Unsupervised Domain Adaptation of Virtual and Real Worlds for Pedestrian Detection

David Vázquez, Antonio M. López, Daniel Ponsa
CVC and C. Sc. Dpt. UAB, Barcelona, Spain
{*david.vazquez, antonio, daniel*}*@cvc.uab.es*

Abstract

Vision-based object detectors are crucial for different applications. They rely on learnt object models. Ideally, we would like to deploy our vision system in the scenario where it must operate. Then, the system should self-learn how to distinguish the objects of interest, i.e., without human intervention. However, the learning of each object model requires labelled samples collected through a tiresome manual process. For instance, we are interested in exploring the self-training of a pedestrian detector for driver assistance systems. Our first approach to avoid manual labelling consisted in the use of samples coming from realistic computer graphics, so that their labels are automatically available [12]. This would make possible the desired self-training of our pedestrian detector. However, as we showed in [14], between virtual and real worlds it may be a dataset shift. In order to overcome it, we propose the use of unsupervised domain adaptation techniques that avoid human intervention during the adaptation process. In particular, this paper explores the use of the transductive SVM (T-SVM) learning algorithm in order to adapt virtual and real worlds for pedestrian detection (Fig. 1).

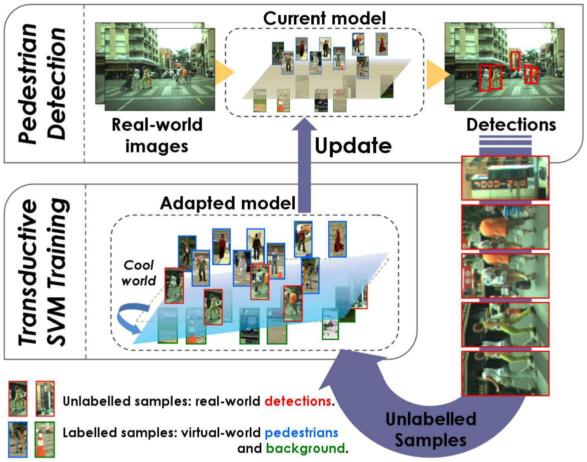
Pedestrian Detection
Real-world images
Current model
Detections
Update
Transductive SVM Training
Adapted model
Cool world
Unlabelled samples: real-world detections.
Labelled samples: virtual-world pedestrians and background.
Unlabelled Samples

Figure 1. We learn a pedestrian model using automatically labelled virtual-world pedestrians and background. With this model we detect pedestrians in real-world images. Some detections will be true positives and others false ones. Since we do not know which ones are of each type and we do not want human intervention, we treat all them as unlabelled data. Next, such unlabelled samples and the virtual-world labelled ones are joined to train a new pedestrian model using the T-SVM.

1 Introduction

Pedestrian detection is a relevant driver assistance functionality. Vision-based pedestrian detection is a current challenge given their variability (pose, clothe, occlusions) and background diversity (scene content and illumination). At the core of any pedestrian detector there is a pedestrian classifier, which decides if a given image window contains a pedestrian. Key components of the classifier are the pedestrian descriptors and the employed machine learning algorithm. Thus, most works on pedestrian detection have focused on these aspects [4, 6, 9]. The initial task of the learning process consists in collecting samples of pedestrians and background, which is important since with poor samples even the best descriptors and learning machines can not provide a good classifier [2]. These labelled samples are collected through a tiresome manual task, where human annotators provide bounding boxes (BBs) framing pedestrians present in a set of training images.

We face the challenge of developing object detectors whose underlying classifier is totally self-trained, *i.e.*, without the intervention of human annotators. In particular, our current interest is to do so for detecting pedestrians. In [12] we did a first step consisting in training a pedestrian classifier using virtual scenarios, *i.e.*, the pedestrian and background samples came from realistic videogame images, with the advantage of obtaining the pedestrian BBs automatically. The method exhibits the desired performance for the Daimler testing set [6].

Therefore, we concluded that training based on virtual worlds can be a manner of avoiding human annotators (note that videogame industry will keep generating realistic scenarios regardless we use them or not). However, in [14, 15] we showed that between virtual and real world it may be a *dataset shift* [13]. However, in [14, 15] we already saw that dataset shift does not come because of virtual-world training data, it can appear due to the use of different cameras and/or operating in different scenarios for training and testing.

Thus, in [14, 15] we applied *domain adaptation* (DA) [1] to virtual and real worlds, where the former is considered as the so-called *source domain* and the latter is the *target domain*. In the DA paradigm it is assumed that we have many labelled samples from the source domain, which in our case are pedestrian and background samples automatically collected (with label) from the virtual world. Regarding the target domain, two main situations are considered: (1) in *supervised DA* (SDA) we collect a small amount of labelled target-domain data; (2) in *unsupervised DA* (UDA) we collect a large amount of target-domain data but without labels. In [14, 15] we proposed a SDA approach based on *active learning*. Obtained results, were totally satisfactory in terms of the performance of the SDA-based pedestrian detectors. However, the use of active learning implies that a human oracle assists the training.

Currently we face the last step towards our self-trained pedestrian detector, *i.e.*, we propose not to involve humans annotators/oracles during the training process. In particular, as we explain in Sect. 2, in this paper we follow an UDA based on transductive SVM (T-SVM) [11]. As we will see in Sect. 3, for our current image acquisition system, the obtained performance is comparable to the one given by a pedestrian detector based on human assisted training. The conclusions of the presented work are summarized in Sect. 4

2 Proposed UDA pedestrian detector

Nowadays, the most relevant baseline pedestrian detector relies on a holistic pedestrian classifier that uses the histograms of oriented gradients (HOG) as visual descriptor, and the linear support vector machines (Lin-SVM) as learning method [3]. New state-of-the-art methods are based on this baseline [5, 8, 16, 17]. Thus, we have started our study using HOG but replacing the Lin-SVM by Lin-T-SVM (SVM ^{light} implementation [10]). If all the provided samples are labelled Lin-T-SVM is equivalent to Lin-SVM. Moreover, as we did in [12, 14, 15], the full pedestrian detector is completed with pyramidal sliding window to provide multi-scale candidate windows to the pedestrian classifier, and with non-maximum suppression to resume the multiple detections coming from the same pedestrian in just one.

Now, let us assume the following inputs. First, our *source domain*: $\Im_{\mathcal{V}}^{tr+}$ denotes a set of virtual-world images with automatically labelled pedestrians, and $\Im_{\mathcal{V}}^{tr-}$ refers to a set of pedestrian-free virtual-world images automatically generated as well. Second, our *target domain*: $\Im_{\mathcal{R}}^{tr}$ is a set of real-world images without labels. Third, a threshold, Thr, such that an image window is said to contain a pedestrian if its classification score is larger than Thr. Then, the steps of our UDA are:

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

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

(S3) *T-SVM learning in cool world*[1] with the virtual-world samples (*i.e.*, $\mathcal{T}_{\mathcal{V}}^{tr+}$ and $\mathcal{T}_{\mathcal{V}}^{tr-}$) and the real-world unlabelled ones (*i.e.*, $\mathcal{T}_{\mathcal{R}}^{tr?}$). Figure 2 illustrates the underlying idea of the T-SVM training. After this new training we obtain the new $\mathcal{C}$ and $\mathcal{D}$.

[1] We used the term *cool world* in [15] 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.

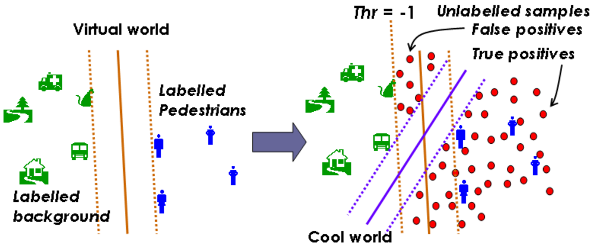

Figure 2. T-SVM training.

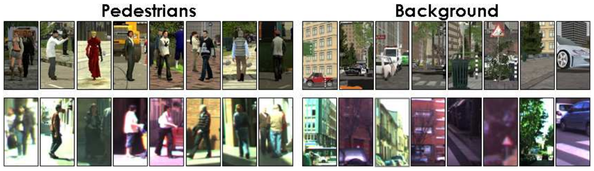

Figure 3. Virtual- and real-world samples from our camera (top & bottom, resp.).

3 Experiments

For our experiments we are using a camera based on a CCD color sensor of VGA resolution, with a lens of 6 mm of focal. The camera is installed in the windshield of a car, forward facing the road. Moreover, to apply our UDA training proposal, we are using 1208 virtual-world pedestrians, and 1219 pedestrian-free virtual-world images to collect background samples. In order to assess the performance of our method we have started with a sequence of 350 real-world images for training (let us denote it by $\Im^{tr}_{cvc}$) and 250 for testing, they do not overlap. In the training sequence there are 581 pedestrians. In the testing sequence there are 290 pedestrians with height equal or larger than 72 pixels, which were manually annotated for validation purposes. We focus on such pedestrians because they are at a distance to the car below 25 m, *i.e.*, the mandatory area of detection [9]. Figure 3 shows some pedestrian and background samples coming from virtual scenarios and our image acquisition system.

In all cases we have set $Thr = -1$ (lower limit of SVM uncertainty area), and the strides for the pyramidal sliding window are $\triangle_x = \triangle_y = 8$ pix with scale step of 1.2. The HOG parameters are those of the original proposal [3]. The performance metric for evaluating the pedestrian detector is the average area under the curve (A-AUC), where the curves are plotting the number of false positives per image (FPPI) *vs* the miss rate. We focus in the range from FPPI=1 to FPPI=0.1, which corresponds to a number of false positives that could be filtered out by a posterior temporal coherence analysis based on tracking [9]. To determine if a detection is right or not during the validation, we use the PASCAL VOC criterion [7]. In summary, we follow the standard evaluation method for object detection.

Figure 4 shows the results of the different experiments we have conducted. Curve *CVC-Train-1* refers to the use of a pedestrian classifier trained only with pedestrian and background samples from $\Im^{tr}_{cvc}$. Of course, for this experiment the pedestrians in $\Im^{tr}_{cvc}$ where manually labelled. Curve *Virtual* refers to the use of only the virtual-world data for training. Curves *UDA* correspond to the results according to the proposed method, *i.e.*, using virtual-world samples and $\Im^{tr}_{cvc}$ (without labels) for building the pedestrian classifier. We iterated the learning process three times, the third was not giving any improvement, so we show just two iterations. From one iteration to the next, we preserved all the detections as unlabelled samples. Curves *SDA* correspond to our approach in [14, 15], iterated also until no improvement was achieved, which happens at third iteration too. Note that SDA involves a human oracle during training.

From results in Fig. 4 we see that *Virtual* performs worse than *CVC-Train-1* (4 A-AUC points) which we hypothesize is due to dataset shift. When virtual and real world samples are combined, A-AUC is around 5 points better than *CVC-Train-1* and 9 points regarding *Virtual*, which are large improvements (see typical performance differences among pedestrian detectors in [4]). Iterations pay back in terms of performance improvement (similar to what happens with bootstrapping). Both approaches, SDA and UDA, perform similarly. Thus, we could say that virtual data is complementing well real one. This viewpoint, which applies to both SDA and UDA, corresponds to label real-world samples and use virtual-world ones to complement them. However, here we are interested in the pure UDA viewpoint, *i.e.*, starting with a classifier only based on virtual-world samples, the proposed T-SVM-based algorithm has been able to adapt it for operating in real-world images without human intervention.

Additionally, to complement the study, we trained a pedestrian classifier fully based on manual annotations. More specifically, we labelled 1016 pedestrians in other images taken with our camera and used 150 pedestrian-free images to collect background samples. Then, we trained the classifier following the same procedure that we used for training the classifier only based on virtual-world samples. The curve is shown as *CVC-Train-2* in Fig. 4. Note, that it gives better results than our DA approaches (around 2 A-AUC points), of course,

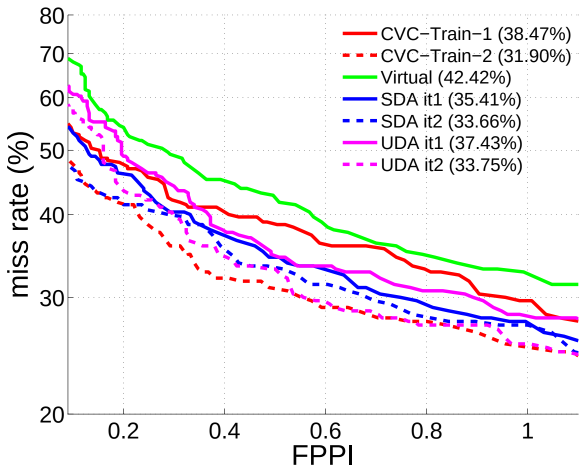

Figure 4. Detection results: in parenthesis the A-AUC of the conducted experiments.

by manually annotating the double of pedestrians than for SDA. However, in the working range from FPPI=1 to FPPI=0.5, UDA (it2) and *CVC-Train-2* shows analogous performance. Now, we can keep improving UDA by showing more sequences to the learning system, while for improving *CVC-Train-2* more manual annotations would be required and it is necessary to follow some sort of active learning procedure (as we do in SDA) to avoid introducing redundant pedestrians.

4 Conclusions

In this paper we have addressed the challenge of developing a method for obtaining self-trained pedestrian detectors, *i.e.*, without human intervention for labelling samples. For that we have proposed an unsupervised domain adaptation procedure based on T-SVM. In the source domain we have automatically labelled samples coming from a virtual-world, while the target domain correspond to real-world images from which the proposed algorithm selects (by detection) unlabelled samples that correspond to pedestrians and background. Obtained results are comparable to traditional learning procedures where the pedestrians samples are collected by tiresome manual annotation. Our proposal can be extrapolated to other objects.

Acknowledgments

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

References

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

[2] 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.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, San Diego, CA, USA, 2005.

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

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

[6] M. Enzweiler and D. Gavrila. Monocular pedestrian detection: survey and experiments. *IEEE Trans. on PAMI*, 31(12):2179–2195, 2009.

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

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

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

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

[11] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, Bled, Slovenia, 1999.

[12] J. Marin, D. Vázquez, D. Gerónimo, and A.M. López. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR*, San Francisco, CA, USA, 2010.

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

[14] D. Vázquez, A.M. López, D. Ponsa, and J. Marin. Virtual worlds and active learning for human detection. In *ICMI*, Alicante, Spain, 2011.

[15] D. Vázquez, A. López, D. Ponsa, and J. Marín. Cool world: domain adaptation of virtual and real worlds for human detection using active learning. In *NIPS-Domain Adaptation Workshop: Theory and Applicatio*, Granada, Spain, 2011.

[16] X. Wang, T.X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, Kyoto, Japan, 2009.

[17] J. Zhang, K. Huang, Y. Yu, and T. Tan. Boosted local structures HOG-LBP for object localization. In *CVPR*, San Francisco, CA, USA, 2010.