Adapting a Pedestrian Detector by Boosting LDA Exemplar Classifiers

Jiaolong Xu[1], David Vázquez[1], Sebastian Ramos[1], Antonio M. López[1,2] and Daniel Ponsa[1,2]
[1]Computer Vision Center [2]Dept. of Computer Science
Autonomous University of Barcelona
08193 Bellaterra, Barcelona, Spain
{jiaolong, dvazquez, sramosp, antonio, daniel}@cvc.uab.es

Abstract

Training vision-based pedestrian detectors using synthetic datasets (virtual world) is a useful technique to collect automatically the training examples with their pixel-wise ground truth. However, as it is often the case, these detectors must operate in real-world images, experiencing a significant drop of their performance. In fact, this effect also occurs among different real-world datasets, i.e. *detectors' accuracy drops when the training data (source domain) and the application scenario (target domain) have inherent differences. Therefore, in order to avoid this problem, it is required to adapt the detector trained with synthetic data to operate in the real-world scenario. In this paper, we propose a domain adaptation approach based on boosting LDA exemplar classifiers from both virtual and real worlds. We evaluate our proposal on multiple real-world pedestrian detection datasets. The results show that our method can efficiently adapt the exemplar classifiers from virtual to real world, avoiding drops in average precision over the* 15%.

1. Introduction

Training a pedestrian detector with synthetic data avoids the need of expensive and tedious manual annotations, keeping decent performance as presented in [9]. This work, tested on a popular automotive dataset from Daimler A.G., suggests that the use of synthetic data for the pedestrian detection task is a research line worth to explore. Similar work has also been reported in [11]. However, as it also happens in real-world datasets, the classifiers' accuracy can drop significantly when the training dataset (source domain) and the application scenario (target domain) have inherent differences [13, 16]. A clear example of this appears when different cameras are used or the most common objects' poses and/or views are substantially different. In the case of classifiers trained with synthetic data (*i.e.* on *virtual worlds*), the drop of performance is presumably due to an appearance difference between virtual- and real-world datasets. Nevertheless, there are sufficient commonalities between these datasets that make possible the application of domain adaptation techniques. Therefore, a classifier trained with virtual-world images that must operate with real-world ones, needs to be adapted, which drives us to the realm of *domain adaptation* as a paradigm to solve tasks like the reuse of ground truth information.

Domain adaptation is an emerging topic in computer vision [12], which has been recently applied for the object detection problem [10, 13, 15, 14]. The objective of the domain adaptation techniques is to face up the task of deploying models that has been built from some fixed source domain, across a different target domain, *i.e.* in our case virtual and real worlds respectively. These different domains should have similarities that are normally measured at feature space. However, the feature distributions of different domain samples may differ tremendously in terms of statistical properties (such as mean and intra-class variance), what makes the domain adaptation task a very challenging one.

Previous work focus on transforming the high dimensional low-level feature space [10, 15], capturing complementary examples from the target domain through a human oracle (active learning) [14], or just considering them as unlabeled examples collected following a self-training idea [15]. Our proposal also applies the strategy of collecting appropriate examples, but in contrast to [14, 15], in our case it is given more relevance to those examples of the source domain that are more *similar* to the ones in the target domain. Motivated by the discriminative exemplar classifier [8], we model the example's similarity based on individual exemplar classifiers. Therefore, the exemplar classifiers trained with the source domain examples that are similar to the target domain ones, are expected to be discriminative in the target domain. We aim at supplementing the power of these source exemplar classifiers with few target domain labeled examples. The questions we confront with are: (1) how to find out the appropriate source domain exemplar classifiers; (2) how to maximize the discriminativity of such classifiers.

The boosting-based transfer learning turns out to be suited for our case, *e.g.* the TrAdaBoost [2] can tune the source classifiers towards the target domain by assigning greater importance to the target examples during the boosting. We propose to use the exemplar classifiers as base learners in the boosting framework, then the learning algorithm selects and combines these base classifiers into a strong classifier that will operate in the target domain. Particularly, we use the recent proposed discriminative decorrelation model to train LDA exemplar classifiers [7]. The LDA classifier can obtain equivalent or even better performance than the SVM classifier while being much faster for training. This property is essential for the boosting-based method since a large number of base learners are involved during the training. We evaluate our proposal on multiple real-world pedestrian detection datasets. The results show that our method can efficiently adapt the exemplar classifiers from virtual to real world, avoiding drops in average precision over the 15%.

The rest of the paper is organized as follows. Section 2 describes the LDA exemplar classifier. In section 3 we explain the boosting algorithm for the domain adaptation task. Section 4 evaluates the performance of the proposed method for pedestrian detection and present corresponding results. Finally, section 5 draws the main conclusions and future research lines.

2. LDA Exemplar Classifier

For domain adaptation, we aim at finding the examples from the source domain that are most *similar* to the target domain ones. These similar examples are then used to train a discriminative classifier in the target domain. In this paper, we do not measure such similarity directly from the low-level feature space (*e.g.* maximum mean discrepancy in the reproducing kernel Hilbert space). Instead of that, we assume that the exemplar classifiers corresponding to the source domain examples, *similar* to the target domain ones, are expected to obtain relatively higher classification accuracy in the target domain. Therefore, we focus our attention of the training process of the exemplar classifiers.

Particularly, we use the recent proposed discrimination decorrelation LDA model [7] to train exemplar classifiers, referred to as the LDA exemplar classifier. The main idea of the LDA method in [7] is to estimate a covariance matrix which captures the statistic properties of nature images. The covariance matrix is then used to remove the natural correlations between foreground and background HOG features, which is called whitening HOG.

The Linear Discriminative Analysis (LDA) is a general model that assumes the classes have a common covariance matrix $\Sigma_k = \Sigma, \forall k$. Given a dataset with features $x \in X$ and class labels $y \in \{0, 1\}$, suppose we model class density as multivariate Gaussian for binary classification,

$$\begin{aligned}\frac{P_r\left(y=1|x\right)}{P_r\left(y=0|x\right)} &= \log\frac{f_{y=1}(x)}{f_{y=0}(x)} + \log\frac{\pi_1}{\pi_0} \\ &= \log\frac{\pi_1}{\pi_0} - \frac{1}{2}\left(\mu_1+\mu_0\right)^T\Sigma^{-1}\left(\mu_1-\mu_0\right) \\ &+x^T\Sigma^{-1}\left(\mu_1-\mu_0\right)\ ,\end{aligned} \tag{1}$$

where μ_1, μ_0 are the vectors with the means of the positive and negative examples respectively, and π_1, π_0 are the class priors. The LDA classification is equivalent to a linear weight $w = \Sigma^{-1}\left(\mu_1 - \mu_0\right)$. For the HOG-based rigid template classifier with N cells in the window, Σ is a $Nd \times Nd$ matrix, where d is the number of bins used to quantize the gradient direction. In [7], it is proposed a simple procedure to learn Σ and μ_0 (corresponding to the background) once, and reuse them for every window size and object class, thus only average positive features are needed to compute the final linear classifier. It has been demonstrated that the LDA classifier can provide similar accuracy to the SVM one, while being much faster for training. In particular, when training large amount of exemplar classifiers, the LDA method needs just a couple of minutes once the covariance matrix has been built, since the main computation is just for computing the mean of the positive features. Instead, the alternative methods (*e.g.* exemplar SVM [8]) may take several days to train these exemplar classifiers.

In this paper, we use the LDA method to train pedestrian exemplar classifiers with virtual-world images. Following [9], we use a video-game to generate a virtual world from which we extract the pedestrian examples. The final exemplar classifiers are computed by simply dot product using

$$h_k = \Sigma^{-1}\left(x_k - \mu_0\right),\ k \in 1, 2, ...K, \tag{2}$$

where x_k is the HOG feature of the k-th example and K is the number of the exemplar classifiers. We use $K = 1267$ in our experiments.

3. Boosting Exemplar Classifiers for Domain Adaptation

We aim at adapting a pedestrian detector trained in a source virtual world to operate in a target real world. The difficulties of the domain adaptation are mainly due to the distribution disparity of the data from both worlds. To understand the relation between the virtual- and the real-world data, we extract HOG features from virtual- and real-world positive examples and apply Principal Component Analysis (PCA) on these features. Figure 1 shows a low-dimension representation of the two different data distributions. Given few real-world target examples, the straightforward way to train an adaptive classifier is the *mix training*, *i.e.* combine all the examples from both domains and retrain. However,

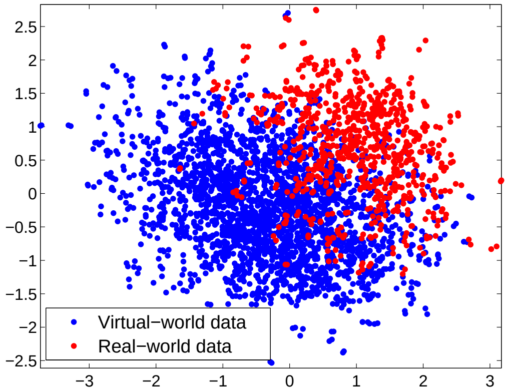


Figure 1. Low dimension distribution of the virtual- and the real-world positive examples.

the model trained from mixed data could be still source-domain-oriented. Instead, the boosting-based domain transfer learning can boost the classifier in the target domain. One example is the TrAdaBoost [2], which prioritizes the misclassified target examples and tunes the classifier towards the target domain data distribution.

In this paper, we use the virtual world exemplar classifiers as base learners and boost them towards high performance in the target real world. Based on TrAdaBoost, a recent study [1] presents a further refined algorithm that integrates a dynamic cost to solve the problem of early convergence to the source examples. In order to boost our exemplar classifiers we employ such a proposal, referred to as D-TrAdaBoost.

Following the terminology of prior literature, we denote the source domain by $\mathcal{D}^{Src}$ and the target domain by $\mathcal{D}^{Tar}$. Given N source examples and M target examples, where $N >> M$, our method proceeds as explained in Algorithm 1. In 3.2 of Algorithm 1, the exemplar classifiers are applied to the training examples and the classification errors are computed. In each iteration, the exemplar classifier with highest accuracy is selected. According to the classification error of each exemplar classifier, the weight of the corresponding example is assigned, *i.e.* the worse an exemplar classifier performs, higher is the weight assigned to the corresponding example. Note that the source examples and the target examples are treated differently (see 3.6 in Algorithm 1). The classification error on target examples will receive more penalty and this way the classifier is tuned towards a higher performance in the target domain. The D-TrAdaBoost [1] incorporates a dynamic cost C_t (3.5 in Algorithm 1) to avoid fast convergence in the source domain, which could further boost the classifiers towards the target domain.

Algorithm 1 Boosting-based domain adaptation with exemplar classifiers

```
Require Source domain training set:
$\mathcal{D}^{Src} = \{(x_i, y_i)\}, i \in (1, N)$
Target domain training set:
$\mathcal{D}^{Tar} = \{(x_i, y_i)\}, i \in (N+1, N+M)$
Maximum number of iterations: $T$
Exemplar classifiers: $h_k, k \in (1, K)$
Output Target classifier: $H = \sum_k \alpha_k^{Tar} h_k$
Procedure
1: Initialize the weight of each example $w_i = \frac{1}{N+M}$
2: Set $\alpha^{Src} = \frac{1}{2} \ln{(1 + \sqrt{2 \ln{\frac{N}{T}}})}$
3: for $t = 1 \to T$
3.1: Normalize the weight vector:
$w_i = \frac{w_i}{\sum_{j=1}^{N+M} w_j}$
3.2: Find the candidate exemplar classifier $h_t$ that minimizes the error for the weighted examples.
3.3: Compute the error on the target examples:
$e_t = \sum_{j=N+1}^{N+M} \frac{w_j [y_j \neq h_t(x_j)]}{\sum_{k=1}^{N+M} w_k}$
3.4: Set $\alpha_t^{Tar} = \frac{1}{2} \ln{(\frac{1-e_t}{e_t})}, e_t < \frac{1}{2}$
3.5: Set $C_t = 2(1 - e_t)$
3.6: Update weights:
$w_i = C_t w_i e^{-\alpha^{Src}[y_i \neq h_t(x_i)]}$, where $x_i \in \mathcal{D}^{Src}$
$w_j = w_j e^{-\alpha_t^{Tar}[y_j \neq h_t(x_j)]}$, where $x_j \in \mathcal{D}^{Tar}$
4: endfor
```

4. Experiments

In this section, we first evaluate our approach on multiple real-world pedestrian datasets. Then, we compare in terms of performance, our method to a classifier retrained using only target domain data or a mix of source and target domain data. Finally, different boosting algorithms are compared and analyzed.

We use INRIA [3] and ETH [17] as real-world datasets and all the experiments are evaluated using the Caltech evaluation framework [4].

4.1. Datasets

4.1.1 Virtual-world Pedestrian Dataset

The virtual-world dataset of [9] was created using the video game Half-Life 2. We follow the same idea to create virtual-world training examples. Our virtual world contains realistic virtual cities under different illumination conditions and with six different object classes, namely road, tree, building, vehicle, traffic sign and pedestrian. The dynamic objects (*i.e.* pedestrian and vehicles) are placed following physical laws. We obtained images containing pedestrians with different poses and backgrounds as well as their corresponding pixel-wise ground truth. The image resolution for this dataset is 640×480 pixels. Figure 2 shows an example of the virtual-world data: image, pedestrians and their corresponding ground truth. We use the virtual-world pedestrian dataset as our source domain. In particular, we used in our experiments 1267 examples from this dataset, for which we had computed the HOG features with a canonical window size of 13×6. Furthermore, we used the same background statistic μ_0 and Σ as in [7] for training the LDA exemplar classifiers.

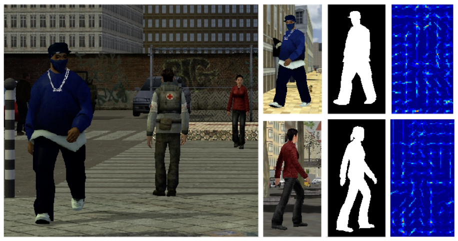


Figure 2. Virtual-world dataset. Pedestrian examples, pixel-wise ground truth and visualization of the respective LDA exemplar classifier model.

4.1.2 Real-world Pedestrian Benchmarks

The considered real-world datasets include INRIA [3] and ETH [17]. The ETH dataset was recorded at a resolution of 640×480 pixels, using a stereo pair mounted on a children stroller. For our particular experiments, only the left images of each image-sequence are used. The ETH dataset contains three sub-sequences, representing three different scenarios, which are denoted as ETH00, ETH01 and ETH02.

4.2. Evaluation

For all the experiments, we use the unified evaluation framework of [4]. We assess detectors' performance using per-image evaluation and plot curves depicting the trade-off between miss rate and the number of false positives per image (FPPI) in logarithmic scale. This is computed by averaging miss rate at nine FPPI rates evenly spaced in log-space in the range from 10^{-2} to 10^{0}. We sample the labeled target data five times and report the average performances. We compare the following classifiers:

SRC: LDA classifier trained with only source examples.

TAR: LDA classifier trained with only selected target examples.

MIX: LDA classifier trained with source and the selected target examples.

ADP: Adaptive classifier trained with source and selected target examples by D-TrAdaBoost.

4.3. Results

4.3.1 Adaptation to INRIA dataset

In the first experiment, we use INRIA pedestrian dataset as target domain. We randomly sample 400 positive examples and 1000 negative images from INRIA training set. The source classifier is trained as mentioned in 4.1.1. Figure 3 (a) depicts the performances on INRIA testing set, using the original HOG-SVM detector [3] ("TAR-FULL-SVM"), which is trained with INRIA full training dataset, and the detectors formulated in 4.2. Our source detector "SRC" has similar performance to "TAR-FULL-SVM", while the adapted detector obtains around 7 points of gain, showing the effectiveness of the adaptation technique. The classifier trained with only a few target examples ("TAR") turns out to have the worst performance due to over-fitting problems. Finally, the "MIX" detector obtains 41.12%, which is still 4 points worse than the proposed adaptation method.

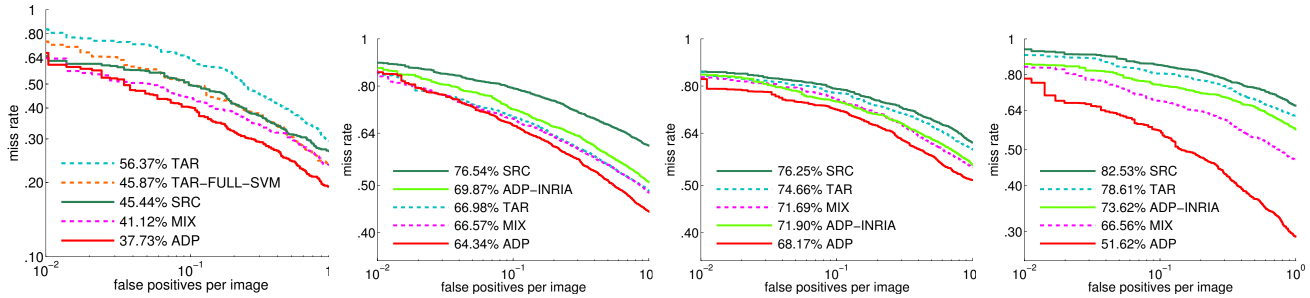
4.3.2 Adaptation to ETH dataset

In the second experiment, we use the three sub-sequences of ETH dataset as different target domains, considering that the three sub-sequences are taken from different scenarios. We use the same setting as before, *i.e.* we use the same source detector trained with virtual-world data and 400 random positives and 1000 negative images from the target domain. As depicted in Figure 3 (b)-(d), the proposed adaptation method obtains again large gains compared to the source classifier and outperforms other training methods.

As can be seen in [4], many detectors are trained with INRIA training set and then tested in other datasets without considering domain adaptation. For the sake of completeness and to evaluate the consequences of doing so, we also include experiments where our "ADP-INRIA" is tested in ETH dataset. Interestingly, the adapted detector "ADP-INRIA" does not work well for ETH data, while the adapted detectors using specific target domain data (ADP in Figure 3), show substantially better performances. This demonstrates the importance of applying the domain adaptation techniques from the source domain to the specific target domain instead of applying them to intermediate domains.

4.4. Evaluation of the boosting algorithms

In this section, we compare different boosting algorithms to investigate their performance on the pedestrian detection task. We evaluate AdaBoost [6], TrAdaBoost [2] and D-TrAdaBoost [1] on INRIA and ETH datasets. Table 4.4 shows the final results obtained with these boosting algorithms on the mentioned datasets. TrAdaBoost and D-TrAdaBoost generally outperform AdaBoost, since the classifiers are *tuned* towards the target dataset. For bet-



(a) Results on INRIA (b) Results on ETH00 (c) Results on ETH01 (d) Results on ETH02

Figure 3. (a) results on INRIA testing set. "TAR-FULL-SVM" represents the original HOG-SVM detector [3] which is trained with INRIA full training dataset. (b)-(d) results on three subsets of ETH dataset. "ADP-INRIA" refers to the adapted detector trained with a few examples from INRIA dataset as target domain.

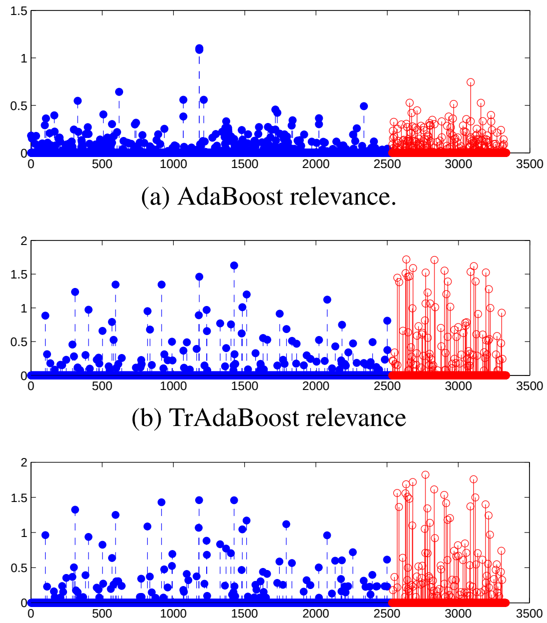
Dataset	AdaBoost	TrAdaBoost	D-TrAdaBoost
INRIA	39.5	37.6	37.7
ETH00	63.5	63.3	64.3
ETH01	69.4	68.7	68.2
ETH02	66.6	53.1	51.6

Table 1. Comparison of boosting algorithms' performances. (Average miss rate %)

ter understanding of the changes in the classifier weight with respect to the boosting algorithms, we visualize the weights of the base classifiers in the final model, which is adapted to INRIA dataset. Figure 4 shows the weights of the exemplar classifiers after the boosting process. The TrAdaBoost and D-TrAdaBoost learning algorithms assign greater weights to the target exemplar classifiers, which indicates that these classifiers are tuned towards the target domain, while the AdaBoost classifiers seem to be still source-domain-oriented.

5. Conclusions

Training vision-based pedestrian detectors using virtual-world data is a useful technique for saving expensive manual labeling. However, when the training dataset (source domain) and the application scenario (target domain) have inherent differences as is often the case, the performance of such a detector can drop significantly. In this paper, we propose a method to reduce the performance drop from the virtual to the real world by boosting LDA exemplar classifiers. This method shows promising results for the cross-domain detection problem, requiring only few labeled examples from the target domain. Overall, our proposal is a promising way to reuse cheap-to-obtain and precise ground truth information. In our future work, we will further consider the adaptation to the more challenging unlabeled real world scenarios, probably using the state-of-the-art part-based model [5].



(a) AdaBoost relevance.

(b) TrAdaBoost relevance

(c) D-TrAdaBoost relevance.

Figure 4. Boosting algorithms' relevance. The blue and red stems are the weights of the source and target exemplar classifiers respectively. Note that the flipped training examples are added, so the source and the exemplar classifiers are 1267 × 2 and 400 × 2 respectively.

Acknowledgments

This work is supported by the Spanish MICINN projects TRA2011-29454-C03-01 and TIN2011-29494-C03-02, the Chinese Scholarship Council (CSC) grant No.2011611023 and Sebastian Ramos' FPI Grant BES-2012-058280.

References

[1] S. Al-Stouhi and C. Reddy. Adaptive boosting for transfer learning using dynamic updates. In *ECML*

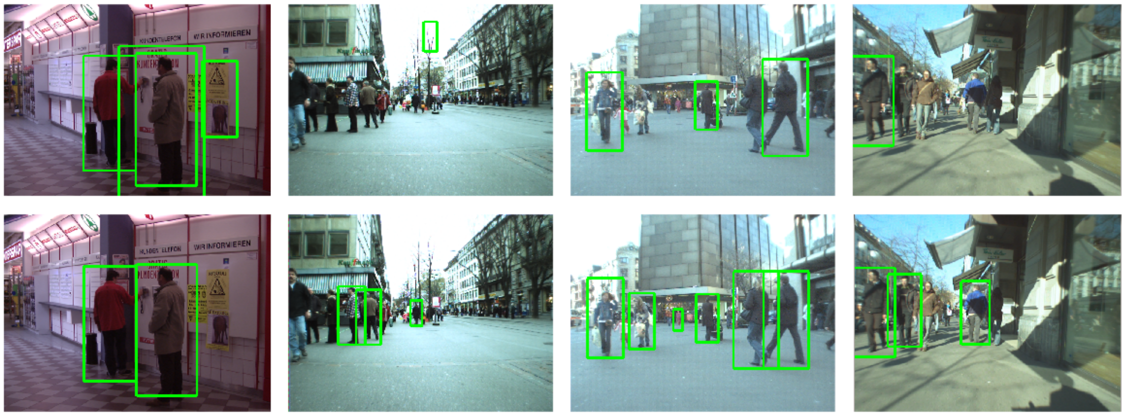


Figure 5. Comparison of detections in the target datasets. The detections are taken at FPPI = 0.1. Top: output of the source domain detector. Bottom: output of the adapted detectors. The images in each column are from INRIA, ETH00, ETH01 and ETH02 datasets respectively.

PKDD, Athens, Greece, 2011.

[2] W. Dai, Q. Yang, G. Xue, and Y. Yu. Boosting for transfer learning. In *International Conference on Machine Learning*, Oregon, USA, 2007.

[3] 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.

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

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

[6] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[7] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *European Conf. on Computer Vision*, Florence, Italia, 2012.

[8] T. Malisiewicz, A. Gupta, and A. a. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, Barcelona,Spain, 2011.

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

[10] J. Pang, Q. Huang, S. Yan, S. Jiang, and L. Qin. Transferring boosted detectors towards viewpoint and scene adaptiveness. *IEEE Trans. on Image Processing*, 20(5):1388–400, 2011.

[11] L. Pishchulin, A. Jain, and C. Wojek. Learning people detection models from few training samples. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011.

[12] K. Saenko, B. Hulis, M. Fritz, and T. Darrel. Adapting visual category models to new domains. In *European Conf. on Computer Vision*, Hersonissos, Heraklion, Crete, Greece, 2010.

[13] D. Vázquez, A. López, and D. Ponsa. Unsupervised domain adaptation of virtual and real worlds for pedestrian detection. In *Int. Conf. in Pattern Recognition*, Tsukuba, Japan, 2012.

[14] 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 *Advances in Neural Information Processing Systems – Workshop on Domain Adaptation: Theory and Applications*, Granada, Spain, 2011.

[15] R. L. Vieriu, A. K. Rajagopal, R. Subramanian, and F. B. Kessler. Boosting-based transfer learning for multi-view head-pose classification from surveillance videos. In *European Signal Processing Conference*, Bucharest, Romania, 2012.

[16] M. Wang and X. Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011.

[17] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009.