeoff between final detection performance and processing time. This procedure has some differences regarding usual pedestrian detectors based in the descriptors tested in this chapter (Sect. 3.2.3). We have experimentally seen (Sect. 3.3) that, in general, the pyramid with anti-aliasing boosts the performance of the pedestrian detectors based on LBP and HOG descriptors.

The CW of a classifier trained with $\mathcal{T}_{\mathcal{I}}^{tr}$ is larger than with $\mathcal{T}_{\mathcal{D}}^{tr}$ (Sect. 3.2.3). Then, if we train with $\mathcal{T}_{\mathcal{D}}^{tr}$ and test with $\mathcal{T}_{\mathcal{I}}^{tt}$, we down-scale the testing images using bilinear interpolation with anti-aliasing. If we train with $\mathcal{T}_{\mathcal{I}}^{tr}$ and test with $\mathcal{T}_{\mathcal{D}}^{tt}$, following [117] advice we up-scale the testing images using bilinear interpolation. $\mathcal{T}_{\mathcal{V}}^{tr}$ can be adapted to any CW (Sect. 3.2.3).

As a result of the pyramidal sliding window, several overlapped positives at multiple scales and positions are usually found around the pedestrians. We apply non-maximum-suppression [62] to (ideally) provide one single detection per pedestrian.

Table 3.1: Summary of descriptors parameters.

Descriptor	**Parameters**	**Descr. dimensionality***	
		INRIA ($\mathcal{I}$)	**Daimler** ($\mathcal{D}$)
HOG	Max-gradient in RGB, 8×8 pix/cell, 2×2 cells/block, block overlap 50%, 9 bins 0° to 180°, L2-Hys normalization	3,780	1,980
LBP	Luminance, 16×16 pix/cell, 1×1 cells/block, block overlap 50%, radius=1 pix, uniform patterns [76] with thr=4, L1-sqrt normalization.	6,195	3,245

($\star$) Virtual ($\mathcal{V}$): as $\mathcal{I}$ for $\mathcal{T}_{\mathcal{I}}^{tt}$ testing, and as $\mathcal{D}$ for $\mathcal{T}_{\mathcal{D}}^{tt}$ testing.

3.2.3 Pedestrian classifier training

In this section we focus on one of the most widespread [27,30,45,112] training methodology within the discriminative paradigm. We refer to the situation in which only randomly annotated training examples are used to learn a classifier. Sets $\mathcal{T}_{\mathcal{X}}^{tr}$, $\mathcal{X} \in \{\mathcal{D}, \mathcal{I}, \mathcal{V}\}$, are of such a type. We term this approach as *passive*.

Discriminative learning of a pedestrian classifier requires the computation of descriptors able to distinguish pedestrians from background. As we introduced in Sect. 3.1, HOG, LBP, ExtHaar and EOH are very well suited for this task. However, the current state of the art is mainly based on HOG and LBP, so, we will focus more on these two. For such descriptors being useful, a canonical size of pedestrian windows must be fixed. This CW size, $w \times h$ pix, depends on the dataset and the descriptor.

In the case of INRIA, we have $w \times h \equiv (32+2f) \times (96+2f)$, where f denotes the thickness (pix) of a background frame around the pedestrian (the so-called *context*). Thus, annotated pedestrians are scaled to 32×96 pix. Analogously, for Daimler $w \times h \equiv (24+2f) \times (72+2f)$. For INRIA and HOG/LBP descriptors, $f = 16$ is of common use in the literature, *e.g.* this f gives rise to the traditional INRIA CW of 64×128 pix [22]. For Daimler and HOG/LBP, $f = 12$, therefore, $w \times h \equiv 48 \times 96$ [30]. In practice, INRIA training pedestrians are provided as 64×128 windows [22] and Daimler as 48×96 ones [30].

In the case of the virtual-world pedestrians, we just consider those larger than 32×96 pix. Then, when training classifiers for testing in $\mathcal{T}_{\mathcal{I}}^{tt}$ we use $w \times h \equiv (32+2f) \times (96+2f)$, while for testing in $\mathcal{T}_{\mathcal{D}}^{tt}$ we use $w \times h \equiv (24+2f) \times (72+2f)$. In both cases we use exactly the same pedestrian annotations for training, but in the case of Daimler we down-scale them more than in the case of INRIA. Hence, we actually have different $\mathcal{T}_{\mathcal{V}}^{tr+}$ sets. However, we avoid a more complex notation for making explicit the differences provided that we have clarified the situation. As

it is done during testing (Sect. 3.2.2) down-scaling uses bilinear interpolation with anti-aliasing.

Collecting the counterexamples to form the $\mathcal{T}_{\mathcal{X}}^{tr-}$ sets, $\mathcal{X} \in \{\mathcal{D}, \mathcal{I}, \mathcal{V}\}$, involves two stages. Conceptually, they can be described as follows. In the first stage, for each example in $\mathcal{T}_{\mathcal{X}}^{tr+}$ we gather two counterexamples by randomly sampling the respective pedestrian-free images (Sect. 3.2.1). For doing such a sampling, the pyramid (Sect. 3.2.2) of each pedestrian-free image is generated and then at random levels and positions two CWs are taken. Since the cardinality of $\mathcal{T}_{\mathcal{X}}^{tr+}$ is 2,416 and to form the initial $\mathcal{T}_{\mathcal{X}}^{tr-}$ we have 1,218 pedestrian-free images, we have approximately the same quantity of examples and counterexamples. In the second step, we follow a *bootstrapping* training methodology [22, 30, 112]. This means that with the initial $\mathcal{T}_{\mathcal{X}}^{tr+}$ and $\mathcal{T}_{\mathcal{X}}^{tr-}$ sets we train a classifier using the desired descriptors and learning machine. Then, the corresponding pedestrian detector (Sect. 3.2.2) is applied on the pedestrian-free training images to extract the so-called *hard* counterexamples, *i.e.* false detections. All these new counterexamples are added to $\mathcal{T}_{\mathcal{X}}^{tr-}$ and, together with $\mathcal{T}_{\mathcal{X}}^{tr+}$, the classifier is trained again. We keep this loop until the number of new hard counterexamples is smaller than 1% of the cardinality of current $\mathcal{T}_{\mathcal{X}}^{tr-}$ set. Following such a stopping rule of thumb and initial 1:1 ratio between examples and counterexamples, we found that one bootstrapping step was sufficient in all the experiments. We forced more bootstrappings in different experiments to challenge the stopping criteria, but the results were basically the same because very few new hard counterexamples were collected. In [112] it is also recommended to follow a strategy such that almost all counterexamples are collected by the bootstrapping.

Table 3.1 summarizes the descriptors parameters. HOG is computed using the parameters of the original proposal [22]. In the case of LBP, we introduce three improvements with respect to the approach in [115]. First, we use a threshold in the pixel comparisons, which increases the descriptor tolerance to noise. Second, we do not interpolate the pixels around the compared central one given that it distorts the texture and can impoverish the results. By doing so we could lose scale-invariance, but in our case it does not matter thanks to the image-pyramid. Third, we perform the computation directly in the luminance channel instead of separately computing the histograms in the three color channels, which reduces the computation time while maintaining the performance. As Lin-SVM implementation we use LibLinear [91], setting $C = 0.01$ and $bias = 100$.

3.2.4 Evaluation methodology

In order to evaluate the performance of the pedestrian detectors we reproduce the procedure proposed in [27]. This means that we use performance curves of *miss rate vs false positives per image*. We focus on the range FPPI=10^{-1} to 10^{0} of such curves, where we provide the *average miss rate* (AMR) by averaging its values taken at steps of 0.01. Accordingly, such an AMR is a sort of expected miss rate when having one false positive per five images. This is an interesting assessment point for our application area, *i.e.* driver assistance, since such a FPPI can be highly reduced by

a temporal coherence analysis. Besides, all annotated INRIA testing pedestrians and the mandatory ones of Daimler must be detected (Sect. 3.2.1).

The evaluation procedure described so far is rather standard. However, according to our daily working experience, even using a good *bootstrapping* method [112] (Sect. 3.2.3) the AMR measure can vary from half to even one and a half points, up or down, due to some random choices during the training process. For instance, initial background samples in standard passive training (Sect. 3.3). Of course, such a small variation does not convert a good detector in bad or the opposite. However, sometimes this is the performance difference among ranked detection algorithms [27]. Thus, it seems reasonable to rely on several training executions per experiment. Nevertheless, this may turn out in an enormous number of costly experiments. Therefore, in this chapter we have opted for repeating only the most representative experiments. Those are the based on INRIA and Daimler datasets trained with HOG and LBP features. For each of these experiments we repeat the training-testing run five times, which is a moderate number of repetitions but, as we will see, it is sufficient to run different statistical tests that will validate our hypothesis of interest. Then, rather than presenting the AMR of a single train-test run, we present the average of five runs and the corresponding standard deviation.

3.3 Experimental results

Table 3.2 shows the performance of the experiments HOG, LBP and HOG+LBP over the INRIA and Daimler datasets: 24 detectors we developed following passive training and evaluated according to Sect. 3.2.4. Figures 3.5 and 3.6 offer a visual insight by plotting the performance curves. Let us remind that in our evaluation process (Sect. 3.2.4) each detector is trained and tested five times. This means that, for each detector, Table 3.2 shows the average AMR and its corresponding standard deviation, and Figures 3.5 and 3.6 plot average curves with their corresponding standard deviation intervals. Thus, these experiments involve 120 train-test runs.

3.4 Discussion

Experiments of table 3.2 sustain our claim that the performance of our pedestrian detectors is boosted by adding a pyramid with anti-aliasing and other improvements for the LBP. Our current HOG implementation gives better results for $\{\mathcal{T}_{\mathcal{I}}^{tr}, \mathcal{T}_{\mathcal{I}}^{tt}\}$ and $\{\mathcal{T}_{\mathcal{D}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$ than the original one [21] (used by us in [73,107]) due to the anti-aliasing in down-scaling operations. Also, our settings for LBP give better results for $\{\mathcal{T}_{\mathcal{I}}^{tr}, \mathcal{T}_{\mathcal{I}}^{tt}\}$ and $\{\mathcal{T}_{\mathcal{D}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$ than the proposal in [115], thanks to the anti-aliasing and the pattern discretization threshold. When using HOG+LBP we obtain an improvement of almost 10 points for $\{\mathcal{T}_{\mathcal{I}}^{tr}, \mathcal{T}_{\mathcal{I}}^{tt}\}$ and around 16 for $\{\mathcal{T}_{\mathcal{D}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$, with respect to [115]. Note that the better the performance when training and testing within the real data, the higher the challenge to reach the same performance when training with virtual data and

Table 3.2: Passive learning results for HOG, LBP and HOG+LBP over INRIA and Daimler datasets. AMR is shown in % for FPPI $\in [10^{-1}, 10^{0}]$. Bold values are the best comparing training sets ($\mathcal{T}^{tr}_{\mathcal{X}}, \mathcal{X} \in \{\mathcal{D}, \mathcal{I}, \mathcal{V}\}$) for fixed descriptor, learning machine, and testing set ($\mathcal{T}^{tt}_{\mathcal{R}}, \mathcal{R} \in \{\mathcal{D}, \mathcal{I}\}$).

Passive Learning	**Training** ($\mathcal{T}^{tr}_{\mathcal{X}}$)	**Testing** ($\mathcal{T}^{tt}_{\mathcal{R}}$)	
		INRIA ($\mathcal{I}$)	**Daimler** ($\mathcal{D}$)
HOG	**Daimler** ($\mathcal{D}$)	38.46±0.45	**30.01**±0.51 35.62±0.33*
	INRIA ($\mathcal{I}$)	**21.27**±0.52 27.86±0.60*	41.12±1.01
	Virtual ($\mathcal{V}$)	32.47±0.47	30.64±0.43
LBP	**Daimler** ($\mathcal{D}$)	39.54±0.55	**35.07**±0.29 50.03±0.36°
	INRIA ($\mathcal{I}$)	**18.42**±0.53 34.53±0.82°	35.40±0.70
	Virtual ($\mathcal{V}$)	28.87±0.70	45.21±0.49
HOG + LBP	**Daimler** ($\mathcal{D}$)	32.28±0.47	**22.48**±0.45 38.04±0.46° 28.85±0.52$^{\bullet}$
	INRIA ($\mathcal{I}$)	**14.35**±0.46 23.92±0.81°	26.22±0.85
	Virtual ($\mathcal{V}$)	23.81±0.53	28.27±0.48

($\star$) Dalal *et al.* implementation [21].
($\circ$) Wang *et al.* impl. [115], without occlusion handling.
($\bullet$) Training with the 15,660 pedestrians (Sect. 3.2.1).

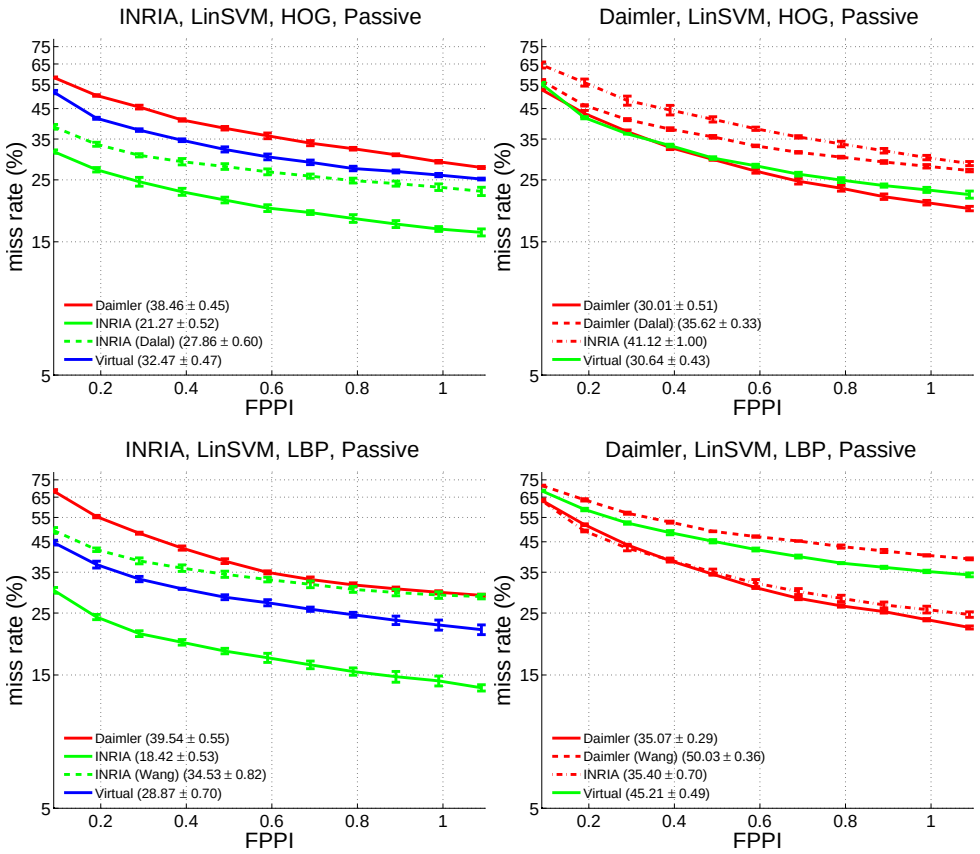

Figure 3.5: Passive learning results for HOG and LBP over INRIA and Daimler datasets.

testing with real data.

We could expect that pedestrian detector trained in a video game could never be applied into the real world but, far from this, it performs fairly good. Table 3.2 reveals that the standard HOG/LinSVM pedestrian detector trained on virtual data and tested on Daimler dataset the performs exactly as its counterpart trained on Daimler training data. However, for INRIA there is a performance gap of 11 points. Even this is a considerable performance drop the virtual detector stills perform fairly well regarding the great challenge of training with virtual data. Nevertheless, using LBP features the performance gap is quite wider 10 points for Daimler and 15 for INRIA. Even that virtual detectors does not perform as well as real world ones they provide promising results.

Table 3.2 shows that using train and test sets of different sources increases the

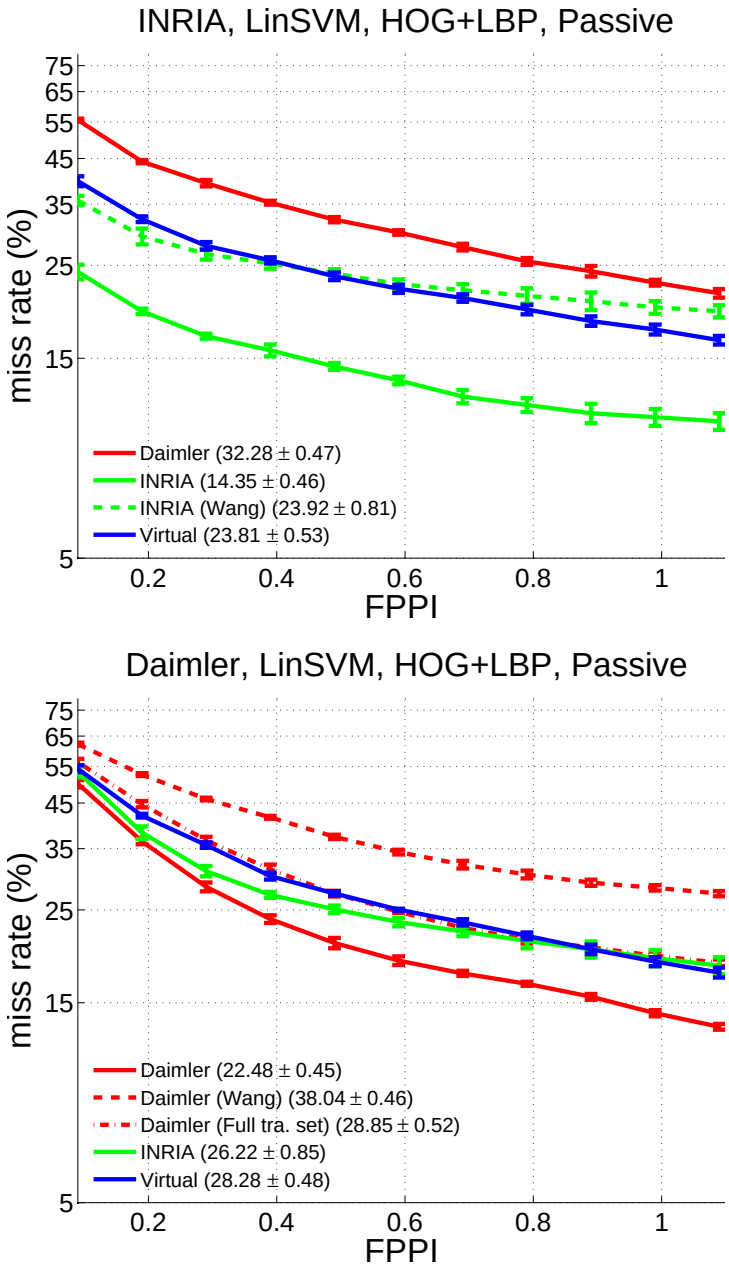

Figure 3.6: Passive learning results for HOG+LBP over INRIA and Daimler datasets.

AMR mean even 15 percentual points depending on the descriptor and dataset. This stands also when training and testing data come from different real world datasets. To asses this claim, we have checked the statistical significance of these results. For each descriptor, we consider all the detectors obtained by the different train-test runs using the two considered real-world training sets. Since we have tested such detectors on both real world test sets, by paring the obtained performances we can apply a *paired Wilcoxon test* [116]. The test reveals that for HOG, LBP and HOG+LBP testing and training with samples/images of the same dataset is better than using different datasets in the 99.9% of the cases (p-value = 0.001, being the null hypothesis that training and testing data are equal). The means of the improvement are 13.62, 10.35 and 10.73 AMR points for HOG, LBP and HOG+LBP, respectively.

We also argue that training with virtual-world data exhibits the same problem, but just as real-world data does. In order to support this claim, we have analysed if detectors trained with $\mathcal{V}$ data behave similarly to detectors trained with real-world data (using $\mathcal{I}$ and $\mathcal{D}$ datasets) when tested on a different real world dataset, again taking into account all performed training-testing runs. In this case, since the compared virtual- and real-world-based detectors use different training data, all feasible pairings between their performances have to be taken into account and an *unpaired Wilcoxon test* (a.k.a *Mann-Whitney U test* [49]) must be applied. This test allows to conclude that when using $\mathcal{I}$ as testing data, detectors trained with $\mathcal{V}$ data provide better results than training with $\mathcal{D}$ data. This is true the 99.6% of the cases (one sided p-value = 0.004). The means of the improvement are 5.94, 10.89 and 8.85 AMR points for HOG, LBP and HOG+LBP respectively. When the testing dataset is $\mathcal{D}$, the analogous analysis reveals that training with $\mathcal{V}$ data is better for HOG than using $\mathcal{I}$ (10.62 points), while for LBP and HOG+LBP training with $\mathcal{I}$ data is better (9.46 and 2.18 points, respectively), with one-sided p-value = 0.004. Therefore, regarding the performance, the virtual-world data is comparable to a real-world one.

Usually there are several (possibly simultaneous) reasons giving this fact. For instance, with HOG, $\{\mathcal{T}_{\mathcal{V}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$ setting offers similar results to $\{\mathcal{T}_{\mathcal{D}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$ one, while $\{\mathcal{T}_{\mathcal{V}}^{tr}, \mathcal{T}_{\mathcal{I}}^{tt}\}$ results are much more distant from $\{\mathcal{T}_{\mathcal{I}}^{tr}, \mathcal{T}_{\mathcal{I}}^{tt}\}$, probably because our virtual-world data comes from urban scenes as Daimler data, but INRIA incorporates other scenarios. For LBP, however, $\{\mathcal{T}_{\mathcal{V}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$ results are much worse than $\{\mathcal{T}_{\mathcal{D}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$ ones. In fact, $\{\mathcal{T}_{\mathcal{V}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$ performance based on HOG is approximately 15 points better than the LBP one, while HOG and LBP show a difference of around 5 points for $\{\mathcal{T}_{\mathcal{D}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$. Thus, the textures of the virtual-world somehow differ more from Daimler images than the shape of the pedestrians. The best performance corresponds to combining HOG and LBP. In this case, for instance, $\{\mathcal{T}_{\mathcal{V}}^{tr}, \mathcal{T}_{\mathcal{I}}^{tt}\}$ setting is around 10 points worse than using $\{\mathcal{T}_{\mathcal{I}}^{tr}, \mathcal{T}_{\mathcal{I}}^{tt}\}$ one. This can be due to the fact that typical background and pose of virtual-world pedestrians do not include all INRIA cases (*e.g.* out-of-city pictures). The result for HOG+LBP and $\{\mathcal{T}_{\mathcal{V}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$ is approximately 2 points[1] worse than for $\{\mathcal{T}_{\mathcal{D}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$, which could come from the pedestrians clothes (texture/LBP) rather than from pedestrian poses (shape/HOG).

[1]The best $\{\mathcal{T}_{\mathcal{V}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$ for HOG+LBP does not correspond to the full set of Daimler pedestrian examples (15,660) but to our selected set (2,416 ones), *i.e.* examples without jittering. Since jittering neither add shape variability nor appearance provided that the

3.5 Additional experiments

For complementing performance assessment, we extend our experiments in two ways: by adding more datasets and by adding more descriptors. We will follow the experimental settings explained in Sect. 3.2. However, as we have seen that the corresponding standard deviations seems to be stable and less than ± 2 to alleviate the number of experiments we do not repeat each experiment.

3.5.1 More datasets

To validate our results we will perform the experiments on the remaining datasets from Fig. 3.3: Caltech, ETH-0,1,2, TUD and CVC02. Next, we point the main characteristics of these datasets.

Caltech [27] is a popular pedestrian dataset. It contains color images of 640×480 pix resolution acquired from a vehicle driven through different urban scenarios at different day times. For training, in [27] it is used INRIA training data but we also use Caltech training data. In particular, from Caltech training videos we selected all the non-occluded pedestrians taller than 72 pix but avoiding the inclusion of the same pedestrian many times. This procedure outputs 790 pedestrians, thus, we have 1580 examples after mirroring. Moreover, to keep the same ratio between positive and negative training data than in previous experiments, we randomly choose 605 pedestrian-free Caltech training frames. We set the CW following INRIA settings, and we use image up-scaling during testing for detecting *reasonable* pedestrians (*i.e.* most representative ones [27]) taller than 50 pix but not reaching the 96 pix of the INRIA CW setting.

The ETH dataset [117] was recorded at a resolution of 640×480 pix resolution, using a stereo pair mounted on a children stroller. For our particular experiments, only the left images of each image-sequence are used. The ETH dataset contains three sub-sequences, representing three different scenarios, which are denoted as ETH0, ETH1 and ETH2. ETH1 corresponds to the 999 frame "BAHNHOF" sequence, ETH1 are the 451 frame "JELMOLI" sequence and ETH1 the 354 frame "SUNNY DAY" sequence. We use the updated annotations of [117] as used in Caltech evaluation framework [27] but we restrict to those pedestrians taller than 72 pix height. Since, these sequences are only for testing as they do not provide training data, we use the INRIA one as in [27].

The TUD-Brussels Pedestrian dataset [117] acquired from a driving car. It contains motion pairs recorded in busy pedestrian zones from a hand-held camera at a resolution of 720×576 pixels. As in [27] we only use the first frame of each motion pair. The training data consists of 1092 images with 1776 annotated pedestrians and

size of the cells we use for HOG and LBP is bigger than the degree of jittering, it may be just introducing some overfitting. Note that in [30] pedestrian classifiers training is based on *local receptive fields* and *neural networks* (LRF/NN), thus, shift invariance must be explicitly introduced (*e.g.* using jittering).

Table 3.3: Passive learning results for HOG, LBP and HOG+LBP over extra datasets: TUD, ETH0-1-2, CVC02 and Caltech (Reasonable).INRIA and Daimler. AMR is shown in % for FPPI $\in [10^{-1}, 10^{0}]$. Bold values are the best for a testing set.

Feature	**Testing** ($\mathcal{T}^{tt}_{\mathcal{R}}$)	**Training** ($\mathcal{T}^{tr}_{\mathcal{X}}$)		Δ
		Real ($\mathcal{R}$)	**Virtual** ($\mathcal{V}$)	
HOG	**ETH0**	**58.74***	65.38	-6.64*
	ETH1	65.51*	69.30	-3.79*
	ETH2	**54.63***	58.97	-4.34*
	TUD	68.84*; **63.34**$^{\circ}$	67.98	-0.86*; -4.64°
	CVC02	45.76*; 32.30$^{\bullet}$	43.85	-1.91*; -11.55$^{\bullet}$
	Caltech	47.88*; 67.53$^{\natural}$	47.81	-0.07*; 19.72$^{\natural}$
LBP	**ETH0**	66.31*	63.75	2.56*
	ETH1	64.01*	63.56	0.45*
	ETH2	73.62*	73.01	0.61*
	TUD	66.24*; 85.66°	79.13	-12.89*; 6.53°
	CVC02	50.66*; 33.93$^{\bullet}$	74.04	-23.38*; 40.11$^{\bullet}$
	Caltech	46.04*; 57.37$^{\natural}$	56.75	-10.71*; 0.62$^{\natural}$
HOG + LBP	**ETH0**	62.70*	60.78	1.92*
	ETH1	**62.88***	60.41	2.47*
	ETH2	56.29*	55.03	1.26*
	TUD	**61.29***; 80.51°	68.77	-7.48*; 11.74°
	CVC02	**41.74***; **24.59**$^{\bullet}$	49.62	-7.88*; -25.03$^{\bullet}$
	Caltech	**42.18***; **53.02**$^{\natural}$	46.94	-4.76*; 6.08$^{\natural}$

($\star$) INRIA training set. ($\circ$) TUD training set.
($\bullet$) CVC02 training set. ($\natural$) Caltech training set.

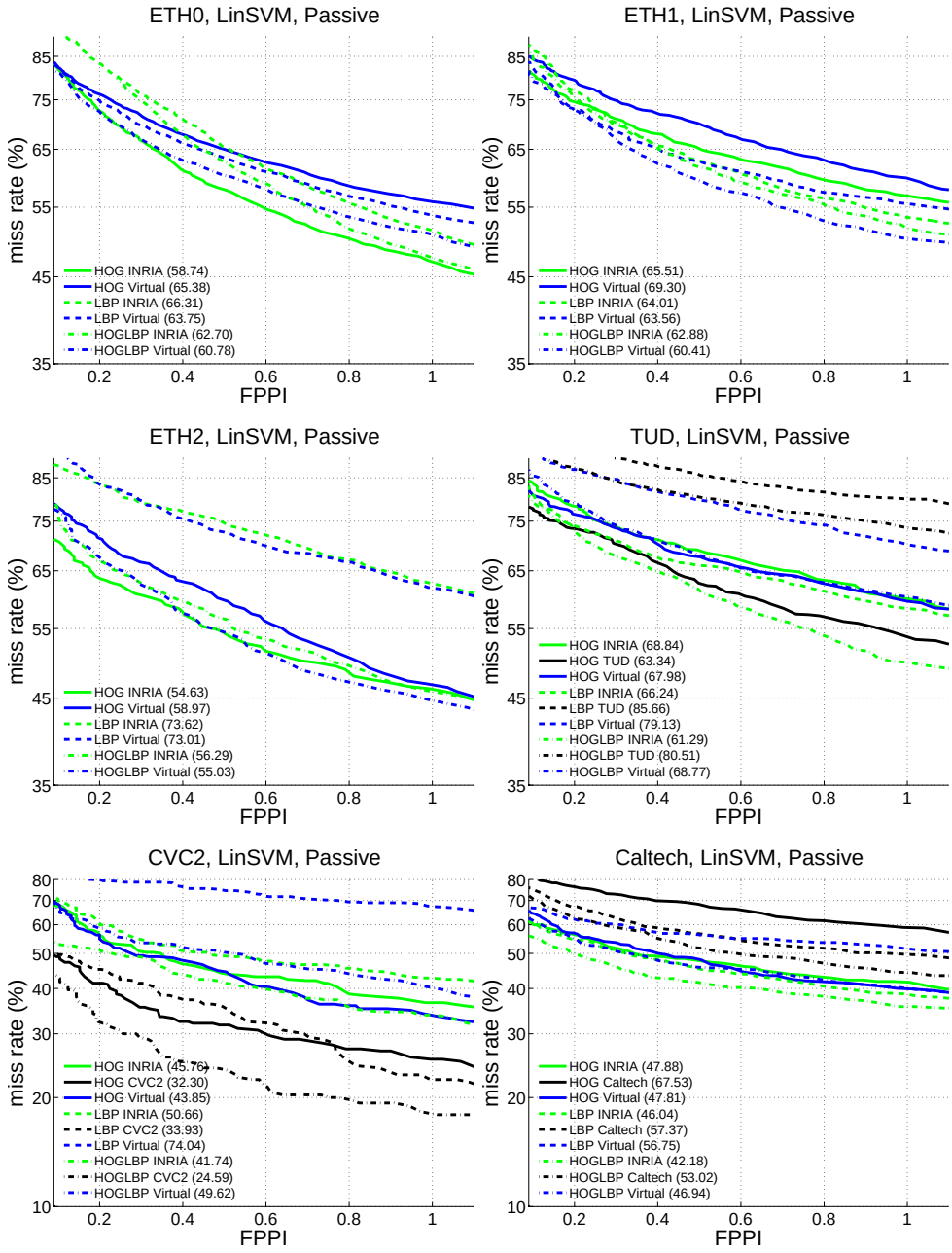

Figure 3.7: Passive learning results over extra datasets: HOG, LBP and HOG+LBP over TUD, ETH-0,1,2, CVC02 and Caltech.

192 negative frames. The test set is recorded with a different camera setting and contains 508 images at a resolution of 640×480 pixels with 1326 annotated pedestrians. The dataset is challenging due to the fact that pedestrians appear from multiple viewpoints, at very small scales, many are partially occluded and the fact of the differences in the acquisition settings from training and testing. As before we restrict our evaluation to those pedestrians taller than 72 pix.

CVC02 [45] it is one of our own datasets. It is recorded using a camera based on a CCD color sensor of of 640×480 pix resolution, with a lens of 6 mm of focal. The camera is installed in the windshield of a car, forward facing the road. From this dataset we use the *classification* sequence. The training data is formed by 1016 cropped pedestrians and 154 negatives frames. Additionally, it has 101 positive frames with 581 annotated pedestrians. From the testing set we use 250 positive frames with 1140 annotations and 150 negative frames where 290 pedestrians taller than 72 pixels.

Table 3.3 shows the performance of HOG, LBP and HOG+LBP for TUD, ETH-0,1,2, Caltech and CVC02: 45 detectors that we developed following the same training and evaluation procedure. Figure 3.7 plots the performance curves. For each detector, Table 3.3 shows its AMR and Fig. 3.7 plots its accuracy curve. For all these experiments we use INRIA training set as done in [27]. Additionally, for the ones that has its own training data, *i.e.* TUD, CVC02 and Caltech, we also train our models using this data. Thus, these experiments involve 45 train-test runs.

From Table 3.3 we can conclude that for ETH-0,1,2 and TUD datsets the LBP feature does not perform well. Thus, when combining LBP with HOG the results d not improve. This fact is mentioned by the datasets authors [27]. As shown in our experiments from Table 3.2 the LBP descriptor is very sensitive to changes in the training/testing data. As for ETH-0,1,2 the training data is INRIA and for TUD the training data comes from a different camera setting than the testing data this could cause the problems for the LBP. Besides, for CVC02 and Caltech where the training/testing data comes from the same setting the LBP performs better. This, the HOG+LBP configuration performs the best. From hereinafter we will restrict our ETH-0,1,2 and TUD experiments to HOG features and the CVC02 and Caltech to HOG+LBP. Classifiers trained and tested on the same dataset gives better results than the ones trained on INRIA. This reinforces our intuition that training with datasets of a different nature than the testing ones can produce a performance drop. In Caltech case, either training with INRIA or virtual-world data performs better than using the Caltech case training data. This may suggest that such data lacks variability. However, this is not important for the purpose of this thesis since we can assume that the baseline performance is the one based on INRIA training data as is usually done [27].

3.5.2 More features

In order to illustrate better the behaviour of training with virtual data versus real one we introduce some extra state-of-the-art descriptors: Haar, EOH and Haar+EOH and a new learning machine: Real-AdaBoost. We follow the experimental settings

explained in Sect. 3.2 but introducing some changes in the pedestrian detector.

In particular, for ExtHaar and EOH, we experimentally found that reducing the margin of the canonical window to $f = 4$ (almost no context) provides better detection performance. Thus, we discard a part of the background frame for training with ExtHaar/EOH. Note that, technically, the set $\mathcal{T}_{\mathcal{I}}^{tr+}$ is different for HOG/LBP than for ExtHaar/EOH. However, since the framed pedestrians are the same and only the f changes, we do not introduce a more complex notation for making explicit such a difference. Analogously for $\mathcal{T}_{\mathcal{D}}^{tr+}$.

Additionally, we use our Real-AdaBoost implementation. For this study we favor performance instead of processing time, thus, we build a single cascade. Weak classifiers are decision stumps. Following the maximum of random variables rule [98], each weak classifier is chosen from a pool of 300 random 1D descriptors, rather than from all 1D possible descriptors (Table 3.1). By using such a pool, it is warranted that the selected 1D descriptor is better than the 99% of the rest with a probability of 95%. We use accuracy (well classified training samples over the total) as the descriptor selection measurement. From all available ExtHaar and EOH 1D descriptors (Table 3.1) we use only 3,780 for building the strong classifier. We selected such a number because it is the dimension of the HOG descriptor when working with INRIA, which gives good results (Sect. 3.3). In fact, setting such a number is not a solved issue, thus, we did several experiments increasing it. However, we did not obtain significative accuracy improvements, while the training and testing computational time increased a lot.

Table 3.4 summarizes the descriptors parameters of ExtHaar and EOH. Bear in mind that we also make use of an image-pyramid (in this case using the so-called integral image representation at each level) maintaining fixed the CW size, which is not the traditional procedure [44] but improves the results.

Table 3.4: Summary of extra descriptor parameters.

Descriptor	**Parameters**	**Descr. dimensionality***	
		INRIA ($\mathcal{I}$)	**Daimler** ($\mathcal{D}$)
ExtHaar	Luminance, 8 filters from [67]: non-rotated edge (2), line (4) and center-surround (1), plus diagonal (1). Contrast normalized CW (by its standard deviation).	$^{\circ}$22,848	$^{\circ}$39,168
EOH	Max-gradient in RGB, 6 interpolated bins (0° to 180°).	$^{\circ}$42,840	$^{\circ}$73,440

($\star$) Virtual ($\mathcal{V}$): as $\mathcal{I}$ for $\mathcal{T}_{\mathcal{I}}^{tt}$ testing, and as $\mathcal{D}$ for $\mathcal{T}_{\mathcal{D}}^{tt}$ testing.
($\circ$) Only 3,780 are selected for the final strong classifier.

Table 3.5 shows the performance of the extra experiments Haar, EOH and Haar+EOH over the main datasets: 42 detectors that we developed following the same training

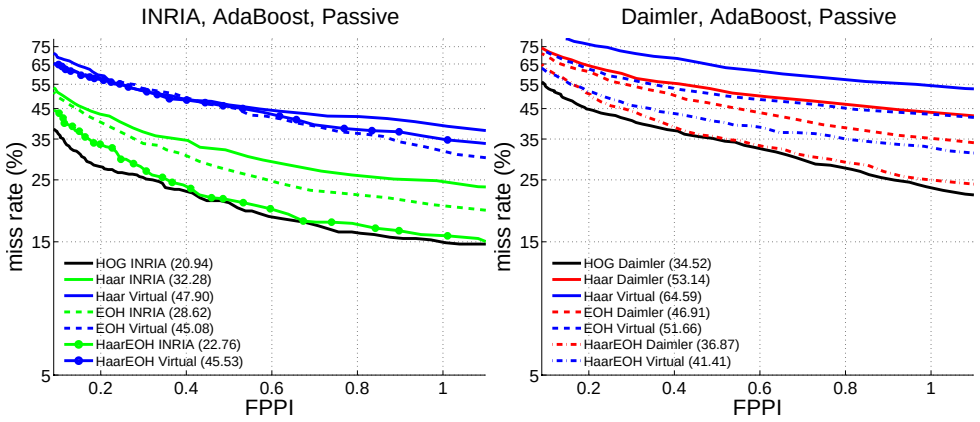

Figure 3.8: Passive learning results for AdaBoost.

and evaluation procedure. Figure 3.8 plots the accuracy curves.

Table 3.5: Passive learning results for ExtHaar, EOH and ExtHaar+EOH over INRIA and Daimler datasets. AMR is shown in % for FPPI $\in [10^{-1}, 10^{0}]$. Bold values are the best for a testing set.

Feature	**Testing** ($\mathcal{T}_{\mathcal{R}}^{tt}$)	**Training** ($\mathcal{T}_{\mathcal{X}}^{tr}$)		Δ
		Real ($\mathcal{R}$)	**Virtual** ($\mathcal{V}$)	
EOH	**Daimler** ($\mathcal{D}$)	46.91	51.66	-4.75
	INRIA ($\mathcal{I}$)	30.25; *76.20	42.75	-12.5
ExtHaar	**Daimler** ($\mathcal{D}$)	53.14	64.59	-11.45
	INRIA ($\mathcal{I}$)	31.19; ▷99.60	46.86	-15.67
EOH + ExtHaar ♮	**Daimler** ($\mathcal{D}$)	**36.87**	41.41	-4.54
	INRIA ($\mathcal{I}$)	**21.54**	37.59	-16.05

(∗) Geronimo *et al.* implementation [44].
(▷) Viola-Jones OpenCV implementation [109].
(♮) 3,780-D, with 7,560-D AMR was only 1% lower.

The largest improvements with respect to common implementations in the literature go to EOH and ExtHaar (*e.g.* 52.28 points and 71.78, resp., for INRIA). For EOH we use 6 bins instead of 4 as in [44]. However, the major advantage is due to the use of a fixed CW and the pyramidal sliding window procedure, instead of scaling the filters (ExtHaar, EOH). Note that filter scaling is the approach giving rise to the usually reported poor performances (*e.g.* as in [27]). In fact, with our approach, EOH+Haar/Real-AdaBoost is similar in performance to HOG+LBP/Lin-SVM. Of

course, scaling the filters turns out in much faster detectors, however, the problem of building fast solutions for pyramidal sliding window is under research [24]. Thus, here we have favored better detection.

Fig. 3.8 plots the performance of real-AdaBoost based pedestrian detectors trained on Real and virtual-word data, applied to INRIA and Daimler testing sets. Comparing the performance of the HOG/linear-SVM and the HaarEOH/real-AdaBoost we realize that is almost the same. Moreover, we show the results of three different pedestrian detectors based on different sets of features: Haar/real-AdaBoost, EOH/real-AdaBoost and HaarEOH/real-AdaBoost. The pedestrian detectors trained on the real datasets clearly outperforms their counterparts trained on the virtual-world one. The gap of performance is over 10 points.

3.6 Summary

In this chapter we have explored how realistic virtual worlds can help in learning appearance-based models for pedestrian detection in real-world images. Ultimately, this would be a proof of concept of a new framework for obtaining low cost precise annotations of objects, whose visual appearance must be learnt.

In order to automatically collect pedestrians and background samples we rely on players/drivers of a photo-realistic videogame borrowed from the entertainment industry. With such samples we have followed a standard passive-discriminative learning paradigm to train a virtual-world based pedestrian classifier that must operate in images depicting the real world (INRIA, Daimler, ETH-0,1,2, TUD and Caltech). Following such a framework we have tested state-of-the-art pedestrian descriptors (HOG/LBP/HOG+LBP) with Lin-SVM and (ExtHaar/EOH/ExtHaar+EOH) with AdaBoost. Within the same pedestrian detection scheme, we have employed virtual-world based classifiers and real-world based ones (Virtual, INRIA, Daimler, CVC02, ETH-0,1,2, TUD and Caltech). In total 203 train-test runs have been performed to assess detection performance. We have reached the conclusion that both virtual- and real-world based training behave in a similar way. This means that virtual-world based training can provide excellent performance, but it can also produce a performance drop as real-world based training when training and testing comes from a different nature. The amount of different pedestrian detectors and datasets allows to take this conclusion as trustworthy. Therefore, we think that the results presented in this chapter are relevant for research in pedestrian detection.

Chapter 4

Virtual and Real World Adaptation

The experiments conducted in previous chapter have shown that virtual-world based training can provide excellent testing performance in real world, but it can also suffer a performance drop, as real-world based training does. This is known as *dataset shift* problem and happens when training and testing data are from a different nature. Accordingly, we have designed a *domain adaptation* framework, V-AYLA (Fig. 4.1), in which we have tested different techniques to collect a few pedestrian samples from the target domain (real world) and combine them with many examples from the source domain (virtual world) in order to train a domain adapted pedestrian classifier that will operate in the target domain. V-AYLA reports the same detection performance than when training with many human-provided pedestrian annotations and testing with real-world images of the same domain. To the best of our knowledge, this is the first work demonstrating adaptation of virtual and real worlds for developing an appearance-based object detector.

4.1 Introduction

Experiments of Chapt. 3 showed that we obtain the similar performance by training with real-world based samples than by using virtual-world ones, which is encouraging from the viewpoint of object detection in general. However, not only good behavior is shared between virtual- and real-world based training, but some undesired effects too. For instance, let us assume that, for learning a pedestrian classifier, we annotated hundred of pedestrians in images acquired with a given camera. Using such camera and classifier we solve our application. Say that later we shall use a camera with a different sensor or we have to apply the classifier in another similar application/context but not equal. This variation can decrease the performance of our classifier because the probability distribution of the training data can be now much different than before with respect to the new testing data. This problem is referred to as *dataset shift* [88] and is receiving increasing attention in the Machine Learning

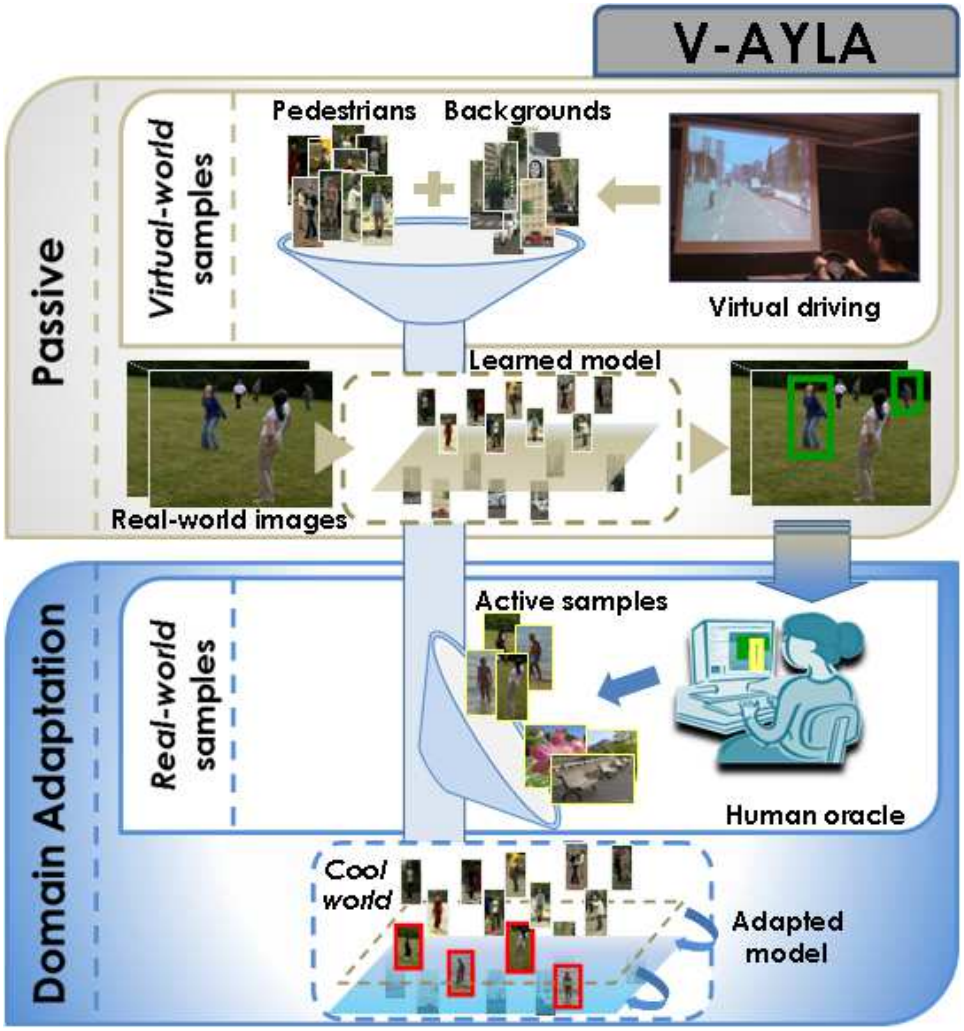

Figure 4.1: V-AYLA: *virtual*-world *a*nnotations *y*et *l*earning *a*daptively. Passive + domain adaptation training.

field [7, 86, 88] due to its relevance in areas like natural language processing, speech processing, and brain-computer interfaces, to mention a few.

Following the same example, the best we can do is to annotate the images from the new sensor/application/context and learn a new classifier. For minimizing the annotation effort, the challenge consists in *adapting* training and testing/application domains. Virtual-world images, although photo-realistic, come from a different *eye* than those acquired with a real camera. Therefore, dataset shift can appear. Thus, our proposal of using virtual worlds to learn pedestrian classifiers is cast in a *domain*

adaptation framework that we call Virtual-AYLA[1], V-AYLA in short, which stands for *virtual*-world *a*nnotations *y*et *l*earning *a*daptively (Fig. 4.1). For reaching the desired performance, V-AYLA combines virtual-world samples with a relatively low number of annotated real-world ones, within what we call *cool world*[2]. To the best of our knowledge, this is the first work demonstrating adaptation of virtual and real worlds for developing an appearance-based object detector.

As proof of concept, in this chapter V-AYLA relies on active learning for collecting a few real-world pedestrians, while each original descriptor space as well as the so-called *augmented descriptor space* will be used as cool worlds. In fact, we will not use the classical sampling within the base-classifier ambiguity region. We borrowed the idea of augmented descriptor space from [86], where it is applied to different problems though no one related to Computer Vision, a field that has largely disregarded dataset shift. Fortunately, this problem has also been explored recently in object recognition [10, 96], although not for a detection task like in this paper. Moreover, in [10, 96] the domains to be adapted are both based on real-world images and the involved descriptors are not the ones used for pedestrian detection.

The remainder of the chapter is organized as follows. Section 4.2 details the domain adaptation technique including the joint domain, the domain exploration and the training. Section 4.3 details the conducted experiments while Sect. 4.4 presents the results and corresponding analysis. Moreover, Sect. 4.5 present additional experiments in terms of extra datasets and features. Finally, Sect. 4.6 draws the main conclusions.

4.2 Domain adaptation

In one-class discriminative learning, samples $\mathbf{s} \in \mathcal{S}$ are randomly collected and an associated label $y \in \mathcal{Y}$ is assigned to each of them, where $\mathcal{S}$ and $\mathcal{Y} = \{-1, +1\}$ are the samples and annotation spaces, resp. The set of annotated samples $\mathcal{T} = \{(\mathbf{s}_k, y_k)|k : 1 \dots n\}$ is divided in two disjoint sets, $\mathcal{T}^{tr}$ and $\mathcal{T}^{tt}$, to train and test a classifier $\mathrm{C} : \mathcal{S} \rightarrow \mathcal{Y}$, resp. It is assumed a joint probability distribution $p(\mathbf{s}, y)$ describing our domain of interest δ. Elements in $\mathcal{T}$ are randomly drawn (i.i.d.) from $p(\mathbf{s}, y)$ and, thus, $\mathcal{T}^{tr}$ and $\mathcal{T}^{tt}$ too. This is the case of settings $\{\mathcal{T}_{\mathcal{I}}^{tr}, \mathcal{T}_{\mathcal{I}}^{tt}\}$ and $\{\mathcal{T}_{\mathcal{D}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$ (Chapt. 3).

[1] *AYLA* evokes the main character, a Cro-Magnon women, of *Earth's Children* saga by *J. M. Auel*. *Ayla* is an icon of robustness and adaptability. During her childhood she is educated by Neanderthals (*The Clan*), whose physical appearance corresponds to *normal humans* for her. However, she recognizes Cro-Magnons as humans too the first time she met them. *Ayla* adapts from Neanderthals to Cro-Magnons customs, keeping the best of both worlds. She is a real survivor in such demanding primitive Earth conditions. Interestingly, *Ayla* is the Hebrew name for *oak tree*. It turns out also that there is a popular videogame that incorporates *Ayla* as character.

[2] *Cool world* term evokes the film with that title. In it, there is a real and a cool world, in the latter real humans live with cartoons.

In practice, there are cases in which samples in $\mathcal{T}^{tr}$ and $\mathcal{T}^{tt}$ follow different probability distributions. As mentioned in Sect. 4.1, *dataset shift* is the generic term to summarize the many possible underlying reasons [88]. We argue that the loss of performance seen in Chapt. 3 when using different sets to test and train pedestrian classifiers is due to some form of dataset shift. For instance, this is our assumption for settings $\{\mathcal{T}^{tr}_{\mathcal{V}}, \mathcal{T}^{tt}_{\mathcal{I}}\}$ and $\{\mathcal{T}^{tr}_{\mathcal{V}}, \mathcal{T}^{tt}_{\mathcal{D}}\}$. Accordingly, in this section we apply *domain adaptation* to overcome the problem.

In domain adaptation, it is assumed a *source* domain, δ_s, and a *target* domain, δ_t, with corresponding $p_s(\mathbf{s}, y)$ and $p_t(\mathbf{s}, y)$, which are different yet correlated distributions since otherwise adaptation would be impossible. Annotated samples from δ_s are available, as well as samples from δ_t that can be either partially annotated or not annotated at all. We focus on the so-called *supervised* domain adaptation [86], where we have a reasonable number of annotations from δ_s and some annotations from δ_t too. Since we aim at reducing manual annotations for building object classifiers, our δ_s is the virtual world $\mathcal{V}$, and δ_t is the real world $\mathcal{R}$ (here $\mathcal{R} \in \{\mathcal{I}, \mathcal{D}\}$).

As in Chapt. 3, we assess domain adaptation for HOG[3] and LBP separately, as well as for HOG+LBP. Also, at the end we we perform some experiments with extra descriptors (Haar, EOH and Haar+EOH) and extra datasets (Caltech, ETH0-1-2, CVC02 and TUD).

4.2.1 Virtual- and real-world joint domain

As described previously, pedestrian classifiers rely on a descriptor extraction process, $\mathbf{D}$, that transforms the samples, $\mathbf{s}$, into their respective descriptors (*i.e.* HOG, LBP, etc.), $\mathbf{x} = \mathbf{D}(\mathbf{s}), \mathbf{x} \in \Omega \subset \Re^d$. Therefore, the learning process holds in Ω. In the domain adaptation framework, some $\mathbf{x}$'s come from δ_s samples and others from δ_t samples. Thus, it arises the question of how to *joint* both types of $\mathbf{x}$'s in order to learn domain adapted classifiers. Since our δ_s is based on virtual-world samples and δ_t in real-world ones, as we mentioned in Sect. 4.1 we call the joint domain *cool world*. In this paper we test two cases.

The first one, called ORG, comes from just treating virtual- and real-world descriptors equally. In other words, from the learning viewpoint, virtual- and real-world samples are just mixed within the original Ω.

The second cool world, AUG, is based on the so-called *feature augmentation* technique proposed in [86]. Instead of working in Ω, we work in Ω^3 by applying the mapping $\Phi : \Omega \rightarrow \Omega^3$ defined as $\Phi(\mathbf{x}) =< \mathbf{x}, \mathbf{x}, \mathbf{0} >$ if $\mathbf{s} \in \mathcal{V}$ and $\Phi(\mathbf{x}) =< \mathbf{x}, \mathbf{0}, \mathbf{x} >$ if $\mathbf{s} \in \mathcal{R}$, where $\mathbf{0}$ is the null vector in Ω, and $\mathbf{x} = \mathbf{D}(\mathbf{s})$. Under this mapping, $< \mathbf{x}, \mathbf{0}, \mathbf{0} >$ corresponds to a common subspace of Ω^3 where virtual and real-world samples (*i.e.* their descriptors) meet, $< \mathbf{0}, \mathbf{x}, \mathbf{0} >$ is the virtual-world subspace, and

[3] In [107,108] we present domain adaptation results for HOG/Lin-SVM. However, only for INRIA and neither using our current HOG (but the original [21]) nor augmented descriptor space. Moreover, in this thesis we reduce the annotation effort in 15 percentual points, *i.e.* from 25% in [107,108] to 10% here.

$< \mathbf{0}, \mathbf{0}, \mathbf{x} >$ the real-world one. The rational is that learning using $\Phi(\mathbf{x})$ descriptors instead of $\mathbf{x}$ ones allows the SVM algorithm to jointly exploit the commonalities of δ_s and δ_t, as well as their differences. We refer to [86] for an explanation in terms of SVM margin maximization.

During training, the full $\Phi(\mathbf{x})$ is used. However, during testing all the samples will come from the real-world (here $\mathcal{I}$ or $\mathcal{D}$), therefore, only the corresponding option of $\Phi(\mathbf{x})$ is used. Let $\mathbf{w} = < \mathbf{w}_{st}, \mathbf{w}_s, \mathbf{w}_t >$ be the weighting vector learnt by the Lin-SVM algorithm, then during testing we have $\mathbf{w} \cdot \Phi(\mathbf{x}) = \mathbf{w}_{st} \cdot \mathbf{x} + \mathbf{w}_t \cdot \mathbf{x} = \mathbf{x} \cdot (\mathbf{w}_{st} + \mathbf{w}_t) = \mathbf{x} \cdot \hat{\mathbf{w}}$ for $\hat{\mathbf{w}} = \mathbf{w}_{st} + \mathbf{w}_t$. Vector $\hat{\mathbf{w}}$ can be computed once and off-line. Thus, the computational cost and memory requirements for using AUG during testing is the same than the one of ORG.

4.2.2 Real-world domain exploration

Let n_t^p be the maximum number of real-world (target domain) pedestrians a human oracle $\mathcal{O}$ is allowed to provide for training. We test four behaviors for $\mathcal{O}$.

Following the first behavior, $\mathcal{O}$ annotates n_t^p pedestrians at *random* (Rnd). The rest of behaviors are based on a sort of *selective sampling* [20]. In particular, there is a first stage consisting in learning a pedestrian classifier, $\mathrm{C}_\mathcal{V}$, by using the the virtual-world samples and passive learning. Such a classifier is used in a second stage to ask $\mathcal{O}$ for *difficult* samples from the real-world data. We will see in Sect. 4.2.3 that such samples jointly with the virtual-world ones will be used in a third stage for retraining.

In the second behavior, *active learning for pedestrians* (Act+), $\mathcal{O}$ annotates n_t^p *difficult-to-detect* pedestrians. Analogously, we term our third behavior as *active learning for background* (Act−) because $\mathcal{O}$ only marks false positives. The idea behind Act− is not to collect the annotated false positives, but the right detections (true positives) as provided by the used pedestrian detector. In other words, in this case, the BB annotations of the n_t^p real-world pedestrians are provided by the pedestrian detector itself. Finally, we term as Act± the fourth behavior since it is a combination of Act+ and Act−. In this case we allow to collect $2n_t^p$ real-world pedestrians because just n_t^p are manually annotated with BBs, which is the task we want to avoid.

Let us define the difficult cases for $\mathrm{C}_\mathcal{V}$. Given a real-world sample $\mathbf{s}_\mathcal{R}$, if $\mathrm{C}_\mathcal{V}(\mathbf{s}_\mathcal{R}) > Thr$, then $\mathbf{s}_\mathcal{R}$ is classified as pedestrian. Accordingly, in the Act+ case, $\mathcal{O}$ will annotate real-world pedestrians, $\mathbf{s}_\mathcal{R}^+$, for which $\mathrm{C}_\mathcal{V}(\mathbf{s}_\mathcal{R}^+) \leq Thr$. In the Act− case, those background samples, $\mathbf{s}_\mathcal{R}^-$, for which $\mathrm{C}_\mathcal{V}(\mathbf{s}_\mathcal{R}^-) > Thr$ must be rejected by $\mathcal{O}$. For the Act± both things hold. In general, selective sampling for SVM focus on samples inside the ambiguity region $[-1, 1]$. However, underlaying such an approach is the assumption of a shared train and test domain. Here, due to dataset shift, wrongly classified samples out of the margins can be important to achieve domain adaptation.

4.2.3 Domain adaptation training: V-AYLA

Assume the following definitions of *training sets*:

- *Source domain.* Let $\Im_{\mathcal{V}}^{tr+}$ be the set of virtual-world images with automatically annotated pedestrians, and $\Im_{\mathcal{V}}^{tr-}$ the set of pedestrian-free virtual-world images automatically generated as well.
- *Target domain.* Let $\Im_{\mathcal{R}}^{tr+}$ be a set of real-world images with non-annotated pedestrians, and $\Im_{\mathcal{R}}^{tr-}$ a set of pedestrian-free real-world images.

Take the following decisions:

- *Classifier basics.* Here we assume Lin-SVM, and $\mathbf{D} \in \{\text{HOG}, \text{LBP}, \text{HOG+LBP}\}$.
- *Cool world.* Choices are ORG and AUG.
- *Oracle.* Choices are $\mathcal{O} \in \{\text{Rnd}, \text{Act}+, \text{Act}-, \text{Act}\pm\}$.

The training method we use for performing domain adaptation can be summarized in the following steps:

(s1) Perform *passive learning in virtual world* using $\{\Im_{\mathcal{V}}^{tr+}, \Im_{\mathcal{V}}^{tr-}\}$ and $\mathbf{D}$ (Sect. 3.2). Let us term as $\text{C}_{\mathcal{V}}$ the passively learnt pedestrian classifier and as $\text{D}_{\mathcal{V}}$ its associated detector (Sect. 3.2.2). Let $\mathcal{T}_{\mathcal{V}}^{tr+}$ be the set of pedestrians used for obtaining $\text{C}_{\mathcal{V}}$ (*i.e.* coming from $\Im_{\mathcal{V}}^{tr+}$, scaled to the CW size and augmented by mirroring), and $\mathcal{T}_{\mathcal{V}}^{tr-}$ the set of background samples (coming from $\Im_{\mathcal{V}}^{tr-}$ after bootstrapping, CW size).

(s2) *Selective sampling in real world.* In order to obtain real-world annotated pedestrians, follow $\mathcal{O}$ by running $\text{D}_{\mathcal{V}}$ on $\Im_{\mathcal{R}}^{tr+}$. If $\mathcal{O} = \text{Act}\pm$, then we collect n_t^p following Act+ style and n_t^p more following Act− style (which does not involve manual pedestrian BB annotations). Otherwise, only n_t^p pedestrians are collected. We term as $\mathcal{T}_{\mathcal{R}}^{tr+}$ the set of such new pedestrian samples scaled to CW size and augmented by mirroring, and as $\mathcal{T}_{\mathcal{R}}^{tr-}$ a set of background samples in CW size, taken from $\Im_{\mathcal{R}}^{tr-}$ as done in the passive learning procedure before bootstrapping (thus, the cardinality of $\mathcal{T}_{\mathcal{R}}^{tr+}$ and $\mathcal{T}_{\mathcal{R}}^{tr-}$ are equal). Note that to follow $\mathcal{O}$ we need to set Thr. For that purpose, we initially select a few images from $\Im_{\mathcal{R}}^{tr-}$ and take a Thr value such that after applying $\text{D}_{\mathcal{V}}$ on the selected images, less than 3 FPPI in average are obtained. We start trying with $Thr = 1$ and decrease the value in steps of 0.5 while such a FPPI holds. This is an automatic procedure.

(s3) Perform *passive learning in cool world.* Map samples in $\mathcal{T}_{\mathcal{V}}^{tr+}$, $\mathcal{T}_{\mathcal{V}}^{tr-}$, $\mathcal{T}_{\mathcal{R}}^{tr+}$, and $\mathcal{T}_{\mathcal{R}}^{tr-}$ to cool world. Next, train a new classifier with them according to $\mathbf{D}$. Then, perform bootstrapping in $\Im_{\mathcal{R}}^{tr-}$. Finally, re-train in cool world to obtain the domain adapted classifier.

When $\mathcal{O} \neq \text{Rnd}$, this is a *batch active learning* procedure [65]. Figure 4.1 summarizes the idea, which, as we introduced in Sect. 4.1, we term as V-AYLA.

Table 4.1: Domain adaptation results for Lin-SVM based pedestrian detectors over Daimler dataset. For FPPI $\in [0.1, 1]$, AMR (%) mean and std. dev. are indicated. Bold values remark the best mean for each real-world testing set.

Daimler	Joint	Oracles			
		Act+	Act∼	Rnd	Act±
HOG	ORG	28.27 ± 0.41	28.59 ± 0.43	28.58 ± 0.36	26.59 ± 0.51
	AUG	$\mathbf{26.13} \pm 0.66$	30.59 ± 1.28	26.30 ± 0.88	27.40 ± 0.65
LBP	ORG	40.25 ± 0.45	41.24 ± 0.51	40.84 ± 0.52	38.54 ± 0.79
	AUG	34.69 ± 1.15	36.27 ± 0.89	34.56 ± 1.23	$\mathbf{33.18} \pm 1.95$
HOG +LBP	ORG	22.85 ± 0.43	24.15 ± 0.73	23.64 ± 0.57	22.18 ± 0.65
	AUG	$\mathbf{21.71} \pm 0.73$	27.43 ± 1.31	22.20 ± 1.26	23.79 ± 1.01

Table 4.2: Domain adaptation results for Lin-SVM based pedestrian detectors over INRIA dataset. For FPPI $\in [0.1, 1]$, AMR (%) mean and std. dev. are indicated. Bold values remark the best mean for each real-world testing set.

INRIA	Joint	Oracles			
		Act+	Act−Act∼	Rnd	Act±
HOG	ORG	25.65 ± 0.48	27.58 ± 0.61 26.99 ± 0.55	27.13 ± 0.71	24.10 ± 0.64
	AUG	22.47 ± 1.01	24.19 ± 0.54 23.95 ± 1.02	22.94 ± 0.88	$\mathbf{21.29} \pm 0.85$
LBP	ORG	23.21 ± 0.52	24.72 ± 0.42 24.98 ± 0.36	23.75 ± 0.73	21.70 ± 0.65
	AUG	22.83 ± 0.92	23.31 ± 0.75 22.08 ± 0.88	19.73 ± 1.19	$\mathbf{18.87} \pm 0.88$
HOG+ LBP	ORG	16.65 ± 0.74	19.34 ± 0.60 19.61 ± 0.51	18.56 ± 0.61	15.10 ± 0.91
	AUG	14.70 ± 0.63	17.46 ± 0.63 15.47 ± 0.89	15.07 ± 1.29	$\mathbf{14.15} \pm 0.58$

Table 4.3: Domain adaptation results comparison. Rows $\mathcal{T}_{\mathcal{V}}^{tr}$ show performance results by training with $\mathcal{V}$ data only (from Tables 4.2 and 4.1). In rows $10\%\mathcal{T}_{\mathcal{I}}^{tr}$ and $10\%\mathcal{T}_{\mathcal{D}}^{tr}$, show results by training with the 10% of the available real-world training data of $\mathcal{I}$ and $\mathcal{D}$, resp. Rows Act+/AUG reproduce domain adaptation results from Table 4.2 and 4.1, where the 10% or real-world pedestrians combined with the virtual-world ones are the same than for the corresponding $10\%\mathcal{T}_{\mathcal{I}}^{tr}$ and $10\%\mathcal{T}_{\mathcal{D}}^{tr}$ rows. Δ_1 rows show the difference between the mean of corresponding $\mathcal{T}_{\mathcal{V}}^{tr}$ and Act+/AUG. Δ_2 rows illustrate differences between $10\%\mathcal{T}_{\mathcal{I}}^{tr}/10\%\mathcal{T}_{\mathcal{D}}^{tr}$ and Act+/AUG.

INRIA ($\mathcal{T}_{\mathcal{I}}^{tt}$)	HOG	LBP	HOG+LBP
$\mathcal{T}_{\mathcal{V}}^{tr}$	32.47 ± 0.47	28.87 ± 0.70	23.81 ± 0.53
$10\%\mathcal{T}_{\mathcal{I}}^{tr}$	30.81 ± 1.51	26.56 ± 1.96	18.89 ± 1.24
Act+/AUG	22.47 ± 1.01	22.83 ± 0.92	14.70 ± 0.63
Δ_1	10.00	06.04	09.11
Δ_2	08.34	03.73	04.19
Daimler ($\mathcal{T}_{\mathcal{D}}^{tt}$)	HOG	LBP	HOG+LBP
$\mathcal{T}_{\mathcal{V}}^{tr}$	30.64 ± 0.43	45.21 ± 0.49	28.27 ± 0.48
$10\%\mathcal{T}_{\mathcal{D}}^{tr}$	34.64 ± 1.31	41.13 ± 1.36	30.96 ± 1.59
Act+/AUG	26.13 ± 0.66	34.69 ± 1.15	21.71 ± 0.73
Δ_1	04.51	10.52	06.56
Δ_2	08.51	06.44	09.25

4.3 Experimental results

In this section we summarize the V-AYLA experiments (300 training-testing runs in total), and we explain how to simulate the application of V-AYLA on INRIA and Daimler data for providing fair performance comparisons with respect to passive learning. Then, as we have done with the passive experiments, we perform experiments with the most promising techniques on the remaining datasets.

First of all, we shall restrict ourselves to a maximum amount of manually annotated pedestrian BBs from the real-world images (target domain). In the supervised domain adaptation framework, the cardinality of annotated target samples is supposed to be much lower than the one of annotated source samples. However, there is no a general maximum since this depends on the application. As rule of thumb, here we want to avoid the 90% of the manually annotated BBs. In particular, since for both INRIA and Daimler we have used 1,208 annotated pedestrians (*i.e.* before mirroring) in the passive learning setting, then we will assume the use of a maximum of 120 manually annotated BBs from real-world images. Thus, we aim to achieve the same performance for the following two scenarios: (1) applying passive training

using training and testing sets from the same domain, with 1,208 annotated real-world pedestrians; (2) applying domain adaptation with a training set based on our virtual-world data plus a set of 120 real-world manually annotated pedestrians and a set of pedestrian-free images, both sets from the same domain in which we are going to perform the testing.

During an actual application of V-AYLA, the 120 real-world pedestrians used for domain adaptation will change from one training-testing run to another. Thus, this is simulated in the experiments conducted in this section. However, although V-AYLA only needs a few manual BB annotations (*i.e.* 120 here), as we have mentioned before, our experiments involve 300 training-testing runs, which would turn out in 36,000 manually annotated BBs. Therefore, in order to reduce overall manual effort, we have simulated the annotation of the 120 pedestrian BBs by just sampling them from the ones available for the passive learning, according to the different oracle strategies. However, note that during the actual application of V-AYLA all such passively annotated pedestrians (*i.e.* the 1,208 ones in each considered real-world data set) are not required in advance for further oracle sampling. Our experiments, without losing generality, use such an approach just for avoiding actual human intervention in each of the 300 training-testing runs. Additionally, in this manner the V-AYLA human annotators are the same than the ones of the passive approach, thus, removing variability due to different human expertise.

Hence, in order to simulate V-AYLA on INRIA for $\mathcal{O}$ = Rnd, we randomly sample $\mathcal{T}_{\mathcal{I}}^{tr+}$ to obtain the 120 real-world pedestrians. For Daimler we do the same using $\mathcal{T}_{\mathcal{D}}^{tr+}$. For simulating the case $\mathcal{O}$ = Act+ on INRIA, we randomly sample the false negatives obtained when applying $\mathrm{C}_{\mathcal{V}}$ on $\mathcal{T}_{\mathcal{I}}^{tr+}$. The desired 120 real-world pedestrians are collected in such a manner. Daimler case is analogous by using $\mathcal{T}_{\mathcal{D}}^{tr+}$.

Rnd and Act+ involve manual annotation of pedestrian BBs. However, in Act− the annotations must be provided by the passively learnt pedestrian detector. INRIA dataset includes the images ($\Im_{\mathcal{I}}^{tr+}$) and annotations from which $\mathcal{T}_{\mathcal{I}}^{tr+}$ is obtained. Thus, we apply $\mathrm{D}_{\mathcal{V}}$ to $\Im_{\mathcal{I}}^{tr+}$ images, and collect the desired number of pedestrian detections following Act− behavior. Note that in these experiments Act+ and Act− take samples from the same original pedestrians in $\Im_{\mathcal{R}}^{tr+}$. Once such pedestrians are scaled to the CW size, the difference between those coming from Act+ and Act− is that, in the former case, the original pedestrians were annotated by a human oracle, while in the latter case it is the own pedestrian detector which annotates them. Simulating Act− in Daimler is not directly possible since $\Im_{\mathcal{D}}^{tr+}$ is not provided, just the corresponding $\mathcal{T}_{\mathcal{D}}^{tr+}$ is publicly available. In this case, instead of applying $\mathrm{D}_{\mathcal{V}}$ to $\Im_{\mathcal{D}}^{tr+}$, we apply $\mathrm{C}_{\mathcal{V}}$ to $\mathcal{T}_{\mathcal{D}}^{tr+}$. Therefore, instead of Act− we term as Act∼ such an $\mathcal{O}$.

For $\mathcal{O}$ = Act±, 240 real-world pedestrians are selected. However, only 120 BBs are annotated by a human oracle (Act+), the others are collected according to either Act− for INRIA or Act∼ for Daimler.

In Sect. 4.2.3 we saw that V-AYLA involves finding a threshold value Thr. Applying the proposed procedure, we found that $Thr = -0.5$ is a good compromise for all descriptors and real-world data under test.

Table 4.2 shows the application of V-AYLA on INRIA for $\mathcal{O} \in \{\text{Rnd}, \text{Act}+, \text{Act}-, \text{Act}\pm\}$, combined with both ORG and AUG. Regarding Daimler (Table 4.1), we show analogous experiments but replacing Act− by Act∼, which propagates to Act±. Figure 4.2 plots the curves corresponding to most interesting results.

4.4 Discussion

For performing domain adaptation, source and target domains must be correlated, *i.e.* the passive learning stage of V-AYLA must not give random results, otherwise, the adaptation stage cannot improve them. Fortunately, such a stage of V-AYLA already offers a good approximation as seen in the results of Table 3.2, *i.e.* virtual-world samples alone help to learn a relatively good pedestrian classifier for real-world images[4]. Thanks to that, the adaptation stage of V-AYLA is able to provide state-of-the-art results by just manually annotating a few real-world pedestrian BBs (max. 120). In order to support this statement we have run different statistical tests to compare the results based on just real-world data with the counterparts based on the analyzed domain adaptation techniques. We explain them in the following.

First, we have compared the different cool worlds, *i.e.* ORG *vs* AUG. In particular, we use a paired Wilcoxon test considering separately the three descriptors times the four oracles, irrespectively of the real-world testing data set. This turns out in 12 tests. For eight of them AUG is better than ORG, while in the rest there is no statistically meaningful difference. In fact, Tables 4.2 and 4.1 shows that in some cases there are large differences (*e.g.* for LBP with Daimler) but for the best detectors (*i.e.* using HOG+LBP) there is not almost performance difference. However, for the sake of reducing the number of remaining statistical tests, in the rest of this subsection we focus on AUG.

Second, we have compared the results of the four oracles using a Friedman test [40]. As intuitively expected from the results in Tables 4.2 and 4.1, among oracles Rnd, Act+ and Act± there is no statistically meaningful performance difference. However, Act− outputs worse results thought still improves the performance of using virtual-world data alone. At this point we chose the use of either Act+ or Act± since they have an advantage with respect to Rnd. In particular, it is worth to mention that the pedestrian examples of both INRIA and Daimler datasets where annotated by Computer Vision experts in proprietary software environments, thus, they present good accuracy and variability. Therefore, the Rnd strategy used here is implicitly assuming good annotators. However, this is not always the case when using modern web-based annotation tools [29,104]. We believe, that active strategies (Act+, Act±) have the potential advantage of teaching the human annotator how good quality annotations should be done, since he/she sees the detections output by the current

[4] *"She [Ayla] knew they were men, though they were the first men of the Others she could remember seeing. She had not been able to visualize a man, but the moment she saw these two, she recognized why Oda had said men of the Others looked like her."* from *The valley of the horses (Earth's Children), J.M. Auel.*

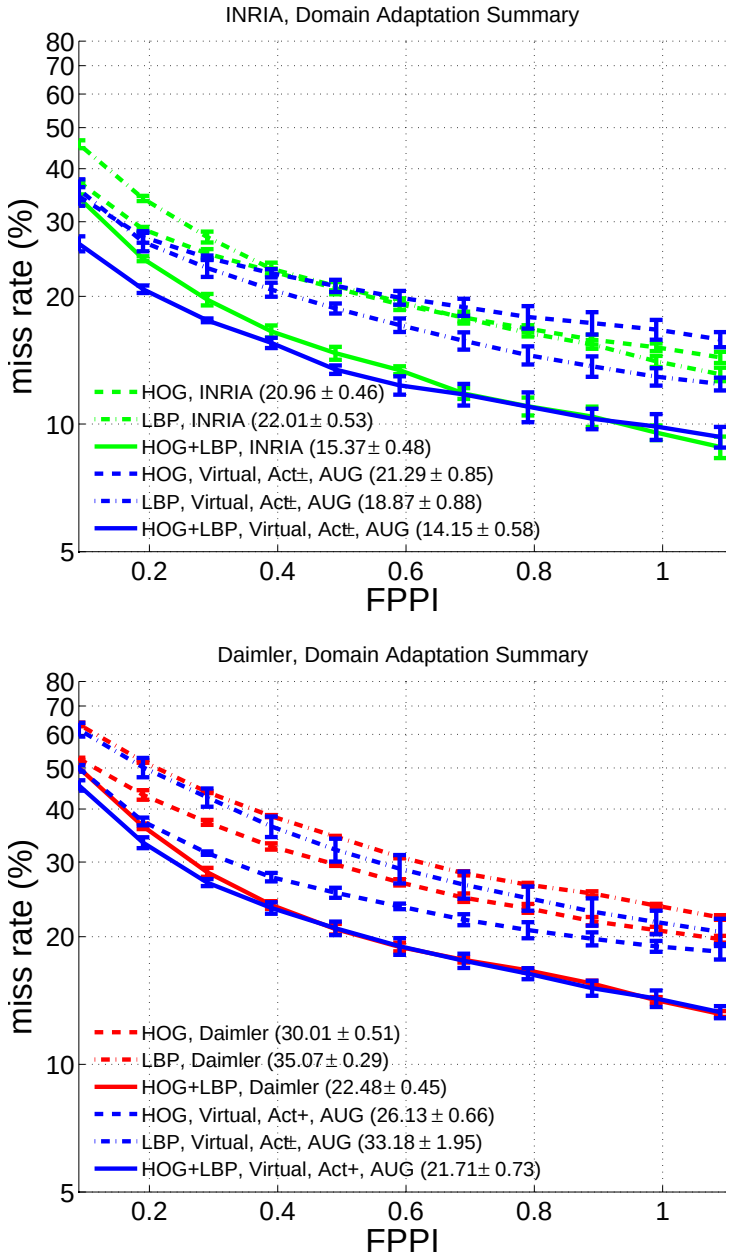

Figure 4.2: Domain adaptation results based on Lin-SVM over core datasets. Results for the best cases in Tables 4.2 and 4.1.

pedestrian detector.

We would like to mention that, although the Act− works worse than the other oracles still is able to provide a step forward in the adaptation (*e.g.* for INRIA setting more than 5 percentual points with respect to virtual-world based training alone) with the advantage of not requiring manual BB annotations. In fact, Act∼ (*i.e.* simulated Act−) drives to an analogous performance even though it is based on manual annotations. Note that, in order to do a fair comparison, for respective V-AYLA train-test runs of Act∼ and Act− the same pedestrians are used, only the BB coordinates framing them are different. Therefore, given the potentiality of even reducing more manual annotation we think that the Act− type of oracles (retrained for adaptation from self-detections) deserves more research in the future.

Using Wilcoxon unpaired test, we assess if V-AYLA (Act+ and Act±) has achieved domain adaptation, *i.e.* the null hypothesis is that classifiers trained according to the passive method and V-AYLA exhibit the same performance. In the case of Act+, for HOG V-AYLA is better in 1.12 percentual points with p-value = 0.9097, for LBP it is worse in 1.89 points with p-value = 0.7337, and for HOG+LBP it is better in 0.35 points with p-value = 0.8501. Therefore, we consider that V-AYLA/Act+ has reached domain adaptation. The analogous analysis for Act± concludes that for HOG V-AYLA is better in 1.25 points with p-value = 0.3847, for LBP the same with 1.65 points and p-value = 0.9097, while for HOG+LBP it is worse in 0.50 points with p-value = 0.3847. Thus, again we consider that V-AYLA/Act+ has reached domain adaptation.

In Table 4.3 we summarize the performance improvement obtained when adding the 10% of real-world data to virtual-world one, and viceversa. Note that adding the 10% of real-world data turns out in improvements from 4.5 percentual points to even 10.5 (Δ_1). An analogous situation is observed regarding the contribution of the virtual data (Δ_2). In the latter case the improvement for HOG (over 8 points for INRIA and Daimler cases) is remarkable since more elaborated models like Latent-SVM part-based ones rely on HOG-style information [34, 35].

In conclusion, V-AYLA allows to significantly save manual annotation effort while providing pedestrian detectors of comparable performance than the obtained by using standard passive training based on a larger amount of manual annotations.

4.5 Additional experiments

For complementing V-AYLA performance assessment, we extend our experiments in two ways: by adding more datasets and by adding more descriptors. To alleviate the number of experiments, as we did in previous chapter, for the remaining experiments we do not repeat the experiments several times. As seen in Table 4.3, classifiers trained with only the 10% of the real data perform clearly worse than our VAYLA ones, so we avoid these experiments with the extra datasets. However, we perform such experiments for the extra features to be sure that this still holds for this features.

Table 4.4: Domain adaptation results for Lin-SVM based pedestrian detectors over extra datasets. For FPPI $\in [0.1, 1]$, AMR (%) is indicated. Bold values remark the best accuracy for each real-world testing set. Δ_1 shows the difference between the corresponding $\mathcal{T}_{\mathcal{V}}^{tr}$ and the best V-AYLA result. Δ_2 illustrates differences between the corresponding real trained experiment and the best V-AYLA result.

Test set	Feature	Joint	Oracles		Δ_1	Δ_2
			Rnd	Act±		
ETH0	HOG	ORG	61.66*	60.41*	9.43	2.79
		AUG	59.40*	**55.95***		
ETH1	HOG	ORG	67.05*	66.64*	4.91	1.12
		AUG	68.67*	**64.39***		
ETH2	HOG	ORG	55.88*	55.24*	6.77	2.43
		AUG	**52.20***	53.90*		
TUD	HOG	ORG	66.41* 68.79°	65.01* 68.19°	5.22	0.58
		AUG	65.64* **62.76**$^\circ$	64.88* 64.68°		
CVC02	HOG+LBP	ORG	42.43* 32.11$^\bullet$	41.52* 36.33$^\bullet$	20.04	-4.99
		AUG	43.68* **29.58**$^\bullet$	45.41* 32.41$^\bullet$		
Caltech	HOG+LBP	ORG	38.45* 36.20$^\natural$	36.18* **32.19**$^\natural$	14.75	9.99
		AUG	42.55* 39.28$^\natural$	47.80* 41.13$^\natural$		

($\star$) Adapted to INRIA training set.
($\circ$) Adapted to TUD training set.
($\bullet$) Adapted to CVC02 training set.
($\natural$) Adapted to Caltech training set.

4.5.1 More datasets

We extend our experiments with the extra datasets also used on Chapt. 3: Caltech, ETH-0,1,2, TUD and CVC02. Table 4.4 shows the application of V-AYLA on these datasets. For all of them we adapt the virtual-world data to the INRIA one. Additionally, for the ones that has its own training data, *i.e.* TUD, CVC02 and Caltech, we also adapt our models using this data. Moreover, we restrict ourselves to just the most promising configurations. Regarding the $\mathcal{O}$ we restrict to Rnd and Act±, combined with both ORG and AUG. From results of Fig. 3.7 we decided to use HOG features for ETH-0,1,2 and TUD datasets and HOG+LBP for CVC02 and Caltech. Figure 4.3 plots the curves corresponding to most interesting results.

Virtual classifiers adapted to the testing datasets give better results than the ones adapted to INRIA. This reinforces our intuition that training with datasets of a different nature than the testing ones can produce a performance drop.

Table 4.4 reveals that our domain adapted pedestrian detectors outperform the counterpart real based pedestrian detectors (except for CVC02). This emphasises

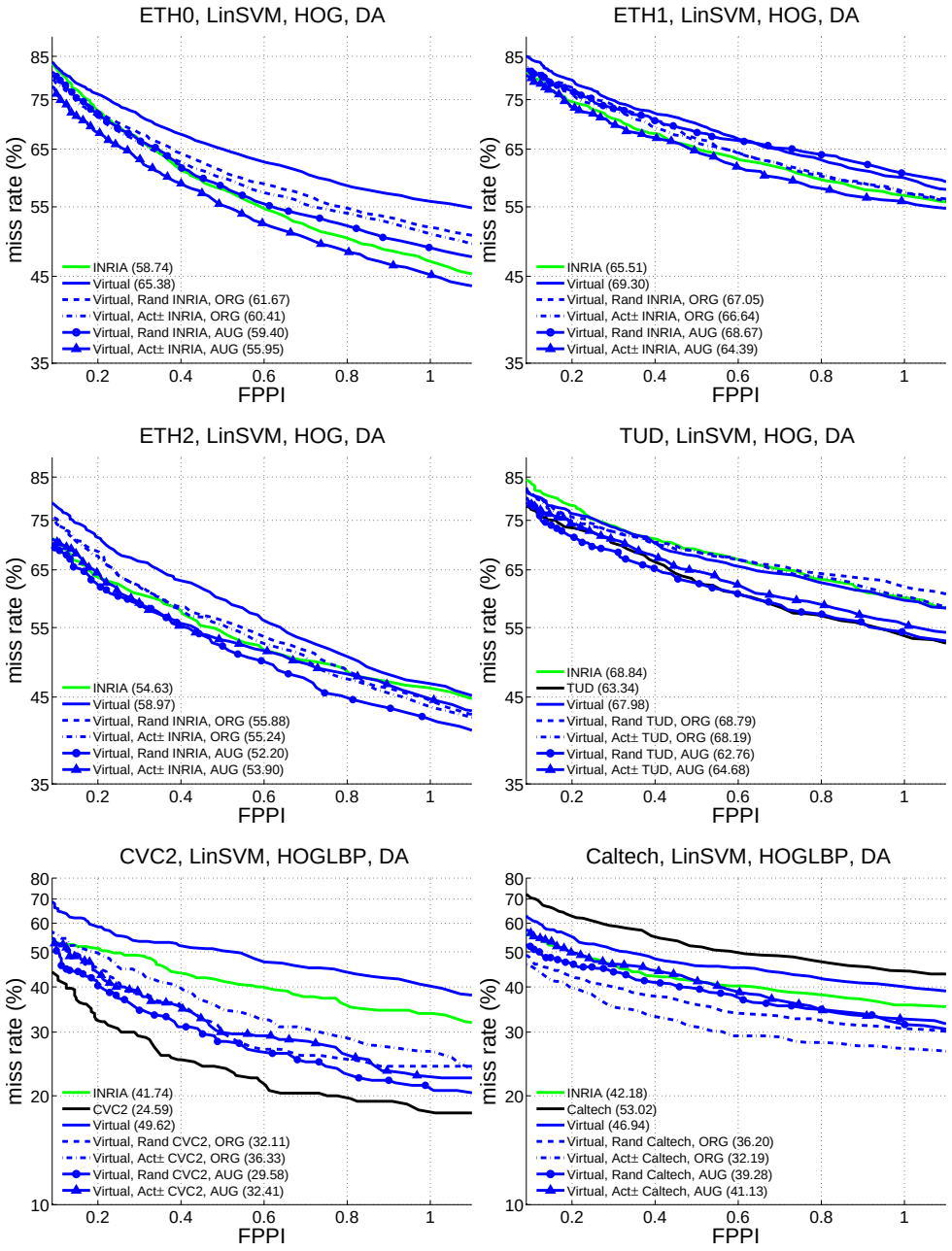

Figure 4.3: Domain adaptation results based over extra datasets: HOG+LBP over TUD, ETH-0,1,2, CVC02 and Caltech.

Table 4.5: Domain adaptation results based on AdaBoost over Daimler dataset. For FPPI $\in [0.1, 1]$, AMR (%) is indicated. Bold values remark the best accuracy for each feature. Δ_1 shows the difference between the corresponding $\mathcal{T}_{\mathcal{V}}^{tr}$ and the best V-AYLA result. Δ_2 illustrates differences between the corresponding real trained experiment and the best V-AYLA result.

Daimler	Joint	Oracles				Δ_1	Δ_2
		Act+	Act$\sim$	Rnd	Act$\pm$		
EOH	ORG	48.47	46.58	47.88	**43.98**	7.68	2.93
ExtHaar	ORG	62.37	57.55	58.75	**54.55**	10.04	-1.41
ExtHaar+EOH	ORG	34.15	36.35	36.21	**30.72**	10.69	6.15

that our methods perform well for a variety of scenarios. Note that the models adapted to its own data are better than the models adapted to INRIA one. Most of the experiments based on models adapted to INRIA with AUG strategy perform better than the ORG ones. This is consistent with the fact that AUG also performed better for the INRIA testing (See Table 4.2). Moreover, the models adapted to its own data following the AUG strategy perform better than the ORG ones. In most of the experiments the Act$\pm$ oracle leads to better results than the Rand one. For all the datasets except CVC02 the performance given by V-AYLA is superior to the real models. V-AYLA performance clearly improves the virtual one by even 13 points in some cases. In the CVC02 dataset the gap between virtual and real detectors is 25 points that was the only case where V-AYLA could not reach so much adaptation.

4.5.2 More features

We introduce the descriptors Haar, EOH and Haar+EOH and the learning machine Real-AdaBoost. Table 4.6 shows the application of V-AYLA on INRIA for $\mathcal{O} \in \{\text{Rnd}, \text{Act}+, \text{Act}-, \text{Act}\pm\}$, combined with ORG. Note that AUG a priori can not be applied to AdaBoost. Regarding Daimler (Table 4.5), we show analogous experiments but replacing Act$-$ by Act$\sim$, which propagates to Act$\pm$. Figure 4.4 plots the curves corresponding to most interesting results.

Fig. 4.4 shows the AdaBoost based experiments in three different plots: Haar, EOH and Haar+EOH. The plots compare the accuracy of the passive trained classifiers with the domain adapted ones. From these experiments we can draw the following observations:

- Reducing the training data of the target domain to the 10% decreases the performance of any trained detector by more than 10 points of AMR.
- All detectors benefit from adding a 10% of random target domain data to the virtual-world set. This benefit varies from 6 to 10 points of AMR.
- Almost all of the tested *Act* experiments outperform the *Rand* and *target 10%* baselines. Note that the trivial procedure of adding random data performs well

Table 4.6: Domain adaptation results based on AdaBoost over INRIA dataset. For FPPI $\in [0.1, 1]$, AMR (%) is indicated. Bold values remark the best accuracy for each feature. Δ_1 shows the difference between the corresponding $\mathcal{T}_{\mathcal{V}}^{tr}$ and the best V-AYLA result. Δ_2 illustrates differences between the corresponding real trained experiment and the best V-AYLA result.

INRIA	Joint	Oracles				Δ_1	Δ_2
		Act+	Act−	Rnd	Act±		
EOH	ORG	35.00	35.66	34.33	**31.89**	10.89	-1.64
ExtHaar	ORG	38.12	38.54	40.33	**33.48**	13.38	-2.29
ExtHaar+EOH	ORG	23.08	27.73	26.16	**20.38**	17.21	1.16

in other contexts and it is usually difficult to outperform. Our proposed active learning procedures clearly outperform the random ones.

- For Haar and EOH *Act+* and *Act-* perform equally but requiring *Act-* less annotation effort. However, for Haar+EOH *Act+* performs better.

- In all the cases *Act±* outperforms the baselines and the other *Act* experiments, even slightly outperforming the target trained pedestrian detector for the most important case, the Haar+EOH.

4.6 Summary

Virtual worlds can help in learning appearance-based models for pedestrian detection in real-world images. This means that even virtual-world based training can provide excellent performance in some cases, it can also suffer the *dataset shift* problem as real-world based training does. Accordingly, in this chapter we have designed a domain adaptation framework, V-AYLA, in which we have tested different techniques to collect a few pedestrian samples from the target domain (real world) and to combine them (cool world) with the many examples of the source domain (virtual world) in order to train a domain adapted pedestrian classifier that will operate in the target domain. Following such a framework, as in previous chapter, we have tested state-of-the-art pedestrian descriptors (HOG/LBP/HOG+LBP) with Lin-SVM and (ExtHaar/EOH/ExtHaar+EOH) with AdaBoost. Within the same pedestrian detection scheme, we have employed virtual-world based classifiers and real-world based ones (Virtual, INRIA, Daimler, ETH-0,1,2, TUD and Caltech). Following V-AYLA we have performed 330 train-test runs to assess detection performance. This assessment shows how V-AYLA reaches the same performance than training and testing with real-world images of the same domain.

Altogether, we consider V-AYLA as a new framework from which more research can be done for improving pedestrian detection results (including new features, new models as multi-view/part ones, and ways of building the cool world and collecting

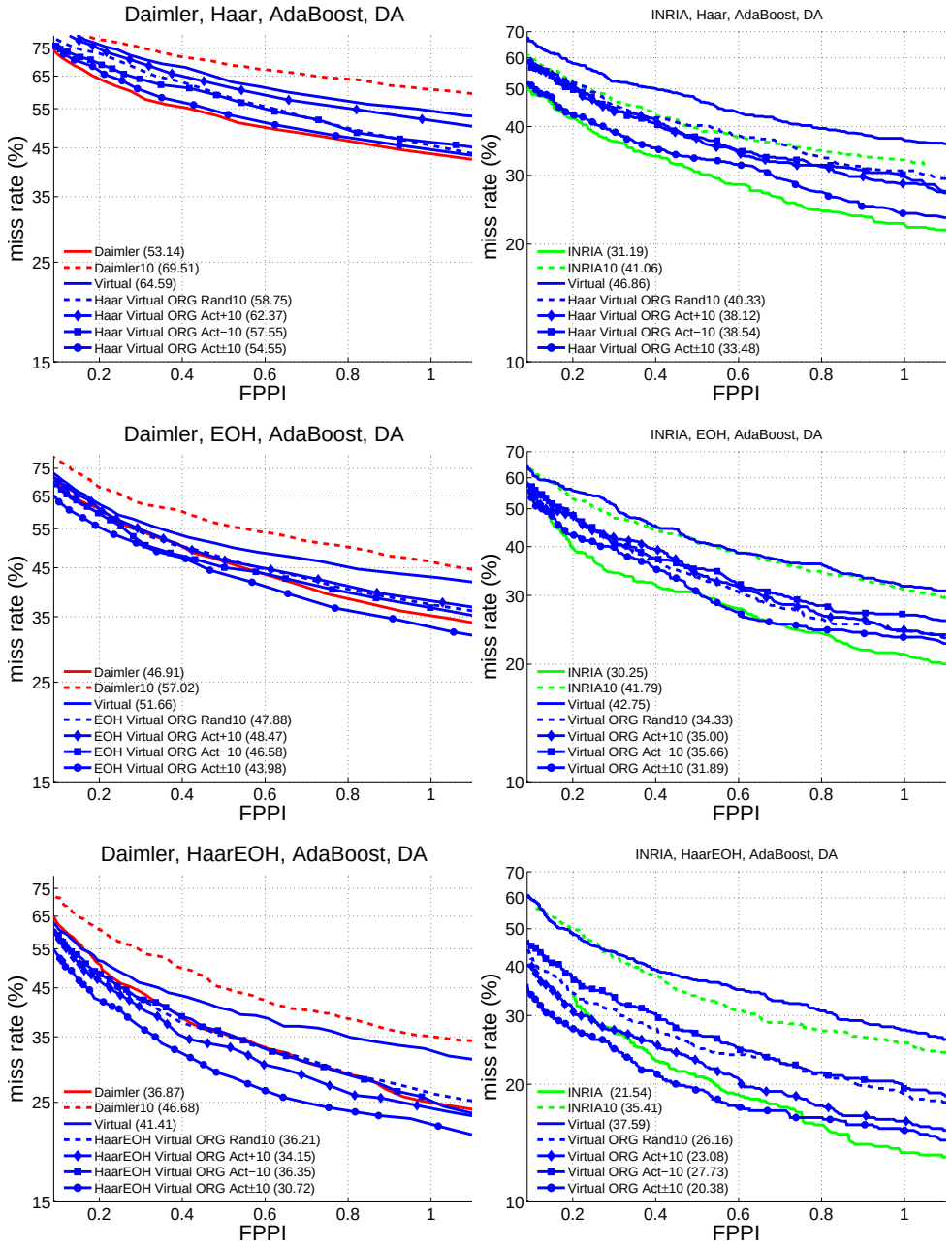

Figure 4.4: Domain adaptation results based on AdaBoost over INRIA and Daimler datasets.

samples), propose unsupervised domain adaptation methods as well as to extend the idea to the detection of other objects.

Chapter 5

Unsupervised Domain Adaptation and Weakly Supervised Annotation

In this chapter we are interested in other alternatives to the supervised domain adaptation. Ideally, we would like to adapt our system without any human intervention. Thus, a first arising question is: ***Can the learnt models automatically adapt to changing situations without human intervention?*** As a proof of concept, we propose the use of unsupervised domain adaptation techniques that avoids human intervention during the adaptation process. We term this system as V-AYLA-U (Fig. 5.1). The last open issue is: ***How can we avoid the dataset shift without performing domain adaptation?*** Accordingly, we assess an strategy to collect samples from the real world and retrain with them, thus avoiding the dataset shift, but in such a way that no BBs of real-world pedestrians have to be provided. Both proposed methods report the same detection accuracy than when training with many human-provided pedestrian annotations and testing with real-world images of the same domain.

5.1 Introduction

Our ideal goal is to remove the humans from the loop and obtain a system which self-learns how to distinguish the objects of interest. Our idea is to remove the oracle of V-AYLA to totally avoid manual labelling. As a first approach, we propose the use of an unsupervised domain adaptation technique. In particular, we explore the use of the transductive SVM (T-SVM) learning algorithm in order to adapt virtual and real worlds for pedestrian detection. We term this system as V-AYLA-U (Fig. 5.1) as it is our unsupervised version of V-AYLA. V-AYLA-U will combine our virtual-world based samples with some real-world based detections to reach the desired performance.

An alternative that we want to explore is the use of the virtual-world data for developing a pedestrian classifier to be used for collecting detections from real-world

images. Then a human oracle validates the detections as right or false. The idea is that at the end of the process we can obtain a large number of real-world pedestrian BBs without manually annotating them, *i.e.* the virtual-world-based pedestrian detector provides BBs for us, while the human oracle just provide easy *yes/no*-feedback to validate such BBs. In this paper we show that accurate BBs can be obtained through this procedure, saving a lot of oracle time. Moreover, the procedure is adaptable to work in crowd-sourcing style but allowing to propose a simpler task less prone to errors.

Section 5.2 presents the V-AYLA-U and its results. Section 5.3 shows the weakly supervised annotation method and results. Finally, section 5.4 draws the main conclusions.

5.2 Unsupervised Domain adaptation

In Sect. 4.2 we applied *domain adaptation* (DA) to virtual and real worlds, where the former is considered as the so-called *source domain* and the latter is the *target domain*. In the DA paradigm it is assumed that we have many labelled samples from the source domain, which in our case are pedestrian and background samples automatically collected (with label) from the virtual world. Regarding the target domain, two main situations are considered: (1) in *supervised DA* (SDA) we collect a small amount of labelled target-domain data; (2) in *unsupervised DA* (UDA) we collect a large amount of target-domain data but without labels. In Sect. 4.2 we proposed a SDA approach based on *active learning*. Obtained results, were totally satisfactory in terms of the accuracy of the SDA-based pedestrian detectors. However, the use of active learning implies that a human oracle assists the training.

Currently we face the last step towards our self-trained pedestrian detector, *i.e.* we propose not to involve humans annotators/oracles during the training process. In particular, as we explain in Sect. 5.2.1, in this chapter we follow an UDA based on transductive SVM (T-SVM) [59]. As we will see in Sect. 5.2.2, for our current image acquisition system, the obtained performance is comparable to the one given by a pedestrian detector based on human assisted training. The conclusions of the presented work are summarized in Sect. 5.2.3

5.2.1 Proposed UDA pedestrian detector

We have started our study using HOG but replacing the Lin-SVM by Lin-T-SVM (SVM ^{light} implementation [58]). If all the provided samples are labelled Lin-T-SVM is equivalent to Lin-SVM.

Now, let us assume the following inputs. First, our *source domain*: $\Im_{\mathcal{V}}^{tr+}$ denotes a set of virtual-world images with automatically labelled pedestrians, and $\Im_{\mathcal{V}}^{tr-}$ refers to a set of pedestrian-free virtual-world images automatically generated as well. Second, our *target domain*: $\Im_{\mathcal{R}}^{tr}$ is a set of real-world images without labels. Third, a threshold,

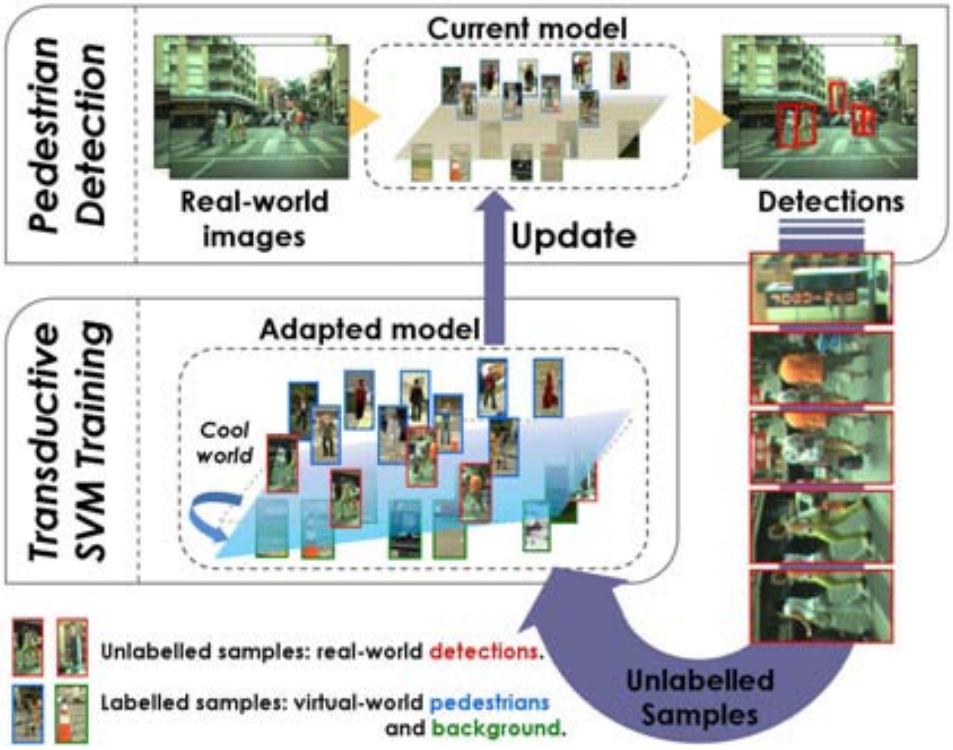

Figure 5.1: Unsupervised domain adaptation general idea. We learn a pedestrian model using automatically labelled virtual-world pedestrians and background. With this model we detect pedestrians in real-world images. Some detections will be true positives and others false ones. Since we do not know which ones are of each type and we do not want human intervention, we treat all them as unlabelled data. Next, such unlabelled samples and the virtual-world labelled ones are joined to train a new pedestrian model using the T-SVM.

Thr, such that an image window is said to contain a pedestrian if its classification score is larger than Thr. Then, the steps of our UDA are:

(S1) *Learning in virtual world* with samples from $\{\Im_{\mathcal{V}}^{tr+}, \Im_{\mathcal{V}}^{tr-}\}$, HOG and Lin-SVM. We term as $\mathrm{C}_{\mathcal{V}}$ the learnt classifier and as $\mathrm{D}_{\mathcal{V}}$ its associated detector. Let $\mathcal{T}_{\mathcal{V}}^{tr+}$ be the set of pedestrians used for obtaining $\mathrm{C}_{\mathcal{V}}$ (*i.e.* coming from $\Im_{\mathcal{V}}^{tr+}$), and $\mathcal{T}_{\mathcal{V}}^{tr-}$ the set of background samples (from $\Im_{\mathcal{V}}^{tr-}$). Samples in $\mathcal{T}_{\mathcal{V}}^{tr+}$ and $\mathcal{T}_{\mathcal{V}}^{tr-}$ are assumed to follow standard training steps of pedestrian classifiers, namely, they are in canonical window (CW) size, $\mathcal{T}_{\mathcal{V}}^{tr+}$ includes mirroring, and $\mathcal{T}_{\mathcal{V}}^{tr-}$ includes bootstrapped hard negatives (previous to bootstrapping, we train the initial classifier with the same number of positive and negative samples). Let C denote the current classifier during our learning procedure, and D its associate detector. Now we provide the initialization

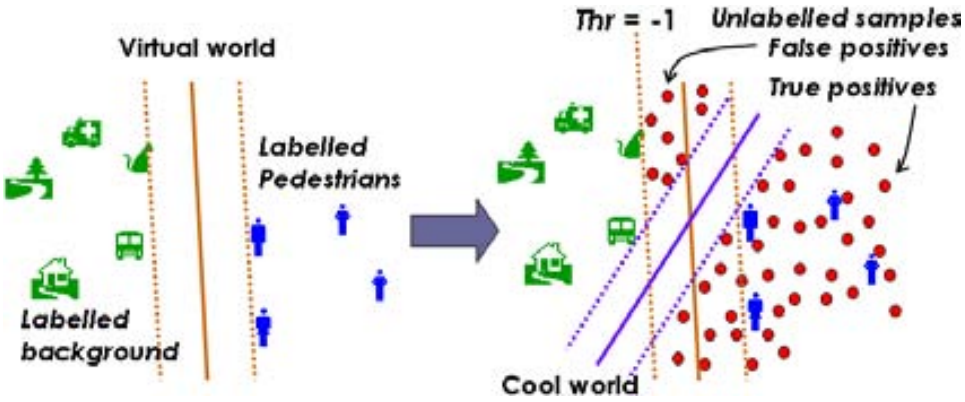

Figure 5.2: T-SVM training.

$C \leftarrow C_{\mathcal{V}}$ (thus, D is $D_{\mathcal{V}}$ at the beginning).

(S2) *Pedestrian detection in real world.* Run D on $\Im_{\mathcal{R}}^{tr}$: only those candidate windows W_c (provided by the pyramidal sliding window) with $C(W_c) > Thr$ are considered for the final non-maximum suppression stage of the detection process. Some of these detections are true positives, while some others are false ones. We do not know, and treat all them as unlabelled samples. Let us term as $\mathcal{T}_{\mathcal{R}}^{tr?}$ the set of such detections and their vertically mirrored counterparts down scaled to CW size.

(S3) *T-SVM learning in cool world* with the virtual-world samples (*i.e.* $\mathcal{T}_{\mathcal{V}}^{tr+}$ and $\mathcal{T}_{\mathcal{V}}^{tr-}$) and the real-world unlabelled ones (*i.e.* $\mathcal{T}_{\mathcal{R}}^{tr?}$). Figure 5.2 illustrates the underlying idea of the T-SVM training. After this new training we obtain the new C and D.

This algorithm can be iterated (with the same or new real-world sequences) by going from **(S3)** to **(S2)**. However, still we must study the best way of combining unlabelled samples from different iterations to avoid an excessive grow up of the number of samples, which could slow down the T-SVM based training. Although, note that only the HOG of samples of the new iteration must be computed since the ones of previous iterations can be stored. Additionally, some stopping criteria should be provided, which could be based on the similarity of consecutive hyperplanes (classifiers).

5.2.2 Experimental results

Experiments are conducted following the experimental setting explained in Sect. 3.2 and we restrict ourselves to our CVC02 dataset explained in Sect. 3.2.1. Figure 5.3 shows the results of the different experiments we have conducted. Curve *CVC-Train-1* refers to the use of a pedestrian classifier trained only with pedestrian and background samples from $\Im_{cvc}^{tr}$. Of course, for this experiment the pedestrians in

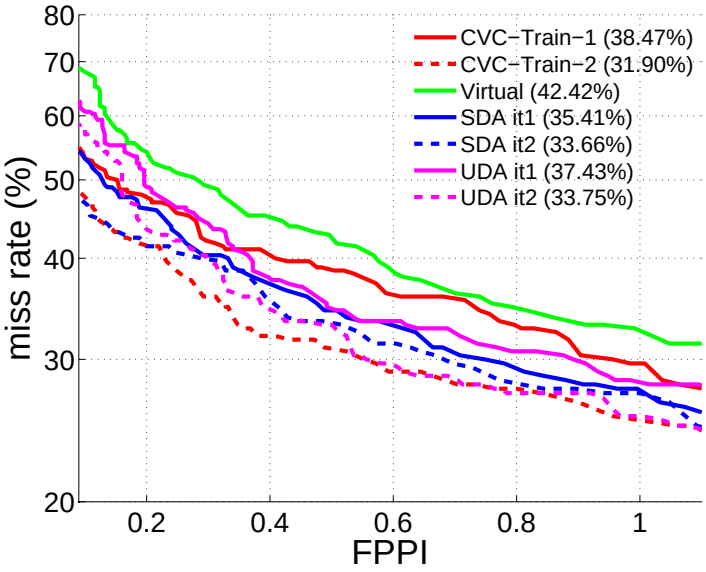

Figure 5.3: Detection results: in parenthesis the AMR of the conducted experiments.

$\Im_{cvc}^{tr}$ where manually labelled. Curve *Virtual* refers to the use of only the virtual-world data for training. Curves *UDA* correspond to the results according to the proposed method, *i.e.* using virtual-world samples and $\Im_{cvc}^{tr}$ (without labels) for building the pedestrian classifier. We iterated the learning process three times, the third was not giving any improvement, so we show just two iterations. From one iteration to the next, we preserved all the detections as unlabelled samples. Curves *SDA* correspond to our approach in Chapt. 4, iterated also until no improvement was achieved, which happens at third iteration too. Note that SDA involves a human oracle during training.

5.2.3 Discussion

From results in Fig. 5.3 we see that *Virtual* performs worse than *CVC-Train-1* (4 AMR points) which we hypothesize is due to dataset shift. When virtual and real world samples are combined, AMR is around 5 points better than *CVC-Train-1* and 9 points regarding *Virtual*, which are large improvements (see typical performance differences among pedestrian detectors in [27]). Iterations pay back in terms of performance improvement (similar to what happens with bootstrapping). Both approaches, SDA and UDA, perform similarly. Thus, we could say that virtual data is complementing well real one. This viewpoint, which applies to both SDA and UDA, corresponds to label real-world samples and use virtual-world ones to complement them. However, here we are interested in the pure UDA viewpoint, *i.e.* starting with a classifier only

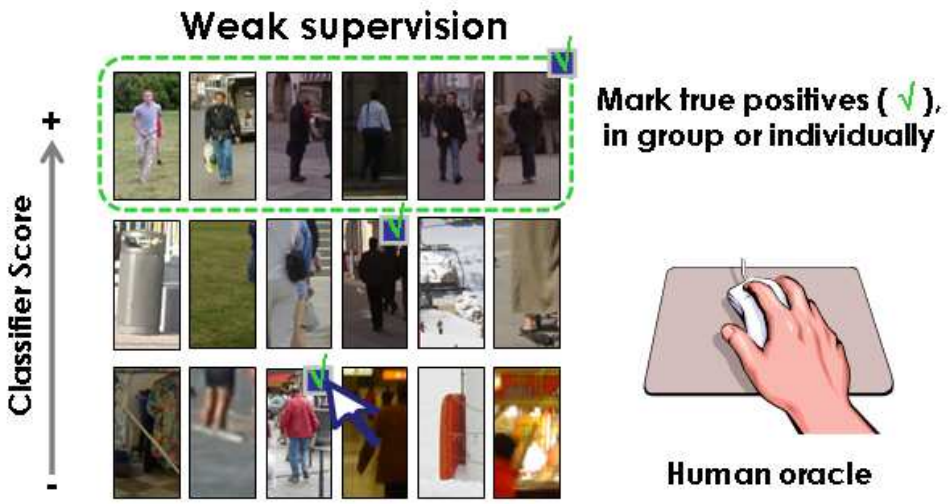

Figure 5.4: Weakly supervised annotation general idea. Detections are presented to the oracle ordered by classifier score and in CW size. The oracle marks right detections individually or in groups indicated by initial and final clicks.

based on virtual-world samples, the proposed T-SVM-based algorithm has been able to adapt it for operating in real-world images without human intervention.

Additionally, to complement the study, we trained a pedestrian classifier fully based on manual annotations. More specifically, we labelled 1016 pedestrians in other images taken with our camera and used 150 pedestrian-free images to collect background samples. Then, we trained the classifier following the same procedure that we used for training the classifier only based on virtual-world samples. The curve is shown as *CVC-Train-2* in Fig. 5.3. Note, that it gives better results than our DA approaches (around 2 AMR points), of course, by manually annotating the double of pedestrians than for SDA. However, in the working range from FPPI=1 to FPPI=0.5, UDA (it2) and *CVC-Train-2* shows analogous performance. Now, we can keep improving UDA by showing more sequences to the learning system, while for improving *CVC-Train-2* more manual annotations would be required and it is necessary to follow some sort of active learning procedure (as we do in SDA) to avoid introducing redundant pedestrians.

5.3 Weakly Supervised Annotation of Pedestrian Bounding Boxes

In this section propose a *weakly supervised* annotation procedure, *i.e.* pedestrian BBs are not manually annotated. We first train a pedestrian classifier using only virtual-world data. Then, such a classifier collects pedestrian examples from real-world images by detection. A human oracle rejects false detections through an efficient procedure (See Fig. 5.4). Thus, at the end of the process we obtain pedestrian examples without requiring manual annotation of BBs. Real-world examples are then used to train the final pedestrian classifier. In fact, this procedure is similar to the oracle Act- but in here we do not retrain with the virtual data and we try to collect as many pedestrian examples as possible.

In order to learn pedestrian classifiers we employ the HOG/LinSVM. Under these settings, we show that our weakly supervised approach provides classifiers analogous to their counterparts trained with examples collected by manually annotating the BBs of all the available pedestrians.

5.3.1 Proposed weakly annotation of BBs

Assume the following definitions of *training sets*:

- *Source domain*. Let $\Im_{\mathcal{V}}^{tr+}$ be the set of available virtual-world images with automatically annotated pedestrians, and $\Im_{\mathcal{V}}^{tr-}$ the set of pedestrian-free virtual-world images automatically generated as well.
- *Target domain*. Let $\Im_{\mathcal{R}}^{tr+}$ be a set of real-world images with non-annotated pedestrians, and $\Im_{\mathcal{R}}^{tr-}$ a set of pedestrian-free real-world images.

Define:

- *Classifier basics*, *i.e.* pedestrian description process ($\mathbf{D}$, *i.e.* features computation) and base learner ($\mathcal{L}$).
- *Detections*, *i.e.* provide a threshold Thr such that an image window is said to contain a pedestrian if its classification score is greater than Thr.

Our weakly supervised training consists of the following steps:

(s1) *Train in virtual world* using $\mathbf{D}$ and $\mathcal{L}$ with samples from $\{\Im_{\mathcal{V}}^{tr+}, \Im_{\mathcal{V}}^{tr-}\}$. Let us term as $\mathrm{C}_{\mathcal{V}}$ the learned classifier and as $\mathrm{D}_{\mathcal{V}}$ its associated detector. Let $\mathcal{T}_{\mathcal{V}}^{tr+}$ be the set of pedestrian examples used for obtaining $\mathrm{C}_{\mathcal{V}}$ (*i.e.* coming from $\Im_{\mathcal{V}}^{tr+}$), and $\mathcal{T}_{\mathcal{V}}^{tr-}$ the set of background examples (*i.e.* coming from $\Im_{\mathcal{V}}^{tr-}$). Examples in $\mathcal{T}_{\mathcal{V}}^{tr+}$ and $\mathcal{T}_{\mathcal{V}}^{tr-}$ are assumed to follow standard steps in the training of pedestrian classifiers, namely, they are in canonical window (CW) size, $\mathcal{T}_{\mathcal{V}}^{tr+}$ includes mirroring, and $\mathcal{T}_{\mathcal{V}}^{tr-}$ includes bootstrapped hard negatives (previous to bootstrapping, the initial classifier

is trained with the same number of positive and negative samples). Let C denote the current classifier during our learning procedure, and D its associate detector. Now we provide the initialization $C \leftarrow C_{\mathcal{V}}$ (thus, D is $D_{\mathcal{V}}$ at the start).

(s2) *Weakly annotating real world*. Run D on $\Im_{\mathcal{R}}^{tr+}$. Show the detections to the human oracle ($\mathcal{O}$) ordered by C score, and let $\mathcal{O}$ to mark the true detections in groups or individually (Fig. 5.4), *i.e.* like when selecting visual items with a graphical interface of many different modern software applications. Equivalently, we could mark false detections, however, usually true detections are quite far less than false ones. We term as $\mathcal{T}_{\mathcal{R}}^{tr+}$ the set of such new pedestrian examples in CW size and augmented by mirroring. Note that we do not annotate BBs here. This means also that miss detections are not provided by $\mathcal{O}$. In order to collect hard false negatives we can just take the false detections in $\Im_{\mathcal{R}}^{tr+}$ (the detections not marked by $\mathcal{O}$). However, for an easier comparison of our proposal with the standard learning methods used in pedestrian detection, we run D on $\Im_{\mathcal{R}}^{tr-}$ in order to collect real-world negative samples. Let us term such set of samples as $\mathcal{T}_{\mathcal{R}}^{tr-}$. Moreover, by doing so it is not necessary to mark all true positives, since not marked detections are not assumed to be false positives.

(s3) *Retrain in real world*. Train a new classifier C with the pedestrian examples collected as validated detections, using **D** and $\mathcal{L}$. The new pedestrian detector D is now based on the new C.

During step **s2**, D is applied for all images in $\Im_{\mathcal{R}}^{tr+}$ and $\Im_{\mathcal{R}}^{tr-}$, then, step **s3** is applied once. During **s2** we take one negative example per each positive one (same cardinality of $\mathcal{T}_{\mathcal{R}}^{tr+}$, and $\mathcal{T}_{\mathcal{R}}^{tr-}$) and leave for step **s3** collecting more hard negatives by training with bootstrapping using the $\Im_{\mathcal{R}}^{tr-}$ pool.

5.3.2 Experimental settings

We follow the experimental settings proposed on Chapt. 3 but restrict ourselves to HOG/LinSVM to learn our pedestrian classifiers. For evaluating our weakly supervised annotation proposal, as training real-world dataset we use the INRIA training set. It is worth to note that the BB annotations of the INRIA training and testing sets are considered as precise [104]. We discard the Daimler dataset [30] because it does not have the training frames that are necessary for this work. As in previous chapters, we use the virtual-world dataset from which training the corresponding classifier. For testing, in addition to INRIA testing set, we use: Caltech-Testing (Reasonable set) [27], ETH-0,1,2 and TUD-Brussels [117]. So in total, six testing sets.

In order to perform fair performance comparison among pedestrian classifiers, for any training we need to rely on the same imaged pedestrians. Thus, we only consider those detections whose BB that actually overlap (PASCAL VOC criterion [33]) with some corresponding INRIA training ground truth BB (manually annotated). Thus, paired results termed as 'Det' and 'GT' correspond to pedestrian detectors whose

Figure 5.5: Alignment comparison between detections and annotations. Ground truth (top row) and detections (bottom row). Left block of five pedestrians contains detections with classifier score in $[-1, 0)$, those in mid block are in $[0, 1)$, and those in left block correspond to values ≥ 1. In our current settings, left and mid blocks are discarded and only the detections of the right block would arrive to the human oracle for validation (along with some hard negatives).

classifiers have been trained with the same pedestrians (INRIA training set), but in one case the BBs of the pedestrians are given by the validated detections ('Det') while in the other such BBs are given by the human oracle ('GT'). Fig. 5.5 compares de 'Det' and 'GT'.

For the experiments presented in Sect. 5.3.3, we also simulate the interaction of the human oracle. In particular, instead of having a person marking the true positives, these are automatically indicated to our system thanks to the training ground truth. This allows to boost the testing of different alternatives at the current stage of our research. However, in Sect. 5.3.3 we evaluate the annotation cost of our proposal by performing some experiments with an actual human oracle in the loop.

Note that we decide if an image window is a detection or not according to the classification score and threshold Thr (Sect. 5.3.1). Here we have set $Thr = -1.0$, *i.e.* the oracle gives *yes/no*-feedback for windows with classification score ≥ -1.0. Note that for SVM classifiers this is in the ambiguity region. Thus, in practice most of the windows presented to the human oracle for *yes/no*-feedback will be pedestrians, but some of them will be hard negatives.

5.3.3 Experimental results

Figure 5.5 provides visual insight about BB localization accuracy for the detections of the virtual-world-based pedestrian detector applied to the INRIA training set. Figure 5.6 plots the results comparing the performance of the pedestrian detectors resulting from manually annotated BBs *vs*the BBs resulting from our method (*i.e.* using validated detections) for the same imaged pedestrians. For sake of completeness, the results of training with both the full INRIA training set and the virtual-world one are plotted as well. Our validated detections reach almost the 80% of the INRIA training

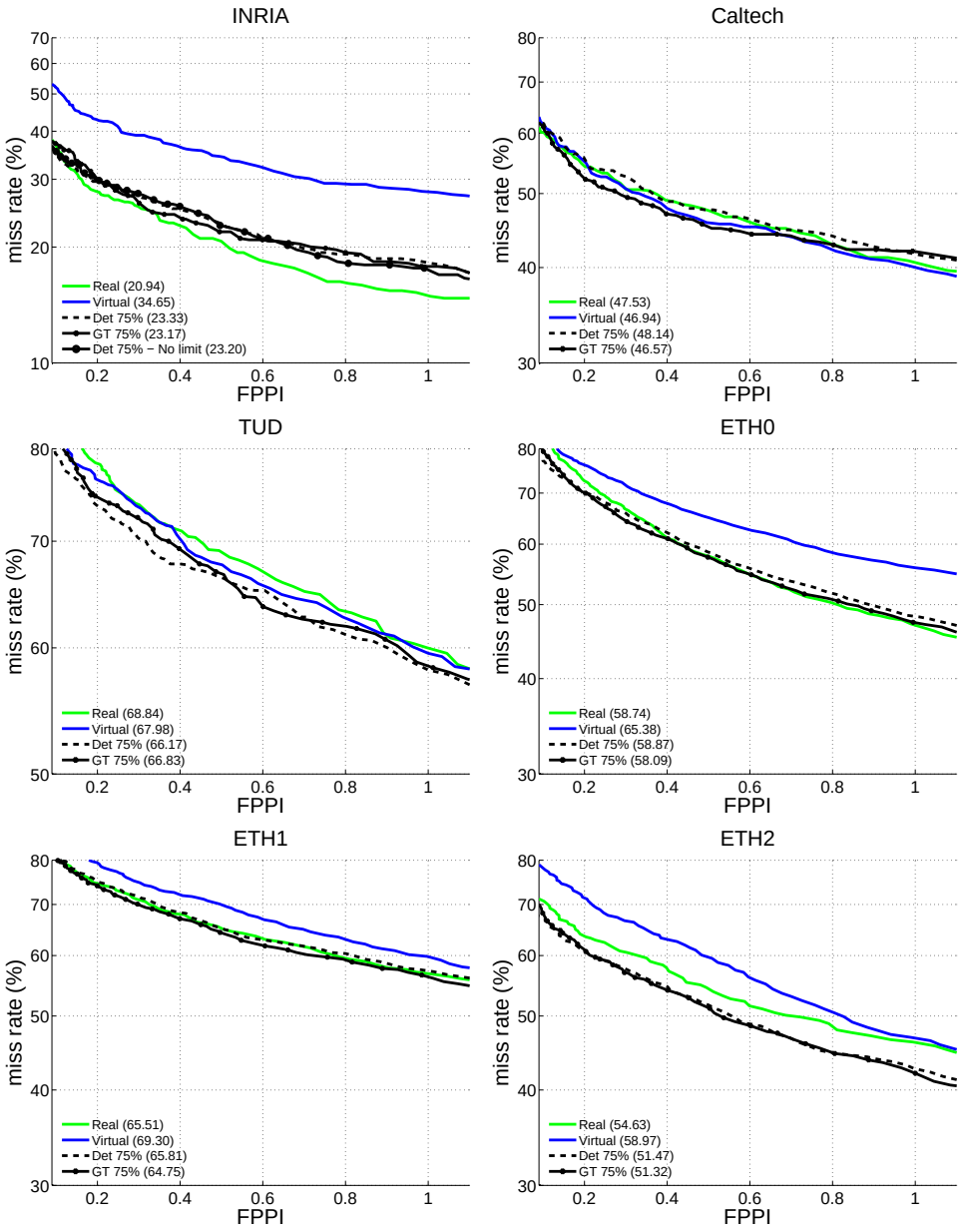

Figure 5.6: Weakly supervision annotation results. Results of detections ('Det' 75%) and corresponding manually annotated BBs ('GT' 75%) from INRIA training set, for different testing sets. 'Real': training with the full INRIA. 'Virtual': training with the virtual-world data.

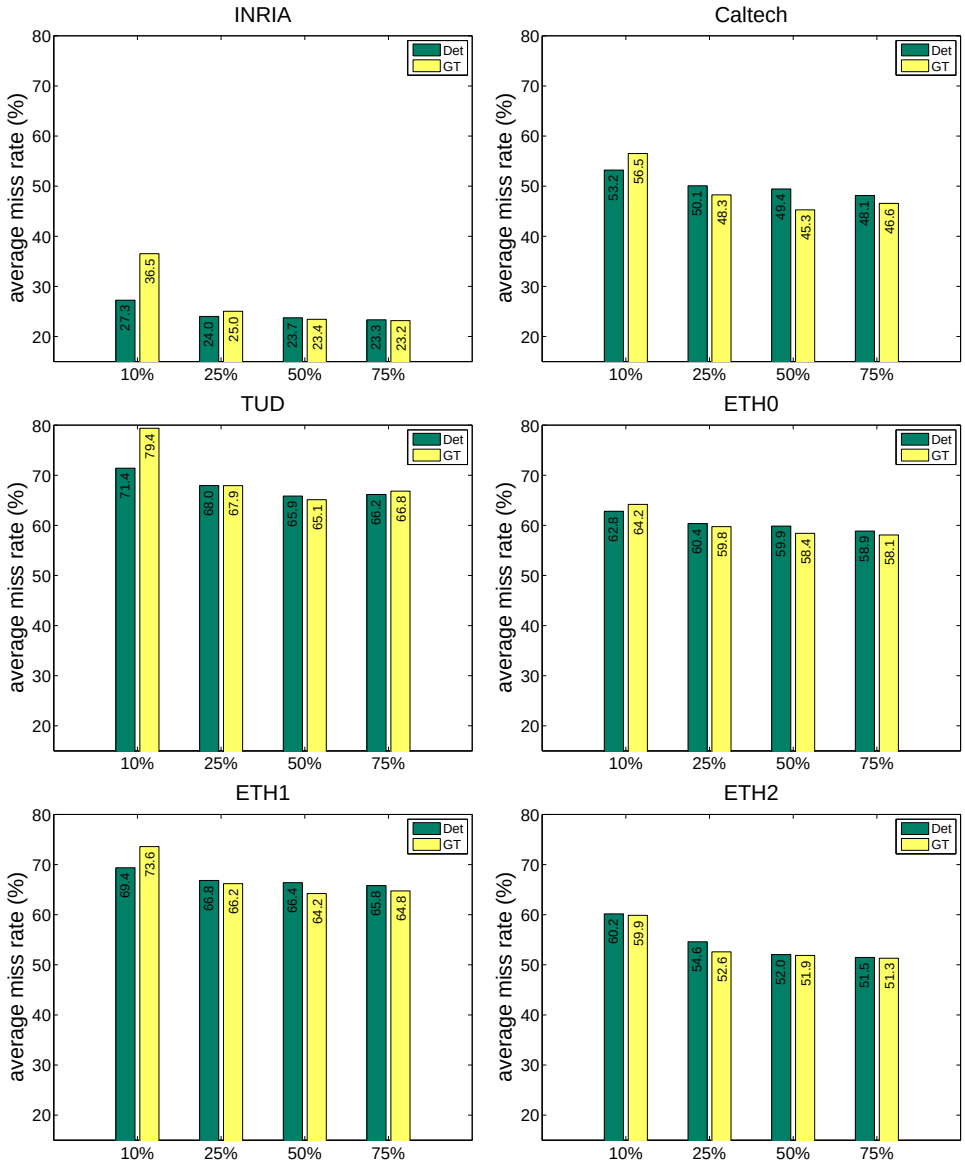

Figure 5.7: Average miss rate at different testing sets when training with different amounts of validated detections ('Det') and corresponding manually annotated groundtruth ('GT') from INRIA training set. Each bar shows its average miss rate (%).

pedestrians, so we decided to set 75% as the limit of our method for such a training set. Figure 5.7 plots the average miss rate of the 'Det' and 'GT' pedestrian detectors

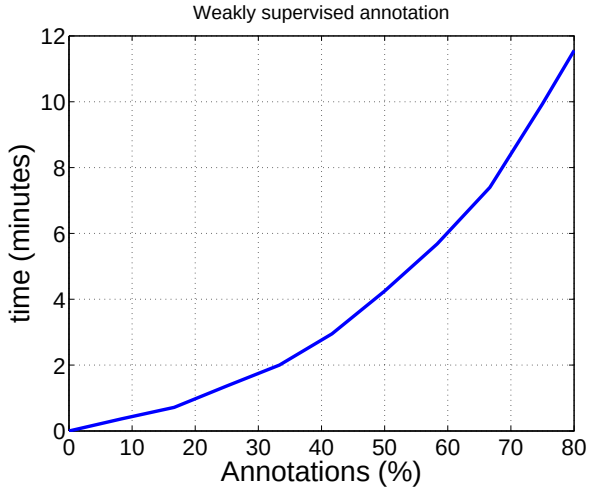

Figure 5.8: Annotation effort with our weakly supervised method.

according to different amounts of training data used, being 75% the maximum.

5.3.4 Discussion

From these results we can draw two main conclusions. On the one hand, the 'Det' and 'GT' performances are so close that we think that BBs from validated detections are as accurate as precise pedestrian BB annotations for developing good classifiers. The difference would be even more negligible by using the HOG/Latent-SVM method for learning deformable part-based models [34], since it is able to refine the annotated BBs provided they are sufficiently precise at the initial stage. On the other hand, the 75% of the annotations seems to already convey the same information than the 100% since the two case give rise to a very similar performance.

In order to quantify the human annotation effort of our weakly supervised method, *i.e.* in comparison with the human annotation of BBs, we provide Fig. 5.8. Note that annotation time is reduced drastically for the human oracle.

For instance, around only 10 minutes are required to annotate the 75% of the pedestrians (906) since no BBs must be provided. We experimented manual annotation of pedestrian accurate BBs and found an average required time of 6 seconds per BB. Thus, annotating the BBs of the 75% would require 90 minutes (9 times more).

V-AYLA-U is a great improvement because it removes the human from the adaptation process allowing the system to self adapt to new scenarios. However, T-SVM training is very slow and moreover it requires as parameter an estimation of the ratio positive/negative samples that it is difficult to adjust. So further research is needed in the area of unsupervised domain adaptation.

5.4 Summary

In this chapter, we presented V-AYLA-U, our first attempt to extended V-AYLA addressing the challenge of developing a method for obtaining self-trained pedestrian detectors, *i.e.* without human intervention for labelling samples. For that we have proposed an unsupervised domain adaptation procedure based on T-SVM. In the source domain we have automatically labelled samples coming from a virtual-world, while the target domain correspond to real-world images from which the proposed algorithm selects (by detection) unlabelled samples that correspond to pedestrians and background. Obtained results are comparable to traditional learning procedures where the pedestrians samples are collected by tiresome manual annotation. However, training a T-SVM is slowly and also requires a parameter to be adjusted.

Moreover, we have presented a method for training pedestrian classifiers without manually annotating their required full-body BBs. The two core ingredients are the use of virtual world data, and the design of a weakly supervised procedure to validate detections by *yes/no* human feedback. Presented results indicate that the obtained classifiers are on pair with the ones based on manual annotation of BBs. Besides, the human intervention is highly reduced in terms of both time and difficulty of the annotation task. Our method can be applied to other objects.

Chapter 6

Conclusions

In this Thesis we presented a pedestrian detection system trained in a virtual world and adapted to operate in the real world. This chapter summarizes the main achievements of our work by revisiting the contributions, strengths and weakness of the proposed methods. Then, we give a perspective of the research line opened with this work. Finally, a brief overview of the future research possibilities in the virtual world generation and domain adaptation techniques are also discussed.

6.1 Summary and contributions

In this Thesis we explored the synergies between modern Computer Animation and Computer Vision in order to *close the circle*: the Computer Animation community has modelled the real world by building increasingly realistic virtual worlds, especially in the field of video games, thus, we learnt our models of interest in such virtual worlds and use them back in real world.

Chapt. 1 introduces the concepts of *virtual world* and *domain adaptation*. Here we specified the Thesis scope, the challenge we aimed to achieve and the questions we wanted to answer.

In Chapt. 2 we reviewed some works that indeed are related to our work, including: pedestrian detection, collecting annotations, engineering examples, and performing domain adaptation.

Chapt. 3 tried to bring light to the following question: ***Can a pedestrian appearance model learnt with virtual-world data work successfully for pedestrian detection in real-world scenarios?***. To answer this question we compared the accuracy of our pedestrian detectors trained with virtual world data with the ones trained with real one. The comparison revealed that, although virtual samples were not specially selected, in some cases both virtual- and real-world based training gave rise to classifiers of similar accuracy while there was a gap for others. We re-

alized that this performance drop happens even when training and testing by using real-world data of different datasets. To the best of our knowledge there is neither previous proposals for pedestrian detection in particular, nor for object detection in general, where annotations coming from a photo-realistic virtual world are used to learn an appearance classifier that must operate in the real-world. This is the first main contribution of this Thesis. The performance drop that happens when training and testing data from a different nature is known as *dataset shift*.

In Chapt. 4 we tried to answer the following question: ***Can we adapt the models learnt in the virtual scenarios to the particularities of the real ones?*** To answer this question we have designed a *domain adaptation* framework, V-AYLA, in which we have tested different techniques to collect a few pedestrian samples from the target domain (real world) and combined them with the many examples of the source domain (virtual world). This in order to train a domain adapted pedestrian classifier that operats in the target domain. V-AYLA reported the same detection accuracy to the one that it is obtained by training with many human-provided pedestrian annotations and testing with real-world images of the same domain. This gives a positive answer to the posed question. To the best of our knowledge, this is the first work that demostrates adaptation of virtual and real worlds for developing an appearance-based object detector. This is the second main contribution to this Thesis.

Chapt. 5 further explored other open issues. Ideally, we would like to adapt our system without any human intervention. Accordingly, we asked ourselves ***Can the learnt models automatically adapt to changing situations without human intervention?***. To answer this question we doted V-AYLA of an unsupervised domain adaptation technique that avoids human intervention during the adaptation process. We term this system as V-AYLA-U. Our preliminary results serveds as a proof of concept that gives a positive answer to the question. The last open issue is ***How can we avoid the dataset shift without performing domain adaptation?*** Our approach consisted on collecting samples by detection with our virtual world classifier from the real world and retrain with them, thus avoiding the dataset shift. After, a human oracle rejected the false detections by an efficiently weak annotation. Finally, a new classifier was trained with the accepted detections. We showed that this classifier is competitive with respect to the counterpart trained with samples collected by manually annotating hundreds of pedestrian bounding boxes.

6.2 Future Perspective

Recently several automobilistic companies have developed their own pedestrian detection systems and started to on-board them in commercial vehicles. Although the technology for developing such systems already exists, the system should work under hight performance requirements in any circustances. To contemplate all these possible scenarios the system has to be trained with an enormous amount of annotated data. Usually, a vehicle fleet is sent to record images during the whole year all around the

world and this data has to be annotated to train and validate the system. Despite acquiring and analysing this data is very expensive and time consuming it presents a major challenge: what happen if the system has to work in a different scenario? *e.g.* changes in the acquisition system, environment, pedestrian cloth style. Then dataset shift problem may appear and domain adaptation techniques can play an important roll to overcome it. Moreover, it arises another question: Could we develop a system that after an initial training it automatically self adapts to new scenarios?

This Thesis is a proof of concept that such systems can self adapt to new scenarios but still further research is required. In the following we point some aspects that has to be studied more in depth.

Improved virtual worlds: Video game industry has been blooming since last decade and is one of the economic activities that has not realized about the economic crisis. The realism of the video games in terms of graphics and artificial intelligence is reaching to be completely or almost indistinguishable from reality. However, in this Thesis we used the Half life 2 video game that was created in 2004. So, we expect that using more realistic video games we could achieve better results. In addition, from these games we could acquire ground truth for different applications such as pixel level segmentations for semantic segmentation and scene understanding algorithms, depth information for stereo algorithms, motion for tracking and optical flow algorithms, etc.

State of the art pedestrian detectors: In this thesis we focused on holistic pedestrian detectors that are the basis of the state-of-the-art pedestrian detectors such as part-based or patch-based ones. We are already working on extending this work to such detectors as can be seen in this Thesis publications.

Domain Adaptation: Domain adaptation is a fundamental problem in machine learning but it only started receiving attention in computer vision applications recently. Accordingly, it is a great research opportunity, specially the unsupervised methods.

Our vision of future ADAS: We expect that self-trained and self-adapted systems will attract much attention in the next years. These systems will automatically learn new concepts and adapt to new environments cutting down the tedious work of human annotation.

El trabajo que puede hacer una máquina es inhumano que lo haga una persona

The work that can be done by a machine is inhuman to be done by a person

Joana S.

Appendix A

Notation

Symbol	Stands for ...
$\mathcal{D}, \mathcal{I}, \mathcal{V}$	Daimler, INRIA, Virtual-world domains, resp.
CW	Cannonical Window
$\mathcal{X}$	Variable used for denoting a virtual- or real-world domain, in particular, $\mathcal{X} \in \{\mathcal{D}, \mathcal{I}, \mathcal{V}\}$.
$\mathcal{R}$	Variable used for denoting a real-world domain, in particular, $\mathcal{R} \in \{\mathcal{D}, \mathcal{I}\}$.
tr, tt	Used as upper indices qualify training and testing sets (of either samples or images), resp.
$+, -$	Used as upper indices qualify pedestrian and background sets (of either samples or images), resp.
$\mathfrak{S}_{\mathcal{X}}^{tr+}$	Training set of images from $\mathcal{X}$, with annotated pedestrians.
$\mathfrak{S}_{\mathcal{X}}^{tr-}$	Training set of pedestrian-free images from $\mathcal{X}$.
$\mathfrak{S}_{\mathcal{R}}^{tr+}$, $\mathfrak{S}_{\mathcal{R}}^{tr-}$	Analogous to $\mathfrak{S}_{\mathcal{X}}^{tr+}$ and $\mathfrak{S}_{\mathcal{X}}^{tr-}$, but restricted to real-world ($\mathcal{R}$) domains.
$\mathcal{T}_{\mathcal{X}}^{tr+}$	Training set of pedestrian cropped windows from $\mathcal{X}$.
$\mathcal{T}_{\mathcal{X}}^{tr-}$	Training set of backg. cropped windows from $\mathcal{X}$.
$\mathcal{T}_{\mathcal{R}}^{tr+}$, $\mathcal{T}_{\mathcal{R}}^{tr-}$	Analogous to $\mathcal{T}_{\mathcal{X}}^{tr+}$ and $\mathcal{T}_{\mathcal{X}}^{tr-}$, but restricted to real-world ($\mathcal{R}$) domains.
$\mathcal{T}_{\mathcal{X}}^{tr}$	Pair $\{\mathcal{T}_{\mathcal{X}}^{tr+}, \mathcal{T}_{\mathcal{X}}^{tr-}\}$.
$\mathcal{T}_{\mathcal{R}}^{tr}$	Pair $\{\mathcal{T}_{\mathcal{R}}^{tr+}, \mathcal{T}_{\mathcal{R}}^{tr-}\}$.
$\mathcal{T}_{\mathcal{R}}^{tt}$	Set of testing images, *i.e.*, with annotated pedestrians (groundtruth) from $\mathcal{R}$.
$\mathrm{C}_{\mathcal{V}}$	Pedestrian classifier (passively) trained with only virtual-world data.
$\mathrm{D}_{\mathcal{V}}$	Pedestrian detector based on $\mathrm{C}_{\mathcal{V}}$.
Rnd	Human oracle that annotates pedestrian bounding

	boxes randomly.
Act+	Human oracle that annotates pedestrian BBs (false negatives from target domain training set).
Act−	Automatic oracle that annotates pedestrian BBs by detection (positives from target domain training set, according to the detection threshold).
Act∼	Substitutes Act− when the original real-world images are not available for pedestrian detection, but cropped windows are available for classification. While Act− can be used in practice, Act∼ is just an approximation used here to work with the publicly available training data of $\mathcal{D}$.
Act±	Combination of Act+ and Act− (Act∼ for Daimler).

List of Publications

This dissertation has led to the following communications:

Journal Papers

- David Vázquez, Javier Marín, Antonio M. López, David Gerónimo, & Daniel Ponsa. (2013). Virtual and Real World Adaptation for Pedestrian Detection (Under minor review). *IEEE Trans. on Pattern Analysis and Machine Intelligence*.

- Javier Marín, David Vázquez, Antonio M. López, Jaume Amores, & Ludmila I. Kuncheva. (2013). Occlusion handling via random subspace classifiers for human detection. *IEEE Trans. on Systems, Man, and Cybernetics (Part B)*.

Book Chapters

- David Vázquez, Antonio M. López, Daniel Ponsa, & David Gerónimo. (2012). Interactive semi-supervised training of human detection. In *Multimodal Interaction in Image and Video Applications*. (Vol. 48, pp. 169–182). Berlin Heidelberg: Springer.

- Javier Marín, David Gerónimo, David Vázquez, & Antonio M. López. (2012). Pedestrian Detection: Exploring Virtual Worlds. In *Handbook of Pattern Recognition: Methods and Application*. iConcept Press.

Conference Contributions

- David Vázquez, Jiaolong Xu, Sebastian Ramos, Antonio M. López, & Daniel Ponsa. (2013). Weakly Supervised Automatic Annotation of Pedestrian Bound-

ing Boxes. In *IEEE Conf. on Computer Vision and Pattern Recognition. Ground Truth Workshop: What is a good dataset?*. Portland, Oregon.

- Jiaolong Xu, David Vázquez, Sebastian Ramos, Antonio M. López, & Daniel Ponsa. (2013). Adapting a Pedestrian Detector by Boosting LDA Exemplar Classifiers. In *IEEE Conf. on Computer Vision and Pattern Recognition. Ground Truth Workshop: What is a good dataset?*. Portland, Oregon.

- Jiaolong Xu, David Vázquez, Antonio M. López, Javier Marín, & Daniel Ponsa. (2013). Learning a Multiview Part-based Model in Virtual World for Pedestrian Detection. In *IEEE Intelligent Vehicles Symposium*. Gold Coast, Australia.

- David Vázquez, Antonio M. López, & Daniel Ponsa. (2012). Unsupervised Domain Adaptation of Virtual and Real Worlds for Pedestrian Detection. In *Int. Conf. in Pattern Recognition*. Tsukuba Science City, Japan.

- Yainuvis Socarrás, David Vázquez, Antonio M. López, David Gerónimo, & Theo Gevers. (2012). Improving HOG with Image Segmentation: Application to Human Detection. In *Advanced Concepts for Intelligent Vision Systems*. Brno, Czech Republic.

- David Vázquez, Antonio M. López, Daniel Ponsa, & Javier Marín. (2011). Cool world: domain adaptation of virtual and real worlds for human detection using active learning. In *NIPS Domain Adaptation Workshop: Theory and Application*. Granada, Spain.

- David Vázquez, Antonio M. López, Daniel Ponsa, & Javier Marín. (2011). Virtual Worlds and Active Learning for Human Detection. In *ACM International Conference on Multimodal Interaction* (pp. 393–400). Alicante, Spain.

- Muhammad Anwer Rao, David Vázquez, & Antonio M. López. (2011). Color Contribution to Part-Based Person Detection in Different Types of Scenarios. In W. Kropatsch A. Berciano H. Molina D. D. P. Real (Ed.), *Int. Conf. on Computer Analysis of Images and Patterns* (Vol. 6855, pp. 463–470). Berlin Heidelberg: Springer.

- Muhammad Anwer Rao, David Vázquez, & Antonio M. López. (2011). Opponent Colors for Human Detection. In J. Vitria, J.M. Sanches, & M. Hernandez (Eds.), *Iberian Conf. on Pattern Recognition and Image Analysis* (Vol. 6669, pp. 363–370). Lecture Notes on Computer Science. Berlin Heidelberg: Springer.

- Javier Marín, David Vázquez, David Gerónimo, & Antonio M. López. (2010). Learning Appearance in Virtual Scenarios for Pedestrian Detection. In *IEEE Conf. on Computer Vision and Pattern Recognition* (pp. 137–144).

Media

- Antonio M. López, & David Vázquez. (2013). *Report at Connexió Barcelona program about Eco-Driver project* In Barcelona Televisió - BTV.

- Antonio M. López, & David Vázquez. (2012). *News report about Eco-Driver project* In Euskal Telebista - ETB.

- Antonio M. López, & David Vázquez. (2012). *News report about Eco-Driver project: Technology can help avoiding accidents produced by human errors* In Televisió de Catalunya - TV3.

Exhibitions

- Antonio M. López, Jiaolong Xu, David Vázquez, Sebastián Ramos, & Germán Ros. (2013). *Barcelona Party [Science + Technology]*

- Antonio M. López, David Vázquez, & Javier Marín. (2013). *MIPRCV Industry day*

- Antonio M. López, & David Vázquez. (2012). *Mobility and transport in the Smart Cities*

- Antonio M. López, & David Vázquez. (2012). *Expoelectric Fórmula-E*

Awards

- Google award for the article of *Cool world: domain adaptation of virtual and real worlds for human detection using active learning* in *NIPS Domain Adaptation Workshop: Theory and Application* (2011)

- ICMI Doctoral Consortim award for the article of *Virtual Worlds and Active Learning for Human Detection* in *ACM International Conference on Multimodal Interaction* (2011)

Bibliography

[1] Y. Abramson and Y. Freund. SEmi-automatic VIsuaL LEarning (SEVILLE): a tutorial on active learning for visual object recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005.

[2] A. Agarwal and B. Triggs. A local basis representation for estimating human pose from cluttered images. In *Asian Conf. on Computer Vision*, Hyderabad, India, 2006.

[3] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[4] A.M. López, J. Hilgenstock, A. Busse, R. Baldrich, F. Lumbreras, and J. Serrat. Nighttime vehicle detection for intelligent headlight control. In *Advanced Concepts for Intelligent Vision Systems*, Juan-les-Pins, France, 2008.

[5] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. LabelMe: a database and web-based tootl for image annotation. *Int. Journal on Computer Vision*, 77(1-3):157–173, 2008.

[6] O. Beijbom. Domain adaptation for computer vision applications. Technical report, University of California, San Diego, 2012.

[7] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J.W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2009.

[8] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, 2006.

[9] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012.

[10] A. Bergamo and L. Torresani. Exploting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, 2010.

[11] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems*, Vancouver, Canada, 2008.

[12] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Int. Conf. on Empirical Methods in Natural Language Processing*, Sydney, Australia, 2006.

[13] A. Broggi, A. Fascioli, P. Grisleri, T. Graf, and M. Meinecke. Model–based validation approaches and matching techniques for automotive vision based pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005.

[14] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[15] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT press, 2006.

[16] C. Chelba and A. Acero. Adaptation of maximum entropy capitalizer: Little data can help a lot. In *Int. Conf. on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004.

[17] Y.-T. Chen and C.-S. Chen. Fast human detection using a novel boosted cascading structure with meta stages. *IEEE Trans. on Image Processing*, 17(8):1452–1464, 2008.

[18] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[19] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[20] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[21] N. Dalal. *Finding People in Images and Videos*. PhD thesis, Institut National Polytechnique de Grenoble, 2006. Advisors: Cordelia Schmid and William J. Triggs.

[22] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005.

[23] D.M. Gavrila. A bayesian exemplar-based approach to hierarchical shape matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(8):1408–1421, 2007.

[24] P. Dollár, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *British Machine Vision Conference*, Aberystwyth, UK, 2010.

[25] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *British Machine Vision Conference*, London, UK, 2009.

[26] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: a benchmark. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009.

[27] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: an evaluation of the state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012.

[28] L. Duan, I.W. Tsang, D. Xu, and S.J. Maybank. Domain transfer svm for video concept detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009.

[29] I. Endres, A. Farhadi, and D. Hoeiemand D.A. Forsyth. The benefits and challenges of collecting richer annotations. In *Advancing Computer Vision with Humans in the Loop, CVPR Workshop*, San Francisco, CA, USA, 2010.

[30] M. Enzweiler and D.M. Gavrila. Monocular pedestrian detection: survey and experiments. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(12):2179–2195, 2009.

[31] M. Enzweiler and D.M. Gavrila. A multi-level mixture-of-experts framework for pedestrian classification. *IEEE Trans. on Image Processing*, 20(10):2967–2979, 2011.

[32] M. Enzweiler and D.M. Gavrila. A mixed generative-discriminative framework for pedestrian classification. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.

[33] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *Int. Journal on Computer Vision*, 88(2):303–338, 2010.

[34] P. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[35] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.

[36] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[37] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal on Computer Vision*, 60(2):91–110, 2004.

[38] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009.

[39] T. Gandhi and M.M. Trivedi. Pedestrian protection systems: issues, survey, and challenges. *IEEE Trans. on Intelligent Transportation Systems*, 8(3):413–430, 2007.

[40] S. García and F. Herrera. An extension on 'statistical comparisons of classifiers over multiple data sets' for all pairwise comparisons. *Journal of Machine Learning Research*, 9:2677–2694, 2008.

[41] D. M. Gavrila, J. Giebel, and S. Munder. Vision-based pedestrian detection: The protector system. In *In IEEE Intelligent Vehicles Symposium*, Lisboa, Portugal, 2004.

[42] D. Gerónimo, A.D. Sappa, A.M. López, and D. Ponsa. Pedestrian detection using adaboost learning of features and vehicle pitch estimation. In *IASTED Int. Conference on Visualization, Imaging and Image Processing*, Palma de Mallorca, Spain, 2006.

[43] D. Gerónimo, A.D. Sappa, A.M. López, and D. Ponsa. Adaptive image sampling and windows classification for on-board pedestrian detection. In *Int. Conf. on Computer Vision Systems*, Bielefeld, Germany, 2007.

[44] D. Gerónimo, A.D. Sappa, D. Ponsa, and A.M. López. 2D-3D based on-board pedestrian detection system. *Computer Vision and Image Understanding*, 114(5):1239–1258, 2010.

[45] D. Gerónimo, A.M. López, A.D. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(7):1239–1258, 2010.

[46] R. Gopalan, R. Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Int. Conf. on Computer Vision*, Barcelona, Spain, 2011.

[47] G.R. Taylor, A.J. Chosak, and P.C. Brewer. OVVV: Using virtual worlds to design and evaluate surveillance systems. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, 2007.

[48] B. Gunyel, R. Benenson, R. Timofte, and L. Van Gool. Stixels motion estimation without optical flow computation. In *European Conf. on Computer Vision*, Florence, Italy, 2012.

[49] H.B. Mann and D.R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[50] J. Heckman. Sample selection bias as a specification error. In *Econometrica*, 1979.

[51] B. Heisele, G. Kim, and A.J. Meyer. Object recognition with 3D models. In *British Machine Vision Conference*, London, UK, 2009.

[52] M. Hussein, F. Porikli, and L. Davis. A comprehensive evaluation framework and a comparative study for human detectors. *IEEE Trans. on Intelligent Transportation Systems*, 10(3):417–427, 2009.

[53] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt. MovieReshape: tracking and reshaping of humans in videos. In *ACM SIGGRAPH Conf. and Exhib. on Computer Graphics and Interactive Techniques in Asia*, Seoul, South Korea, 2010.

[54] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 5(6):429–450, 2002.

[55] I. Jhuo, B. Liu, D.T. Lee, and Shih-Fu Chang. Robust visual domain adaptation with low-rank reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012.

[56] J. Jiang. A literature survey on domain adaptation of statistical classifiers. Technical report, School of Information Systems, Singapore Management University, 2008.

[57] J. Jiang and C. Zhai. Instance weighting for domain adaptation in nlp. In *Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 2007.

[58] T. Joachims, editor. *Making large-scale SVM learning practical*. Advances in Kernel Methods - Support Vector Learning. The MIT Press, 1999.

[59] T. Joachims. Transductive inference for text classification using support vector machines. In *Int. Conf. on Machine Learning*, Bled, Slovenia, 1999.

[60] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2011.

[61] L. Kuncheva. *Combining Pattern Classifiers. Methods and Algorithms*. Wiley, 2004.

[62] I. Laptev. Improving object detection with boosted histograms. *Image and Vision Computing*, 27(5):535–544, 2009.

[63] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *Int. Journal on Computer Vision*, 77(1-3):259–289, 2008.

[64] K. Levi and Y. Weiss. Learning object detection from a small number of examples: the importance of good features. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2004.

[65] M. Li and I.K. Sethi. Confidence-based active learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(8):1251–1261, 2006.

[66] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3D feature maps. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.

[67] R. Lienhart and J. Maydt. An extended set of Haar-like features for rapid object detection. In *IEEE Int. Conf. on Image Processing*, Rochester, NY, USA, 2002.

[68] Z. Lin and L. Davis. Shape-based human detection and segmentation via hierarchical part-template matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(4):604–618, 2010.

[69] L.von Ahn and L. Dabbish. Labeling images with computer game. In *ACM Int. Conf. on Human Factors in Computing Systems*, Vienna, Austria, 2004.

[70] L.von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *ACM Int. Conf. on Human Factors in Computing Systems*, Montreal, Quebec, Canada, 2006.

[71] S. Maji, A.C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.

[72] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: learning bounds and algorithms. In *Conference on Learning Theory*, Montreal, Quebec, 2009.

[73] J. Marin, D. Vázquez, D. Gerónimo, and A.M. López. Learning appearance in virtual scenarios for pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010.

[74] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.

[75] S. Munder and D.M. Gavrila. An experimental study on pedestrian classification. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(11):1863–1868, 2006.

[76] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[77] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Juan, PR, USA, 1997.

[78] S.J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.

[79] J. Pang, Q. Huang, S. Yan, S. Jiang, and L. Qin. Transferring boosted detectors towards viewpoint and scene adaptiveness. *IEEE Trans. on Image Processing*, 20(5):1388–400, 2011.

[80] C. Papageorgiou and T. Poggio. A trainable system for object detection. *Int. Journal on Computer Vision*, 38(1):15–33, 2000.

[81] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *European Conf. on Computer Vision*, Crete, Greece, 2010.

[82] M. Pedersoli, A. Vedaldi, and J. Gonzàlez. A coarse-to-fine approach for fast deformable object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011.

[83] D. Ponsa and A.M. López. Cascade of classifiers for vehicle detection. In *Advanced Concepts for Intelligent Vision Systems*, Delf, The Netherlands, 2007.

[84] T. Pouli, D.W. Cunningham, and E. Reinhard. Image statistics and their applications in computer graphics. In *European Computer Graphics Conference and Exhibition*, Norrköping, Sweden, 2010.

[85] Amazon Mechanical Turk. www.mturk.com.

[86] H. Daumé III. Frustratingly easy domain adaptation. In *Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 2007.

[87] H. Daumé III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26(1):101–126, 2006.

[88] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence, editors. *Dataset shift in machine learning*. Neural Information Processing. The MIT Press, 2008.

[89] M. Yeh Q. Zhu, S. Avidan and K. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005.

[90] F. Qureshi and D. Terzopoulos. Towards intelligent camera networks: a virtual vision approach. In *Joint IEEE Int. Works. on Visual Surveillance and Performance Evalutation of Tracking and Surveillance*, Beijing, China, 2005.

[91] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: a library for large linear lassification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[92] R.E. Schapire and Y. Singer. Improved boosting using confidencerated predictions. *Machine Learning*, 37(3):297–336, 1999.

[93] E. Reinhard, P. Shirley, M. Ashikhmin, and T. Troscianko. Second order image statistics in computer graphics. In *Symposium on Applied Perception in Graphics and Visualization*, New York, NY, USA, 2004.

[94] E. Reinhard, P. Shirley, and T. Troscianko. Natural image statistics for computer graphics. Technical Report UUCS-01-002, School of Computing, University of Utah, March 26th, 2001.

[95] J. Ross, L. Irani, M.Six Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers? shifting demographics in Mechanical Turk. In *ACM Int. Conf. on Human Factors in Computing Systems*, Atlanta, GA, USA, 2010.

[96] K. Saenko, B. Hulis, M. Fritz, and T. Darrel. Adapting visual category models to new domains. In *European Conf. on Computer Vision*, Hersonissos, Heraklion, Crete, Greece, 2010.

[97] S. Satpal and S. Sarawagi. Domain adaptation of conditional probability models via feature subsetting. In
PPKDD, Warsaw, Poland, 2007.

[98] B. Scholkopf and A. Smola. *Learning with kernels support vector machines, regularization, optimization and beyond*. MIT Press, Cambridge, MA, 2002.

[99] G. Shakhnarovich, P. Viola, and T. Darrel. Fast pose estimation with parameter sensitive hashing. In *Int. Conf. on Computer Vision*, Nice, France, 2003.

[100] W. Shao. *Animating Autonomous Pedestrians*. PhD thesis, Dept. C. S., Courant Inst. of Mathematical Sciences, New York Univesity, 2006. Advisor: Dimitri Terzopoulos.

[101] H. Shimodaira. Improving predictive inference under covariance shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[102] S. Sivaraman and M.M. Trivedi. A general active-learning framework for on-road vehicle recognition and tracking. *IEEE Trans. on Intelligent Transportation Systems*, 11(2):267–276, 2007.

[103] D. Tang, Y. Liu, and T.-K. Kim. Fast pedestrian detection by cascaded random forest with dominant orientation templates. In *British Machine Vision Conference*, Surrey, UK, 2012.

[104] T.L Berg, A. Sorokin, G. Wang, D.A. Forsyth, D. Hoeiem, I. Endres, and A. Farhadi. It's all about the data. *Proceedings of the IEEE*, 98(8):1434–1452, 2010.

[105] A. Torralba and A.A. Efros. Unbiased look at dataset bias. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011.

[106] O. Tuzel, F. Porikli, and P. Meer. Region covariance: a fast descriptor for detection and classification. In *European Conf. on Computer Vision*, Graz, Austria, 2006.

[107] D. Vázquez, A.M. López, D. Ponsa, and J. Marin. Cool world: domain adaptation of virtual and real worlds for human detection using active learning. In *Advances in Neural Information Processing Systems Workshop on Domain Adaptation: Theory and Applications*, Granada, Spain, 2011.

[108] D. Vázquez, A.M. López, D. Ponsa, and J. Marin. Virtual worlds and active learning for human detection. In *ACM International Conference on Multimodal Interaction*, Alicante, Spain, 2011.

[109] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, HI, USA, 2001.

[110] P. Viola and M. Jones. Robust real-time face detection. *Int. Journal on Computer Vision*, 57(2):137–154, 2004.

[111] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *Int. Journal on Computer Vision*, 63(2):153–161, 2005.

[112] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010.

[113] Meng Wang and Xiaogang Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011.

[114] Meng Wang and Xiaogang Wang. Transferring a generic pedestrian detector towards specific scenes. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012.

[115] X. Wang, T.X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *Int. Conf. on Computer Vision*, Kyoto, Japan, 2009.

[116] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[117] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009.

[118] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Int. Conf. on Machine Learning*, Banff, Canada, 2004.

[119] J. Zhang, K. Huang, Y. Yu, and T. Tan. Boosted local structured HOG-LBP for object localization. In *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010.

[120] L. Zhao and C. Thorpe. Stereo- and neural network-based pedestrian detection. In *IEEE Int. Conf. on Intelligent Transportation Systems*, Toronto, Canada, 1999.

[121] Q. Zhu, S. Avidan, M.-C. Yeh, and K.-T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *IEEE Conf. on Computer Vision and Pattern Recognition*, New York, NY, USA, 2006.