Transfer Pedestrian Detector Towards View-Adaptiveness and Efficiency*

Junbiao Pang[†,‡] Qingming Huang[†,‡] Shuqiang Jiang[†] Zhipeng Wu[‡]

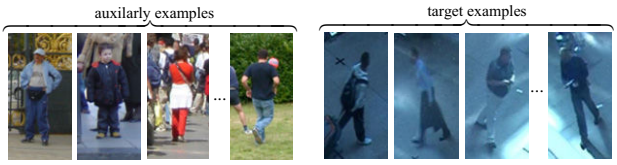
[†]Key Lab. of Intell. Info. Process, Inst. of Comput. Tech., Chinese Academy of Sciences, China
[‡]Graduate University of Chinese Academy of Sciences, China
{jbpang,qmhuang,sqjiang,zpwu}@jdl.ac.cn

Abstract

The distribution disparity is often inevitable between the pedestrian training examples and the test data from a specific application scenario, which may result in unsatisfactory detection accuracies. In this paper, we investigate how to efficiently adapt a generic boosting-style detector for a new scenario, e.g.*, with a distinctive capture view-angle, with only very limited examples (*e.g.*, $\sim$ 200). The basic notation is to* transfer *the auxiliary knowledge encoded within the well-trained detector to a new scenario. When specific to boosting-style detectors, this auxiliary prior knowledge includes the selected features and the weights for the weak classifiers. For a new scenario, these features are reused and shifted to the most discriminative positions and scales, and the weights are further adapted by covariate shift, which introduces the covariate loss. Extensive experiments on cross-view detector adaption show the encouraging detection accuracy improvements brought by our proposed algorithm with very limited new examples.*

1. Introduction

Many practical solutions have been presented to visual object detection/location problems [31, 9, 24, 19]. Although the typical objects to be detected are frontal human face [31] and pedestrian [24], it has been shown that these approaches are general and can be extended for other objects, *e.g.*, automobile, profile face. By inheriting these successes, construction of new types of object detectors may be straightforward, namely, to collect sufficient training examples, and then select a proper detection approach, finally train the detector. Many object detection tasks are still beyond the capability of the state-of-the-arts, but even for those nearly solved tasks, the *high initial cost*, *i.e.*, the cost in acquiring sufficient training examples, may prohibit building practical systems for different application scenarios.



(a) Frontal view-angle (b) Overhead view-angle

Figure 1. Distribution disparity problem illustrated by examples selected from the different view-angles.

The *high initial cost* inherently arises from the fact that most current solutions are based on statistical learning techniques, which boost building trainable object detectors. Consequently, several thousand positive examples are typically required to train a detector. The requirement of large number of examples is also aggravated by the large variation of object's appearance. Furthermore, the cost in collecting the positive data is often very high, since each example needs to be located, and even aligned manually. On the other hand, several thousand negative images which do not contain positive instances are also required. Because that in the training phase, the detector bootstraps the hard negative examples from these images to ensure that the false positive rate is very low, *e.g.*, 10^{-4}.

In practice, the data distribution disparity is often inevitable between the training data and those from a specific application scenario as view-angle change. Here, let's take the pedestrian detection as an example, the appearance of a pedestrian may be substantially changed when the capture view-angle is changed (see Figure 1). If we take the examples captured from the overhead view-angle as test data while use the detectors trained with data from the frontal view-angle, the performance would be far from satisfaction. Otherwise, the detectors are needed to be re-trained with examples collected from the new view-angle. The *high initial cost* issue is un-avoided for this naïve approach. Rather than discarding examples from the frontal view-angle, one natural question is whether we can obtain a new detector by 1)

*This work is supported in part by National Natural Science Foundation of China: 60773136 and 60833006, in part by National Hi-Tech Development Program (863 Program) of China: 2006AA01Z117.

2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops
978-1-4244-4441-0/09/$25.00 ©2009 IEEE

utilizing the knowledge within the well-trained detector and examples from the frontal view-angle, and 2) only requiring very limited examples from the overhead view-angle. The purpose is efficiently alleviating the *high initial cost* issue. We give a positive answer in this paper.

As we elaborate in this paper, the proposed approach is to transfer the pedestrian detector cross view-angles. Although the examples captured from the different view-angles have great appearance difference, there exists certainly close relationship between them. In other words, the detector learned from the frontal view-angle can provide many valuable clues in training a new detector for the overhead view-angle. To determine which part in the old detector is still useful, we utilize a small amount of labeled new data captured from the new view-angle, called *target-distribution* training data. We instead call the data from the frontal view-angle as *auxiliary-distribution* data, since some of the data might be still useful for the target task. We transfer the detector from the auxiliary task into the target task by exploiting the relationship between the auxiliary data and the target data, which naturally leads to an instance of classical transfer learning [3, 25, 7, 28, 8].

The key assumption of our transfer learning is that the shared features may handle the overall appearance change/distortion caused by view-angle change. The shared features means that these local features may be semantically identical, but are observed in both auxiliary data and target examples possibly at different positions and scales (see Figure 2(a)). Therefore, it is desirable to find the correspondence of local features between the different view-angles – utilizing the auxiliary data yet transferring the discriminative power of the auxiliary detectors.

In this paper, we approach this in boosting-style detector [31, 30] for view-adaptiveness. The reason is that the boosting-style detector has been successfully applied in the detection of various objects, *e.g*., face [31] and pedestrian [30]. Boosting-style detector constructs weak classifiers based on local image patches, which may only be partly changed in positions and scales cross view-angles. Therefore, we can say that the patch-level features may be shared cross view-angles although the overall appearances are different. The states (left-top and right-bottom coordinates) of the image patches, need be transferred and tailored to the target data, and thus we propose the feature shift process. By feature shift, we stochastically search for the shared image patches in target data with the maximal margin rule. The advantages of such a process include that: 1) the stochastic searching can quickly locate the image patches in the target examples; and 2) the maximal margin criterion can retain the discriminating ability of the old detectors.

Further in grouping these shifted local pathes into the target detectors, the covariate shift [27] is applied in the exponential loss to optimize the weights of the weak classifiers. By covariate shift, the auxiliary data can be re-utilized with the target data via the covariate loss in the hope of retaining the generalization ability of the derived detectors. Rather than adopting the batch optimization (which is computation-intensive, because the weights are optimized by iterative approximation) [16, 34], we instead propose a fast computation scheme by additive approach to calculate the weights sequentially.

The rest of this paper is structured as follows. The related work is further summarized in Section 2. Section 3 first reviews the loss function for the classical boosting-style detector, followed by the details of our proposed solution to transferring pedestrian detector cross view-angles. The comparison experiments are discussed in Section 4. The conclusive remarks are made in Section 5.

2. Related Work

The possible solutions to the *high initial cost* problem are partly related to the three popular research topics, *i.e*., co-training, multi-task learning and transfer learning.

Co-training: A family of the related works are the co-training [6] based methods to utilize the unlabeled examples for the detector. In co-training, multiple independent detectors are applied to automatically label the same examples. If some of the detectors have high confidences on a particular example, the label of the example can be obtained to retrain the remaining ones. The detectors are iteratively improved with all the labeled examples [20]. To avoid the time-costing retraining process, the seminal idea [20] inspires the research in [18][33] to combine the co-training with the online method [23] for updating the detector. These ideas can be concluded as re-training new detector via co-training to label enough target examples, rather than utilizing the auxiliary data. However, co-training requires different visual cues to build the independent detectors.

Multi-task learning: Learning for multiple related tasks simultaneously can be advantageous, in terms of performance relative to learning for these tasks independently [7, 15]. There have also been various attempts to theoretically study multi-task learning [4, 5]. In computer vision, the representative work is the jointBoost [29], which simultaneously trains several object detectors behaving well than independently learned ones. Recently, Ahmed *et al*. [2] also propose to learn shared feature simultaneously from pseudo (auxiliary) tasks and target task, with convolutional neural networks (CNN) for building the several visual object classifiers. Multi-task learning can partly solve the the deficiency of training examples. However, multi-task learning requires that new task has sufficient examples to simultaneously learn with other related tasks. The *high initial cost* cannot be avoid within the context of multi-task learning.

Transfer learning: Transferring knowledge across re-

lated tasks is a known phenomenon in human learning [25]. The related research can be roughly divided into three categories according to the level of transferred knowledge. Model-level transfer category first estimates the hyper prior of model's parameters from several tasks, and then this hyper prior is transferred to similar tasks, *e.g.*, hierarchical Bayesian models with hyper priors constrained for similar tasks [13, 4, 32, 26]. However, the priors are often difficult to build on the discriminative classifiers [14]. Data-level transfer category instead discovers the useful examples from the auxiliary tasks, and then uses them along with the target data for learning [8] [27]. The third category is the feature-level transfer, which searches for the shared features with satisfactory performance cross domains. To uncover these shared features, one might introduce some related auxiliary tasks [3], or learn a distance function which behaves well to transfer knowledge [28]. For instance, Farhadi *et al*. [12] propose to construct the stable features for recognizing activities from different view-angles.Our proposed algorithm in this paper is a hybrid of the data-level transfer and feature-level transfer, and Figure 3(a) illustrates the mechanism of our proposed algorithm.

The seemingly most promising approach for view-angle adaptiveness problem may be the online method [23]. The online boosting adopts essentially i.i.d assumption, while the different view-angles make the i.i.d barely hold. Therefore, we believe that the online boosting is unsuited for view-angle adaptiveness.

3. Transfer Detector Cross View-angles

3.1. Boosting Detector and Basic Notations

The general approach of object detection is to learn a classifier, which predicts the class label for a sub-window, *e.g.*, 1 for *yes* and -1 for *no*. Within the context of boosting-style detector, the strong classifier $H(x)$ can be obtained by minimizing the exponential loss $\mathcal{L}$

$$\mathcal{L} = \int_{\Omega} p(x,y)e^{-yH(x)}d(x,y), \tag{1}$$

where Ω is the definition field of example and label pair (x,y) with the distribution as $p(x,y)$, and $y \in \{-1,+1\}$ is the class label for example x. The strong classifier $H(x) : x \rightarrow y$ is the additively obtained from a set of weak classifiers as

$$H(x) = \sum_{m=1}^{M} \alpha_m h_m(x),$$

where $h_m(x)$ is the weak classifier selected by a boosting process, and $\alpha_m \in \mathbb{R}$ is the weight characterizing the importance of the weak classifier $h_m(x)$. Within the context of boosting-style detector [31, 30], the $h_m(x)$ essentially consists of a local image patch and the classifier's parameters.

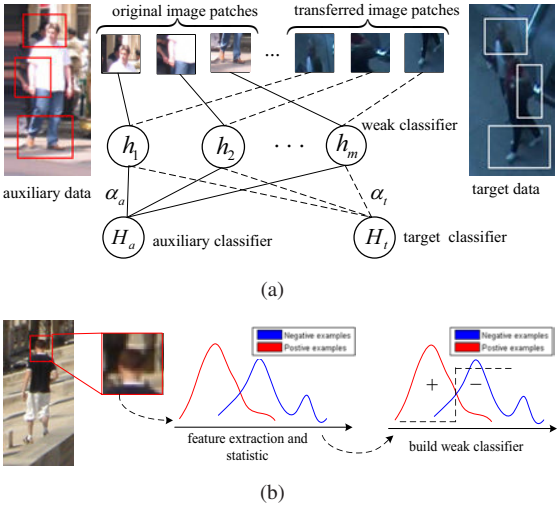


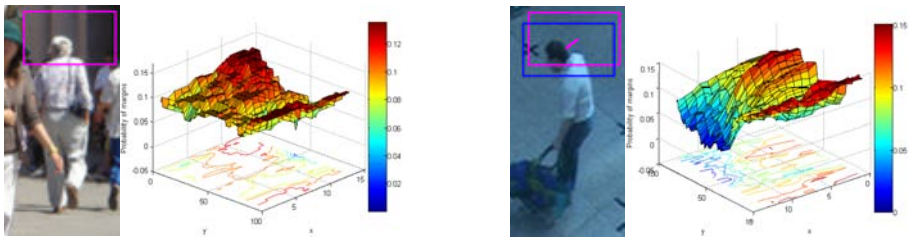
Figure 3. Mechanism illustration of the transfer learning for pedestrian detection across view-angles. (a) Transfer learning in boosting, the image patches are transferred from the auxiliary task to the target task. The weight of each image patch is also re-learned. Note that the image patches are all normalized into the same size to facilitate visualization purpose. (b) A weak classifier is built upon an image patch, and the image patches with various scales and locations generate the weak classifier pool.

That is, the weak classifier $h_m(x)$ first extracts feature from the predefined image patch, and then gives the decision with the learned parameters. Figure 3(b) illustrates this relationship between the image patch and the parameters. The final object detector D is then built as a cascaded strong classifier $H(x)$ [31].

For view-angle adaptiveness, let $\mathcal{T}_t = \{(x_i^t, y_i^t)\}_{i=1}^n$ be the target examples, where $x_i^t \in \mathcal{X}_t$ is *i.i.d* drawn from the *target-distribution* probability $p_t(x)$. Let $\mathcal{T}_a = \{(x_i^a, y_i^a)\}_{i=1}^m$ be the auxiliary examples, where $x_i^a \in \mathcal{X}_a$ is sampled from *auxiliary-distribution* probability $p_a(x)$[1]. For the scenario of detector transfer cross view-angles, $\mathcal{T}_a$ represents the examples collected from the frontal view-angle, and $\mathcal{T}_t$ denotes the data captured from other view-angle, *e.g.*, overhead view-angle (see Figure 1).

Intuitively, if the different view-angles are related (with overlapping areas), there should exist shared parts between the auxiliary data $\mathcal{T}_a$ and the target data $\mathcal{T}_t$. Therefore, the detector D_a learned from the frontal view-angle should have shared features with the detector for the target view-angle. To transfer the shared features, it is desirable to find the correspondence of shared features for different view-angles.

[1]Hereafter, the notation t and a generally represent the target data and the auxiliary data, respectively.



(a) Auxiliary example and margin gain map. (b) Target example and margin gain map.

Figure 2. The feature drifts from the old (pink) position to a new (blue)position. Comparing the maximal points in the margin gain map of (a) and (b), we can see that the feature shifts to the maximum point, which corresponds to the highest discriminating capability.

3.2. Transfer Weak Classifiers by Feature Shift

Denote the state of a image patch as $\theta = (l, t, r, b)$, where the l, t, r, b are the left-top corner and the right-bottom corner coordinates. Based on the above analysis, the state of image patch should be transferred to the target status θ_t based on the target data $\mathcal{T}_t$ – in other words, the new state θ_t should be found by using old state as "prior" knowledge.

To exploit the correlation between different view-angles, the dependence between the target status θ_t and the auxiliary status θ_a is assumed to be

$$p(\theta_t|\theta_a) \propto \mathcal{N}(\theta_a, \sigma^2 I), \tag{2}$$

where $\mathcal{N}(\theta_a, \sigma^2 I)$ is a Gaussian distribution with mean as θ_a and variance matrix as $\sigma^2 I$. The σ is empirically set to be 10 pixel in this work. The Gaussian dependence in Eq. (2) means that the target state θ_t deviates from the the old state θ_a locally. It is also reasonable for overlapped view-angle change.

The state θ can be estimated by computing the probability $p(\theta_t|\mathcal{T}_t, \mathcal{T}_a)$ in terms of Bayesian inference. However, the conditional probability cannot be computed directly. $p(\theta_t|\mathcal{T}_t, \mathcal{T}_a)$ would be simplified with the conditional independency between the $\mathcal{T}_a$ and $\mathcal{T}_t$, that is,

$$p(\mathcal{T}_a, \mathcal{T}_t|\theta_t) = p(\mathcal{T}_a|\theta_t)p(\mathcal{T}_t|\theta_t). \tag{3}$$

Using the Bayes's rule twice and Eq. (3), we have

$$p(\theta_t|\mathcal{T}_t, \mathcal{T}_a) \propto p(\mathcal{T}_t|\theta_t)p(\theta_t|\mathcal{T}_a). \tag{4}$$

Then based on the Bayesian rule, we have

$$\begin{aligned} p(\theta_t|\mathcal{T}_t, \mathcal{T}_a) &\propto p(\mathcal{T}_t|\theta_t)p(\theta_t|\mathcal{T}_a) \\ &= p(\mathcal{T}_t|\theta_t) \int p(\theta_t|\theta_a)p(\theta_a|\mathcal{T}_a)d\theta_a, \end{aligned} \tag{5}$$

where

$$p(\mathcal{T}_t|\theta_t) \propto \frac{1}{\sum_i \exp(-y_i^t h^a(x_i^t))} \tag{6}$$

be the margin gain of weak classifier $h^a(x)$, and the $p(\theta_a|\mathcal{T}_a)$ be the probability that the weak classifier occurs at the state θ_a. $\exp(-y_i^t h^a(x_i^t))$ in Eq. (6) measures the classification ability of the weak classifier $h^a(x)$ on the target data $\mathcal{T}_t$, that is, the ability to separate apart the positive and negative examples. Therefore, the essence of Eq. (5) is to search for the target status θ_t by following the largest margin criterion.

The optimal target status θ_t can be estimated via uniform sampling [21] as follows. A set of samples $\{s_l\}_{l=1,\ldots,L}$ are generated by repeating θ_a. After the state s_l transits to s_l^t with Eq. (2), the state s_l^t is associated with the weights

$$\pi_l^t \propto p(\mathcal{T}_t|s_l^t)p(s_l^t|s_l) \text{ with } \sum_l^L \pi_l^t = 1. \tag{7}$$

We use the Monte Carlo approximation of the expectation $\hat{\theta}_t = \sum_{l=1}^L s_l^t \pi_l^t$ as the optimal target state θ_t. In this paper, we set L to be 300. To give a better display of how feature shift, Figure 7 shows the correlation between feature shift and the corresponding classification ability.

Here we would like to highlight that the particle filter (PF) [17, 35] for visual object tracking, which is quite different to feature shift in motive. Both of them use the sampling method for the global maximum of the object function. However, PF is originally designed to sequentially search for the targets, while feature shift is for improving classification ability.

3.3. Transfer Classifier Weights by Covariate Shift

Although the auxiliary-distribution $p_a(x)$ is generally different from the target-distribution $p_t(x)$ ($p_a(x) \neq p_t(x)$), the conditional probability distribution can be considered equal, namely $p_a(y|x) = p_t(y|x)$. Therefore, *covariate shift* [27] can be used for reweighting the weak classifiers. Applying covariate shift into loss Eq. (1), we have covariate loss

$$\widetilde{\mathcal{L}} = \sum_{i=1}^{n} e^{-y_i H_t(x_i)} + \sum_{j=1}^{m} r_j e^{-y_j H_t(x_j)}, \tag{8}$$

where the $r_j = \frac{p_t(x_j, y_j)}{p_a(x_j, y_j)}$ is the ratio of the target data and the auxiliary data densities. Essentially, the second item

in Eq. (8) uses the $p_a(x_j, y_j)$ as proposal density in importance sampling to reuse the auxiliary data. Rather than estimating $p_t(x, y)$ and $p_a(x, y)$ with non-parameter probability estimation (Parzen windows [10]) or cross-validation [27], we reformulate the density ratio r_j with conditional probability by using Bayesian rules twice:

$$\begin{aligned} r_j &= \frac{p_t(x,y)}{p_a(x,y)} \\ &= \frac{p(t|x,y)p(x,y)}{p(t)} \frac{1}{\frac{p(a|x,y)p(x,y)}{p(a)}} \quad (9) \\ &= \frac{p(t|x,y)p(a)}{p(a|x,y)p(t)} \quad (10) \end{aligned}$$

Assuming the $p(a) = p(t)$, Eq. (9) can only be estimated by ratio of condition probability, which can be modeled as a logistic function

$$p(t|x, y) = \frac{1}{1 + e^{-yH(x)}}. \quad (11)$$

Therefore, we have

$$r_j = \frac{1 + e^{-y_j H_a(x_j)}}{1 + e^{-y_j H_t(x_j)}}. \quad (12)$$

The covariate loss can be written as

$$\widetilde{\mathcal{L}} = \sum_{i=1}^{n} e^{-y_i H_t(x_i)} + \sum_{j=1}^{m} \frac{1 + e^{-y_j H_a(x_j)}}{1 + e^{y_j H_t(x_j)}}. \quad (13)$$

The loss $\widetilde{\mathcal{L}}$ comes from the two different data: the auxiliary data and the target data. Rather than mixture training in multi-task learning, we weight every auxiliary examples (x_i^a, y_i^a). Literature [16, 34] weight the "old distribution" (corresponds the auxiliary distribution in our paper) and target distribution, that is, $p(x) = p_a(x) + \lambda p_t(x)$ (or $p(x) = (1 - \lambda)p_a(x) + \lambda p_t(x)$). Table 1 further summaries the difference between multi-task learning, approaches [16, 34] and our method.

3.4. Boosting the Covariate Loss

A directly method to optimize the hybrid loss can be gradient based batch optimization – concatenating all $\{\alpha_m^t\}$ as a vector. But besides iterative optimization discussed in section 1, it is also particularly inefficient because total examples must be used to compute the gradient and the loss at each iteration optimization.

To avoid the inefficiency of batch optimization, our method adopt the step-wise optimization method (a comparison will be illustrated in subsection 4.1). Following the AnyBoost [22], we select the weak classifier $h(x)$ to minimize a first-order expansion of Eq. (13) around $h(x) = 0$ by Taylor expansion

$$\sum_{x \in \mathcal{T}_t} e^{-yH_t(x)} y h_m^t(x) + \sum_{x \in \mathcal{T}_a} \frac{(1 + e^{-yH_a(x)})e^{yH_t(x)}}{(1 + e^{yH_t(x)})^2} y h_m^t(x). \quad (14)$$

Method	Loss function	Comments
Multi-task learning	$\mathcal{L}_a + \mathcal{L}_t$	Learning all tasks simultaneously and indifferently.
Method[16, 34]	$(1 - \lambda)\mathcal{L}_a + \lambda\mathcal{L}_t$ [16] or $\mathcal{L}_a + \lambda\mathcal{L}_t$ [34]	The $\lambda(0 \le \lambda_i \le 1)$ controls the degree of adaption. If $\lambda = 0$, there is no adaption process. If $\lambda = 1$, only target data is used to learn. The optimal λ can be estimated via cross-validation technique.
Our method	$\sum_i \lambda_i \text{Loss}(x_i^a, y_i^a) + \mathcal{L}_t$	The $\lambda_i(0 < \lambda_i < +\infty)$ acts as example "selector". If $\lambda_i \approx 0$, (x_i^a, y_i^a) will be useless for classifier adaption; Otherwise, (x_i^a, y_i^a) will contribute to adaption. λ_i can be estimated via Eq. (9).

Table 1. A comparison among different methods. Note that $\text{Loss}(x_i^a, y_i^a)$ is the loss of every example, which corresponds to $e^{-yH(x)}$ in Eq. (1) in our work.

In this paper, we use gradient based method to find the α_m^t. The gradient of Eq. (13) with respect to α_m^t can be computed as:

$$\begin{aligned} \frac{\partial \widetilde{\mathcal{L}}}{\partial \alpha_m^t} &= -\sum_{(x,y) \in \mathcal{T}_t} y h_m^t e^{-y(H_t + \alpha_m^t h_m^t)} \\ &\quad - \sum_{(x,y) \in \mathcal{T}_a} \frac{y h_m^t e^{y(H_t + \alpha_m^t h_m^t)} \left(1 + e^{-yH_a}\right)}{\left(1 + e^{y(H_t + \alpha_m^t h_m^t)}\right)^2} \quad (15) \end{aligned}$$

The algorithm can be computed efficiently by recording the weight $D \leftarrow e^{yH_t}$ iteratively as follows. At the k-th round, the weight for target data is updated $D_t \overset{(x,y) \in \mathcal{T}_t}{\longleftarrow} e^{yH_t(x)}$, and the weight for auxiliary data is also updated $D_a \overset{(x,y) \in \mathcal{T}_a}{\longleftarrow} e^{yH_a(x)}$. The k-th weak classifier h_k and α_t is only computed by using the weight D_t and D_a. The algorithm can be formulated as in Algorithm 1.

3.5. Cross-view Pedestrian Detector Transfer

In this work, we implement the above transfer learning algorithm based on the cascaded detector proposed by Viola and Jones in [31]. To transfer cascaded AdaBoost detector,

Algorithm 1. Covariate shift boost (CovBoost).

1: **Given:** The target examples $\mathcal{T}_t$, the auxiliary examples $\mathcal{T}_a$, and the learned auxiliary classifier $H_a(x)$.
2: Initialize weight $D_t(x_i) = 1$ for each target example, the weight $D_a(x_j) = 1$ for every auxiliary examples, and compute the $V(x_j) = 1 + e^{-yH_a(x_j)}$ for every auxiliary example.
3: **For** $m = 1, \ldots, M$
4: Find the weak classifier $h_m^t(x)$ from the shifted feature set by maximizing the weighted loss

$$\sum_i D_t(x_i) h_m^t(x_i) + \sum_j \frac{V(x_j) D_a(x_j)}{(1 + D_a(x_j))^2} h_m^t(x_j).$$

5: Find coefficient α_m^t that minimize the weighted loss

$$-\sum_i D_t(x_i) e^{-y_i \alpha_m^t h_m^t} - \sum_j \frac{V(x_j)}{1 + D_a(x_j) e^{-y_j \alpha_m^t h_m^t}}$$

via gradient based optimization method.
6: Update the weights by

$$D_t(x_i) = D_t(x_i) e^{-y_i \alpha_m^t h_m^t(x_i)}, D_a(x_j) = D_a(x_j) e^{-y_j \alpha_m^t h_m^t(x_j)}$$

7: **End for**
8: **Output**: target strong classifier $H_t = \sum_t \alpha_m^t h_m^t(x)$.

the classifiers of every stage in the cascade structure are updated sequentially, and the target data are filtered through all the cascade stages with zero false negative rate. The detailed transfer learning algorithm is formalized as Algorithm 2.

Algorithm 2. Cross-view Detector Transfer.

1: **Given:** The target positive examples $\mathcal{T}_t^+$, the target negative images $\mathcal{T}_t^-$, the learned auxiliary cascaded detector $D_a = \{H_a^1(x), \ldots, H_a^K(x)\}$, where $H_a^k(x)$ is k-th stage auxiliary classifier and the auxiliary examples$\{\mathcal{T}_a^k\}$ from k-th stage training process.
2: **For** $k = 1, \ldots, K$
- $H_t^k(x) =$ Transfer in Boosting$(H_a^k, \mathcal{T}_t^+, \mathcal{T}_t^-, \mathcal{T}_a^k)$, via feature shift and CovBoost.
- Bootstrap the hard negative examples from the target negative images set $\mathcal{T}_t^-$.

End For
3: **Output**: The target detector $D_t = \{H_t^1(x), \ldots, H_t^K(x)\}$.

4. Experiments

The proposed algorithm has been thoroughly tested on both synthetic and real data sets. In both case we illustrate the effects of CovBoost and feature shift.

4.1. Synthetic Data Experiments

Following the second toy data in [27], we show experiments on two dimensional (2-D) data to emphasize the efficacy of usage of auxiliary data in Figure 4. The synthetic data consists of two parts: auxiliary data and target data, where 2000 auxiliary data and 30 target data are generated from the distributions in Figure 4(a). The result of Figure 4(c) is trained on target data via AdaBoost. While, Figure 4(d) is obtained by trained on auxiliary data and target data via CovBoost. Comparing the decision planes in (c) and (d), (d) is very close to the ground truth. One can immediately see that the use of auxiliary data can help to obtain more accuracy classifier. The source code is available at http://www.jdl.ac.cn/user/jbpang/adaption.htm.

Further in comparing the efficacy of "batch optimization" [16, 34] with our method, we only select 80 weak classifiers to optimize the covariate loss Eq. (13). For the batch optimization, the $\{\alpha_m^t\}_{m=1,\ldots,80}$ are concatenated as a vector, and optimized with toolbox supplied by Matlab (under PC with 2GB RAM, and 2.66GHz intel CPU). 500 examples are generated from the target distribution to evaluate the accuracy of different optimization methods. The comparisons are given in Table 2. Our approach achieves approximate 400 times faster than bath optimization without damaging the accuracy.

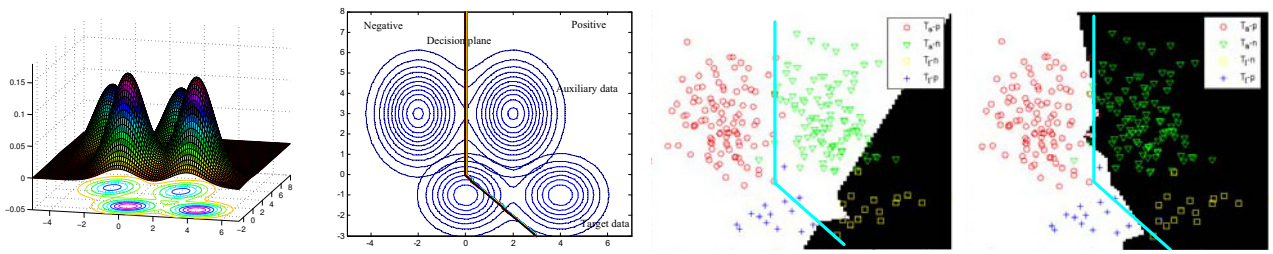
Method	Running Time	Accuracy
Batch Optimization	349.26± 6.63 (sec.)	0.56± 0.046
Our method	*0.76± 0.035* (sec.)	*0.77±0.083*

Table 2. A comparison among the different optimization methods. The results are the averages of 10 random repeats, as well as their standard deviations. The accuracy is evaluated as: $1 - \frac{\#\text{the miss-classified}}{\#\text{total examples}}$.

4.2. Real Data Experiments

In this subsection, we evaluate the effectiveness of the proposed algorithm for transferring pedestrian detector between two distinctive view-angles. The testset is obtained from the PETS 2007 [1], which was captured from the real environment with different humans activities. We manually labeled this dataset, and divide it into the training set and test set. In every labeled frame, each pedestrian is marked with a hand-drawn box around the whole human body. For the training set, only 220 positive target examples (with reflection images) from the *Dataset S7 view3* are randomly selected and normalized into the size of 64×128 pixels, and 150 negative frames without pedestrians are used as negative examples. The *Dataset S8 view3* is labeled at every 10 frame as the test set, which finally includes 300 frames with 973 pedestrian instances. The auxiliary data is borrowed from INRIA pedestrian dataset [9].

The evaluation protocol of PASCAL Visual Object Classes challenge [11] is adopted in this work for measuring algorithmic performance. A correct detection is recognized,



(a) Synthetic data (b) Decision plan (c) AdaBoost (d) CovBoost

Figure 4. Synthetic data. Subfigure (a) and (b) show the toy data and corresponding decision plan. Subfigure (c) and (d) show the result of classifiers trained on different data set respectively. In Subfigure (c) and (d), the legend "T_a-p","T_a-n", "T_t-n",and "T_t-p" represents the positive auxiliary data, the negative auxiliary data, the negative target data, the positive target data respectively. Note that only 30% auxiliary data is displayed for better illustrating the decision plane.

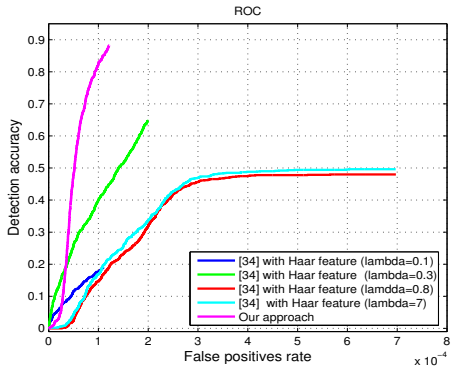


Figure 5. Comparison with other methods. For better viewing, please see original color pdf file.

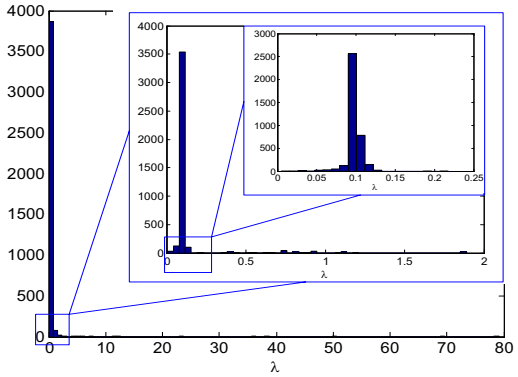


Figure 6. The distribution of λ_i. The mean of λ_i is 0.334 with standard deviation (std) 3.115.

when the rate of the overlapping a_o between the predicted bounding box B_p and ground truth bounding box B_{gt} exceeds 0.5, *i.e.*, $a_o = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} > 0.5$.

4.2.1 Analysis of the Algorithm

First, we compare our transfer learning algorithm with the Taylor expansion based method in [34]. Because it is the most related work for detector adaption. For a fair comparison, we use the Haar features as in [34], and tune all the listed values of the parameter λ, which is the relative importance of the target data in [34]. Report [34] only updates the weight α_m by optimizing the hybrid loss as listed in Table 1. The transferred detector achieves approximate 84% accuracy at 10^{-4} false positive rate, which is close to the result on INRIA data, $84\% - 89\%$, learned with linear support vector machines (SVMs) and HOG features in literature [9]. However, SVMs is trained on several thousand examples, while our result is achieved with very small *initial cost*, and only few hundred examples are required.

Note that the performance of the auxiliary detector is not plotted in Figure 5, because the auxiliary detector rejects all image patches as the negative examples in this experiment, namely the performance of the auxiliary detector stays at the origin point.

The reusability of auxiliary data: To analysis the reusability of auxiliary data, Figure 6 shows the distribution of λ_i at different scales. It is obvious that the λ_i of the most of auxiliary examples concentrate around 0.1, which means that auxiliary data does contribute to detector adaption. Interestingly, the mean of λ_i, 0.334, is consistent with the best performance value in [34]. (In Figure 5, best performance is obtained at $\lambda = 0.3$). However, our method doesnot need cross-validation to select the best λ_i, without lowering the accuracy. The large std, 3.115, shows that a few examples in auxiliary data may live in the target distribution. Because that larger λ_i means that $p_t(x, y)$ is larger than $p_a(x, y)$, that is, example (x, y) is more close to the target data than auxiliary data.

Feature shift: An analysis is done to study the efficacy of feature shift: remove or keep "feature shift"process at Step 2 in Algorithm 2. Figure 7 shows that feature shift gives near 10% improvement in accuracy.

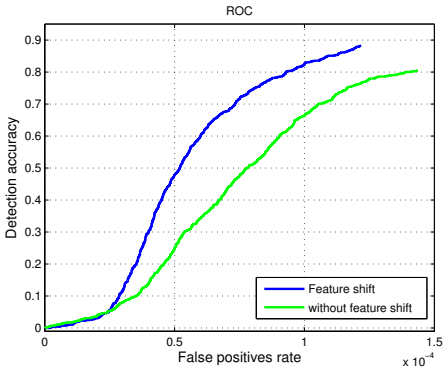


Figure 7. The efficacy of feature shift.

5. Conclusions and Future Work

In practice, there exist both necessity and feasibility for transferring generic pedestrian detectors to a new scenario. The difficulties to collect universal training data make a well-trained detector easy to fail in specific scenarios with disparate view-angles. The possibilities to reuse the generic detector come from the fact that the new samples in new scenarios may still share common local patches with those data used for training the generic pedestrian detector.

In this paper, we investigate how to transfer boosting-style detector for new view-angle. The underlying truth is that the weak classifiers correspond to the local image patches. This makes the shared local patches very suitable for transferring weak classifiers cross view-angles via feature shift. Then the covariate shift is utilized to transfer the auxiliary data for updating the weights for the weak classifiers. In addition, the efficiency in the covariate shift step is guaranteed by the step-wise optimization method.

Currently the detector transfer is founded on the assumption that there exists sufficient shared features cross view-angles, but when the change of view-angle is huge, the appearance of the pedestrian may vary sharply. Thus, there may exist only very few image patches shared with the auxiliary data. How to transfer pedestrian detector to a new scenario with huge view-angle disparity is one of our future research directions. A possible solution is to utilize 3D pedestrian model for estimating the underlying variation mechanism of the local image patches cross view-angles, and then perform the transfer learning by allowing for variations between the shared structures or local patches.

6. Acknowledgement

This work is supported in part by National Basic Research Program of China (973 Program): 2009CB320906, and in part by Beijing Natural Science Foundation: 4092042. We would thank Qianqian Xu and Jie Zhang for their help on labeling dataset. We would also thank the anonymous reviewers for their valuable comments.

References

[1] http://pets2007.net.

[2] A. Ahmed, K. Yu, W. Xu, Y. Gong, and E. Xing. Training hierarchal feed-forward visual recognition models using transfer learning from pseudo-taks. In *European Conference on Computer Vision*, 2008.

[3] R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learing Research*, 2005.

[4] T. Bakker. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4, 2003.

[5] S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *Proceedings of computational learning theory*, 2003.

[6] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *11th Annual Conference on Computational Learning Theory*, 1998.

[7] R. Caruana. Multi-task learning. *Machine learning*, 28, 1997.

[8] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *ICML*, 2007.

[9] N. Dalal and B. Triggs. Histogams of oriented gradients for human detection. In *CVPR*, 2005.

[10] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern classification. In *John Wiley and Sons, Inc., 2nd edition*, 2001.

[11] M. Everingham, A. Zisserman, C. K. I. Williams, and L. V. Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf.

[12] A. Farhadi and M. Tabrizi. Learning to recognize activities from the wrong view point. In *ECCV*, 2008.

[13] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 2006.

[14] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proc. Euro. Conf. on Computational Learning Theory*, 1995.

[15] T. Heskes. Empirical bayes for learning to learn. *ICML*, 2000.

[16] C. Huang, H. Ai, T. Yamashita, S. Lao, and M. Kawade. Incremental learning of boosted face detector. In *Proc. of International Conference Computer Vision*, 2007.

[17] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Int. Journal Computer Vision*, 29:5–28, 1998.

[18] O. Javed, S. Ali, and M. Shah. Online detection and classification of moving objects using progressively improving detectors. In *CVPR*, 2005.

[19] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow. In *CVPR*, 2008.

[20] A. Levin, P. viola, and Y. Freund. Unsupervised improvement of visual detector using co-training. In *ICCV*, 2003.

[21] D. J. Mackay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[22] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting algorithm as gradient descent. In *NIPS*, 2003.

[23] N. Oza and S. Russel. Online bagging and boosting. In *Artifical Intelligence and Statistics*, 2001.

[24] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *International Conference on Computer Vision*, 1998.

[25] G. Perkins. Transfer of learning. In *2nd edn, International Encyclopedia of Education*, 1992.

[26] L. Rosenstein and Z. Marx. To transfer or not to transfer. In *Tenique Report*, 2005.

[27] M. Sugiyama, M. Krauledat, and K. R. Muller. Covariate shift adaption by importance weighted cross validation. *Journal of Machine Learning research*, 8:985–1005, 2007.

[28] S. Thrun. Is learning the n-th thing any easier than learning the first? *NIPS*, 1996.

[29] A. Torralba, k. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*, 2004.

[30] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *CVPR*, 2007.

[31] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.

[32] A. Wilson, A. Fern, S. Ray, and P. tadepalli. Multi-task reinforecement learning: a hierarchical bayesian approach. In *ICML*, 2007.

[33] B. Wu and R. Nevatia. Improving part based object detection by unsupervised, online boosting. In *Proc. Computer society Conference on Computer Vision and Pattern Recognition*, 2007.

[34] C. Zhang, R. Hamid, and Z. Zhang. Taylor expansion based classifier adaption: application to person detection. In *Proc. Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[35] X. Zhang, W. Hu, S. Maybank, X. Li, and M. Zhu. Sequential particle swarm optimization for visual tracking. In *CVPR*, 2008.