Virtual and Real World Adaptation for Pedestrian Detection

David Vázquez, Antonio M. López, *Member, IEEE,* Javier Marín, Daniel Ponsa, David Gerónimo

Abstract—Pedestrian detection is of paramount interest for many applications. Most promising detectors rely on discriminatively learnt classifiers, *i.e.*, trained with annotated samples. However, the annotation step is a human intensive and subjective task worth to be minimized. By using virtual worlds we can automatically obtain precise and rich annotations. Thus, we face the question: *can a pedestrian appearance model learnt in realistic virtual worlds work successfully for pedestrian detection in real-world images?*. Conducted experiments show that virtual-world based training can provide excellent testing accuracy in real world, but it can also suffer the *dataset shift* problem as real-world based training does. Accordingly, 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 combine them 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. V-AYLA reports the same detection accuracy 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 object detector.

Index Terms—pedestrian detection, photo-realistic computer animation, dataset shift, domain adaptation

1 Introduction

Pedestrian detection is of paramount interest in fields such as driving assistance, surveillance and media analysis [1], [2], [3], [4]. The main component of a pedestrian detector is its *pedestrian classifier*, since it decides if an image window contains a pedestrian or not. Most promising classifiers have appearance as core feature and are learnt discriminatively, *i.e.* from annotated windows; where such pedestrian and background windows are the *examples* and the *counterexamples*, respectively, to feed the learning machine. Having sufficient variability in the *samples* (examples and counterexamples) is decisive to train classifiers able to generalize properly [5]. Obtaining such a variability is not straightforward since annotated samples are obtained through a subjective and tiresome manual task. 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 learning.

Accordingly, in the last years different web-based tools have been proposed for manually collecting image annotations [6]. Among them, Amazon Mechanical Turk (MTurk) [7] centralizes nowadays the most powerful annotation force. However, *how to collect data on the internet* is a non-trivial question that opens a new research area [6], [8] involving *ethical questions* too [9] since human paid work is at the core.

• *Vázquez, López, Ponsa and Gerónimo are with the Computer Vision Center (CVC) and the Computer Science Dpt. at the Univ. Autònoma de Barcelona. Marín is with the CVC. E-mail: david.vazquez@cvc.uab.es*

In order to collect useful samples and reduce manual annotations, the so-called *active learning* [10] has also been explored in object detection. For instance, SEVILLE (*se*mi-atomatic *vi*sua*l* *le*arning) [11] and ALVeRT (*a*ctive-*l*earning-based *ve*hicle *r*ecognition and *t*racking) [12] systems develop a pedestrian and a vehicle detector, respectively. Both systems use AdaBoost as base classifier. In SEVILLE weak rules are decision stumps based on descriptors referred to as YEF. In ALVeRT the decision stumps are based on Haar descriptors. In such systems active learning consists of a stage to obtain an initial classifier (*passive learning*), followed by a loop in which the current classifier is plugged in a detector that is applied to unseen images, a *human oracle* performs *selective sampling* (*i.e.* annotation of image windows falling into the classifier *ambiguity region*) and then previous and new annotations are used for re-training the classifier. Active learning is especially interesting to collect informative examples (pedestrians/vehicles) since *hard* counterexamples (background) are satisfactorily collected by having example-free images and properly using *bootstrapping* [1]. SEVILLE uses 215 passively annotated pedestrians, while 2,046 new ones are actively annotated. ALVeRT annotates 7,500 vehicles passively and 10,000 actively.

In the approaches mentioned so far, examples are annotated from existing image captures. Alternatively, we can *engineer* them as in [13], where pedestrians are synthesized by transforming their original shape, texture, and surrounding background. Aiming at increasing pedestrian variability, such a transform is applied according to selective sampling within an active learning framework. Both NNs with LRFs and SVM with

Haar are used as base classifiers. The reported results show the same accuracy than when using a human oracle. However, much of the improvement comes from enlarging the training set by slightly perturbing each pedestrian bounding box (BB) location (so-called *jittering*) as well as gathering counterexamples with selective sampling on pedestrian-free images (a sort of bootstrapping). Moreover, for synthesizing the pedestrians it was not sufficient to annotate them with BBs, but costly manual silhouette delineation was required. 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 [14].

Instead of developing pedestrian models from real-world images, rough textureless pedestrian templates were used in [15] for building a template-matching-based pedestrian detector for far infrared images (*i.e.* capturing relative temperature). However, the authors admit poor results due to the lack of realism in the synthetic templates. Similarly, in [16] a human renderer is used to randomly generate synthetic human poses for training an appearance-based pose recovery system. However, 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 field-of-view is filled by a human.

In this paper 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 back successfully in the real world?*. Note that modern videogames, simulators and animation films, are gaining photo-realism. In fact, all the ingredients for creating *soft artificial life* are being improved: visual appearance both global (3D shape, pose) and local (texture, where involved Computer Graphics aim at approaching the power spectrum of real images [17]), kinematics, perception, behavior and cognition. For instance, see [18] 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 controlled manner. Within this global context we are currently focused on a more specific question, namely, *can a pedestrian appearance model learnt in realistic virtual scenarios work successfully for pedestrian detection in real images?* (Fig. 1).

There are different possibilities to assess the posed question. We can learn a holistic pedestrian classifier using dense descriptors [1], [2], [4], or the pedestrian silhouette [19]. Analogously, we can learn a part-based pedestrian classifier with dense descriptors [20], or using the pedestrian silhouette instead [21]. In all

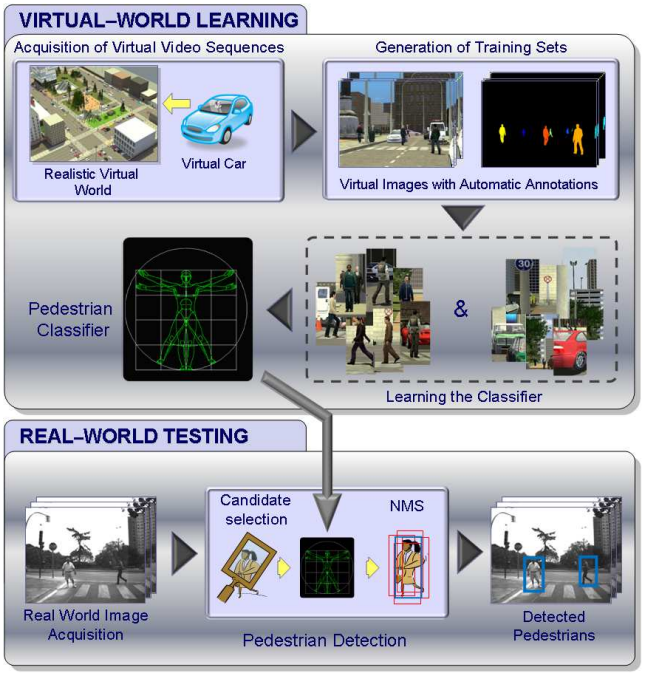


Fig. 1. Can a pedestrian appearance model learnt in realistic virtual scenarios work successfully for pedestrian detection in real images?

cases, different learning machines can be tested as well. Thus, given such an amount of possibilities, in [22] we followed the popular wisdom of *starting from the beginning*. In particular, using virtual-world samples, we trained a holistic pedestrian classifier based on histograms of oriented gradients (HOG) and linear SVM (Lin-SVM) [23]. We tested such a classifier in a dataset made available by Daimler AG [2] for pedestrian detection benchmarking in the driver assistance context, our main field of interest. The results were compared with a pedestrian detector whose classifier was trained using real-world images. The comparison revealed that virtual and real-world based training give rise to similar classifiers.

In this paper we present a more in depth analysis by introducing new descriptors and datasets used in the context of pedestrian detection, so that we can better appreciate the effect of employing virtual world for training. We use an improved virtual world too. Training with Lin-SVM we assess the behavior of HOG, and of cell-structured local binary patterns (LBP) [24]. Since HOG is more related to overall shape and LBP to texture, following [24] 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 behavior 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 [25]. HOG and LBP are key descriptors of state-of-the-art part-based object detectors [20], [26].

Our experiments will show that we obtain the same accuracy 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 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 problem is referred to as *dataset shift* [27] and is receiving increasing attention in the Machine Learning field [27], [28], [29] due to its relevance in areas like natural language processing, speech processing, and brain-computer interfaces, to mention a few.

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 requires *adapting* training and testing/application domains. For that we propose a *domain adaptation* framework called Virtual-AYLA[1], V-AYLA in short, which stands for *virtual*-world *a*nnotations *y*et *l*earning *a*daptively. For reaching the desired accuracy, 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 paper 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. Note that in SEVILLE and ALVeRT active learning is not used for doing domain adaptation since training and testing sets come from the same camera and scenarios. We borrowed the idea of augmented descriptor space from [29], 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 [30], [31], although not for a detection task like here. Moreover, in [30], [31] the domains to be adapted are both based on real-world images and the involved descriptors are not the ones used for pedestrian detection.

Section 2 details the datasets, pedestrian detector stages, and evaluation methodology. Section 3 reports the results of the detectors developed without domain adaptation. Section 4 presents V-AYLA and its results. Section 5 draws the main conclusions.

2 Experimental settings

2.1 Datasets

In order to illustrate the domain adaptation problem and our proposal, we start working with two real-world datasets and our virtual-world one. As generic real-world dataset we have chosen the INRIA ($\mathcal{I}$) one [32] since it is very well-known and still used as reference [1], [4], [24], [33]. 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 ($\mathcal{D}$) [2], 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 [22]. 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. Virtual-world data is only for training.

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

A priori training with more pedestrians could lead to better classifiers. For avoiding such a potential effect, the cardinality of the smallest pedestrian training set (*i.e.* 1,208) 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

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.

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.

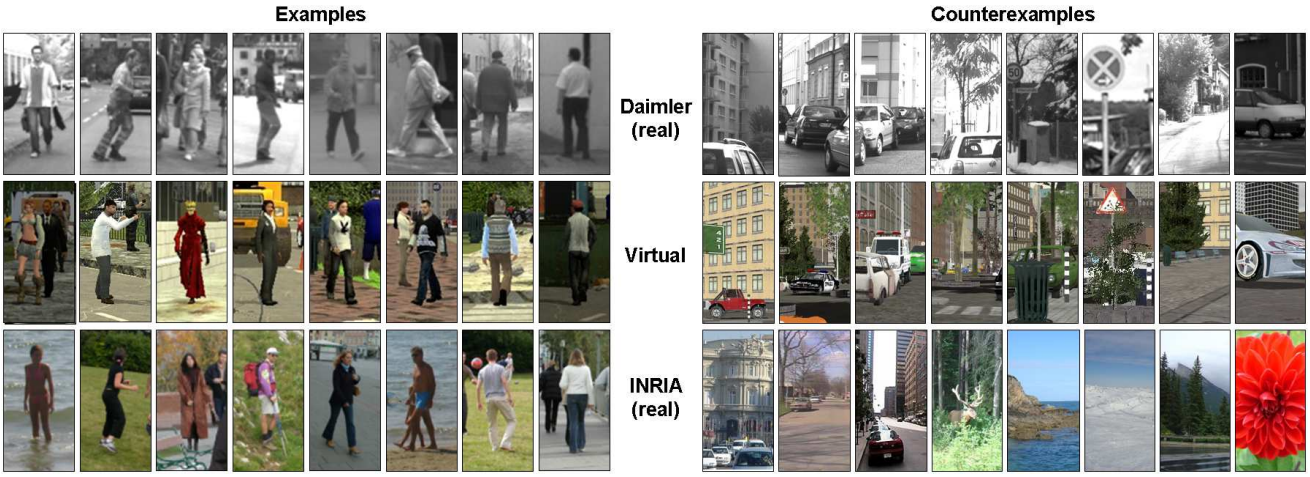


Fig. 2. Some of the samples used to train with real-world images (Daimler and INRIA) and virtual-world ones.

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.

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 randomly collected 1,218 pedestrian-free images. The final number of used counterexamples from each dataset depends on *bootstrapping* (Sect. 3.1). 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 $\mathcal{T}_{\mathcal{V}}^{tr-}$, respectively. 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. 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).

Figure 2 shows virtual- and real-world samples. Moreover, Table 5 summarizes the notation used along the paper, *e.g.* as the one in this subsection.

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 1 illustrates the idea for a pedestrian classifier learnt with virtual-world data. We will describe the learning of such classifiers in Sect. 3.1. In the following we briefly review the employed scanning and selection procedures.

The scanning uses a pyramidal sliding window [32]. It consists in constructing a pyramid of scaled images. The bottom (higher resolution) is the original image, while the top is limited by the size of the so-called *canonical window* (CW, Sect. 3.1). 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 [32], for building levels of lower resolution we perform down-sampling by using standard bilinear interpolation with anti-aliasing, as in [34]. 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 set $< s_x, s_y, s_p > := < 8, 8, 1.2 >$ as a good tradeoff between detection accuracy and processing time. LBP and HOG are usually computed without anti-aliasing. However, we have experimentally seen (Sect. 3.2) that, in general, the pyramid with anti-aliasing boosts the accuracy of the pedestrian detectors based on them.

The CW of a classifier trained with $\mathcal{T}_{\mathcal{I}}^{tr}$ is larger than with $\mathcal{T}_{\mathcal{D}}^{tr}$ (Sect. 3.1). 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 [35] advice we up-scale the testing images using bilinear interpolation. $\mathcal{T}_{\mathcal{V}}^{tr}$ can be adapted to any CW (Sect. 3.1).

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 [36] to (ideally) provide one single detection per pedestrian.

2.3 Evaluation methodology

In order to evaluate the pedestrian detectors we reproduce the proposal in [4]. Thus, 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. 2.1).

The evaluation procedure described so far is rather standard. However, according to our daily working experience, even using a good *bootstrapping* method [1] (Sect. 3.1) 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), or samples from real world in domain adaptation training (Sect. 4). Thus, in this paper we repeat each 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. Overall, this turns out in 520 train-test runs done for this paper.

3 Passive Training

In this section we focus on the situation in which a prefixed set of annotated data is used to learn a classifier. We term this approach as *passive*. Sets $\mathcal{T}_{\mathcal{X}}^{tr}$, $\mathcal{X} \in \{\mathcal{D}, \mathcal{I}, \mathcal{V}\}$, are of such a type, and passive learning is the most widespread in pedestrian detection [3].

3.1 Pedestrian classifier training

Discriminative learning of a pedestrian classifier requires the computation of descriptors able to distinguish pedestrians from background. As we introduced in Sect. 1, HOG and LBP are very well suited for this task. 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.

For 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. 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 [23]. For Daimler and HOG/LBP, $f = 12$, therefore, $w \times h \equiv 48 \times 96$ [2].

We just consider virtual-world pedestrians larger than 32×96 pix. 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. 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. 2.1). For doing such a sampling, the pyramid (Sect. 2.2) of each 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 apply *bootstrapping* [1], [2], [23]. Thus, 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. 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 [1] it is also recommended to follow a strategy such that almost all counterexamples are collected by the bootstrapping.

Table 1 summarizes the descriptors parameters. HOG ones are the originals [23]. In the case of LBP, we introduce three improvements with respect to the approach in [24]. 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.

TABLE 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 [37] with thr=4, L1-sqrt normalization.	6,195	3,245

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

TABLE 2
Passive learning results. For FPPI $\in [0.1, 1]$, AMR (%) mean and std. dev. are indicated. Bold style remarks the lowest mean comparing training sets ($\mathcal{T}_{\mathcal{X}}^{tr}$) for fixed descriptor and testing set ($\mathcal{T}_{\mathcal{R}}^{tt}$).

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

(⋆) Dalal *et al.* implementation [32].
(○) Wang *et al.* impl. [24], without occlusion handling.
(●) Training with the 15,660 pedestrians (Sect. 2.1).

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 accuracy. Finally, as Lin-SVM implementation we use LibLinear [38], setting $C = 0.01$ and $bias = 100$.

3.2 Experimental results and discussion

Table 2 shows the results of the 24 detectors obtained by passive training. We see that using train and test sets of different *domain* increases the AMR mean even 15 points depending on the descriptor and dataset. We argue that we are facing a *dataset shift* problem. 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 the training domain (using the corresponding testing set) and the other one, by paring the obtained performances we can apply a *paired Wilcoxon test* [39]. The test reveals that for HOG, LBP and HOG+LBP testing and training with samples/images of the same domain is better than using different domains in the 99.9% of the cases (p-value = 0.001, being the null hypothesis that source and target domains 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 dataset shift, but just as real-world data does. In order to support this claim, we have analyzed if detectors trained with $\mathcal{V}$ data behave similarly to detectors trained with real-world data (using $\mathcal{I}$ and $\mathcal{D}$ domains) when tested on a different domain, 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* [40]) must be applied. This test allows to conclude that when using $\mathcal{I}$ as testing domain, 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 domain 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 dataset shift, the virtual-world domain is comparable to a real-world one.

Usually there are several (possibly simultaneous) reasons giving rise to domain shift. 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, $\{\mathcal{T}_{\mathcal{V}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$ result 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 result 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 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). We do not want to analyze every single possible reason producing domain shift, instead we want to treat all them simultaneously by applying domain adaptation techniques.

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 [32] (used by us in [22], [41]) 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 [24], 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 [24]. Note that the better the accuracy when training and testing within the same domain, the higher the challenge to reach the same result when using different domains.

4 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 $\{\mathcal{T}_{\mathcal{D}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$ (Sect. 3).

In practice, there are cases in which samples in $\mathcal{T}^{tr}$ and $\mathcal{T}^{tt}$ follow different probability distributions. As mentioned in Sect. 1, *dataset shift* is the generic term to summarize the many possible underlying reasons [27]. We argue that the loss of accuracy seen in Sect. 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 $\{\mathcal{T}_{\mathcal{V}}^{tr}, \mathcal{T}_{\mathcal{I}}^{tt}\}$ and $\{\mathcal{T}_{\mathcal{V}}^{tr}, \mathcal{T}_{\mathcal{D}}^{tt}\}$. 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. In this paper, we focus on *supervised* domain adaptation [29], where we have a reasonable number of annotations from δ_s and some ones from δ_t too. In particular, our δ_s is the virtual world $\mathcal{V}$, and δ_t is the real world $\mathcal{R}$ (here $\mathcal{R} \in \{\mathcal{I}, \mathcal{D}\}$). We assess domain adaptation for HOG[3] and LBP separately, as well as for HOG+LBP.

4.1 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. 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 [29]. 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 $< \mathbf{0}, \mathbf{0}, \mathbf{x} >$ 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 [29] for an explanation in terms of SVM margin maximization.

4.2 Real-world domain exploration

Let $n_{t.p.}$ be the maximum number of target domain (real-world) 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* [10]. 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. 4.3 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.

3. In [42], [41] we presented domain adaptation results for INRIA with HOG/Lin-SVM. However, we used the implementation proposal in [32] instead of our current one.

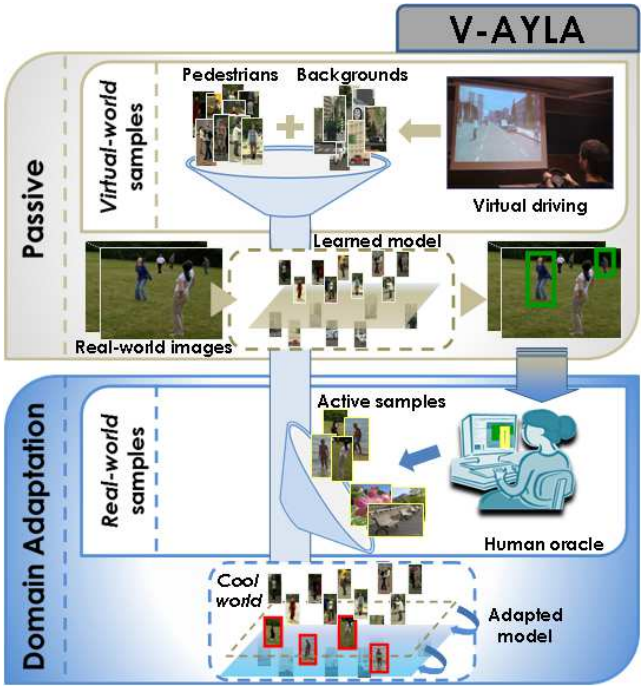


Fig. 3. V-AYLA: passive + domain adaptation training.

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.

Finally, we would like to clarify that in this paper we are interested in assessing the type of real-world data, which changes with $\mathcal{O}$, needed for complementing the virtual-world one for domain adaptation. Our next future work will use the obtained conclusions for devising unsupervised domain adaptation methods, while the supervised domain adaptation results obtained here will act as baseline. Thus, in this paper we do not focus on evaluating the human annotation effort of active oracles *vs* Rnd. In our work the significant saving of human annotations comes from the use of the virtual world.

4.3 Domain adaptation training: V-AYLA

Assume the following definitions of *training sets*:

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

Take the following decisions:

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

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

(s1) Perform *passive learning in virtual world* using $\{\Im^{tr+}_\mathcal{V}, \Im^{tr-}_\mathcal{V}\}$ and $\mathbf{D}$ (Sect. 3.1). Let us term as $\mathrm{C}_\mathcal{V}$ the passively learnt pedestrian classifier and as $\mathrm{D}_\mathcal{V}$ its associated detector (Sect. 2.2). Let $\mathcal{T}^{tr+}_\mathcal{V}$ be the set of pedestrians used for obtaining $\mathrm{C}_\mathcal{V}$ (*i.e.* coming from $\Im^{tr+}_\mathcal{V}$, scaled to the CW size and augmented by mirroring), and $\mathcal{T}^{tr-}_\mathcal{V}$ the set of background samples (coming from $\Im^{tr-}_\mathcal{V}$ after bootstrapping, CW size).

(s2) *Selective sampling in real world*. In order to obtain real-world annotated pedestrians, follow $\mathcal{O}$ by running $\mathrm{D}_\mathcal{V}$ on $\Im^{tr+}_\mathcal{R}$. If $\mathcal{O} = \mathrm{Act{\pm}}$, 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 $\mathcal{T}^{tr+}_\mathcal{R}$ the set of such new pedestrian samples scaled to CW size and augmented by mirroring, and as $\mathcal{T}^{tr-}_\mathcal{R}$ a set of background samples in CW size, taken from $\Im^{tr-}_\mathcal{R}$ as done in the passive learning procedure before bootstrapping (thus, the cardinality of $\mathcal{T}^{tr+}_\mathcal{R}$ and $\mathcal{T}^{tr-}_\mathcal{R}$ are equal). Note that to follow $\mathcal{O}$ we need to set Thr. For that purpose, we initially select a few images from $\Im^{tr-}_\mathcal{R}$ and take a Thr value such that after applying $\mathrm{D}_\mathcal{V}$ 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 $\mathcal{T}^{tr+}_\mathcal{V}$, $\mathcal{T}^{tr-}_\mathcal{V}$, $\mathcal{T}^{tr+}_\mathcal{R}$, and $\mathcal{T}^{tr-}_\mathcal{R}$ to cool world. Next, train a new classifier with them according to $\mathbf{D}$. Then, perform bootstrapping in $\Im^{tr-}_\mathcal{R}$. Finally, re-train in cool world to obtain the domain adapted classifier.

When $\mathcal{O} \neq$ Rnd, this is a *batch active learning* procedure [43]. Figure 3 summarizes the idea, which, as we introduced in Sect. 1, we term as V-AYLA.

TABLE 3
Domain adaptation results. 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.

HOG				
INRIA ($\mathcal{T}_{\mathcal{I}}^{tt}$)	Act+	Act− Act∼	Rnd	Act±
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	$\mathbf{21.29} \pm 0.85$
Daimler ($\mathcal{T}_{\mathcal{D}}^{tt}$)	Act+	Act∼	Rnd	Act±
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				
INRIA ($\mathcal{T}_{\mathcal{I}}^{tt}$)	Act+	Act− Act∼	Rnd	Act±
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	$\mathbf{18.87} \pm 0.88$
Daimler ($\mathcal{T}_{\mathcal{D}}^{tt}$)	Act+	Act∼	Rnd	Act±
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				
INRIA ($\mathcal{T}_{\mathcal{I}}^{tt}$)	Act+	Act− Act∼	Rnd	Act±
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	$\mathbf{14.15} \pm 0.58$
Daimler ($\mathcal{T}_{\mathcal{D}}^{tt}$)	Act+	Act∼	Rnd	Act±
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
Rows $\mathcal{T}_{\mathcal{V}}^{tr}$ show results by training with $\mathcal{V}$ data only (from Table 2). 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 3, 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

4.4 Experimental results

This section summarizes the V-AYLA experiments, and we explain how to simulate the application of V-AYLA on INRIA and Daimler data for providing fair comparisons with respect to passive learning.

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 result 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 $n_{t.p.} = 120$ real-world manually annotated pedestrians. In both cases it is assumed a set of real-world pedestrian-free images.

During an actual application of V-AYLA, the 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, since the five repetitions we apply lead to 300 training-testing runs, although V-AYLA only needs 120 BB annotations here, this would turn out in 36,000 manually annotated BBs. Therefore, in order to reduce overall manual effort, we have simulated the annotation of the 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 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. 4.3 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 3 shows the application of V-AYLA on INRIA for $\mathcal{O} \in \{\text{Rnd}, \text{Act+}, \text{Act−}, \text{Act±}\}$, combined with both ORG and AUG. Regarding Daimler, we show analogous experiments but replacing Act− by Act∼, which propagates to Act±. Figure 4 plots the curves corresponding to most interesting results.

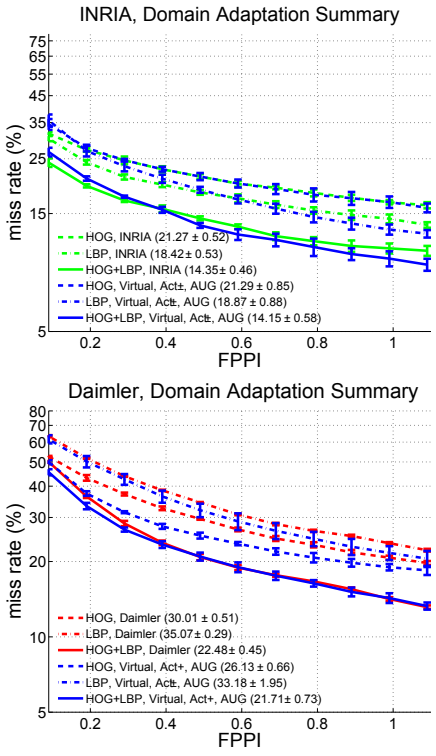


Fig. 4. Results for the best cases in Table 3.

4.5 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 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 the desired results by just manually annotating a few real-world pedestrian BBs (*i.e.* 120 here). 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.

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, irrespective 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, Table 3 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 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 [44]. As intuitively expected from the results in Table 3, 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

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.*

is not always the case when using modern web-based annotation tools [6], [8]. 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 pedestrian detector.

We would like to mention that, although the Act− works worse than the other oracles still provides a large adaptation (*e.g.* for INRIA setting more than 5 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 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 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 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], [20].

In conclusion, V-AYLA allows to significatively 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.

4.6 Additional experiments

For complementing V-AYLA performance assessment, we rely on the popular pedestrian dataset of Caltech

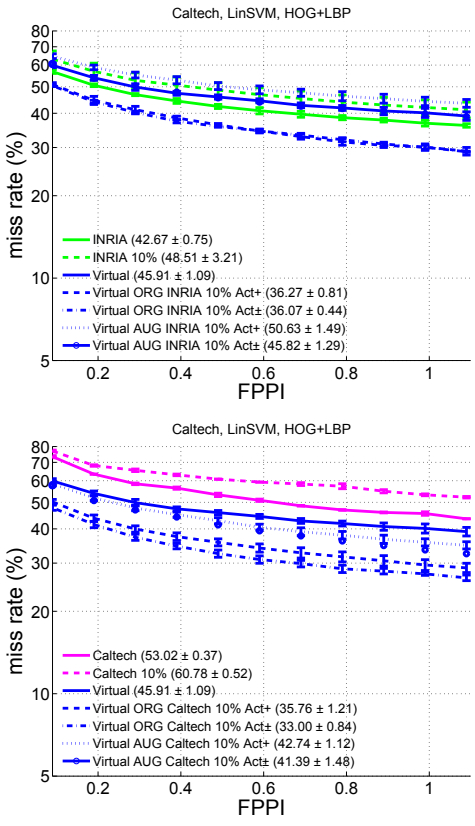


Fig. 5. Adaptation for Caltech *reasonable* testing set, using INRIA (top) and Caltech (bottom) training data.

[4]. It contains color images of 640×480 pix resolution acquired from a vehicle driven through different urban scenarios at different day times. We focus on the best performance curves, *i.e.* those provided by the combination HOG+LBP as well as Act+ and Act±. For this set of features both AUG and ORG provided rather close performance (Table 3), thus, we test both.

For training, in [4] it is used INRIA training data. Here we perform a set of experiments following such approach, thus, we will adapt the virtual-world data to INRIA training one and then test in Caltech data. In addition, we also use Caltech training data to perform another set of experiments where the virtual-world data is directly adapted to Caltech one, *i.e.* instead of doing it through INRIA 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. In both types of experiments we set the CW following INRIA settings, and we use image up-scaling during testing for detecting *reasonable* pedestrians (*i.e.* most representative ones [4]) taller than 50 pix but not

reaching the 96 pix of the INRIA CW setting.

Results are plotted in Fig. 5. Both training with INRIA and virtual-world data performs better than using the Caltech training data our automatic procedure has collected. This may suggest that such data lacks variability. In fact, it can be thought as a random human annotation of 790 pedestrians taller than 72 pix. However, this is not important for the purpose of this paper since we can assume that the baseline performance is the one based on INRIA training data as is usually done [4]. Note also that using only the 10% of the real-world training data, either Caltech or INRIA, drops the performance in more than 6 points. However, combining such an amount of real-world data with the virtual-world one improves more than 6 points the baseline when using the ORG cool world, for both the Act+ and Act$\pm$ oracles. Let us remind that in the Act$\pm$ case it is used an additional amount of 10% real-world data collected by V-AYLA without requiring manual annotation of pedestrian BBs. The best performance is given by V-AYLA based on ORG/Act$\pm$ and Caltech training data. In terms of manually annotated pedestrian BBs only 79 are provided for training such a pedestrian detector. In these experiments ORG approach clearly outperforms AUG, which opens a question for our future work, namely, how to determine a priori the best training cool world, if possible.

Figure 6 shows the last set of experiments. We progressively increase the amount of target domain (real-world) pedestrians combined with the virtual-world ones for training. Again we focus on HOG+LBP and both ORG and AUG. Since the total amount of real-world pedestrians available for training is fixed, the more we add the less differences would be among oracles. Thus, we just run the Rnd one and each experiment is run just once. We can see that for Daimler and INRIA testing, ORG and AUG do not show significant differences, while for Caltech again ORG outperforms AUG. We see that by incorporating the real-world samples the AMR is reduced. However, there is a point where the improvement stops, which we think is because the limit of the holistic model based on HOG+LBP/Lin-SVM is reached.

5 Conclusion

In this paper we have explored how 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 and Caltech). Following such a framework we have tested state-of-the-art pedestrian descriptors (HOG/LBP/HOG+LBP) with Lin-SVM. Within the same pedestrian detection scheme, we have employed virtual-world based classifiers and real-world based ones (Virtual, INRIA, Daimler). In total 120 train-test runs have been performed to assess detection performance. We have reached the conclusion that both virtual-world and real-world based training behave equally. This means that virtual-world based training can provide excellent performance, but it can also suffer the *dataset shift* problem as real-world based training does.

Accordingly, 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 V-AYLA we have performed 400 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 (in the case of Caltech it is clearly improved).

Our next future work is twofold. On the one hand, we will extend the V-AYLA concept to deformable part models [20] with the aim of going beyond state-of-the-art results. On the other hand, we want to achieve unsupervised domain adaptation, being the challenge how to reproduce Act$\pm$ oracle without human intervention. Finally, we also plan to extend our work for detecting other types of objects (*e.g.* vehicles).

Acknowledgments

This work is supported by Spanish MICINN projects TRA2011-29454-C03-01 and TIN2011-29494-C03-02.

References

[1] 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.

[2] M. Enzweiler and D.M. Gavrila, "Monocular pedestrian detection: survey and experiments," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2179–2195, 2009.

[3] 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*, vol. 32, no. 7, pp. 1239–1258, 2010.

[4] 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*, vol. 34, no. 4, pp. 743–761, 2012.

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

[6] 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*, vol. 98, no. 8, pp. 1434–1452, 2010.

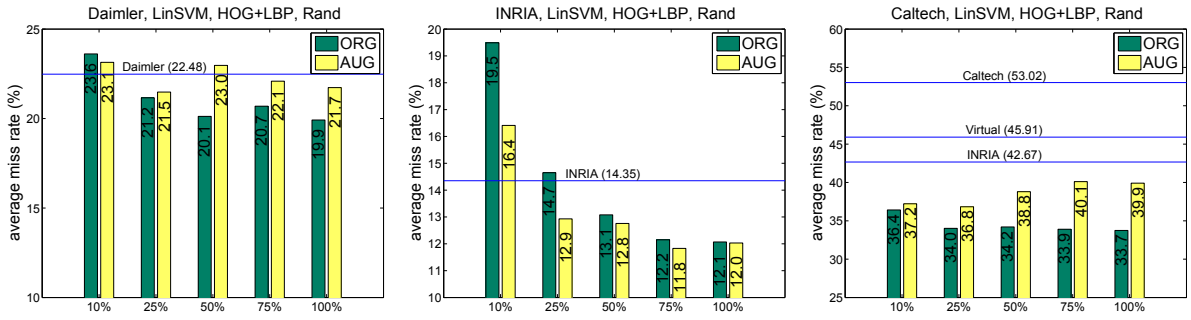


Fig. 6. AMRs of progressively adding more $(10\% - 100\%)$ target domain (real-world) training pedestrians to the virtual-world ones. Lines mark the AMR for detectors trained with the 100% of the indicated training set.

TABLE 5
Notation summary.

Symbol	Stands for ...
$\mathcal{D}, \mathcal{I}, \mathcal{V}$	Daimler, INRIA, Virtual-world domains, resp.
$\mathcal{X}$	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.
$\Im_{\mathcal{X}}^{tr+}$	Training set of images from $\mathcal{X}$, with annotated pedestrians.
$\Im_{\mathcal{X}}^{tr-}$	Training set of pedestrian-free images from $\mathcal{X}$.
$\Im_{\mathcal{R}}^{tr+}$, $\Im_{\mathcal{R}}^{tr-}$	Analogous to $\Im_{\mathcal{X}}^{tr+}$ and $\Im_{\mathcal{X}}^{tr-}$, 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).
$n_{t.p.}$	Amount of target domain pedestrians manually annotated (BB).
ORG, AUG	*Cool-world* types (Sect. 4.1). ORG refers to a direct mixing of virtual and real-world examples. AUG refers to feature augmentation [29].

[7] Amazon Mechanical Turk, www.mturk.com.

[8] I. Endres, A. Farhadi, and D. H. 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.

[9] 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.

[10] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[11] 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.

[12] S. Sivaraman and M.M. Trivedi, "A general active-learning framework for on-road vehicle recognition and tracking," *IEEE Trans. on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 267–276, 2007.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] W. Shao, "Animating autonomous pedestrians," Ph.D. dissertation, Dept. C. S., Courant Inst. of Mathematical Sciences, New York Univesity, 2006, Advisor: Dimitri Terzopoulos.

[19] D.M. Gavrila, "A bayesian exemplar-based approach to hierarchical shape matching," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1408–1421, 2007.

[20] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[21] Z. Lin and L. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 604–618, 2010.

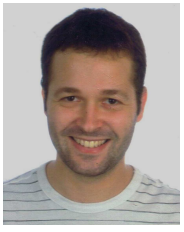
[22] J. Marin, D. Vázquez, D. Gerónimo, and A.M. López, "Learning appearance in virtual scenarios for pedestrian detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, 2010.

[23] 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.
[24] 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.
[25] M. Enzweiler and D.M. Gavrila, "A multi-level mixture-of-experts framework for pedestrian classification," *IEEE Trans. on Image Processing*, vol. 20, no. 10, pp. 2967–2979, 2011.
[26] 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.
[27] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence, Eds., *Dataset shift in machine learning*, ser. Neural Information Processing. The MIT Press, 2008.
[28] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1, pp. 151–175, 2009.
[29] H. D. III, "Frustratingly easy domain adaptation," in *Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 2007.
[30] 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.
[31] 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.
[32] N. Dalal, "Finding people in images and videos," Ph.D. dissertation, Institut National Polytechnique de Grenoble, 2006, Advisors: Cordelia Schmid and William J. Triggs.
[33] Y.-T. Chen and C.-S. Chen, "Fast human detection using a novel boosted cascading structure with meta stages," *IEEE Trans. on Image Processing*, vol. 17, no. 8, pp. 1452–1464, 2008.
[34] 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.
[35] 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.
[36] I. Laptev, "Improving object detection with boosted histograms," *Image and Vision Computing*, vol. 27, no. 5, pp. 535–544, 2009.
[37] 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*, vol. 24, no. 7, pp. 971–987, 2002.
[38] 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*, vol. 9, pp. 1871–1874, 2008.
[39] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.
[40] 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*, vol. 18, no. 1, pp. 50–60, 1947.
[41] 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.
[42] ——, "Virtual worlds and active learning for human detection," in *ACM International Conference on Multimodal Interaction*, Alicante, Spain, 2011.
[43] M. Li and I.K. Sethi, "Confidence-based active learning," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1251–1261, 2006.
[44] 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*, vol. 9, pp. 2677–2694, 2008.

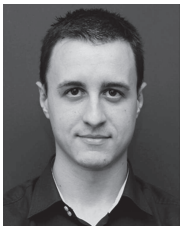
David Vázquez received the B.Sc. degree in Computer Science from the Universitat Autònoma de Barcelona (UAB). He received his M.Sc. in Computer Vision and Artificial Intelligence at the Computer Vision Center (CVC/UAB) in 2009. As member of the CVC's Advanced Driver Assistance Systems (ADAS) group, he is currently working to obtain the Ph.D. degree. His research interests include pedestrian detection, virtual worlds, domain adaptation and active learning.

Antonio M. López received the B.Sc. degree in Computer Science from the Universitat Politècnica de Catalunya (UPC) in 1992 and the Ph.D. degree in Computer Vision from the UAB in 2000. Since 1992, he has been giving lectures in the UAB, where he is now Associate Professor. In 1996, he participated in the foundation of the CVC, where he has held different institutional responsibilities. In 2003 he started the CVC's ADAS group, presently being its head.



Javier Marín received the B.Sc. degree in Mathematics from the Universitat de les Illes Balears (UIB) in 2007. He received his M.Sc. in Computer Vision and Artificial Intelligence at the UAB in 2009. As member of the CVC's ADAS group, he is currently working to obtain the Ph.D. degree. His research interests include pedestrian detection, virtual worlds, and random subspace methods.

Daniel Ponsa received the B.Sc. degree in Computer Science and the Ph.D. degree in Computer Vision from the UAB in 1996 and 2007, respectively. He did his Ph.D. as member of the CVC's ADAS group. He is currently an Assistant Professor at the UAB as well as member of the CVC's ADAS group. His research interests include tracking, motion analysis, and machine learning, for driver assistance and unmanned aerial vehicles.



David Gerónimo received the B.Sc. degree on Computer Science and the Ph.D. degree in Computer Vision from the UAB in 2004 and 2010, respectively. He did his Ph.D. as member of the CVC's ADAS group, where he was post-doc until 2013. His research interests include object detection and tracking, action recognition and scene understanding in the contexts of driver assistance and surveillance video mining.