Virtual Worlds and Active Learning for Human Detection

David Vázquez
Computer Vision Center and
Computer Science Dpt. UAB,
Campus UAB, Spain
08193 Bellaterra, Spain
david.vazquez@cvc.uab.es

Antonio M. López
Computer Vision Center and
Computer Science Dpt. UAB,
Campus UAB, Spain
08193 Bellaterra, Spain
antonio@cvc.uab.es

Daniel Ponsa
Computer Vision Center and
Computer Science Dpt. UAB,
Campus UAB, Spain
daniel@cvc.uab.es

Francisco J. Marín
Computer Vision Center,
Campus UAB, Spain
jmarin@cvc.uab.es

ABSTRACT

Image based human detection is of paramount interest due to its potential applications in fields such as advanced driving assistance, surveillance and media analysis. However, even detecting non-occluded standing humans remains a challenge of intensive research. The most promising human detectors rely on classifiers developed in the discriminative paradigm, *i.e.*, trained with labelled samples. However, labelling is a manual labor intensive step, especially in cases like human detection where it is necessary to provide at least bounding boxes framing the humans for training. To overcome such problem, some authors have proposed the use of a *virtual world* where the labels of the different objects are obtained automatically. This means that the human models (classifiers) are learnt using the appearance of rendered images,*i.e.*, using realistic computer graphics. Later, these models are used for human detection in images of the *real world*. Indeed, the results of this technique are surprisingly good. However, these are not always as good as the classical approach of training and testing with data coming from the same camera, or pretty similar ones. Accordingly, in this paper we address the challenge of using a virtual world for gathering (while *playing a videogame*) a large amount of automatically labelled samples (virtual humans and background) and then training a classifier that performs equal, in real-world images, than the one obtained by training from manually labelled real-world samples. For doing that, we cast the problem as one of *domain adaptation*. Thus, we assume that a small amount of manually labelled samples from real-world images is required. To collect these labelled samples we propose a non-standard *active learning* technique. Therefore, ultimately our human model is learnt by the combination of virtual and real world labelled samples (Fig. 1), something not done before. We present quantitative results showing that this approach is valid.

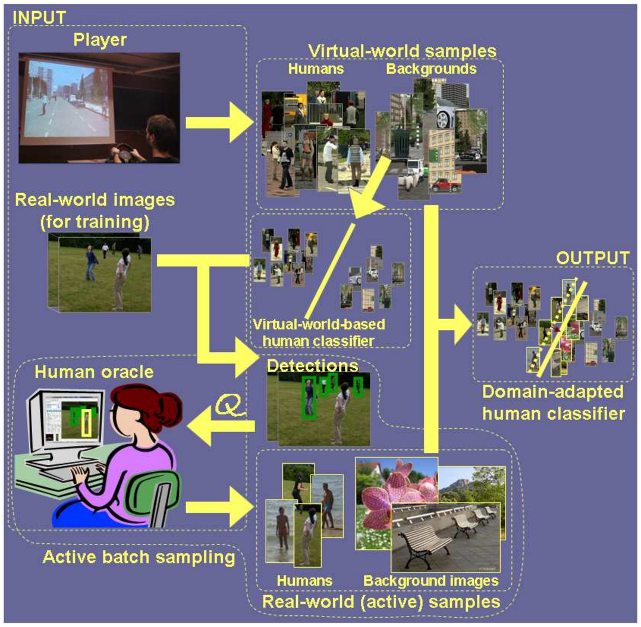


Figure 1: Our proposal in a nutshell. By playing a videogame we automatically gather labelled samples of virtual humans (pedestrians in a city) and backgrounds (buildings, cars, trees, etc.). Following the discriminative learning paradigm, we learn a human classifier from such virtual samples. Images of the real world are used to challenge the classifier, collecting all human detections. Then we start an active learning procedure of batch type. For each real-word image the query (Q) to be answered by the human oracle is: (1) *is it a human-free image?* (2) if the answer is *no*, then *label (with bounding box) non-detected humans.* Finally, a new classifier is learnt by using the labelled humans coming from virtual and real world as examples of the *human class*, and labelled backgrounds from both worlds are used as examples of the *non-human class*. Thus, virtual world is adapted to real world by active learning.

1. INTRODUCTION

Image based human detection is of paramount interest due to its potential applications in fields such as advanced driving assistance, video surveillance and media analysis. However, by reading some recent surveys of the field [10, 13] we see that even detecting non-occluded standing humans remains challenging. This is not surprising due to the great variety of backgrounds (scenarios, illumination) in which humans are present, as well as their intra-class variability (pose, clothe, occlusion). Nowadays, the most relevant baseline human detector relies on a (holistic) human classifier that uses the so-called histograms of oriented gradients (HOG) as features, and the support vector machines (SVMs) as learning method [7, 6]. New methods have been developed on top of this baseline in order to take into account relative pose of human parts [12], to handle occlusions [20], for taking advantage of color [19], etc.

One can deduce, from the state-of-the-art proposals in this field, that the most promising human detectors rely on classifiers developed by following the discriminative paradigm, *i.e.*, trained with labelled samples. Being, HOG and SVM key ingredients. However, labelling is a manual labor intensive step, especially, in cases like human detection where labelling objects (humans) means to provide at least bounding boxes. Note that this is more costly for a *human labeller* than just answering to *yes/no*-questions like *is there any human in this image?* (*i.e.*, without specifying *where* in the affirmative cases). In addition, it is well accepted that having sufficient variability in the labelled samples is decisive to train classifiers able to generalize properly [4]. However, traditional (passive) manual labelling do not evaluate the degree of variability achieved by the labelled samples. A common approach is assuming that the larger the set of labelled samples the higher the variability. However, just subjectively adding more examples does not guarantee higher variability, *e.g.*, it can happen that we are just adding human samples too similar to the ones we already collected. Accordingly, different authors have worked in the problem of reducing the labelling burden.

In [5] a hierarchy of synthesized (non-realistic) *pedestrian*[1] templates are used for pedestrian detection in far infrared images, *i.e.*, images capturing relative temperature. However, the authors admit poor performance and high computational cost.

In [9] a set of pedestrians is first manually segmented, and then different types of transformations (jittering, mirroring, shape deformations, texture variations, etc.) are applied to obtain joint pedestrian and background variability. A classifier is then learned following the discriminative paradigm. Local receptive fields with neural networks, and so-called Haar filters with SVM are tested. Since the mentioned transformations encode a generative 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 samples 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 samples with new manually labelled ones. However, the authors show that much of the improvement comes from enlarging the training set by applying jittering to the pedestrian samples as well as by introducing more background ones. Note that jittering does not involve synthesizing pedestrians since it only requires shifting their framing bounding box (assuming a little background frame around the pedestrian), *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 labelling much more labor intensive than standard bounding box framing of pedestrians.

In [1] a pure *active learning* technique is used. In particular, starting by 215 passively (arbitrary) labelled pedestrians and sufficient background samples, it is constructed a pedestrian classifier using an AdaBoost cascade, where the weak rules are single-feature decision stumps, and the features are 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 collected (labelled) 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 (determined by hold-out validation in the labelled data, 2/3 for training and 1/3 for testing).

Finally, in [14] it is proposed the use of a realistic videogame in order to capture labelled samples of pedestrians and background by *playing*. More specifically, a *driver* moves a virtual car equipped with a forward facing virtual camera along the road of a virtual city, and all the pedestrians appearing in the image are automatically extracted up to a precise pixel-level segmentation (from which it is trivial to obtain a bounding box). The pixels of the image not labelled as pedestrian pixels are considered background. The challenge then is to see if the appearance of the virtual pedestrians and background is sufficiently realistic to lead to a pedestrian model that can be successful applied in real images. For that, again, a discriminative paradigm is followed. In particular, HOG features and linear SVM are used. The presented results show that, when using HOG/linear-SVM, the pedestrian classifier trained with only virtual data is totally equivalent, in terms of performance, to its counterpart trained using real images.

In this paper, we are not interested only in pedestrian detection as in [5, 9, 1, 14], but in detecting humans out of cities as well. On the other hand, with these works we share the interest of reducing manual labelling. In fact, collecting good data at low cost for training appearance-based classifiers is a new research area as reviewed in [3]. Roughly speaking,

[1] We use the term *pedestrian* to refer to a human as a traffic participant.

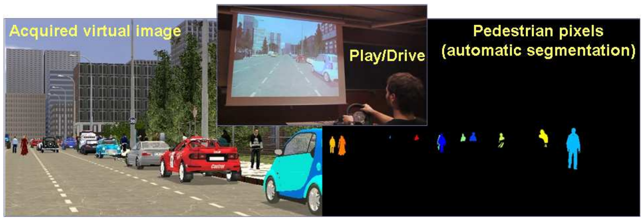


Figure 2: Our framework for acquiring virtual images with pixel-level groundtruth of pedestrians.

we establish three levels of labelling from the point of view of object detection: (1) LIL: image level (*is there an object inside this image? yes/no*); (2) LBBL: bounding-box level (*e.g.*, *framing objects inside the images with rectangles*); (3) LPL: pixel level (*e.g.*, *delineating the silhouette of the objects present in the images*). In [3] there are reviewed several techniques to automatically collect labelled samples from the internet using context (*e.g.*, the captions of the figures). This is not trivial and it is focused only on LIL, which is useful for applications like image retrieval. However, for applications that require locating objects inside an image only manual-based labelling of type LBBL and LPL is reported (*e.g.*, using the web-based Amazon's Mechanical Turk[2]). Note, that LBBL is far more costly than LIL, and LPL much more expensive than LBBP. In fact, in order to reduce such a tedious tasks, some approaches [18] try to transform it in a web-based interactive game.

According to the exposition done so far, we bet for the approach proposed in [14]. There are several reasons for that, which we summarize in the following. Videogames are reaching very high degrees of realism, not only at the objects level but even reproducing the characteristic spectral slope of natural images [15]. Besides, videogame industry is one of the most powerful worldwide, so players are a legion. In addition, the near future will be internet-based playing of realistic videogames, thus, there would exist the possibility of collecting labelled data from many virtual scenes in centralized special sites (game servers). In fact, the own computers could play the games, so that human players would not be really needed. Note, that videogames have generative models behind (scene formation, character design, physic laws, artificial intelligence, etc.). Thus, precise contextual information can be obtained, not only noisy context like when using the internet for collecting the data [3].

Currently, like in [14], we have a software that can obtain a continuous labelling of pedestrian pixels by driving through a virtual city (Fig. 2), *i.e.*, groundtruth of type LPL, the most difficult to collect manually. This is done frame by frame, thus not only appearance-based features can be used, but also temporal features as well as temporal processes (*e.g.*, tracking as in [16]). Note, that such temporal information can not be obtained by approaches like [5, 9, 1] because it is totally prohibitive in practice. In the future we will use more than one virtual camera, so that even stereo data acquisition can be simulated (similar to [17]).

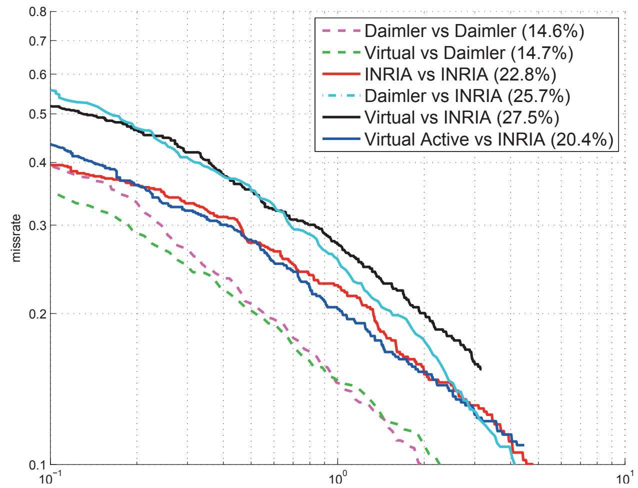


Figure 3: Per-image evaluation of different pedestrian detectors. The notation *DB1 vs DB2*, means that the corresponding classifier was learnt using *DB1* training data, and evaluated in *DB2* testing data. Daimler refers to the data sets used in [10, 14] for pedestrian detection (videos taken from a car inside a city). INRIA refers to a widely used data set for human detection first introduced in [7, 6] (it contains photographic images). Virtual refers to the data sets we have collected in our virtual city (with automatic labelling of pedestrians). In all cases the percentage inside the parenthesis indicates the missrate for one false positive per image. Virtual active refers to the approach in this paper (Fig. 1).

In order to address our problem of human detection, we followed the approach in [14], *i.e.*, we developed a pedestrian detector as follows. We collected a data set of images acquired at urban scenarios with labelled pedestrians. Then we learnt a HOG/linear-SVM classifier with such virtual labelled samples. For building the whole pedestrian detector, the pedestrian classifier must be completed with a previous stage that selects windows to be classified from the image (pyramidal windows search) and with a posterior stage that eliminates redundant detections due to shift-invariance of the classifier (non-maximum-suppression) [13]. For such pre- and post-classifier stages we also followed the suggestions in [14]. Finally, we tested the learnt pedestrian model in the video sequences of Daimler A.G. [10] as done in [14]. The obtained *per-image*[3] performance can be seen in Fig. 3 compared to the one obtained by learning the pedestrian classifier using manually labelled pedestrians of Daimler real-world images. In both types of training, virtual and real, we used the same amounts of samples. Basically, the conclusion

[2] www.mturk.com

[3] In the so-called *per-image* evaluation it is usually plotted the number of false positives (pedestrians) per image vs the missrate (ratio of undetected pedestrians). The overlapping measure between detections and groundtruth, required for such evaluation, is the proposed in the PASCAL VOC challenge [11]. It is worth to mention that one false positive per image is an interesting point of such performance curve for pedestrian detection in the driving assistance context, since, if such noise is not correlated from frame to frame, then it could be easily removed by temporal coherence analysis.

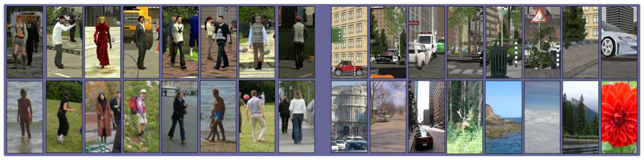


Figure 4: Top: virtual pedestrians and city scenarios. Bottom: INRIA photographs with humans and diversified scenarios as city, countryside, beach, etc. Humans appear also in such scenarios. Domain adaptation by batch active learning (Fig. 1) will bring together virtual samples and difficult real ones to learn real-world human classifiers.

is the same than in [14], namely, both pedestrian models lead to pedestrian detectors of analogous performance.

For the more general problem of human detection we followed the same scheme than before. In this case our experiments where based on the widespread INRIA dataset for human detection [7, 6]. The differences between Daimler and INRIA datasets is that while the former is composed of video sequences of urban scenarios, the latter is composed of photographic pictures of people in different environments (city, countryside, beach, etc.). Besides, humans in INRIA were captured with more resolution than pedestrians in Daimler. Thus, using our virtual samples, we trained a new pedestrian classifier for INRIA using a canonical pedestrian window of higher resolution than for Daimler. Then, we applied the corresponding pedestrian detector to the testing set of INRIA dataset. Again, we compared the obtained results with the counterpart human classifier learnt from the training data of INRIA dataset, *i.e.*, using images of the real world. As can be appreciated in Fig. 3, the performance provided by the virtual-based classifier is significatively worse than the provided by the real-based one (the real-based one is giving the performance reported in [7, 6]). The doubt here is whether the difference comes from the virtual-vs-real training style, or just because human detection is different than pedestrian detection in the sense that the former must deal with more types of environments (not only cities) and with more pose variability than the one of pedestrians (most can be catalogued as side/frontal/rear-views while walking). In other words, the doubt is if the HOG/linear-SVM scheme fails to be robust to world changes or if we have a problem of *domain adaptation* [2], or both. In order to asses this question, we also learnt a pedestrian classifier adapted to INRIA resolution by using Daimler training data (upscaling the images was required). The results of applying the corresponding detector in the INRIA training set are plotted in Fig. 3 too. Note that they are analogous to those obtained with the virtual-based pedestrian detector. It is worth to mention that the number of pedestrians/humans used for training was the same independently of the training data (virtual, INRIA and Daimler), and the same for background examples (Fig. 4 shows samples from virtual and INRIA datasets). Thus, we argue that, in fact, we are facing a problem of domain adaptation.

Accordingly, in the context of virtual worlds we have the option of developing other environments out of cities in which to capture virtual humans with corresponding labelling. However, if the HOG/linear-SVM scheme is not totally world invariant we still could have troubles to reach the performance of classifiers based on real-world images for training. The same can happen with other features, since HOG/linear-SVM scheme still remains as a state-of-the-art baseline [8, 19]. Thus, in this paper we propose to face the domain adaptation problem. Therefore, we assume that a small amount of manual annotations from real-world images is required. In particular, in order to transform virtual-world learnt pedestrian classifiers into real-world human classifiers, we explore an *active learning* scheme (summarized in Fig. 1) that brings together virtual-world automatically labelled data and difficult-to-classify real-world data actively labelled by a human oracle. As far as we know, such across-world training has not been done before in the field of human detection.

Comparing our approach to previously mentioned works we observe the following. In [9] the major benefit came from jittering, as we have mentioned, including jittering and mirroring is always easy (in fact, training done for obtaining Fig. 3 already incorporate such operations). Besides, further work using temporal features is not possible, while with virtual data is. In addition, in [9] it is not tested the state-of-the-art HOG/linear-SVM baseline. Moreover, [9] requires LPL for a set of pedestrians in order to initialize the proposed generative model, while virtual-world-based approach does not require initial manual labelling. In [1] such baseline is neither tested. Besides, neither the used video sequences nor the code are publicly available. In addition, the proposed active learning scheme can lead to a high number of *yes/no*-questions. Moreover, some *yes*-answers can be given for pedestrians just roughly aligned in the detected window (extrem jitter) which can significantly drop further training based on densely computed features as widespread HOG and Haar, *i.e.*, the most promising ones [8, 19]. As we will see, our proposal does not suffer from that. We share with [1] the use of a batch approach during active learning, for computation efficiency. However, as we will see, our selective sampling scheme is different since in [1] the training and testing domains are the same, *i.e.*, images captured from the same camera, which is not our case. Again, [1] requires LBBL for a set of pedestrians in order to learnt the initial pedestrian classifier, while virtual-world-based approach does not require initial manual labelling. With respect to [14], the approach we present here incorporates domain adaptation by batch active learning in order to transform pedestrian detectors in human detectors by jointly considering virtual-world labelled data and a small amount of real-world manually labelled data. In fact, we will show how just by using the 25% of the labelled data of the INRIA training set, together with the virtual training set, we achieve the same or better performance than by using the whole INRIA training set alone (Fig. 3). Additionally, note that in our proposal, the input of the human users (those contributing in the process of training classifiers) is of multimodal nature: player/driver and oracle roles.

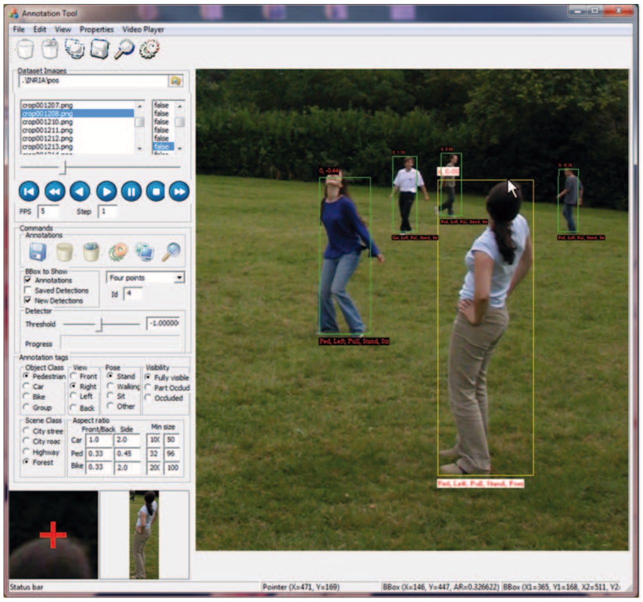
The rest of the paper is organized as follows. In Sect. 2 we provide more details of our proposal. In Sect. 3 we draw our experiments, discussing the corresponding results. Finally, section Sect. 4 summarizes the main conclusion of our work.

2. FROM PEDESTRIAN TO HUMAN CLASSIFIERS USING ACTIVE LEARNING

Let us start by introducing some notation and concepts. We denote by $\mathcal{D}_s$ and $\mathcal{D}_t$ two domains from which we observe samples. We refer to $\mathcal{D}_s$ as the *source* domain, while $\mathcal{D}_t$ is the *target* domain. Our problem is that given a sample $x_t \in \mathcal{D}_t$, we want to know if $x_t \in w_t$, using w_t to denote the samples in $\mathcal{D}_t$ with a particular property in which we are interested in. We want to face this problem by learning a classifier $\mathcal{C}$ able to answer if $x_t \in w_t$. To learn $\mathcal{C}$ we want to follow a discriminative paradigm, *i.e.*, learning from labelled samples. If $x_t \in \mathcal{D}_t$, its corresponding label ℓ_{x_t} equals $+1$ if $x_t \in w_t$ and -1 otherwise. It turns out that we have very few labelled samples drawn from $\mathcal{D}_t$ as to learnt a reliable classifier. However, we have sufficient labelled samples drawn from $\mathcal{D}_s$. If the distributions of the samples in $\mathcal{D}_s$ and $\mathcal{D}_t$ are uncorrelated, then we have nothing to do. However, if they have a sufficient correlation, then we are facing a problem of *domain adaptation* [2]. More specifically, we can use the large amount of labelled data from $\mathcal{D}_s$ and a low amount of labelled data from $\mathcal{D}_t$ to learn a $\mathcal{C}$ with chances of succeeding in the task of classifying unseen samples from $\mathcal{D}_t$. Roughly speaking, our $\mathcal{D}_s$ is the set of image windows cropped from virtual images, and our $\mathcal{D}_t$ the set of image windows cropped from the real-world images in which we want to detect humans. A sample x_t is just an image window, w_t is the property of imaging a human (*human* class), and $\mathcal{C}$ a human classifier.

Since we can collect in a cheap way as many examples as we need from our virtual cities, the setting for $\mathcal{D}_s$ holds. However, we assume that we start with no labelled samples from $\mathcal{D}_t$. As we have seen in Sect. 1 (Fig. 3), a pedestrian classifier trained on virtual samples works pretty well when applied to real-world video sequences of city driving. Then we can assume that there is sufficient correlation between $\mathcal{D}_s$ and $\mathcal{D}_t$, at least to the *eyes* of the features and base learning machine we use, *i.e.*, as we have already mentioned in Sect. 1, HOG features and linear SVM. Of course, as we deduce also from results in Fig. 3, $\mathcal{D}_s$ and $\mathcal{D}_t$ are not equal at all. In our case, $\mathcal{D}_t$ is more general (*i.e.*, human detection is more general than pedestrian detection) because more types of scenarios are faced ($\mathcal{D}_s$ is urban like).

Therefore, our problem reduces to obtain some labelled samples from $\mathcal{D}_t$, in a cheap way. For that, our proposal consists in following an *active learning* procedure using a *human oracle* to label *difficult samples* from $\mathcal{D}_t$. Usually, the difficult samples are defined as those falling in the ambiguity region of the base classifier at hand. For instance, in the case of a SVM this may correspond to the area inside the margins. However, in these cases, $\mathcal{D}_s$ and $\mathcal{D}_t$ are, in fact, the same distribution and the aim is to label as few samples as possibles but being meaningful. Our case, however, is different. Let us say that $\mathcal{C}_s$ has been learnt from $\mathcal{D}_s$ (using HOG and linear SVM) and that $x_t \in \mathcal{D}_t \wedge x_t \in w_t$. If $\mathcal{C}_s(x_t)$ is a negative value, large in magnitude, it turns out that from the viewpoint of $\mathcal{D}_s$, x_t is far from being in w_t, from imaging a human in our case. In our domain adaptation proposal, we do not consider such x_t as an outlier. On the contrary, these are the informative samples for adapting the domains, *i.e.*, the samples that must label the human oracle.



.\INRIA\pos

crop001207.png	false
crop001208.png	false
crop001210.png	false
crop001211.png	false
crop001212.png	false
crop001213.png	false

FPS 5 Step 1

☑ Annotations
☐ Saved Detections
☑ New Detections

Four points
Id 4

Threshold -1.00000
Progress

Object Class: ● Pedestrian ○ Car ○ Bike ○ Group
View: ○ Front ● Right ○ Left ○ Back
Pose: ● Stand ○ Walking ○ Sit ○ Other
Visibility: ● Fully visible ○ Part Occlud ○ Occluded
Scene Class: ○ City stree ○ City roac ○ Highway ● Forest

	Front/Back	Side	Min size	
Car	1.0	2.0	10(	50
Ped	0.33	0.45	32	96
Bike	0.33	2.0	20(	100

Pointer (X=471, Y=169) BBox (X=146, Y=447, AR=0.326622) BBox (X1=365, Y1=168, X2=511, Y2

Figure 5: Labelling tool. For each displayed image, the human oracle (Fig. 1) does the following task: (1) if there are not humans, it marks the image as *human-free*; (2) if there are humans, some of them have been detected by the previous classifier (green bounding box), but others may not (not framed). The undetected humans must be manually framed by the human oracle (yellow bounding box).

Accordingly, given a collection of real-world images it is processed using $\mathcal{C}_s$ to detect pedestrians. Detections are kept. By detections we consider those image windows x_t for which $\mathcal{C}_s(x_t) > th$. For our SVM, $|\mathcal{C}_s(x_t)| \geq 1$ means to be out of the learnt margins. Then, it is started a working session in which such images and detections are presented to the human oracle. The responsibility of the oracle is to say if a given image contains no humans (*yes/no*-question) and to label missed humans with a rectangular bounding box (Fig. 5). Once the whole sequence is processed by the oracle, a new classifier is trained using the labelled samples that where used to build $\mathcal{C}_s$ (virtual-world ones) as well as the new collected difficult samples (real-world ones). This type of active learning is termed as *batch mode*, because a set of images is processed before re-training. The overall approach is summarized in Fig. 1. We think that a noticeable fact is to use virtual- and real-world samples to train a human classifier, something not done before up to the best of our knowledge. This kind of process can be iterated.

Some additional details are that: (1) each real-world sample labelled by the oracle is mirrored to duplicate the number of positives; (2) for each new positive we collect ten negative ones (because everything not being a human is background) from the images labelled as *human-free*, we call this $1:10$ ratio and it is pretty common in human detection [7, 10, 14]. If our system must sample N negatives, it selects the N closest to th (and larger) according to the classification score. The initial $\mathcal{C}_s$ is learnt following such ratio as well.

3. EXPERIMENTAL RESULTS

Datasets. In this section we conduct a series of experiments for assessing the goodness of the proposal sketched in Sect. 2. However, instead of actually having a human oracle *working actively*, we will use a dataset passively labelled beforehand by a human oracle. In this way we can compare fully passive labelling with *simulated* active one. In particular, as we pointed out in Sect. 1, we will use the widespread INRIA dataset for human detection [7, 6]. This dataset is divided in separated sets of null intersection for training and testing, say $\mathcal{I}^{\text{train}}$ and $\mathcal{I}^{\text{test}}$, resp. The training set contains 2,416 *positive* samples consisting in image windows (original plus vertical mirror, *i.e.*, 1,208 manually labelled samples), each one containing a human framed by certain amount of background. We term this set of windows as $\mathcal{I}_+^{\text{train}}$. For collecting *negative* samples, *i.e.*, image windows that do not contain humans, there are 1,218 human-free available images. We term this set of images as $\mathcal{I}_-^{\text{i,train}}$. Windows are randomly collected from $\mathcal{I}_-^{\text{i,train}}$ to fulfil a ratio of ten negatives per one positive sample (1 : 10 ratio), we term $\mathcal{I}_-^{\text{train}}$ the collected negative windows. All positive and negative windows are down-scaled to a canonical window size. After this, $\mathcal{I}^{\text{train}} = \mathcal{I}_+^{\text{train}} \cup \mathcal{I}_-^{\text{train}}$. The testing set consists of: (1) $\mathcal{I}_-^{\text{i,test}}$: 453 human-free images; (2) $\mathcal{I}_+^{\text{i,test}}$: 288 images containing 563 labelled humans (ground truth). Then, $\mathcal{I}^{\text{test}} = \mathcal{I}_-^{\text{i,test}} \cup \mathcal{I}_+^{\text{i,test}}$.

Passive learning. As we mentioned in Sect. 2 we use HOG features and Linear SVM learning machine for training human/pedestrian classifiers, in both cases with the parameters identified in [7, 6] as the best, applying also the mirroring technique. Accordingly, we train the human classifier using $\mathcal{I}^{\text{train}}$ and the pedestrian one using $\mathcal{V}^{\text{train}} = \mathcal{V}_+^{\text{train}} \cup \mathcal{V}_-^{\text{train}}$. The cardinality of $\mathcal{I}_+^{\text{train}}$ and $\mathcal{I}_-^{\text{train}}$ equals the one of $\mathcal{V}_+^{\text{train}}$ and $\mathcal{V}_-^{\text{train}}$, resp. During training, *bootstrapping* is used, *i.e.*, appending the respective negative training sets with hard negative samples and re-training. Hard negatives are collected from the corresponding negative training images by applying the initially learnt classifier. The process is iterated until very few new negatives are incorporated. In practice, these training sets saturate with a single step. Let us refer by $\mathcal{C}_{\mathcal{I}}^{\text{pas}}$ to the passively learnt classifier based on $\mathcal{I}^{\text{train}}$, and by $\mathcal{C}_{\mathcal{V}}^{\text{pas}}$ to the equivalent one based on $\mathcal{V}^{\text{train}}$.

Active learning. Given $\mathcal{C}_{\mathcal{V}}^{\text{pas}}$, we have conducted several experiments following our proposal (Sect. 2), where $\mathcal{I}^{\text{train}}$ is used as real data set for performing the (simulated) active learning. We denote by $\mathcal{C}_{\mathcal{V}}^{\text{act}}$ to any classifier learnt by using $\mathcal{V}^{\text{train}}$ and samples actively collected from $\mathcal{I}^{\text{train}}$.

Discussion. Experiments have been conducted to give different insights about our proposal (*e.g.*, dependence of the results on th, on the $p : n$ ratio, etc.), always using $\mathcal{I}^{\text{test}}$ for testing. Performance curves of Fig. 6 summarize the results obtained in our experiments, and draw us to the following observations:

- According to Fig. 6-(1), even introducing a 2% ($th = -2$, 27 new manually labelled humans) of samples from $\mathcal{I}^{\text{train}}$ (ratio 1 : 10, *i.e.*, one from $\mathcal{I}_+^{\text{train}}$ and ten from $\mathcal{I}_-^{\text{train}}$) provides a $\mathcal{C}_{\mathcal{V}}^{\text{act}}$ of better performance than $\mathcal{C}_{\mathcal{V}}^{\text{pas}}$. Using $th \leq -1$ (*i.e.*, negative score out of the uncertainty area of $\mathcal{C}_{\mathcal{V}}^{\text{pas}}$) implies to improve performance (the best is at the margin border (*i.e.*, in $th = -1$), which implies to consider the 20% of samples from $\mathcal{I}^{\text{train}}$ (manually labelling 246 new humans[4]).

- It is also worth to mention that we trained a classifier using only the actively collected real-world samples from $\mathcal{I}^{\text{train}}$ ($th = -1$, 1 : 10 ratio). The performance was quite poor (100% missrate at one FPPI), thus, denoting that such real samples are a complement of the virtual ones, but they are not useful on their own.

- According to Fig. 6-(2,3), introducing either only positive or only negative samples from $\mathcal{I}^{\text{train}}$ (*i.e.*, from $\mathcal{I}_+^{\text{train}}$ and $\mathcal{I}_-^{\text{train}}$, resp.) can improve a little the performance of $\mathcal{C}_{\mathcal{V}}^{\text{pas}}$, again if $th \leq -1$, but still remaining far from $\mathcal{C}_{\mathcal{I}}^{\text{pas}}$.

- According to Fig. 6-(4), it seems that the ratio 1 : 10 for introducing actively obtained samples from $\mathcal{I}_+^{\text{train}}$ and $\mathcal{I}_-^{\text{train}}$, is a good compromise. The ratio 1 : 1 already offers significative performance improvements with respect to $\mathcal{C}_{\mathcal{V}}^{\text{pas}}$.

- According to Fig. 6-(1,5), it seems that by randomly sampling $\mathcal{I}^{\text{train}}$ in equivalent amounts to the considered active selection of samples (*i.e.*, for $th \in \{-2.0, -1.5, -1.0, -0.5, 0.0\}$), we obtain similar performance improvements. In fact, the active method is slightly better. In this context, such a random sampling would correspond to the human oracle labelling humans from $\mathcal{I}^{\text{train}}$ on his own. However, it is unclear if by doing so the human oracle would actually provide random samples, it may be the case that he/she is biased to cases that likes more. Thus, altogether led us to recommend the active paradigm, after all, the labelling effort is the same.

- Finally, Fig. 6-(6) simulates the case in which the human oracle only labels a certain percentage of the undetected humans. We see that even only labelling a 25% (61 new labelled humans) of the cases already provides a $\mathcal{C}_{\mathcal{V}}^{\text{act}}$ clearly better than $\mathcal{C}_{\mathcal{V}}^{\text{pas}}$. Missing the 25% the the labelling suggested by the active procedure is unnoticeable (185 humans would be labelled).

After this analysis there is still a remaining question: *what type of humans leaves active learning for manual labelling?*. To answer, we classified all the humans in $\mathcal{I}^{\text{train}}$ according to the scenario where they are. The defined scenarios are: city, beach, countryside, indoor, and snow. There are, resp., 916, 50, 138, 87, and 17 humans. Thus, 292 humans are not from the city. Therefore, most of the humans are of the type used for training $\mathcal{C}_{\mathcal{V}}^{\text{pas}}$ (*i.e.*, pedestrians) but the 32% are not. This suggests that $\mathcal{C}_{\mathcal{V}}^{\text{pas}}$ should not be very far away from $\mathcal{C}_{\mathcal{I}}^{\text{pas}}$ in performance, as it happens, but it is also to be expected $\mathcal{C}_{\mathcal{I}}^{\text{pas}}$ being better. Of course, we are assuming

[4] Experiments with the software of Fig. 5 reveal that manually labelling 250 human bounding boxes is a matter of around 25 minutes.

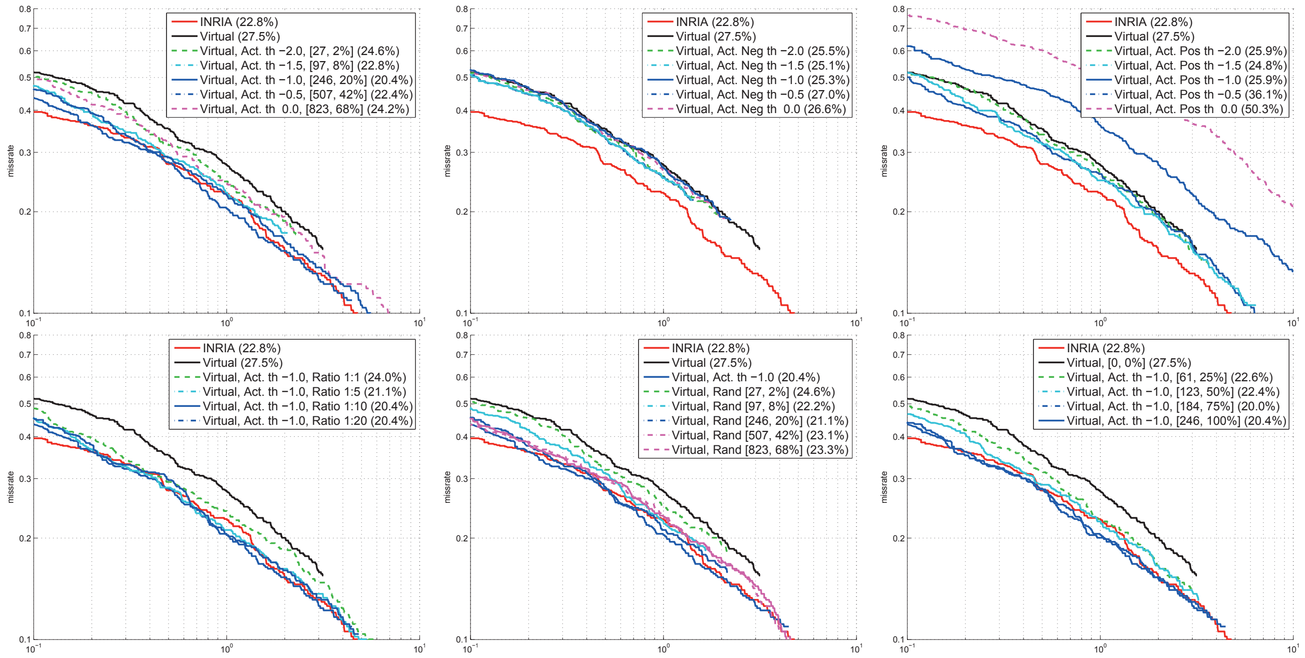


Figure 6: In all cases the percentage inside the parenthesis corresponds to the missrate at one FPPI. All the curves correspond to testing in $\mathcal{I}^{\text{test}}$. In all cases labelled as 'Virtual', the full $\mathcal{V}^{\text{train}}$ set is used during training. The different 'Active' curves correspond to different ways of complementing $\mathcal{V}^{\text{train}}$ by actively taking examples from $\mathcal{I}^{\text{train}}$. *Top, from left to right*: (1) Results for different th, $[N, P]$ refers to the number of samples (N) chosen from $\mathcal{I}_{+}^{\text{train}}$ given th, and P is the corresponding percentage. From $\mathcal{I}_{-}^{\text{train}}$ we take $10N$, *i.e.* $1:10$ ratio. (2) Analogous to '(1)' but only considering the samples taken from $\mathcal{I}_{-}^{\text{train}}$. (3) Only considering the samples from $\mathcal{I}_{+}^{\text{train}}$. *Bottom, from left to right*: (4) Analogous to '(1)' but changing the ratio of positive vs negative samples actively taken from $\mathcal{I}_{+}^{\text{train}}$ and $\mathcal{I}_{-}^{\text{train}}$, resp., running from $1:1$ to $1:20$. (5) Analogous to '(1)' but instead of using the th rule we just sample randomly $\mathcal{I}_{+}^{\text{train}}$ and $\mathcal{I}_{-}^{\text{train}}$, taking the same percentage of samples than for the considered th in '(1)'. (6) Results for $th = -1.0$ but rather than taken all the samples of $\mathcal{I}_{+}^{\text{train}}$ fulfilling such threshold condition, we only take the percentage in brackets. Of course, we take also the number of samples from $\mathcal{I}_{-}^{\text{train}}$ that keeps the $1:10$ ratio.

that the data distributions of $\mathcal{I}^{\text{train}}$ and $\mathcal{I}^{\text{test}}$ are highly correlated as is to be expected too. We have checked that the active learning procedure suggests to label the following distribution of humans, according to the same scenarios: 156, 14, 24, 46, and 6. These numbers correspond to the 17.03%, 28.00%, 17.39%, 52.87%, and 35.29% of each scenario, resp. Note, that 90 humans to be labelled are not from the city (30% of the ones to be labelled). This analysis confirms that some pedestrians still where not well represente by the $\mathcal{C}_{\mathcal{V}}^{\text{pas}}$ model, however, most of the badly represented in percentage correspond to indoor, snow, beach, countryside and, finally, city. We think that these results reinforce our approach of casting the performance gap between $\mathcal{C}_{\mathcal{I}}^{\text{pas}}$ and $\mathcal{C}_{\mathcal{V}}^{\text{pas}}$ as a domain adaptation problem, and that batch active learning can be effective to address it.

As summary we include Fig. 7, which plots the per-image performance of $\mathcal{C}_{\mathcal{I}}^{\text{pas}}, \mathcal{C}_{\mathcal{V}}^{\text{pas}}$ and the best $\mathcal{C}_{\mathcal{V}}^{\text{act}}$ ($th = -1$, ratio $1:10$). Additionally, we include the performance of a classifier obtained by passively learning using all the examples in $\mathcal{I}^{\text{train}}$ and $\mathcal{V}^{\text{train}}$. Note, that just mixing samples from different worlds in a blind way just harms the performance of both $\mathcal{C}_{\mathcal{I}}^{\text{pas}}$ and $\mathcal{C}_{\mathcal{V}}^{\text{pas}}$.

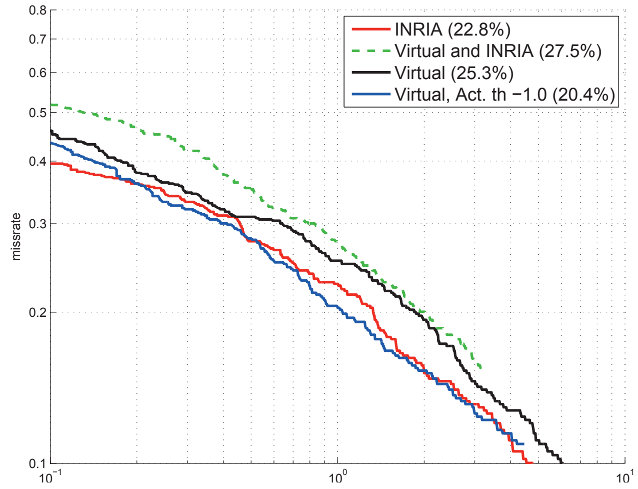


Figure 7: Final comparative results.

4. CONCLUSION

In this paper we have addressed a core problem in the field of human detection, namely, the acquisition at low cost of good samples to train. In order to collect most of the human and background samples we rely on players/drivers of a videogame, *i.e.*, we automatically collect labelled samples while enjoying a game. With them we learn a virtual-world based pedestrian classifier that must work as a human classifier in images depicting the real world. In city scenarios, those where the pedestrian classifier is initially trained, the exhibited performance is equivalent to the one obtained by a counterpart real-world based classifier, *i.e.*, requiring costly manual labelling. However, in scenarios out of the city, both types of classifiers can not reach the performance of a classifier learnt using data manually labelled for training in such new scenarios. In order to keep the advantage of the cost-free labelling in virtual-worlds, we have cast the problem of transforming the virtual-world based pedestrian classifier into a human classifier for real world images of general scenarios, as a domain adaptation problem. To perform the adaptation, we have proposed a batch active learning technique that, with just a few manually labelled humans from the real images, is able to reach the same performance than a human classifier entirely trained from a much large amount of manually labelled data. Ultimately, or human classifier has been trained by using samples from virtual and real worlds, which is totally new in the field of appearance based human detection up to the best of our knowledge. We observe that, in a way, we have adopted a multimodal approach from two view points: (1) using two different types of raw data (virtual and real), and (2) collecting the data by playing in the one hand and by working on the other. Finally, we would like to mention that our proposal can be extended in the future in several ways, *e.g.*, detecting other targets and incorporating spatio-temporal features, just to mention a few. Using incremental learning machines, *e.g.* incremental SVM, will be also assessed with the aim of eliminating the batch mode from the active learning to see if the manual labelling can be even more reduced.

Acknowledgments

This work was supported by the Spanish MICINN, in particular by projects TRA2010-21371-C03-01 and Consolider Ingenio 2010: MIPRCV (CSD200700018).

5. REFERENCES

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

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

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

[5] 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.

[6] N. Dalal. *Finding people in images and videos*. PhD Thesis, Institut National Polytechnique de Grenoble / INRIA Rhône-Alpes, 2006.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] J. Marín, 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.

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

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

[17] W. van der Mark and D. M. Gavrila. Real-time dense stereo for intelligent vehicles. *IEEE Trans. on Intelligent Transportation Systems* , 7(1):38–50, 2009.

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

[19] 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.

[20] 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.