# Cool world: domain adaptation of virtual and real worlds for human detection using active learning

**David Vázquez**
CVC and C. Sc. Dpt. UAB, Barcelona, Spain
david.vazquez@cvc.uab.es

**Antonio M. López**
CVC and C. Sc. Dpt. UAB, Barcelona, Spain
antonio@cvc.uab.es

**Daniel Ponsa**
CVC and C. Sc. Dpt. UAB, Barcelona, Spain
daniel@cvc.uab.es

**Javier Marin**
CVC, Barcelona, Spain
jmarin@cvc.uab.es

## 1 Introduction

Image based human detection is of great interest due to its potential applications. However, even detecting non-occluded standing humans remains challenging [4]. This is not surprising due to the large 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 linear support vector machines (Lin-SVM) as learning method [2]. New methods have been developed on top of this baseline, most of them following a discriminative learning paradigm. This means that human classifiers are trained from labelled samples. Labelling is performed by a human oracle. On the one hand, the oracle must select human-free images from which negative samples can be taken, *i.e.*, background windows. On the other hand, the oracle must draw a bounding box (BB) per each human sample of interest within non-human-free images, *i.e.*, positive samples are image windows framing humans. In practice, this implies that a core issue as having good samples to train, relies on a subjective and tiresome manual task.

In order to automatize and control the labelling process, we proposed the use of a realistic videogame in [5], *i.e.*, to capture labelled samples of pedestrians[1] 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 labelled. 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 successfully applied in real images. For that, we used the HOG/Lin-SVM baseline and Daimler real-world images [3]. The presented results show that the pedestrian classifier trained with only virtual-world samples is totally equivalent, in terms of *per-image average miss rate*[2] performance, to its counterpart trained using real-world ones. This is illustrated in Fig. 1, where *Daimler vs Daimler* (*i.e.*, training and testing sets are from Daimler real-world images) shows similar performance to *Virtual vs Daimler* (*i.e.*, training is based on our virtual-world images while testing is done with Daimler real-world ones). We use the same number of pedestrian samples for training, irrespective of coming only from real world or just from virtual world.

Afterwards, we were not only interested in pedestrian detection, but in detecting humans out of outdoor city scenarios as well. Then, we followed the same approach. However, in this case we based our experiments on the widespread INRIA dataset for human detection [2]. A major difference be-

[1] Hereinafter we use the term *pedestrian* to refer to a human as a traffic participant of an urban area.

[2] *Per-image* refers to curves of false positives (pedestrians) per image *vs* miss rate (ratio of undetected pedestrians). In pedestrian detection the FPPI range $[10^{-1}, 10^{0}]$ is of especial interest since a further temporal coherence analysis can reduce the FPPI. Thus, we compute the average miss rate for such a range.

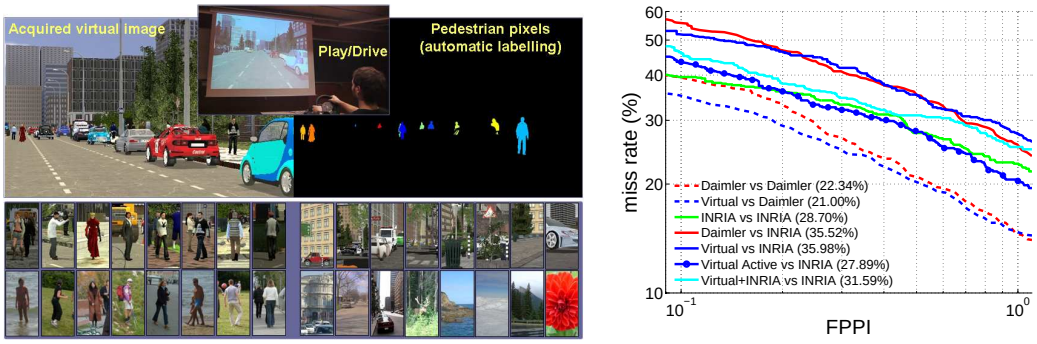


Figure 1: *Top-Left*: our framework for acquiring virtual images with pixel-level labelling of pedestrians. *Bottom-Left*: pedestrians and city background samples from virtual world; and real-world INRIA samples with humans within diversified scenarios as city, countryside, and beach. *Right*: per-image performance curves (average miss rate in parenthesis). 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 [3] 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 [2] (it contains photos of people and places). *Virtual* refers to the data sets we have collected in our virtual city (with automatic labelling of pedestrians). *Virtual Active* refers to the approach in Fig. 2.

tween Daimler and INRIA datasets is that while the former is composed of video sequences of urban scenarios, the latter is composed of photos of people in different environments (city, countryside, beach, etc.). Figure 1 shows comparative results between the cases *INRIA vs INRIA* (INRIA training and testing sets) and *Virtual vs INRIA* (virtual-world training, INRIA testing). Training with virtual data is significatively worse now (7.28 points of average miss rate). The question is if the difference comes from the virtual-*vs*-real training-testing style, or if it is because human and pedestrian detection are different in the sense that the former deals with more environments (not only cities) and poses than the latter (pedestrian poses are just side/frontal/rear-views while walking). In other words, if the HOG/linear-SVM scheme fails to be robust to *world changes* or if we have a problem of *domain adaptation* [1], or both. In order to assess this issue, we also learnt a pedestrian classifier using Daimler training data and the corresponding detector was tested in INRIA. Figure 1 shows that such *Daimler vs INRIA* experiment has the same problem than the *Virtual vs INRIA* one since it is 6.82 points below *INRIA vs INRIA*, *i.e.*, performances of *Daimler vs INRIA* and *Virtual vs INRIA* are similar (such 0.46 points of difference are not significative in pedestrian detection). Thus, we argue that, in fact, we are facing a problem of domain adaptation. We recently addressed this problem in [6]. Here we summarize such work, but using average miss rate metric to evaluate the performance.

## 2 Domain adaptation based on active learning and cool world

Let $\mathcal{D}_s$ and $\mathcal{D}_t$ be two different domains from which we observe samples. Where $\mathcal{D}_s$ is the *source* domain and $\mathcal{D}_t$ 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 $\ell_{x_t}$ 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 learn 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 *supervised domain adaptation* [1]. 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-world 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.

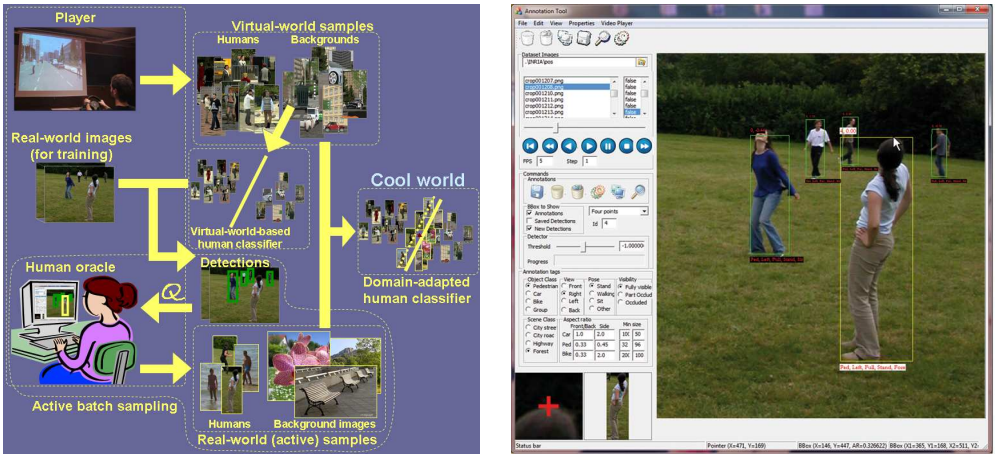


Figure 2: By playing a videogame we gather labelled samples of virtual humans and backgrounds, which are used to learn a human classifier. This classifier is applied to real-world images. In particular, for each real-word image, we ask a human oracle if it is a human-free one; if not, non-detected humans are labelled by him/her (right: green boxes are detections, yellow one is a miss detection being labelled). After, we create the *cool world* by combining the virtual-world samples and the actively collected real-world ones. Finally, a new classifier is learnt in such cool world. Thus, virtual world is adapted to real world by active learning, highly reducing the human-based labelling effort.

In order to perform the supervised domain adaptation, we follow an *active learning* procedure using a *human oracle* that labels *difficult samples* of $\mathcal{D}_t$. 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 $\mathcal{D}_t$ follow 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 $\mathcal{D}_s$ (with HOG/Lin-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 out of the ambiguity region, it turns out that from the viewpoint of $\mathcal{D}_s$, $x_t$ is far from being in $w_t$, *i.e.*, 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.

Accordingly, given a collection of real-world images it is processed using $\mathcal{C}_s$ to detect humans. 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, a working session is started 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 (Y/N-$\mathcal{Q}$uestion) and to label missed humans (BB-$\mathcal{Q}$uestion). From Y/N-$\mathcal{Q}$uestions we can collect false positives (*i.e.*, background windows classified as containing a human), while from BB-$\mathcal{Q}$uestions we collect false negatives (*i.e.*, non detected humans). Using these difficult real-world samples and the original virtual-word ones we can build what we term HOG-based *cool world*[3], *i.e.*, to learn a new (domain adapted) classifier, we combine the virtual- and real-world samples without origin distinction. Since a set of images is processed before re-training, this is a type of *batch mode* active learning. The overall approach is summarized in Fig. 2. This process can be iterated in order to incrementally obtain better classifiers.

Results in Sect. 1 lead us to focus the domain adaptation challenge on $\mathcal{D}_t \equiv$ INRIA. We have seen that *INRIA vs INRIA* performs 7.28 points of average miss rate better than *Virtual vs INRIA*. This can be expected since $\mathcal{D}_t$ includes more scenarios and human poses than $\mathcal{D}_s$. However, *Virtual vs INRIA* still gives only a $35.98\%$ of miss rate. Thus, $\mathcal{D}_s$ and $\mathcal{D}_t$ are somehow correlated for the HOG/Lin-SVM setting. Hence, the proposed supervised domain adaptation procedure is worth to be applied. For doing so, instead of actually having a human oracle *working actively*, we will use the INRIA training set passively labelled beforehand by a human oracle. In this way we can com-

[3]We use the *cool world* term as a tribute to the 1992 movie with that title. In this movie, there is a real world and a cool world, the latter shared by real humans and cartoons.

pare fully passive labelling with *simulated* active one. By setting $th$ to 1, we roughly bound the number of actively collected humans to the $20\%$ of the total number (all the collected passively, which includes the simulated active ones). Actively collected background samples follow the same proportion (before bootstrapping). The resulting performance is shown in Fig. 1 as *Virtual Active vs INRIA*, which is similar to *INRIA vs INRIA* (the former 0.81 points better than the latter), thus, achieving domain adaptation. It is worth to mention that *INRIA vs INRIA* requires the labelling of 1,208 real-world humans, while *Virtual Active vs INRIA* only 246 (1,208 virtual-world pedestrians + 246 real-world humans are used in this case for training). We checked that active labelling introduces humans from city, countryside, beach, snow, and indoor scenarios. In particular, the $17.03\%$, $17.39\%$, $28.00\%$, $35.29\%$, and $52.87\%$, resp., of those available per scenario. Thus, confirming that real-world humans out of outdoor city scenarios are complementing the virtual-world pedestrians within cool world. Fig. 1 shows as *Virtual+INRIA vs INRIA* the case of using all available virtual-world pedestrians and INRIA humans (*i.e.*, 2,416 labelled positives) to train. Note, that the result is worse than *Virtual Active vs INRIA* (3.7 points) and than *INRIA vs INRIA* (2.89 points). Which means that just blindly adding more real-world samples does not guarantee a better performance. Of course, a sort of domain adaptation holds since it improves 4.39 points regarding *Virtual vs INRIA*, but in this setting it has no sense to use the virtual world since all real-world samples are available.

## 3 Conclusion

We have addressed a core problem in the field of human detection, namely, the acquisition at low cost of good samples to train. We have build a cool world composed of automatically labelled virtual-world samples and a few actively collected real-world ones. Within such cool world we have learnt a human classifier that shows the same performance that its counterpart totally trained with real world images from the same domain than the testing images. Thus, through our cool world we have obtained a domain adapted classifier based on a relatively low human labelling effort. We consider our current framework for training human detectors as a *proof of concept* that can be further enhanced by new domain adaptation techniques, always looking for the best performance while minimizing the number of manual labelling.

## Acknowledgments

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

## References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

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