Transferring a Generic Pedestrian Detector Towards Specific Scenes

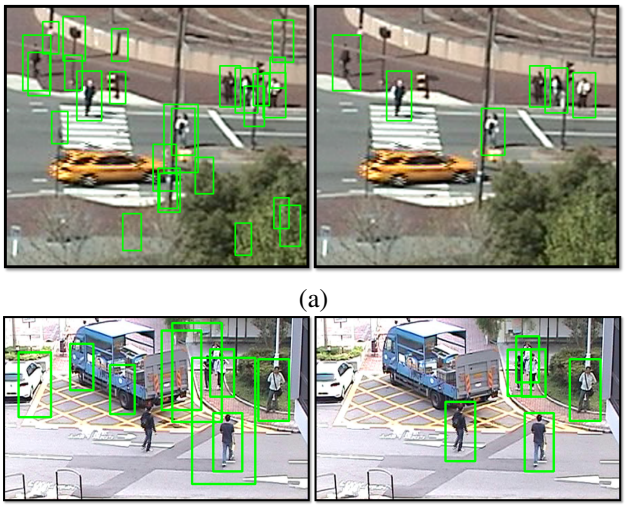
Meng Wang Wei Li Xiaogang Wang
Department of Electronic Engineering, The Chinese University of Hong Kong
{mwang,liwei,xgwang}@ee.cuhk.edu.hk

Abstract

The performance of a generic pedestrian detector may drop significantly when it is applied to a specific scene due to mismatch between the source dataset used to train the detector and samples in the target scene. In this paper, we investigate how to automatically train a scene-specific pedestrian detector starting with a generic detector in video surveillance without further manually labeling any samples under a novel transfer learning framework. It tackles the problem from three aspects. (1) With a graphical representation and through exploring the indegrees from target samples to source samples, the source samples are properly re-weighted. The indegrees detect the boundary between the distributions of the source dataset and the target dataset. The re-weighted source dataset better matches the target scene. (2) It takes the context information from motions, scene structures and scene geometry as the confidence scores of samples from the target scene to guide transfer learning. (3) The confidence scores propagate among samples on a graph according to the underlying visual structures of samples. All these considerations are formulated under a single objective function called Confidence-Encoded SVM. At the test stage, only the appearance-based detector is used without the context cues. The effectiveness of the proposed framework is demonstrated through experiments on two video surveillance datasets. Compared with a generic pedestrian detector, it significantly improves the detection rate by 48% and 36% at one false positive per image on the two datasets respectively.

1. Introduction

Significant progress has been made on pedestrian detection in the past decade [4, 7, 2, 15]. However, it is still a challenging task to train a generic pedestrian detector which works reliably on all kinds of scenes. It not only requires a huge training set to cover a very large variety of view points, resolutions, lighting conditions, motion blur effects and backgrounds observed under numerous conditions, but also a very complex model to handle so many variations. It has been observed that the performance of state-of-the-art pedestrian detectors trained on a general dataset drops significantly when they are applied to some video sequences taken from specific scenes [5]. However, for a particular application, the variety of scenes is likely to be small. For example, most cameras are stationary in video surveillance. The video sequences captured from a single camera only have limited variations on view points, resolutions, lighting conditions and backgrounds. Therefore, it is much easier to train a pedestrian detector with high accuracy using the samples from the target scene (see examples in Figure 1). However, this often requires a lot of extra labeling effort and it is not practical in many scenarios. In recent years, some research efforts [10, 18, 23, 19, 21] have been made to automatically train a scene-specific detector starting with a generic detector without labeling data from the target scene. However, many existing approaches are based on ad hoc rules and their detectors have the risk of drifting during the training process. In this work, we tackle this problem by



(a) (b)

Figure 1: Detection results on the MIT Traffic dataset (a) and the CUHK Square dataset (b). The left is the results of a generic pedestrian detector using HOG+SVM [4]. The right is the results of a scene-specific detector automatically trained by our approach.

proposing a transfer learning framework. Transfer learning has been used to solve various domain adaptation problems with great achievement in both theories and applications. However, there is only very limited work [16] on applying it to pedestrian detection and many important issues regarding this problem are not well studied yet.

1.1. Motivations and Contributions

The motivations of the proposed framework can be explained from three aspects as below.

(1) The distribution of the source dataset used to train the generic detector usually does not match that of the samples from the target scene. As examples shown in Figure 2 (c)(d), some source samples are more similar to the target samples, because they are taken under similar view points, resolutions and lighting conditions, or the negative samples come from the same background categories (trees, streets and buildings). Therefore, it is desirable to assign larger weights to such samples during training rather than treating all the samples equally. In this work, we build a graph according to the visual similarities between source samples and target samples, and re-weight the source samples according to their indegrees from target samples. The indegree can detect the boundary between the distributions between the source samples and target samples. It was not well studied in previous work on transfer learning.

(2) Context information, such as motions and scene structures, provides useful cues to guide transfer learning. These cues can be used to selected confident positive/negative samples from the target scene to train the appearance-based scene-specific detector. The context information is complementary to the image appearance and they are used to compute the confidence scores of target samples in our framework. The confidence scores are well incorporated into our proposed Confidence-Encoded SVM, in which target samples with small confidence scores have little influence on the training of the scene-specific detector. Confidence-Encoded SVM provides a more principled and reliable way to utilize the confidence information than existing approaches [14, 10, 21] which selected target samples by hard-thresholding the confidence scores and caused the problems of drifting or inefficient training.

(3) According to the context information, only a small portion of target samples have high confidence scores and it may predict wrong labels. Our approach propagates the confidence scores among target samples along a graph and correct wrong labels according to underlying visual structures of samples (see examples in Figure 3). It improves the efficiency of transfer learning.

All the considerations are well integrated under a single objective function of the proposed Confidence-Encoded SVM. The effectiveness of the proposed framework is demonstrated through experiments on two video surveillance datasets. It significantly improves the performance compared with the generic pedestrian detector as well as other domain adaptation methods previously applied to pedestrian detection. Surprisingly, it even outperforms the scene-specific detector trained on more than 400 frames (including more than $1,500$ positive samples) from the target scene with manually labeled ground truth.

1.2. Related Work

A typical way of adapting a generic pedestrian detector to a specific scene is to automatically select positive and negative samples from the target scene to re-train the scene-specific detector iteratively. Existing approaches were mainly based on ad hoc rules. Rosenberg et al. [18] selected samples which were confidently classified by the appearance-based generic detector. Nair et al. [14] labeled the target samples according to the background subtraction results. Wang et al. [21] integrated multiple cues of motions, path models [22], locations, sizes and appearance to select confidence positive and negative samples from the target scene. In [10, 23], co-training was used to train two detectors based on different types of features iteratively. In these approaches, target samples were selected by hard-thresholding the confidence scores obtained from the appearance-based detectors or context information. It is unreliable and discards some useful information. An aggressive threshold makes the detector drift, while a conservative threshold makes the training inefficient and results in many rounds of re-training to converge.

Transfer learning provides a more principled way to solve the domain adaptation problems. In the recent years, it has been successfully applied to object recognition[9], scene categorization [17], action recognition[11] and retrieval [25]. In cross-domain SVM [8] and TrAdaBoost [3], samples in the source dataset and the target dataset are re-weighted differently. Wu and Srihari [24] introduced a weighted soft margin SVM to incorporate prior knowledge in the training process. However, not much work has been done on pedestrian detection by transfer learning yet. Pang et al. [16] proposed a transfer learning approach to transfer features and to adapt weights of weak classifiers learned from the source dataset to the target dataset to handle the variation of view points. It assumed that some samples in the target set were manually labeled. As discussed in Section 1.1, many important issues on transfer learning in pedestrian detection are to be studied yet.

2. Method

2.1. Overview

Our proposed transfer learning framework is summarized in Algorithm 1 and relevant notations are summarized in Table 1. It starts with a generic pedestrian detector $\mathbf{\Theta}$

Table 1: Notation

$(\mathbf{w}, b)$	parameters of SVM weights and bias
$\mathcal{D}^s$	source dataset
$\mathcal{D}^t$	target dataset
$\mathcal{V}$	a video sequence from the target scene
$\mathbf{\Theta}$	the pedestrian detector
$\mathbf{c}_0$	initial confidence estimation on $\mathcal{D}^t$
$\mathbf{\Phi}$	assigns $\mathbf{c}_0$ to $\mathcal{D}^t$ according to scene context information
$\mathbf{c}$	propagated confidence on $\mathcal{D}^t$
$\boldsymbol{\nu}$	confidence on $\mathcal{D}^s$
$\mathbf{\Psi}$	assigns $\boldsymbol{\nu}$ on $\mathcal{D}^s$
G	objective function of Confidence-Encoded SVM

using HOG+SVM [4] trained on a general source dataset $\mathcal{D}^s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$[1]. An unlabeled video sequence $\mathcal{V}$ is captured from the target scene. $\mathbf{x}_i^s$ is a source sample and $y_i^s = \pm 1$ is its label. $+1$ indicates positive samples and -1 indicates negative samples. $\mathbf{\Theta} = (\mathbf{w}_0, b_0)$ is parameterized by the weights and the bias term of (linear) SVM. Once $\mathbf{\Theta}$ is applied to $\mathcal{V}$, a target dataset $\mathcal{D}^t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$ is obtained by selecting the target samples whose detection scores are positive, i.e. $\mathbf{\Theta}(\mathbf{x}_i^t) > 0$. Since the generic detector is far from perfect, there are many false positives in $\mathcal{D}^t$. The context information helps to assign a label y_i^t and a confidence score $c_i \in [-1, +1]$ to each target samples $\mathbf{x}_i^t$. $c_i = \pm 1$ indicates the highest confidence on the predicted label y_i^t and $c_i = 0$ indicates no confidence. Each source sample $\mathbf{x}_i^s$ is re-weighted by ν_i according to its visual distance to the target samples. A new scene-specific detector $(\mathbf{w}_r, b_r)$ is trained on both $\mathcal{D}^s$ and $\mathcal{D}^t$ given $\boldsymbol{\nu} = \{\nu_i\}$ and $\mathbf{c}_0 = \{c_{0i}\}$ under the proposed Confidence-Encode SVM in Eq (4). In the Confidence-Encoded SVM, the initial confidence estimation $\mathbf{c}_0$ propagates to confidence scores $\mathbf{c}$ which is jointly estimated with $(\mathbf{w}_r, b_r)$. Once the detector $\mathbf{\Theta}$ is updated with $(\mathbf{w}_r, b_r)$, it is applied to $\mathcal{V}$ again to start the next round of training. Experiments on two different datasets show that our approach is very efficient in training and it quickly converges after one or two rounds. The details of the framework are given in the subsections below.

2.2. Initial Estimation of Confidence Scores

In [21], the labels of target samples were predicted and their confidence scores $\mathbf{c}_0$ were computed from the context information, which included multiple cues of motions, path models, locations and sizes. In this work, we adopt the same approach with slight difference. In [21], mean shift clustering based on image appearance was used to further exclude unreliable positive or negative samples. It required carefully controlling the bandwidth of mean shift to reject a proper portion of outliers. Without this step, the detector trained by [21] will drift. However, this step is not required by our approach, since confidence propagation and Confidence-Encoded SVM adopted at later stages make our approach quite robust to the existence of outliers. It is more convenient to use. Also, note that our proposed method is a general framework, which can well integrate with other ways of computing the initial confidence score $\mathbf{c}_0$ depending on the application scenarios.

[1]The INRIA dataset [4] is used in this work.

Algorithm 1 The Proposed Transfer Learning Framework

Input:
The generic detector $(\mathbf{w}_0, b_0)$
The source dataset $\mathcal{D}^s$
A video sequence $\mathcal{V}$ from the target scene
The target dataset $\mathcal{D}^t \leftarrow \emptyset$
Output:
The scene-specific detector $(\mathbf{w}, b)$
for $r = 1, \cdots, R$ **do**
$(\mathbf{w}_r, b_r) \leftarrow (\mathbf{w}_{r-1}, b_{r-1})$
$\mathcal{D}_r^t \leftarrow \mathbf{\Theta}(\mathbf{w}_r, b_r, \mathcal{V})$
$\mathcal{D}^t \leftarrow \mathcal{D}^t \cup \mathcal{D}_r^t$
$\mathbf{c}_{r,0} \leftarrow \mathbf{\Phi}(\mathcal{D}_r^t)$
$\boldsymbol{\nu}_r \leftarrow \mathbf{\Psi}(\mathcal{D}^s, \mathcal{D}^t, \mathbf{c}_{r,0})$
$(\mathbf{w}_{r,0}, b_{r,0}) \leftarrow (\mathbf{w}_r, b_r)$
`/* Optimize the Confidence-Encoded SVM */`
$k = 0$
repeat
$k \leftarrow k + 1$
$\mathbf{c}_{r,k} \leftarrow \underset{\mathbf{c}}{\operatorname{argmin}}\, G(\mathbf{c}, \mathbf{w}_{r,k-1}, b_{r,k-1}; \mathbf{c}_{r,0}, \boldsymbol{\nu}_r, \mathcal{D}^s, \mathcal{D}^t)$
$(\mathbf{w}_{r,k}, b_{r,k}) \leftarrow \underset{\mathbf{w},b}{\operatorname{argmin}}\, G(\mathbf{c}_{r,k}, \mathbf{w}, b; \mathbf{c}_{r,0}, \boldsymbol{\nu}_r, \mathcal{D}^s, \mathcal{D}^t)$
until Converge
$(\mathbf{w}_r, b_r) \leftarrow (\mathbf{w}_{r,k}, b_{r,k})$
end for
$(\mathbf{w}, b) \leftarrow (\mathbf{w}_R, b_R)$

2.3. Re-weighting Source Samples

As examples shown in Figure 2, some source samples better match the target dataset than others and therefore they should gain larger weights during training. To re-weight source samples, a graph between $\mathcal{D}^t$ and $\mathcal{D}^s$ is built, where nodes are samples and edges are created based on K-nearest-neighbors (KNNs). Under the L2 distance, $d_{j,i} = ||\mathbf{x}_j^t - \mathbf{x}_i^s||^2$, for each target sample $\mathbf{x}_j^t$, there is an edge pointing from j to each of its KNNs in the source dataset as shown in Figure 2. The weight of the edge is

$$w_{j,i} = \exp(-\frac{d_{j,i}^2}{\sigma^2}), j = 1, \cdots, n_t, i = 1, \cdots, n_s \quad (1)$$

The indegree of a source sample is defined as

$$\text{indegree}(\mathbf{x}_i^s) = \sum_{\mathbf{x}_i^s \in \text{KNN}(\mathbf{x}_j^t)} w_{j,i} \quad (2)$$

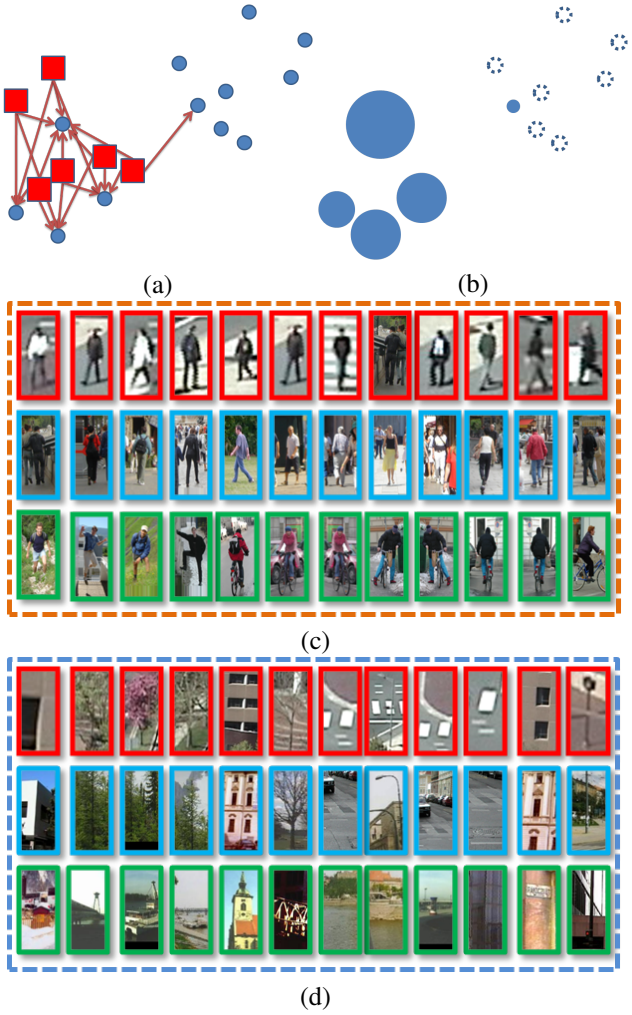


Figure 2: (a) Red squares indicate target samples and blue points indicate source samples. Each target sample has K ($K = 3$) directed edges pointing towards it K nearest neighbors in the source set. If a source sample is outlier of the target set, it has a small indegree. (b) The sizes of points indicate the indegrees of source samples. Some source samples have zero indegree and they are denoted as dashed circles. (c) Positive target samples (first row), positive source samples with large indegrees (second row), and positive source samples with zero indegree (third row). (d) Negative target samples (first row), negative source samples with large indegrees (second row), and negative source samples with zero indegree (third row). The source set is the INRIA dataset and the target set is the MIT Traffic dataset.

As shown in Figure 2, if a source sample is an inlier of the target set, there are many edges pointing to it and therefore it has a large indegree. If it is an outlier of the target set, its indegree is small and could be zero. Indegree has been widely studied in complex network [13]. In our application scenario, it can better detect the boundary between the distributions of the target dataset and the source dataset. Indegree has not been studied in transfer learning yet. Most transfer learning algorithms [8] directly use KNN to estimate the distance between a source sample and the target dataset.

Figure 2 (c) and (d) show source samples with large indegrees and zero indegree. It is observed that positive source samples with large indegrees have similar view points as the target samples, and negative source samples with large indegrees are from the same background categories (trees, buildings, roads and poles) as the target samples.

The confidence score ν_i of a source sample is computed as a sum of indegrees weighted by the initial confidence scores of the target samples,

$$\nu_i = \sum_{\mathbf{x}_i^s \in \text{KNN}(\mathbf{x}_j^t)} w_{j,i} c_{0j}. \tag{3}$$

The confidence scores of all the source samples are further normalized to the range of $[-1, +1]$.

2.4. Confidence-Encoded SVM

The proposed Confidence-Encoded SVM is an extended version of L2-regularized L2-loss SVM. Its objective function is shown in Eq (4).

$$\begin{aligned} \min_{\mathbf{c},\mathbf{w},b} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n_s}(\nu_i\xi_i^s)^2 + C\sum_{j=1}^{n_t}(c_j\xi_j^t)^2 + \\ & \frac{\mu}{2}\mathbf{c}^T\mathbf{L}\mathbf{c} + \frac{\lambda}{2}(\mathbf{c}-\mathbf{c}_0)^T\mathbf{A}(\mathbf{c}-\mathbf{c}_0) \\ \text{s.t.} \quad & y_i^s(\mathbf{w}^T\mathbf{x}_i^s + b) \geq 1 - \xi_i^s, i = 1, \cdots, n_s \\ & y_j^t(\mathbf{w}^T\mathbf{x}_j^t + b) \geq 1 - \xi_j^t, j = 1, \cdots, n_t \\ & \xi_i^s \geq 0, i = 1, \cdots, n_s \\ & \xi_j^t \geq 0, j = 1, \cdots, n_t. \end{aligned} \tag{4}$$

C, μ and λ are pre-set parameters. $\mathbf{c} = (c_1, \ldots, c_{n_t})$ are the propagated confidence scores on the target dataset. They are jointly estimated with SVM weights and bias. The slack penalty of misclassifying a source (target) sample $\mathbf{x}_i^s(\mathbf{x}_j^t)$ is proportional to its confidence score ν_i (c_j). The lower confidence a sample has, the smaller penalty it is imposed on if misclassified and the smaller influence it has on training SVM. Many approaches such as [21] selected positive/negative target samples by hard-thresholding the confidence scores and treated these samples equally when training SVM. It is special case of ours, considering c_j can only be ± 1 or 0. Our approach certainly has advantages, since it does not require thresholding which causes errors. If the threshold is aggressive, some wrongly labeled samples are used to train SVM and cause the drifting problem. If the threshold is conservative, not enough samples are selected and the performance of the detector improves slowly even after many rounds of training. It also does not make sense to treat all the training samples with different confidences

equally after thresholding. Experiments on the MIT Traffic dataset show that the approach proposed in [21] converges after 10 rounds of iterations, while our approach converges after 2 rounds of iterations with much higher efficiency.

2.4.1 Confidence Propagation

Utilizing context information alone, only a small portion of target samples have high confidence scores and some predicted labels may be wrong (see examples in Figure 3). Image patches from the same scene form clusters and manifolds based on their visual similarities. If two image patches are visually similar, they should have the same label. This inspired us to propagate confidence scores to obtain more samples with high confidence and reduce the confidence of samples with wrong labels.

The estimation of confidence scores $\mathbf{c}$ depends on three terms in Eq (4). $\mathbf{c}^T\mathbf{L}\mathbf{c}$ comes from graph Laplacian and requires that visually similar samples have similar confidence scores. Graph Laplacian was used for label propagation in semi-supervised learning [26, 27] and image retrieval [20, 12]. To the best of our knowledge, not much light is shed on this approach in detection works. A pairwise weight matrix $\mathbf{W}$ is calculated from $\mathcal{D}^t$ by

$$w_{i,j} = \exp(-\frac{||\mathbf{x}_i^t - \mathbf{x}_j^t||^2}{\sigma^2}). \quad (5)$$

It is further sparsified by setting $w_{ij} = 0$, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are not the K nearest neighbors of each other. A diagonal matrix $\mathbf{D}$ is defined by $\mathbf{D}_{ii} = \sum_{j=1}^{n_t} w_{ij}$. Then the graph Laplacian constructed from the above sample set is $\mathbf{L} = \mathbf{D} - \mathbf{W}$. Although our work only considers visual distances, other cues which characterize the structures of samples can also be used to compute the graph Laplacian. For example, temporal consistency of samples can be exploited if tracking is available [1].

$\mathbf{A}$ is a diagonal matrix where $\mathbf{A}_{jj} = |c_{0j}|$. Therefore,

$$(\mathbf{c} - \mathbf{c}_0)^T\mathbf{A}(\mathbf{c} - \mathbf{c}_0) = \sum_{j=1}^{n_t}(c_j - c_{0j})^2|c_{0j}| \quad (6)$$

is used to regularize $\mathbf{c}$ from its initial estimation $\mathbf{c}_0$. It is assumed that if c_{j0} is low, which means that the context information does not has a strong opinion on the label of $\mathbf{x}_j^t$, then its confidence score can be easily influenced by other samples with less penalty. Otherwise, its confidence score can be changed only when there is strong evidence from other samples because the context information has a strong opinion on its label.

The third term $\sum_{j=1}^{n_t}(c_j\xi_j^t)^2$ tends to assign small confidence scores to samples misclassified by SVM (with large ξ_j^t), since the context information and appearance-based classifier have disagreement on them.

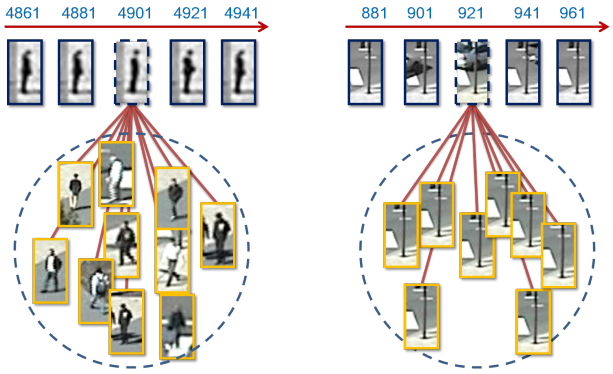


Figure 3: Example on the left: a pedestrian is stationary for a long period and therefore is labeled as a negative sample with an initial high confidence score according to the motion cue. Its confidence score gets close to zero after confidence propagation because many other samples with similar visual appearance to it are labeled as positive samples with high confidence scores Therefore it will not have bad influence on training. Example on the right: a background patch is labeled as a negative sample with a low initial confidence score because a vehicle happens to pass by and causes motions. Its confidence score becomes high after confidence propagation because some similar background patches are labeled as negative samples with high confidence scores.

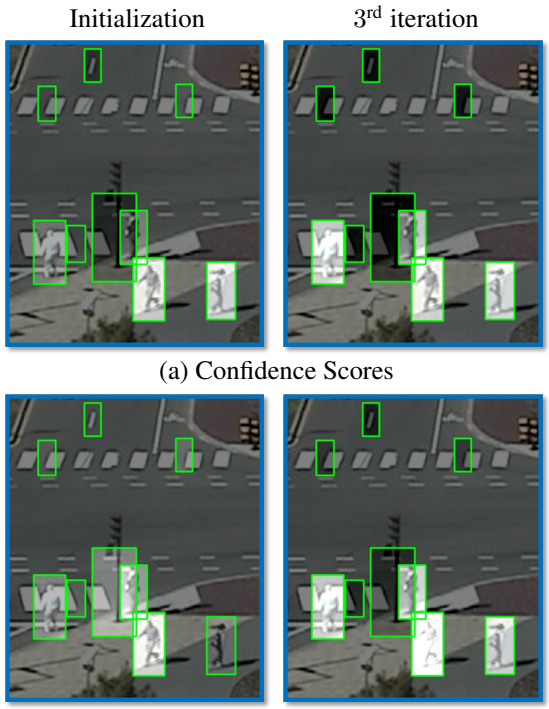
2.4.2 Optimization

We optimize Equation 4 in an iterative manner. Denote the objective function by $G(\mathbf{c}, \mathbf{w}, b)$. Optimization starts with an initial model $(\mathbf{w}_0, b_0)$. At each iteration k, let $\mathbf{c}_k$ minimize $G(\mathbf{c}, \mathbf{w}_{k-1}, b_{k-1})$. Since it is a convex quadratic function, finding an optimal $\mathbf{c}_k$ is straightforward by setting its derivative to be 0. We obtain the parameters $(\mathbf{w}_k, b_k)$ of a new model by minimizing $G(\mathbf{c}_k, \mathbf{w}, b)$ using a modified version of LIBLINEAR [6], which is based on the Trust Region Newton Method (TRON). This optimization algorithm converges since the objective function monotonically decreases after each step. According to our experimental results, it usually converges within five iterations. Figure 4 shows an example of how the confidence scores and detection scores by SVM change after three iterations.

3. Experiments

Experiments are conducted on the MIT Traffic dataset [21] and the CUHK Square dataset which is constructed by us. The two scenes are shown in Figure 5[2]. We adopt the PASCAL "50% rule", i.e. the overlapping region between the detection window and the ground truth must be at least 50% of the union area. Recall Rate versus False Positive Per Image (FPPI) is used as the evaluation metric.

[2]The dataset can be downloaded from `http://www.ee.cuhk.edu.hk/~xgwang/CUHK_square.html`.



(a) Confidence Scores

(b) Detection Scores by SVM

Figure 4: The confidence scores (a) and detection scores by SVM (b) change after three iterations when optimizing the Confidence-Encoded SVM. Green windows indicate image patches in $\mathcal{D}^t$. A bright window indicates that the score is close to $+1$ and a dark window indicates that the score is close -1. At initialization, there are large differences between the confidence scores and detection scores. After three iterations, they look more consistent and correct. The experiment is on the MIT Traffic dataset.

(a) (b)

Figure 5: MIT Traffic dataset (a) and CUHK Square dataset (b).

3.1. Data sets

MIT Traffic dataset. It is a 90-minute long surveillance video sequence at 30 fps. The video captures a traffic scene at a street intersection. It includes walking pedestrians and moving vehicles in a far field. Occlusions and varying illumination conditions apply. In [21], 420 frames were uniformly sampled from the first 45 minutes video and were used to train the scene-specific pedestrian detector. 100 frames were uniformly sampled from the last 45 minutes video and were used for testing. We follow the same convention.

CUHK Square dataset. Similar to the MIT Traffic dataset, it is also captured by a stationary camera from a bird-view. It is a 60-minutes long video at 25 fps. Since the camera was placed at a location much lower than that in the MIT Traffic dataset, perspective deformation is much more challenging. 350 frames are uniformly sampled from the first 30 minutes video and used to train the scene-specific detector. 100 frames uniformly sampled from the last 30 minutes video are used for testing.

Note that when our approach trains the scene-specific detector, it does not use any labeled samples from the videos. The test setup is identical to [21]. The context cues are computed in the same way as [21]. In the test stage, only the appearance-based detector without context cues are used.

3.2. Parameter Setting

In Eq (4), we choose $C = 1/(\frac{1}{n_s+n_t}(\sum_{i=1}^{n_s}||\mathbf{x}_i^s|| + \sum_{i=1}^{n_t}||\mathbf{x}_i^t||))^2$ as recommended by SVMLight[3], and set $\mu = \lambda = 1$. Actually, the performance of our approach is stable when μ and λ change in a relatively large range. σ in Eq (1) is defined by

$$\sigma^2 = \frac{1}{n_t \cdot n_s}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} d_{ji}^2$$

where d_{ji} is given by the L2 distance between a source example $\mathbf{x}_i^s$ and a target example $\mathbf{x}_j^t$. In Eq (5), σ is given by

$$\sigma^2 = \frac{1}{(n_t-1)^2}\sum_{j=1}^{n_t}\sum_{i=1}^{n_t} d_{ij}^2.$$

The experiments on the two datasets use the same fixed-value parameters for μ and λ, and compute parameters in the same way.

3.3. Results

We compare with the following approaches:

- A generic HOG+SVM detector trained on the INRIA dataset (Generic).
- The approach of automatically adapting a generic detector to a specific scene utilizing multiple cues proposed in [21] (Wang CVPR'11).
- The approach of automatically adapting a generic detector to a specific scene using background subtraction to select samples (it is similar to [14], but its detector is HOG+SVM not boosting) (Nair CVPR'04).
- A scene-specific HOG+SVM detector trained on N manually labeled frames from the target scene

[3] `http://svmlight.joachims.org/`. In our implementation, we used LIBLINEAR, but the parameter setting follows the suggestion of SVMLight.

(Manual(N)). Negative examples are firstly randomly sampled, and then bootstrapped using a typical training method in [4].

In the discussions below, when we talk about detection rates, it is assumed that FPPI $= 1$. Figure 7 (c) and (f) show that the scene-specific detector obtained by our approach significantly outperforms the generic detector. On the MIT Traffic test dataset, it improves the detection rate from 21% to 69%. On the CUHK Square test dataset, it improves the detection rate from 15% to 51%. It even outperforms the scene-specific detectors trained with manually labeled frames from the target scenes. When all the 420 and 350 training frames of the two datasets are used[4], the detection rates are 66% and 45% respectively. It may be partially due to the fact that our training includes the source dataset. It is also crucial to properly re-weight the source dataset, since [21] has shown that directly combining the source dataset and the labeled target dataset to train a detector could not beat the detector trained on the labeled target dataset alone.

Figure 7 (a)(b) and (d)(e) compare with the other two automatic scene adaptation approaches (Wang CVPR'11 [21] and Nair CVPR'04 [14]) on both training and testing sets. Our approach clearly outperforms them in both efficiency and accuracy. [14] converges after four rounds on the two datasets and its accuracy is much lower than [21] and ours. Compared with [21], our approach converges after much fewer rounds (2 versus 10 on the MIT Traffic dataset, and 1 versus 7 on the CUHK Square dataset), and at the same time leads to a higher performance (about 7% improvement on detection rate). This is because that [21] is based on ad hoc rules and hard-thresholding, which greatly reduce its efficiency. At the first round of training, both [21] and ours have the same target set and initial confidence scores $\mathbf{c}_0$, since they utilize the same context information. The Confident-Encoded SVM achieves a 48% detection rate after the first round training, while [21] only achieves 30%.

In Figure 6, we further investigate the effectiveness of (1) including target samples for training, (2) confidence propagation, and (3) re-weighting source samples using indegrees, on the MIT Traffic dataset. The inclusion of target samples for training is essential. Only re-weighting source samples without including target samples (denoted as "Source Only"), the detection rate is only marginally improved from 21% to 26% compared with the generic detector. Without confidence propagation (denoted as "No Propagation"), it takes two more rounds to converge and the detection rate drops by 11%. If source samples are not re-weighted (denoted as "No Source Re-weighting"), the detection rate drops by 6%. If source samples are re-weighted directly using KNN as [8] (denoted as "Outdegree"), the detection rate drops by 5%.

[4]The corresponding numbers of positive examples are 1573 and 956, respectively.

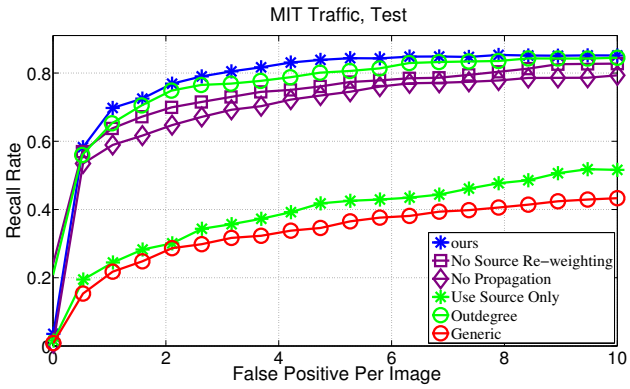


Figure 6: Investigate the effectiveness of different factors in our approach on the MIT Traffic test data.

4. Conclusions and Discussions

In this paper, we propose a new transfer learning framework to automatically train a scene-specific pedestrian detector starting with a generic detector without manually labeling any data in the target scene. The source dataset, the context information, and the visual structures of target samples are well integrated under the proposed Confidence-Encoded SVM. It significantly outperforms not only the generic pedestrian detector, but also the scene-specific detector trained on manually labeled frames. It quickly converges after one or two rounds of training. It is well applied to two different scenes without tuning parameters.

Although HOG+SVM is used as the pedestrian detector in this work, the proposed framework may also be extended to other more advanced detectors. For example, it is feasible to extend Confidence-Encoded SVM to Latent SVM [7] by adding more terms on parts and geometry in Eq (4) with the same optimization strategy. This will be our future work. However, on the two datasets used in this paper, Latent SVM and other advanced part-based models may not be good choices because pedestrians in these datasets are very small in size and part-based models usually require higher resolutions. Also HOG+SVM is much faster and thus more suitable for some online surveillance applications.

5. Acknowledgement

This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project No. CUHK417110 and CUHK417011) and National Natural Science Foundation of China (Project No. 61005057).

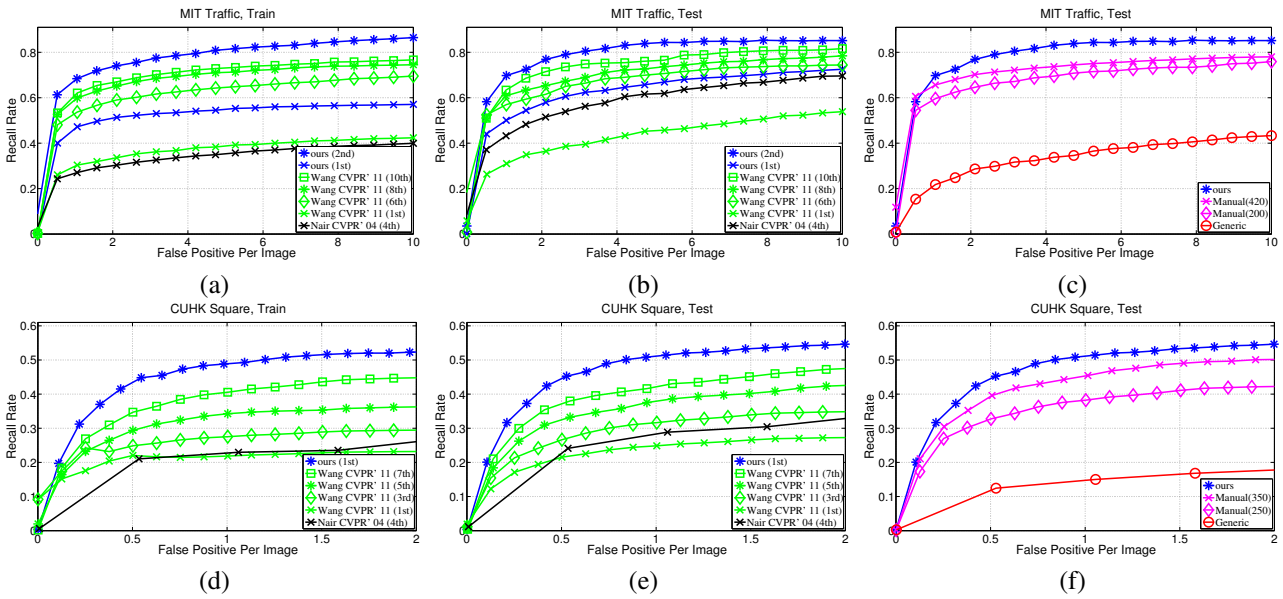


Figure 7: Results on MIT Traffic dataset (a)-(c) and CUHK Square dataset (d)-(f). (a)-(b) and (d)-(e) compare with two automatic scene adaptation approaches (Wang CVPR'11 [21] and Nair CVPR'04 [14]) on both training and testing sets after different rounds of training. (c) and (f) compare with the generic detector and the scene-specific detector trained on different numbers of manually labeled frames.

References

[1] K. Ali, D. Hasler, and F. Fleuret. FlowBoost — Appearance learning from sparsely annotated video. In *Proc. CVPR*, 2011.

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Proc. ICCV*, 2009.

[3] W. Dai, Q. Yang, and G. Xue. Boosting for transfer learning. In *Proc. ICML*, 2007.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. on PAMI*, 2011.

[6] R. Fan, K. Chang, and C. Hsieh. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. on PAMI*, 32:1627–1645, 2010.

[8] W. Jiang, E. Zavesky, S. Chang, and A. Loui. Cross-domain learning methods for high-level visual concept classification. In *Proc. of ICIP*, 2008.

[9] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proc. CVPR*, 2011.

[10] Levin, Viola, and Freund. Unsupervised Improvement of Visual Detectors using Co-Training. In *Proc. ICCV*, 2003.

[11] J. Liu, M. Shah, B. Kuipers, and S. Savarese. Cross-view action recognition via view knowledge transfer. In *Proc. CVPR*, 2011.

[12] W. Liu, Y.-G. Jiang, J. Luo, and S.-F. Chang. Noise resistant graph ranking for improved web image search. In *Proc. CVPR*, 2011.

[13] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. ACM SIGCOMM*, 2007.

[14] V. Nair. An unsupervised, online learning framework for moving object detection. In *Proc. CVPR*, 2004.

[15] W. Ouyang and X. Wang. A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling. In *Proc. CVPR*, 2012.

[16] J. Pang, Q. Huang, S. Yan, S. Jiang, and L. Qin. Transferring Boosted Detectors Towards Viewpoint and Scene Adaptiveness. *IEEE Trans. on Image Processing*, 20:1388–1400, 2011.

[17] G. Qi, C. Aggarwal, Y. Rui, Q. Tian, S. Chang, and T. Huang. Towards cross-category knowledge propagation for learning visual concepts. In *Proc. CVPR*, 2011.

[18] C. Rosenberg, M. Hebert, and H. Schneiderman. Semi-supervised self-training of object detection models. In *Proc. of IEEE Workshop on Application of Computer Vision*, 2005.

[19] P. Roth, S. Sternig, H. Grabner, and H. Bischof. Classifier grids for robust adaptive object detection. In *Proc. CVPR*, 2009.

[20] J. Wang, Y.-G. Jiang, and S.-F. Chang. Label diagnosis through self tuning for web image search. In *Proc. CVPR*, 2009.

[21] M. Wang and X. Wang. Automatic Adaptation of a Generic Pedestrian Detector to a Specific Traffic Scene. In *Proc. CVPR*, 2011.

[22] X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Trans. on PAMI*, 31:539–555, 2008.

[23] B. Wu and R. Nevatia. Improving part based object detection by unsupervised, online boosting. In *Proc. CVPR*, 2007.

[24] X. Wu and R. Srihari. Incorporating prior knowledge with weighted margin support vector machines. In *Proc. of ACM SIGKDD*, 2004.

[25] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proc. of ACM Multimedia*, 2007.

[26] D. Zhou and B. Schölkopf. Semi-supervised learning on directed graphs. In *Proc. NIPS*, 2005.

[27] X. Zhu. Learning from labeled and unlabeled data with label propagation. In *Technical Report CMU-CALD-02-107, Carnegie Mellon University*, 2002.