Interactive Training of Human Detectors

David Vázquez, Antonio M. López, Daniel Ponsa, and David Gerónimo

Abstract. Image based human detection remains as a challenging problem. Most promising detectors rely on classifiers trained with labelled samples. However, labelling is a manual labor intensive step. To overcome this problem we propose to collect images of pedestrians from a virtual city, *i.e.*, with automatic labels, and train a pedestrian detector with them. The resulting detector performs correctly when such virtual-world data are similar to testing one, *i.e.*, real-world pedestrians in urban areas. When testing data is acquired in different conditions than training ones, *e.g.*, human detection in personal photo albums, dataset shift appears. In previous work, we treat this problem as one of domain adaptation and solve it with an active learning procedure. In this work, we focus on the same problem but evaluate a different set of faster to compute features, *i.e.*, Haar, EOH and their combination. In particular, we train a classifier with virtual-world data, using such features and Real AdaBoost as learning machine. This classifier is applied to real-world training images. Then, a human oracle interactively corrects the wrong detections, *i.e.*, few miss detections are manually annotated and some false ones are pointed out too. A low amount of manual annotation is fixed as restriction. Real- and virtual-world difficult samples are combined within what we call *cool world* and we retrain the classifier with this data. Our experiments show that this adapted classifier is equivalent to the one trained with only real-world data but requiring 90% less manual annotations.

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

David Vázquez · Antonio M. López · Daniel Ponsa · David Gerónimo
Computer Vision Center (CVC) and the Computer Science Dept.
at the Autonomous University of Barcelona (UAB)
e-mail: {david.vazquez,antonio,daniel,dgeronimo}@cvc.uab.es

A.D. Sappa and J. Vitrià: *Multimodal Interact. in Image & Video Appl.*, ISRL 48, pp. 169–184.
DOI: 10.1007/978-3-642-35932-3_10

analysis. However, by reading some recent surveys of the field [8, 11, 7] 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 algorithm [5, 4]. New methods have been developed on top of this baseline in order to take into account relative pose of human parts [9], to handle occlusions [27], to take advantage of color [26], etc.

One important aspect of a human detector is its computational cost. HOG features are very effective but expensive to compute. Some works tried to speed up its computation by using integral histograms [18] or specific hardware [20], being the former the most promising one. Haar features combined with AdaBoost [23] were one of the first proposals for *pedestrian*[1] detection, and also made use of integral features. It was extended with Edge Orientation Histograms (EOH) features too [10]. More recently [6] presented a detector based on different integral features, such as color and gradient orientations, which is one of the best performing ones in the state of the art. This work was extended by [2] proposing the fastest pedestrian detector to the date, running at more than 100 frames per second.

One can deduce that the most promising human detectors rely on classifiers developed by following the discriminative paradigm, *i.e.*, trained with labelled samples, being integral features and AdaBoost key ingredients. However, labelling is a manual labor intensive step, especially in cases such as human detection in which 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 [3]. However, traditional (passive) manual labelling does not evaluate the degree of variability achieved by the labelled samples. A common approach is to assume that the larger the set of labelled samples the higher the variability. However, just subjectively adding more examples does not guarantee a higher variability, *e.g.*, it can happen that we are just adding human samples too similar to the ones we already collected.

In order to obtain good samples to train as well as significantly reducing human labelling effort, in [16] we used a video game to collect city images with automatically labelled pedestrians (Fig. 1). By using such virtual-world data we trained a pedestrian classifier that was used within a pedestrian detector operating in real-world images. We employed HOG as pedestrian descriptor and linear SVM as learning machine. The results provided by the virtual-world based pedestrian detector were equivalent to a counterpart detector which pedestrian classifier was trained on real-world images.

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

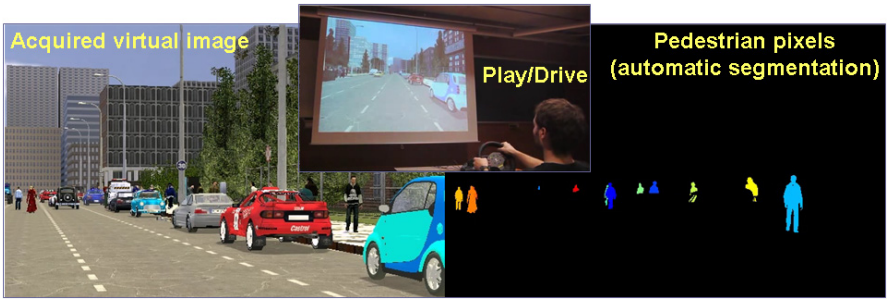

Fig. 1 Virtual-world images with pixel-level groundtruth of pedestrians

In [22, 21] we applied the same procedure for detecting humans in more general images, for instance, in holiday photos. In this case the performance shown by the virtual-world based pedestrian detector was far worse from the one obtained by its real-world counterpart. In fact, we illustrated how the problem remains the same when training and testing with real-world data coming from different domains. In other words, we were suffering dataset shift, but not because the training data was from virtual-world but just because it came from a domain different than the one where the pedestrian detector operated, *e.g.*, training with urban sequences and testing in landscape or indoor scenarios. Accordingly, we casted the problem in a domain adaptation framework based on active learning, *i.e.*, we followed a semi-supervised domain adaptation approach (Fig. 2). Such a framework allowed us to obtain the desired performance by combining our virtual-world data with just a few real-world one (25%) actively labelled.

In [16, 22, 21] we focused on HOG/linear-SVM. In this chapter we extend our study to Haar, EOH and Haar with EOH descriptors, employing AdaBoost as learning algorithm. This is an important setting since, as we mentioned before, it can lead to fast pedestrian detectors thanks to the use of integral images and decision cascades. We will see that Haar/EOH/HaarEOH with Adaboost also presents dataset shift. Fortunately, as for HOG/linear-SVM, we will show how our semi-supervised domain adaptation proposal provides the desired results in this case. We restrict more the number of allowed real-world pedestrian annotations, i.e., those done manually. In [16, 22, 21] we allowed the 25% of the virtual-world pedestrians; here we have reduced it to 10%. User interaction comes in the form of a human oracle who actively annotates some difficult samples from real-world images.

The rest of the chapter is organized as follows. Sect. 2 presents the HaarEOH AdaBoost human detection method. In Sect. 3 we show the details of the proposed semi-supervised domain adaptation algorithm. In Sect. 4 we present and discuss the obtained results. Finally, Sect. 5 summarizes our conclusions.

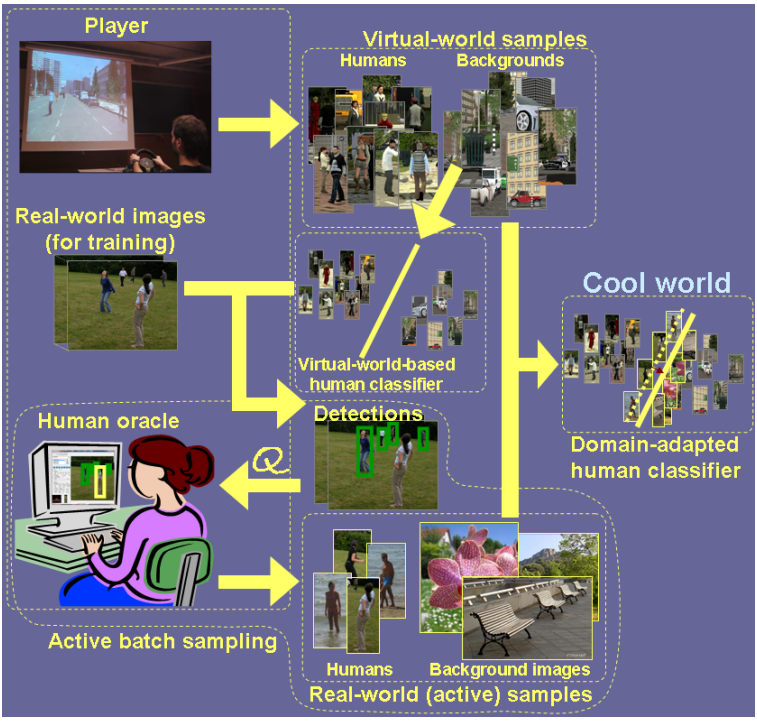

Fig. 2 Our proposal in a nutshell: domain adaptation based on active learning

2 HaarEOH-Based Pedestrian Detection

2.1 Detection Architecture

A pedestrian detector, *scans* an image with a window determining if it contains a pedestrian (*positive*) or not (*negative*) by using a learnt pedestrian classifier which comes from a *learning machine* process. The classifier gets the features computed over each window as input and its class, *i.e.*, positive or negative, as output. Since multiple positive windows can be detected for a single pedestrian, we must *select* a representative one, *i.e.*, the window *detecting* the pedestrian. Let us briefly review the features, learning machine, scanning, and selection.

2.1.1 Features

We use two different features that can be computed using the so-called integral image which has been demonstrated that speeds up object detection [25] and that recently is attracting much interest [6, 2].

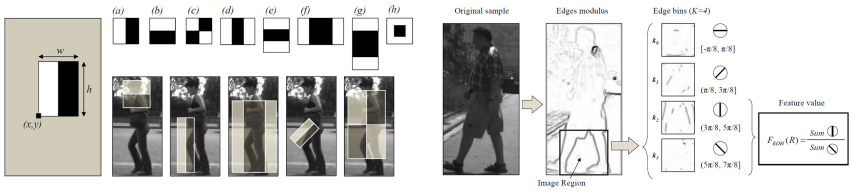

Fig. 3 *Left: Haar filters.* Example of a filter with parameters (x,y,w,h) with basic forms of the Extended Haar set and examples of filters that give high response in regions containing pedestrians. *Right: EOH features.* The feature is defined as the relation between two orientations of a region. In this case, vertical orientations are dominant with respect to the diagonal orientations (k_3), so the feature will have a high value.

Haar filters were introduced in [15] to detect pedestrians using a static camera. A single feature of this set is defined as the difference of illumination between two areas (white and black, see Fig. 3 left). The sum of the pixel values of a given region, $E(R)$, can be efficiently computed by only four accesses to the integral image. The feature value is:

$$Feature_{Haar}(R,f) = \frac{E(R_{white})E(R_{black})}{wh\sqrt{(E(R))^2 - E^2(R) + \varepsilon}},$$

where R_{white} and R_{black} are the white and the black rectangles of the filter f, and w, h refer to the width and height of rectangle R. $E^2(R)$ is the sum of the square of the pixels of a given region. Note that the denominator is a contrast normalization factor that depends on the region size and standard deviation. Originally, [15] presents three basic filters, namely (a)(b)(c) in Fig. 3. Posteriorly, [24] add filters (d) and (e) to the previous set in order to achieve face detection and for pedestrian detection using a static camera in [25]. This set is referred in this chapter as Simple Haar set. In our work, we use filters from (a) to (h), coming to use the Extended Haar set described in [14] to detect faces.

EOH features were proposed in [13] for face detection and used for pedestrian detection in [10] as well. These features are based on gradient information, which not only maintains invariance to global illumination changes, but is also able to extract shape information difficult to capture by Haar filters. This feature extracts similar information to the HOG feature, which is the standard pedestrian detection descriptor, but EOH can be easily computed by using integral images. Features are computed as follows (see Fig. 3). First, the image derivative is computed with a Sobel mask to get the edge orientation. Then, the derivative image pixels are classified according to its edge orientation into K (in our case $K = 6$) images corresponding to K orientation bins. Therefore, a pixel in bin $k_n \in K$ contains its gradient magnitude

if its orientation is inside k_n range, otherwise is null. Integral images are now used to store the accumulation image of each of the edge bins. Finally, the feature value is defined as the ratio between two orientations, k_i and k_j, of region R:

$$Feature_{EOH}(k_i, k_j, R) = \frac{E_{k_i}(R) + \varepsilon}{E_{k_j}(R) + \varepsilon}.$$

If the feature value is greater than a given threshold, then the orientation k_i is dominant to orientation k_j at the region R, which can be exploited as a weak hypothesis too. The small value ε is added to the factors for smoothing purposes.

2.1.2 Learning Machine

The feature descriptors we use are of high dimensionality. Accordingly, we need a machine learning algorithm able to work in such spaces. Boosting algorithms are the most suitable ones as they automatically select a subset of the features that best characterize the problem to learn. From the different boosting proposals, we use real AdaBoost [19], more specifically, we use the implementation of [17]. The key idea is to build a strong classifier by combining a set of weak classifiers.

In an iterative manner, real AdaBoost chooses the weak classifiers that best classify the training set. In the algorithm, each sample has a weight depending on prior classifications; this value is increased in case it has been missclassified by previous rules. Hence, at each iteration, the algorithm focuses its efforts on previously missclassified samples. Finally, the strong classifier is composed of n weak classifiers, where n is defined by the user and usually is much lower than the total number of features of the samples. In our case we collect a total of 22,848 features per window for the Haar descriptor and 42,840 for the EOH one. We restrict the strong classifier to select the same amount of weak classifiers as the dimension of the HOG descriptor we used in [22, 21], which is 3780.

In fact, the learning process could continue until constructing a cascade of strong classifiers (as in [24] for face detection), where the first layer discards clear non-pedestrians, the second layer would discard less clear non-pedestrians and so on, being pedestrians those windows that are not rejected at any layer. This cascade procedure speeds up detection as most of the windows are rejected at the early stages of the cascade. However, in this chapter we are more interested in showing that the pedestrian detector based on Haar and EOH can be learnt using virtual-world samples and domain adaptation techniques. Therefore, we only present results based on training a single layer with the real AdaBoost algorithm.

2.1.3 Scanning

As scanning procedure we apply the pyramidal sliding window [4]. It consists in constructing a pyramid of scaled images, for the range of scales in which we want to

detect the pedestrians. The bottom of the pyramid (higher resolution) is the original image, while the top is limited by the size of the smaller pedestrian to detect. 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 or the original image ($i = 0$), and s_p a provided parameter. We down-sample the image using bilinear interpolation with anti-aliasing as in [9] for building the lower resolution levels of the pyramid. Then, a canonical window (CW) of fixed size scans each pyramid level according to strides s_x and s_y, in x and y axes, respectively. We set $< s_x, s_y, s_p > := < 8, 8, 1.2 >$ like in [4] as it is a good tradeoff between processing time and final detection performance. This scanning procedure is not the standard for pedestrian detectors based on descriptors as Haar or EOH. Haar and EOH features are usually scaled themselves instead of using an image pyramid. However, we have experimentally seen that, in general, the pyramid with anti-aliasing boosts the performance of the pedestrian detectors based on self-scaled descriptors.

2.1.4 Selection

Detection over multiple scales and different positions usually yields several detections which frequently refer to a single object. In order to obtain a unique detection per object (pedestrian), we apply the non-maximum-suppression approach proposed in [12].

2.2 Datasets: Real- and Virtual-World Samples

To perform our experiments we use the generic real-world dataset INRIA [5, 4] and our virtual-world one [16]. The widespread INRIA dataset for human detection contains color images of different resolution (320×240 pix, 1280×960 pix, etc.) with persons photographed in different scenarios (urban, nature, indoor). INRIA data includes a set of training images with the bounding box annotation of 1,208 humans (that can be vertically mirrored to obtain 2,416 positive samples). In addition, 1,218 human-free images are provided for training. For testing, INRIA includes a dataset consisting of 563 annotated humans in 288 images, and 453 human-free images.

The virtual-world dataset [16] is generated with Half Life 2 videogame by city driving. It is composed of color images of 640×480 pix. From the provided virtual-world data we mimic INRIA settings for fair comparison. Thus, we use 1,208 virtual-world humans that are vertically mirrored to obtain 2,416 ones, as well as 1,218 human-free virtual-world images. Virtual-world data is only used for training, *i.e.*, for being domain adapted to the training data of INRIA.

2.3 Training with Virtual- vs Real-World Samples

As we mentioned in Sect. 2.1.1, in this chapter we use Haar, EOH and HaarEOH features and real AdaBoost learning machine for training human/pedestrian

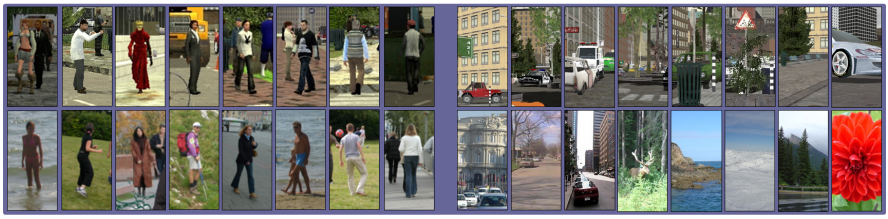

Fig. 4 Top: virtual pedestrians and city scenarios. Bottom: INRIA photographs with humans and diversified scenarios as city, countryside, beach, etc. Humans appear also in such scenarios. Domain adaptation by batch active learning (Fig. 2) will bring together virtual-world samples and difficult real ones to learn real-world human classifiers.

classifiers. Accordingly, we train the INRIA human classifier using the INRIA training set and the virtual-world pedestrian one using virtual-world training data. During training, *bootstrapping* is used, *i.e.*, enriching the respective negative training sets with hard negative samples and then 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 particular training sets saturate with a single step.

2.4 Testing with Real-World Images: Dataset Shift

In order to evaluate the performance of the pedestrian detectors we follow the procedure proposed in [7] for this purpose. This means that we use performance curves of *miss rate vs false positives per image*. We focus on the range FPPI=$[0.1, 1]$ 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.

Fig. 5 plots the performance of human detectors based on INRIA and virtual-word training data, applied to INRIA testing set. Comparing the performance of the HOG/linear-SVM and the HaarEOH/real-AdaBoost we realize that is almost the same. Moreover, we show the results of three different pedestrian detectors based on different sets of features: Haar/real-AdaBoost, EOH/real-AdaBoost and HaarEOH/real-AdaBoost. The pedestrian detectors trained on the INRIA dataset clearly outperforms their counterparts trained on the virtual-world one. The gap of performance is over 10 points. We argue, as in [22, 21] that this gap is due to dataset shift. In the next section we will explain how to solve this problem by using a semi-supervised domain adaptation technique.

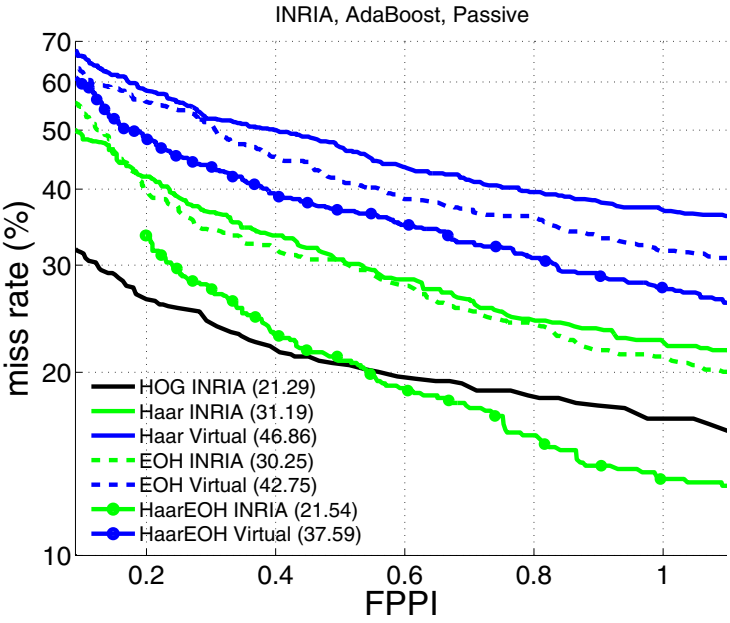

Fig. 5 Per-image evaluation of different pedestrian detectors. The notation *Feature DB* means that the corresponding classifier was learnt using *Feature*, and *DB* training data. In all cases the number inside the parenthesis indicates the average miss rate (AMR) in percentage, for the plotted FPPI range.

3 Semi-supervised Domain Adaptation

The dataset shift problem can be solved with a domain adaptation technique. In this case we use an active learning procedure similar to the one we employed in [22, 21] but using other kind of sample selection methods to reduce the amount of needed supervision during the annotation of the active samples. Additionally, we reduce the amount of allowed actively collected samples from the 25% in [22, 21] to 10%, which is a much more restrictive scenario.

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 $\mathcal{D}_t$ to learnt a reliable classifier. However, we have enough labelled samples drawn from $\mathcal{D}_s$. This scenario is called semi-supervised. If the distributions

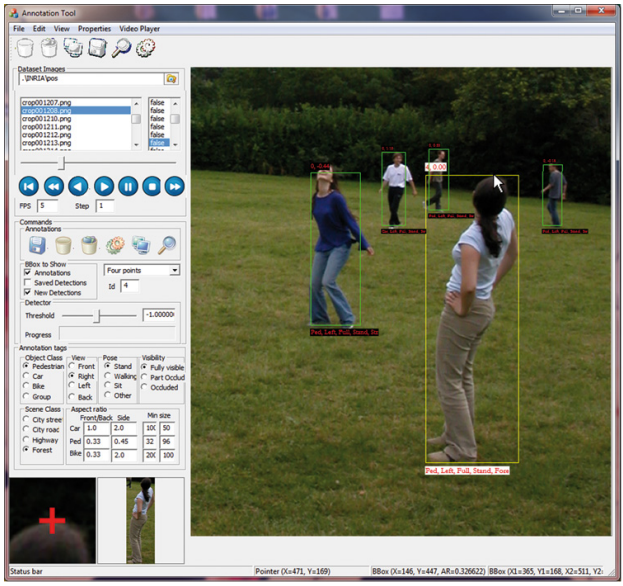

Fig. 6 Labelling tool. For each displayed image, the human oracle (Fig. 2) performs as follows: (1) if there are not humans, it marks the image as *human-free*; (2) if there are humans, some of them have been detected by the previous classifier (green bounding box), but others may not (not framed). The undetected humans must be manually framed by the human oracle (yellow bounding box).

of the samples in $\mathcal{D}_s$ and $\mathcal{D}_t$ are uncorrelated, then the task would be impossible. However, if they have a sufficient correlation, then we can cast the problem as one of *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.

Since we can collect in a cheap way as many samples 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. 2.4, a pedestrian classifier trained on virtual-world samples does not perform as good as we expect when applied to some real-world images. However, the obtained performance allows us to assume that there is sufficient correlation between $\mathcal{D}_s$ and $\mathcal{D}_t$, to the *eyes* of the features and base learning machine we use. Of course, as we deduce also from results in Fig. 5, $\mathcal{D}_s$ 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 is reduced to obtaining some labelled samples from $\mathcal{D}_t$, in a cheap way. Our proposal consists in an extension of the *active learning* procedure proposed in [22, 21] using a *human oracle* to label *difficult samples* and to confirm *easy samples* as right classified. All these samples are coming from $\mathcal{D}_t$. Usually, the difficult samples are defined as those falling in the ambiguity region of the base classifier at hand, *i.e.*, the area close to the decision boundary. 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 possible but being meaningful. Our case, however, is different. Let us say that $\mathcal{C}_s$ has been learnt from $\mathcal{D}_s$ 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. As an extension we also include easy samples, *i.e.*, those ones falling out of the ambiguity region.

Accordingly, a given 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 $|\mathcal{C}_s(x_t)| \geq th$. For our real AdaBoost, $|\mathcal{C}_s(x_t)| \geq 1$. 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), to label missed humans with a rectangular bounding box (Fig. 6) and to confirm the correct detections. Once the whole sequence is processed by the oracle, a new classifier is trained using the labelled samples that were used to build $\mathcal{C}_s$ (virtual-world ones) as well as the new collected difficult samples (real-world ones). We call *cool world*[2] to the joint space of virtual- and real-world samples. This type of active learning is termed as *batch mode*, because a set of images is processed before re-training. We think that a noticeable fact is to use virtual- and real-world samples to train a human classifier. This kind of process can be iterated. The overall approach is summarized in Fig. 2

4 Experimental Results

Plots of Fig. 5 show the effect of the dataset shift on the performance. To solve this problem we employed the semi-supervised domain adaptation technique proposed in section 3. Fig. 7 and 8 show these experiments in three different plots: Haar, EOH and HaarEOH. The plots compare the performance of the passive trained classifiers shown in Fig. 5 with the active ones explained in section 3. Four experiments are tested following the active learning procedure. Each experiment relies on a different manner of collecting the samples. As a reference we show two baselines that differ in the data used to train their classifiers: *INRIA 10%* one uses a 10% of the INRIA training dataset and the *Rand* one uses the virtual-world training set plus a 10% of the INRIA training dataset. *Act* refers to the active learning experiments explained

[2] 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.

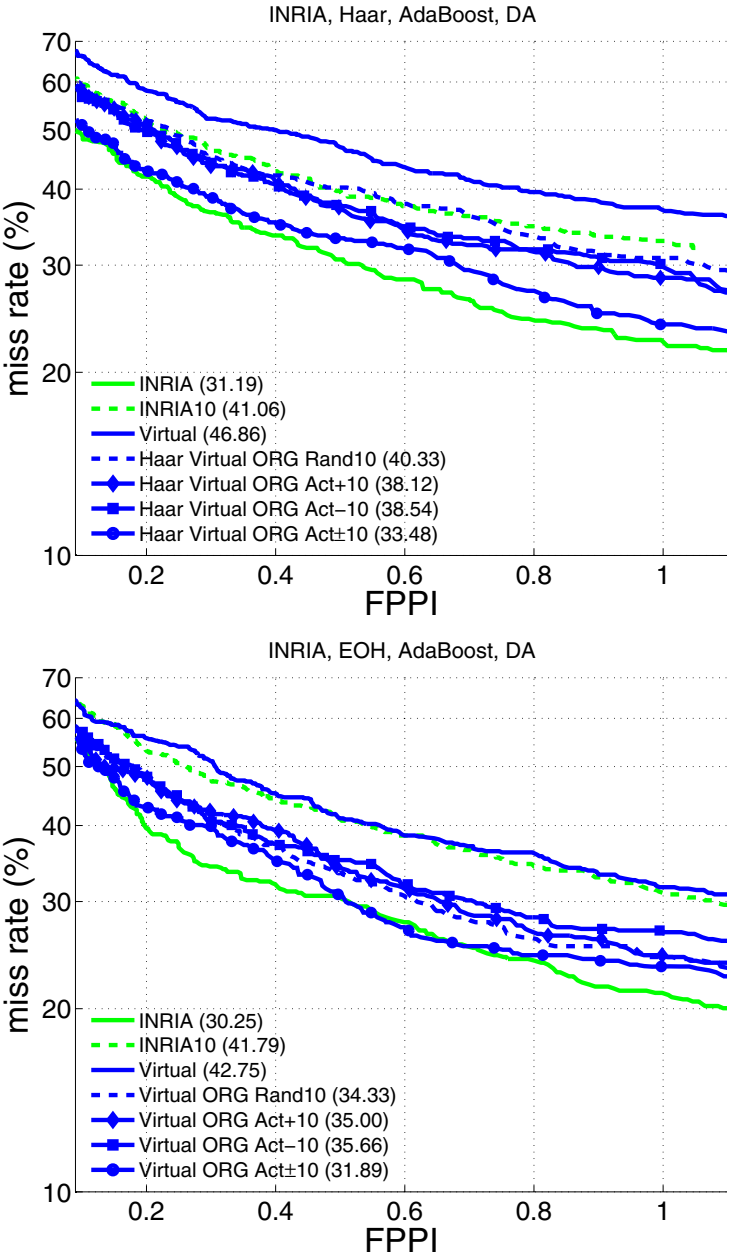

Fig. 7 Semi-supervised domain adaptation experimental results

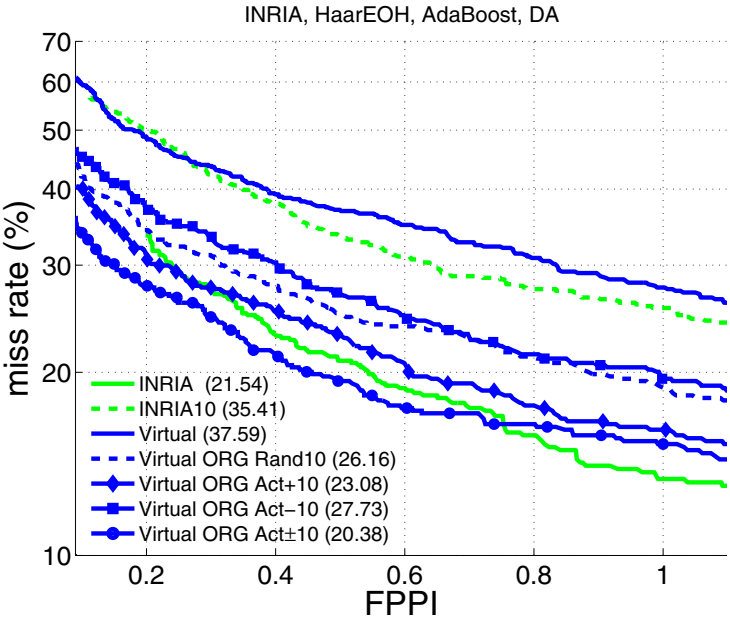

Fig. 8 Semi-supervised domain adaptation experimental results

in section 3. They differ on the oracle annotation procedure. In *Act+* the oracle labels the difficult samples, *i.e.*, the not detected humans, by framing them with a bounding box. In *Act-* the oracle annotates the easy samples, *i.e.*, the correctly detected humans, by just eliminating false positives. Note that *Act-* requires much less annotation effort. Accordingly, *Act*± refers to the combined annotation effort of *Act+* and *Act-*.

From the experiments we can draw the following observations:

- Reducing the training data of INRIA to the 10% decreases the performance of any trained detector by more than 10 points of AMR.
- All detectors benefit from adding a 10% of random INRIA data to the virtual-world set. This benefit varies from 6 to 10 points of AMR.
- Almost all of the tested *Act* experiments outperform the *Rand* and *INRIA 10%* baselines. Note that the trivial procedure of adding random data performs well in other contexts and it is usually difficult to outperform. Our proposed active learning procedures clearly outperform the random ones.
- For Haar and EOH *Act+* and *Act-* perform equally but requiring *Act-* less annotation effort. However, for HaarEOH *Act+* performs better.
- In all the cases *Act*± outperforms the baselines and the other *Act* experiments, even slightly outperforming the INRIA trained pedestrian detector for the most important case, the HaarEOH.

5 Conclusion

In this chapter we have addressed a core problem in the field of human detection, namely, the acquisition of good samples to train at low cost. 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 INRIA images, the virtual-world trained pedestrian classifier cannot reach the performance of a classifier learnt using data manually labelled for training in such dataset. 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 used a batch active learning technique that, with just a few manually labelled humans from the real-world images, is able to reach the same performance than a human classifier entirely trained from a much large amount of manually labelled data. Ultimately, our human classifier has been trained by using HaarEOH features that can be computed fast using integral images. 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. Besides, user interaction is key as we need a human oracle to annotate real-world samples. 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.

Acknowledgements. This work was supported by the Spanish MICINN, in particular by projects TRA2011-29454-C03-01, TIN2011-29494-C03-02 and Consolider Ingenio 2010: MIPRCV (CSD200700018).

References

1. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine Learning 79(1), 151–175 (2009)
2. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. In: IEEE Conf. on Computer Vision and Pattern Recognition, Providence, RI, USA (2012)
3. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
4. Dalal, N.: Finding People in Images and Videos. PhD thesis, Institut National Polytechnique de Grenoble, Advisors: Cordelia Schmid and William J. Triggs (2006)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA (2005)
6. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: British Machine Vision Conference, London, UK (2009)

7. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. on Pattern Analysis and Machine Intelligence 34(4), 743–761 (2012)
8. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: survey and experiments. IEEE Trans. on Pattern Analysis and Machine Intelligence 31(12), 2179–2195 (2009)
9. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, AK, USA (2008)
10. Gerónimo, D., Sappa, A.D., López, A.M., Ponsa, D.: Pedestrian detection using adaboost learning of features and vehicle pitch estimation. In: IASTED Int. Conference on Visualization, Imaging and Image Processing, Palma de Mallorca, Spain (2006)
11. Gerónimo, D., López, A.M., Sappa, A.D., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. IEEE Trans. on Pattern Analysis and Machine Intelligence 32(7), 1239–1258 (2010)
12. Laptev, I.: Improving object detection with boosted histograms. Image and Vision Computing, 27(5), 535–544 (2009)
13. Levi, K., Weiss, Y.: Learning object detection from a small number of examples: the importance of good features. In: IEEE Conf. on Computer Vision and Pattern Recognition, Washington, DC, USA (2004)
14. Lienhart, R., Maydt, J.: An extended set of Haar-like features for rapid object detection. In: IEEE Int. Conf. on Image Processing, Rochester, NY, USA (2002)
15. Sinha, P., Osuna, E., Oren, M., Papageorgiou, C., Poggio, T.: Pedestrian detection using wavelet templates. In: IEEE Conf. on Computer Vision and Pattern Recognition, San Juan, PR, USA (1997)
16. Marin, J., Vázquez, D., Gerónimo, D., López, A.M.: Learning appearance in virtual scenarios for pedestrian detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, San Francisco, CA, USA (2010)
17. Ponsa, D., López, A.: Cascade of Classifiers for Vehicle Detection. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2007. LNCS, vol. 4678, pp. 980–989. Springer, Heidelberg (2007)
18. Yeh, M., Zhu, Q., Avidan, S., Cheng, K.: Fast human detection using a cascade of histograms of oriented gradients. In: IEEE Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA (2005)
19. Schapire, R.E., Singer, Y.: Improved boosting using confidencerated predictions. Machine Learning 37(3), 297–336 (1999)
20. Sudowe, P., Leibe, B.: Efficient Use of Geometric Constraints for Sliding-Window Object Detection in Video. In: Crowley, J.L., Draper, B.A., Thonnat, M. (eds.) ICVS 2011. LNCS, vol. 6962, pp. 11–20. Springer, Heidelberg (2011)
21. Vázquez, D., López, A.M., Ponsa, D., Marin, J.: Cool world: domain adaptation of virtual and real worlds for human detection using active learning. In: Advances in Neural Information Processing Systems. Domain Adaptation Workshop: Theory and Application, Granada, Spain (2011)
22. Vázquez, D., López, A.M., Ponsa, D., Marin, J.: Virtual worlds and active learning for human detection. In: ACM International Conference on Multimodal Interaction, Alicante, Spain (2011)
23. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conf. on Computer Vision and Pattern Recognition, Kauai, HI, USA (2001)
24. Viola, P., Jones, M.: Robust real-time face detection. Int. Journal on Computer Vision 57(2), 137–154 (2004)

25. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. Int. Journal on Computer Vision 63(2), 153–161 (2005)
26. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, San Francisco, CA, USA (2010)
27. Wang, X., Han, T.X., Yan, S.: An HOG-LBP human detector with partial occlusion handling. In: Int. Conf. on Computer Vision, Kyoto, Japan (2009)