# Adaptive Structural Model for Video Based Pedestrian Detection

Junjie Yan, Bin Yang, Zhen Lei, Stan Z. Li

Center for Biometrics and Security Research & National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{jjyan,bin.yang,zlei,szli}@nlpr.ia.ac.cn

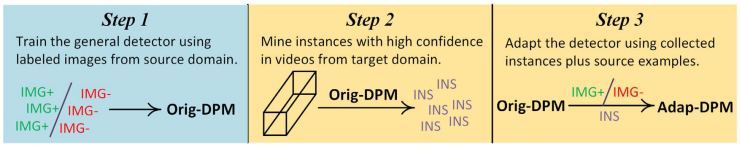
**Abstract.** The performance of generic pedestrian detector usually declines seriously for videos in novel scenes, which is one of the major bottlenecks for current pedestrian detection techniques. The conventional works improve pedestrian detection in video by mining new instances from detections and adapting the detector according to the collected instances. However, when treating the two tasks separately, the detector adaptation suffers from the defective output of instance mining. In this paper, we propose to jointly handle the instance mining and detector adaption using an adaptive structural model. The regularization function of the model is applied on detector to prevent overfitting in adaption, and the loss function is designed to evaluate the combination of mined instances set and detector. Particularly, we extend the Deformable Part Model (DPM) to adaptive DPM, where an adaptive feature transformation defined on low-level HOG cell is learned to reduce the domain shift, and the regularization function for the detector is conducted on the transformation. The loss of the instance set and detector is measured by a cost-flow network structure which incorporates both the appearance of frame-wise detections and their spatio-temporal continuity. We demonstrate an alternating minimization procedure to optimize the model. The proposed method is evaluated on ETHZ, PETS2009 and Caltech datasets, and outperforms baseline DPM by 7% in terms of mean miss rate.

## 1 Introduction

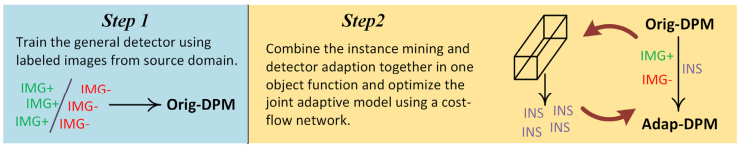
Pedestrian detection has been a hot research topic for decades. Benefitting from the advances in low-level feature and high-level model, static image based pedestrian detection has achieved impressive progresses in both effectiveness [1–8] and efficiency [9–13]. With well-designed feature and model, current detectors trained on a large set can handle some occlusions, pose and viewpoint variations. However, the performance on novel scenes may drop disastrously due to the domain shift. For example, according to the evaluation in [14], the state-of-the-art pedestrian detector Crosstalk [12] achieves 19% mean miss rate on INRIA test set, while increases to 54% on Caltech Pedestrian Benchmark.

To handle the domain shift, one promising solution is the automatic adaption of the generic detector to the target scenes, as recently explored in [15–22]. Most of the works followed an unsupervised paradigm since the annotations in novel scenes are often unavailable. These works usually considered two tasks in detector adaptation. The

first is to mine new positive and negative instances of the target video in an unsupervised manner, and the second is to adapt the the generic detector when the training instances of the target video are collected. The standard paradigm in these works is as follows (Fig. 1 (a)), (1) conduct frame-wise detection on video; (2) use various information (e.g., tracking, background substraction and optical flow) to mine instances from the detection result; (3) take the mined instances as online training samples to update the detector.




**(a) Conventional detector adaption paradigm**




**(b) The proposed joint adaptive model**

**Fig. 1.** Different paradigms of detector adaptation for video based pedestrian detection. The conventional methods take the training instance mining and detector adaptation as two separate tasks. In this paper, to explore the benefit from each other, we propose to optimize the instance mining and detector adaptation jointly in a structural model.

The motivation of our approach is that the instance mining and detector adaptation procedures should be explored jointly. For example, the detector can benefit from the confidently mined clear instances in adaptation, and the well adapted detector can further improve the quality of mined instances. Given the frame-wise detection result, we build a joint structural model to find an optimal combination of adapted detector and new instances from the video (Fig. 1 (b)). Particularly, we build a frame-wise detector with DPM. To avoid the complexity in shifting the high dimensional DPM parameters directly, we propose to use a linear transformation to capture the domain shift on low-level HOG cell, which can effectively capture the variations in different conditions with much less parameters. The loss function in the structural model is built on a cost-flow network to capture the structure in video, where both the frame-wise appearance and the video continuity among frames are encoded. We show that when the instance set is fixed, the optimal detector can be solved by standard quadratic programming, and when the detector is fixed, the model can be solved by efficient successive shortest-path algorithm. In optimizing the structural model, we conduct an alternating scheme to conduct frame-wise detection and structural model adaptation iteratively.

We validate the detection performance following the protocol provided in [14] on challenging videos from ETHZ, Caltech, and PETS2009. Our structural model de-

creases mean miss rate by more than $7\%$ compared to the baseline DPM (version 5), and outperforms the best published results by a $2\%$ margin.

The rest of the paper is organized as follows. We discuss the related work in Section 2 and provide the background of DPM and cost-flow network based data association in Section 3. The joint structural model and the corresponding optimization method are described in Section 4. We demonstrate the experimental results in Section 5 and finally conclude the paper in Section 6.

## 2 Related work

There are numerous works on pedestrian detection, and we refer readers to [23, 24, 14] for the detailed survey. Our work is most related to the work on adapting the generic pedestrian detector to videos, which is recently explored in [16–22, 25, 26]. These works differ in the training instances mining and detector adaptation methods. Methods presented in [16, 18, 19] mined new instances from detections according to a pre-defined threshold . To reduce the noise in the detected results, context cues were applied to refine the detections. [19, 20, 25] explored context cues on background to remove the detections with high scores. [16] used multiple target tracking to associate detections with trajectories, and took the non-associated detections as negative instances. [18] conducted KLT tracking to collect positive instances with low detection scores. [17] proposed an unsupervised tree coding method to cluster the detections. [20] proposed a confidence score SVM to encode the confidence scores in the model updating. The learning algorithms largely depend on the models used in the generic detector. For example, [25, 16, 21, 26] were built on boosting detectors, [17–20] were built on SVM detectors, [22] was built on the deep neural network. To the best of our knowledge, this is the first work to jointly consider the instance mining and detector adaptation in one objective function.

The problem setting is also related to the domain adaptation, which has been studied extensively in computer vision. We learn a feature transformation between the source and the target domain, as explored in [27–29] in an unsupervised manner. The most similar models as ours are the unsupervised approaches proposed in [30, 31]. However, these unsupervised approaches were based on generative models, making themselves unsuitable for real world detection tasks where discriminative models are always adopted. In addition, these work are designed for images instead of videos.

Our detector is built on DPM (Deformable Part Model) [6], which is one of the state-of-the-art detectors for generic detection tasks on static images. However, as evaluated in [14], its performance is unsatisfying for videos in real applications (e.g., Caltech benchmark [14]). Our method can be seen as an extension of DPM to adaptive-DPM for videos, where the detector is adapted automatically and the spatio-temporal continuity in videos is explored. Since the feature dimension in DPM is very high (more than 20K), it's infeasible to adapt it directly. Instead we introduce a feature transform on the cell level of HOG features, which can effectively capture the holistic low-level appearance change caused by domain shift. In the description of the model, we use a bilinear form of DPM used in [32] to simplify our notation.

We employ cost-flow network as part of our structural model, which is related to detection based multiple-target tracking [33–39]. Previous detection based multiple-target tracking methods rely on a fixed detector, while we adapt the detector automatically. The tracker outputs the trajectories of detections of multiple targets, while we focus on improving the detection performance. Due to the noticeable improvement of object detection in videos, our work can serve as a more reliable initialization for detection based tracking.

## 3 Preliminaries

In this section, we briefly introduce the bilinear form of DPM [6] and the cost-flow based data association [33]. The former model is the basis of our pedestrian detector, and the latter one is the structure on which we define the objective function with both appearance and spatio-temporal constraints.

### 3.1 Bilinear DPM

The popular DPM provides a hierarchical representation for pedestrians (as well as other objects). It contains a root filter and a set of deformable part filters. Without loss of generalization, we take the root as a special part here. Given an image $I$ and a configuration of parts $\zeta = \{l_0, l_1, \cdots, l_m\}$ in the detection window, we define the detection score of the configuration with respect to the DPM detector $\mathcal{F}$ as

$$score(\mathcal{F}, I, \zeta) = \sum_{i=0}^{m} w_i^T \phi_a(I, l_i) + w_s^T \phi_s(\zeta), \tag{1}$$

where $w_i$ is the filter of the $i^{th}$ part, and $\phi_a(I, l_i)$ is the HOG [1] feature vector extracted at $l_i$. $w_s$ is the shape prior which prefers a particular configuration, and $\phi_s(\zeta)$ is the spatial feature vector of the configuration $\zeta$. Here $l_0$ is the location of the root, and $l_i$ is the location of the $i^{th}$ part. It is straightforward to introduce the mixture components in Eq. 1, so we leave them out to simplify the exposition.

In this paper, we use the bilinear form of DPM originally introduced in [32]. It equals to the standard DPM, but can simplify the notation of our adaptive DPM introduced in the next section. Similar formulation is also proposed in [8]. The HOG feature of the $i^{th}$ part is denoted as a $n_f \times n_k$ dimensional matrix $\phi_a(I, l_i)$ , where $n_k$ is the number of cells in the part, and $n_f$ is the dimension of gradient histogram feature vector for a cell. Each column in $\phi_a(I, l_i)$ is a feature vector of a cell. $\phi_a(I, l_i)$ are further concatenated to be a large matrix $\Phi_a(I, \zeta) = [\phi_a(I, l_0), \phi_a(I, l_1), \cdots \phi_a(I, l_m)]$. The appearance filters in the detector are concatenated to be a matrix $W_a$ in the same way. With these notations, the detection score of DPM [6] equals

$$score(\mathcal{F}, I, \zeta) = Tr(W_a^T \Phi_a(I, \zeta)) + w_s^T \phi_s(\zeta), \tag{2}$$

where $Tr(\cdot)$ is the trace operation which is defined as summation of the elements on the main diagonal of a matrix. The bilinear form DPM detector is parameterized by the

appearance parameter matrix $W_a$ and spatial parameter matrix $w_s$. For a scanning window in detection, only the root location $l_0$ is given, and all the part locations are taken as latent variables which are optimized at runtime. The detection score of the sliding window is denoted as $score(I, \zeta^*)$, where $\zeta^*$ is the best possible part configuration when the root location is fixed to be $l_0$. Using quadratic function to model the spatial deformation of each part, the problem can be effectively solved with linear complexity [6].

### 3.2 Cost-flow based Data Association

Cost-flow based data association is proposed in [33] to associate detections in a video to be long trajectories. Finding the globally optimal trajectories for detections in video is reformulated as finding the min-cost flow in a network. Let us define the detection set in a video as $\mathcal{D} = \{d_1, \cdots, d_n\}$, where $d_i = \{\zeta_i, \sigma_i, t_i\}$ and $n$ is the number of detections. $\zeta_i$, $\sigma_i$, and $t_i$ stand for the location, scale, and frame index respectively. The detection $d_i$ corresponds to an edge from node $u_i$ to $v_i$ in the network. The $c_i$ is the weight to represent the cost for $d_i$ to be a pedestrian activation. For detections between different frames that have the possibility to belong to the same trajectory, an edge $(v_i, u_j)$ is created and $c_{i,j}$ is used to represent the cost for the transition between $v_i$ and $u_j$ in one trajectory. To start the flow, source node $s$ and sink node $t$ are added to the network, where $s$ links to all the $u_i$ with cost $c_{s,i}$ and all the $v_i$ are linked to the sink node $t$ with cost $c_{t,i}$. The cost $c_{s,i}$ and $c_{t,i}$ are used to punish the number of trajectories. For each edge in the flow, there is an indicator to represent whether the edge is included in one trajectory, which is denoted as $y_i$, $y_{i,j}$, $y_{s,i}$ and $y_{t,i}$ for the edge $(u_i, v_i)$,$(v_i, u_j)$,$(s, u_i)$ and $(v_i, t)$, respectively. To interpret the network flow as no overlap trajectory, the model uses the following constraints

$$y_{s,i} + \sum_{j=1}^{n} y_{j,i} = y_i = y_{t,i} + \sum_{j=1}^{n} y_{i,j}, \forall i \tag{3}$$
$$y_i, y_{s,i}, y_{t,i}, y_{i,j} \in \{0, 1\},$$

where $y_i$ is 1 when the detection $d_i$ is included in current trajectory, and otherwise 0. The above constraints guarantee that no paths share a common edge. The flow in the network is specified by $\mathcal{Y} = \{y_i, y_{i,j}, y_{s,i}, y_{t,i}\}$. Given the network and a configuration of $\mathcal{Y}$, the total cost is

$$L(\mathcal{D}, \mathcal{Y}) = \sum_{i=1}^{n} c_i y_i + \sum_{i=1}^{n} c_{s,i} y_{s,i} + \sum_{i=1}^{n} c_{t,i} y_{t,i} + \sum_{i=1}^{n} \sum_{j=1}^{n} c_{i,j} y_{i,j}, \tag{4}$$

where costs of all activated edges are summarized. When the cost terms are properly defined, finding the globally optimal trajectories is equivalent to solving the min-cost flow problem, where the cost is defined in Eq. 4 with constrains in Eq. 3. The optimization problem has been well explored. For example, [34] has shown an efficient successive shortest-paths algorithm with the complexity of $O(knlogn)$, where $k$ is the number of trajectories and $n$ is the number of detections. It is efficient enough in real applications (e.g., fewer than 10 seconds for a $10^3$-frame video with $10^6$ detections).

## 4 Adaptive Structural Model

Conventional methods consider the instance mining and detector adaptation as two separate tasks. In this way, the errors in instance mining could result in the drift of the detector, and the drift in detector adaptation could further harm the instance mining. To avoid the vicious circle, in this work, we propose to capture instance mining and detector adaptation jointly, where the instance mining and detector adaptation are handled in one objective function, and the joint optimization procedure outputs a combination of new instances and adapted detector.

Given the target video, we first conduct frame-wise detection and denote the detection result as $\mathcal{D}$. Due to the noise in detection, the detections in $\mathcal{D}$ cannot be taken as ground truth, instead we take them as latent variables, and label $\mathcal{D}$ by $\mathcal{Y}$ via an instance mining module. Here the indicator set $\mathcal{Y} = \{y_1, \cdots, y_n\}$ and $y_i \in \{0, 1\}$, where $y_i = 1$ indicates that the detection $d_i$ is taken as the true positive, otherwise $y_i = 0$. Given the original detector $\mathcal{F}^0$ and the detection set $\mathcal{D}$, we propose to find the optimal adapted detector $\mathcal{F}^*$, and the new indicator set $\mathcal{Y}^*$ with our designed objective function as

$$(\mathcal{F}^*, \mathcal{Y}^*) = \arg\min_{\mathcal{F},\mathcal{Y}} R(\mathcal{F}, \mathcal{F}^0) + \eta L(\mathcal{D}, \mathcal{F}, \mathcal{Y}), \tag{5}$$

where $R(\mathcal{F}, \mathcal{F}^0)$ is used to regularize the new detector $\mathcal{F}$ by the original detector $\mathcal{F}^0$, and $L(\mathcal{F}, \mathcal{D}, \mathcal{Y})$ is the loss term to measure the fitness between the adapted model $\mathcal{F}$ and the final detection result, which is specified by $\mathcal{D}$ and its indicator set $\mathcal{Y}$. The loss function is designed to encode the structural information in the target video. Two kinds of information can be encoded, the first is the appearance in detections, and the second is the spatio-temporal continuity in the video. The objective function in Eq. 5 naturally combines the instances mining and detector adaptation in a unified framework, and enables two tasks to benefit from each other. In the following parts, we define the regularization and loss function, and show how to optimize the objective function.

### 4.1 Adaptive DPM and Regularization

We use the generic DPM detector [6] as the initial detector $\mathcal{F}^0$ and aim to find an optimized DPM detector $\mathcal{F}$ on the target video. To avoid the direct adaptation of parameters of high dimensionality in DPM, we introduce a simple but effective adaptive DPM and show how to regularize it in the structural model defined below.

In DPM based representation, pedestrian consists of a number of local parts, and each part is represented by HOG cells. In applying the generic pedestrian detector to a novel scene, we only consider the domain shift in appearance (e.g., illumination, imaging condition) and ignore the variations in viewpoint, since the viewpoint variations could be naturally handled by the DPM mixture model. Under this assumption, we argue that the structure of parts and HOG spatial relationship between different parts should remain unchanged in detector adaptation process, while domain shift can be captured at feature level. Particularly, we use a linear transformation $P$ to model the mapping between the source and target domain in HOG cell level, which is a $n_f \times n_f$

dimensional matrix. When the transformation matrix $P$ is given, the detection score in the *adaptive DPM* for a part configuration $\zeta$ is defined as

$$score(\mathcal{F}, I, \zeta) = Tr(W_a^T P \Phi_a(I, \zeta)) + w_s^T \phi_s(\zeta), \tag{6}$$

where an additional feature transformation $P$ is conducted before the feature $\Phi_a(I, \zeta)$, which is then fed into the appearance filter $W_a$. Here the model $\mathcal{F}$ is specified by $W_a$, $w_s$ and $P$. The Eq. 2 can be taken as a special case of Eq. 6 when the transformation matrix $P$ is the identity matrix. To avoid the overfitting in model adaptation, we use the identity matrix $I$ to regularize $P$ by

$$R(\mathcal{F}, \mathcal{F}^0) = \|P - I\|_F^2, \tag{7}$$

where the Frobenius norm $\|\cdot\|_F$ is defined as the square root of the sum of the absolute squares of the elements in a matrix. It is of particular importance for video based detection since the feature vector is high dimensional while the number of mined instances is usually about a few hundred. The number of variables needed for adaptation is $n_f \times n_c$ in the original DPM, where $n_c$ is the number of HOG cells in all parts. In adaptive DPM, we only need to adapt $n_f \times n_f$ parameters, which brings about more efficiency (in typical DPM models, $n_c$ is one order larger than $n_f$).

### 4.2 Loss Function

The loss function is used to measure the detector $\mathcal{F}$ and indicator set $\mathcal{Y}$ on the target video. In this paper, two kinds of information are considered. The first is the frame-wise detection information, which means that detections activated in $\mathcal{Y}$ should have low appearance loss (i.e. high detection score). The second is that the activated detections should satisfy the video continuity, for example a stand alone detection in video is very likely to be a false positive and should be indicated as a false detection in $\mathcal{Y}$. To capture the above two types of information, we borrow the idea from cost-flow based data association, and measure the loss of indicator set $\mathcal{Y}$ and detector $\mathcal{F}$ jointly with the following function

$$\begin{aligned} L(\mathcal{F}, \mathcal{D}, \mathcal{Y}) &= \sum_{i=1}^{n} (c_{t,i} y_{t,i} + c_{s,i} y_{s,i} + c_i y_i) \\ where \quad & c_i = \max(\xi_1, \xi_2 - score(\mathcal{F}, I, \zeta_i)) \\ s.t. \quad & y_{s,i} + \sum_{j=1}^{n} y_{j,i} = y_i = y_{t,i} + \sum_{j=1}^{n} y_{i,j}, \forall i \\ & and \quad y_i, y_{i,j}, y_{s,i}, y_{t,i} \in \{0, 1\}, \end{aligned} \tag{8}$$

where the indicator $\mathcal{Y}$ now includes auxiliary variables. The $score(\mathcal{F}, I, \zeta_i)$ is defined in adaptive DPM as Eq. 6. The above problem can be seen as an instantiation of the general cost-flow based data association problem introduced in Eq. 4. The appearance cost $c_i$ is defined as a generalized hinge loss of adaptive DPM detection score. The intuition inside the definition is that detections with high appearance scores should be activated

with a negative loss value, while the detections with low appearance scores should be suppressed with a positive value. The parameter $\xi_1$ and $\xi_2$ can be tuned according to the range of detection scores. The costs $c_{s,i}$ and $c_{t,i}$ which involve the source and sink nodes are fixed to be positive constraint. They can be considered as a punishment to the number of trajectory, which can help to remove the discontinuous false positives.

### 4.3 Adaptive Optimization

In video based detection, we need to determine the new detector $\mathcal{F}$ specified by $P$, the frame-wise detection set $\mathcal{D}$, and the indicator set $\mathcal{Y}$ of the detection set. Advocated by recent latent structural learning works, we adopt the alternating minimization procedure to optimize them.

**Algorithm 1** Adaptive Structural Optimization for Video based Pedestrian Detection.

**Input:**
The video $V$, and the generic detector $\mathcal{F}_0$.
Set $\mathcal{F} = \mathcal{F}_0$, i.e. $P = I$.
**for** i=1 **to** $T_1$ **do**
  Conduct the frame-wise adaptive DPM detection procedure with detector $\mathcal{F}$ by Eq. 6, and get the detection set $\mathcal{D}$ of the video.
  **for** j=1 **to** $T_2$ **do**
    Fix the $P$, and solve the optimal indicator set $\mathcal{Y}$ by minimizing $L(P, \mathcal{D}, \mathcal{Y})$ with the successive shortest-paths algorithm.
    Fix the $\mathcal{Y}$, and solve the optimal $P$ in Eq. 11 with standard quadratic programming procedure.
  **end for**
**end for**
**return** $(\mathcal{F}, \mathcal{D}, \mathcal{Y})$.

The whole optimization procedure is shown in Algorithm 1. In the outer loop, we conduct the adaptive DPM detector for frame-wise detection to get the detection set $D$. When $\mathcal{D}$ is fixed, the detection indicator $\mathcal{Y}$ and the adapted detector $\mathcal{F}$ are jointly optimized by the following problem

$$
\begin{aligned}
(\mathcal{Y}^*, P^*) &= \arg\min_{\mathcal{Y}, P} \|P - I\|_F^2 + \eta L(P, \mathcal{D}, \mathcal{Y}) \qquad (9)\\
s.t. \quad & y_{s,i} + \sum_{j=1}^{n} y_{j,i} = y_i = y_{t,i} + \sum_{j=1}^{n} y_{i,j}, \forall i \\
and \quad & y_i, y_{s,i}, y_{t,i}, y_{i,j} \in \{0, 1\},
\end{aligned}
$$

where the $L(P, \mathcal{D}, \mathcal{Y})$ is exactly the $L(\mathcal{F}, \mathcal{D}, \mathcal{Y})$, since $F$ can be specified by $P$. The filters $W_a$ and $w_s$ in $\mathcal{F}$ are from the generic detector and fixed in the whole procedure. It is a difficult mixed programming non-convex problem when both $\mathcal{Y}$ and $P$ are free. We therefore resort to an iterative algorithm based on the fact that solving $\mathcal{Y}$ given $P$,

and solving $P$ given $\mathcal{Y}$ are convex problems, and there exists off-the-shelf solvers. In detail, we solve the following two problems.

**Fix $\mathcal{Y}$ to solve $P$** When $\mathcal{Y}$ is given, the constrains in Eq. 9 can be removed, and the problem becomes to be

$$\arg\min_{P} \|P - I\|_F^2 \tag{10}$$
$$+\eta \sum_{i=1}^{n} y_i \cdot max(\xi_1, \xi_2 - (Tr(W_a^T P \Phi_a(I, \zeta_i)) + w_s^T \phi_s(\zeta_i))),$$

Since $Tr(W_a^T P \Phi_a(I, \zeta_i))$ is equal to $Tr(P \Phi_a(I, \zeta_i) W_a^T)$, the above problem equals

$$\arg\min_{P} \|vec(P) - vec(I)\|^2 \tag{11}$$
$$+\eta \sum_{i=1}^{n} y_i \cdot max(\xi_1, \xi_2 - (vec(P)^T vec(\Phi_a(I, \zeta_i) W_a^T) + w_s^T \phi_s(\zeta))),$$

where $vec(\cdot)$ is the operator to reshape the matrix to be a vector in a column-wise manner. The above problem can be solved effectively by standard quadratic programming solvers [40].

**Fix $P$ to solve $\mathcal{Y}$** When the transformation matrix $P$ is given, Eq. 5 becomes $\arg\min_{\mathcal{Y}} L(P, \mathcal{D}, \mathcal{Y})$ under the cost-flow constraint, which can be effectively solved by successive shortest-paths algorithm described in [34]. In the algorithm, we iteratively find the minimum-cost parts $\gamma$ from the source to the sink in the residual graph, and update the flow by pushing the unit-flow along $\gamma$ if the total cost of the path is negative.

Since the objective value is reduced in both of the two subproblems of the inner loop, it can be easily proved that the inner loop will converge to a local minima. We set the loop number $T_1$ to be 5, the loop number $T_2$ to be 8 and validate the convergence of the whole optimization procedure in experiments.

## 5 Experiment

Experiments are conducted on challenging videos from ETHZ [41], Caltech [10] and PETS2009 [1] pedestrian datasets. The ETHZ and Caltech are captured from moving camera, while the PETS2009 is captured from stationary camera. Particularly, the Bahnhof sequences from ETHZ, S2-L2 from PETS2009, and 8 sequences from Caltech with most people are selected for evaluation. The ETHZ-Bahnhof sequences contain 999 frames and 8467 pedestrians; the PETS2009-S2-L2 sequences contain 436 frames and 8927 pedestrians; the 8 sequences from Caltech testset are with the length of about 1800 frames. These videos are challenging for cluttered background, large illumination variations, and heavy occlusion. The DPM detector (Version 5) trained on the INRIA dataset [1] is taken as the baseline. Since the detector can only detect pedestrians of above 120 pixels in height, we resize every video frame with a scale of 2.5, and only measure

[1] http://www.cvg.rdg.ac.uk/PETS2009/

pedestrians of above 50 pixels in height as suggested in [14]. For all the experiments, $\xi_1$, $\xi_2$, $c_{s,i}$ and $c_{t,i}$ used in Eq. 8 are fixed to be -1, 0.2, 10 and 10, respectively.

We follow the publicly available evaluation protocol in [14], except that the evaluations are conducted for each video separately. Full ROC curve and mean miss rate[2] are used to compare different algorithms. In the following parts, we compare different detector adaptation methods and examine the convergence of the outer loop in Algorithm 1, and finally compare the detection performance with other state-of-the-art detectors.

### 5.1 Different Methods for Video based Detection

In this part, we compare four different approaches for video based pedestrian detection on PETS2009-S2-L2, which is challenging for appearance variations in illumination and occlusion. These approaches include: (1) Generic DPM, the baseline DPM detector (version 5) learned on INRIA; (2) DPM + Adaptation, which iteratively adds new training instances according to the frame-wise detection score, and then adapts the DPM detector using the collected instances; (3) DPM + Tracking + Adaptation, which uses the tracked detections as the new instances to adapt the DPM detector, where the tracking is solved by cost-flow network; (4) The proposed method that jointly considers instance mining and detector adaptation.

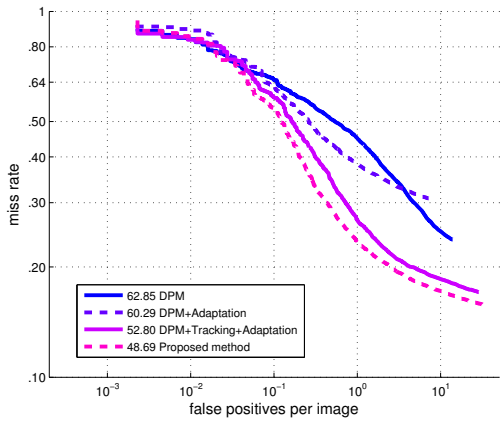



**Fig. 2.** Pedestrian detection results of different methods on PETS2009-S2-L2.

ROC curves and mean miss rates of the four methods are demonstrated in Fig. 2. Due to the noise in training instances used, direct adaptation could cause the drift problem, and in this experiment it only improves a small margin over the original generic

[2] The mean miss rate defined in P. Dollár's toolbox is used here, which is the average miss rate at 0.0100, 0.0178, 0.0316, 0.0562, 0.1000, 0.1778, 0.3162, 0.5623 and 1.0000 false-positive-per-image.

detector. Since a lot of false positives can be removed by optimizing the cost-flow network, the instances used in adapting the detector are clear enough in the DPM + Tracking + Adaptation procedure, and it improves the performance with quite a large margin. Benefiting from the joint model, our approach achieves the best performance and outperforms the baseline generic DPM with 14% reduction in mean miss rate. Compared with the sequential instance mining and detector adaptation approach, the proposed joint learning method further reduces the mean miss rate by 4%.

### 5.2 The Convergence

In this part, we validate the convergence of the proposed method. 8 videos from Caltech testset with most pedestrians are selected for evaluation. Since the inner loop in Algorithm 1 is sure to converge to a local stable point, in this part, we only validate the convergence of the outer loop, which iteratively conducts frame-wise detection and optimizes the structural model. The selected video ID and the mean miss rate at each loop are reported in Fig. 3.

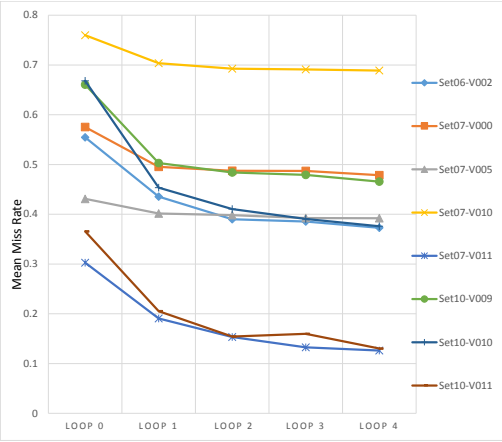


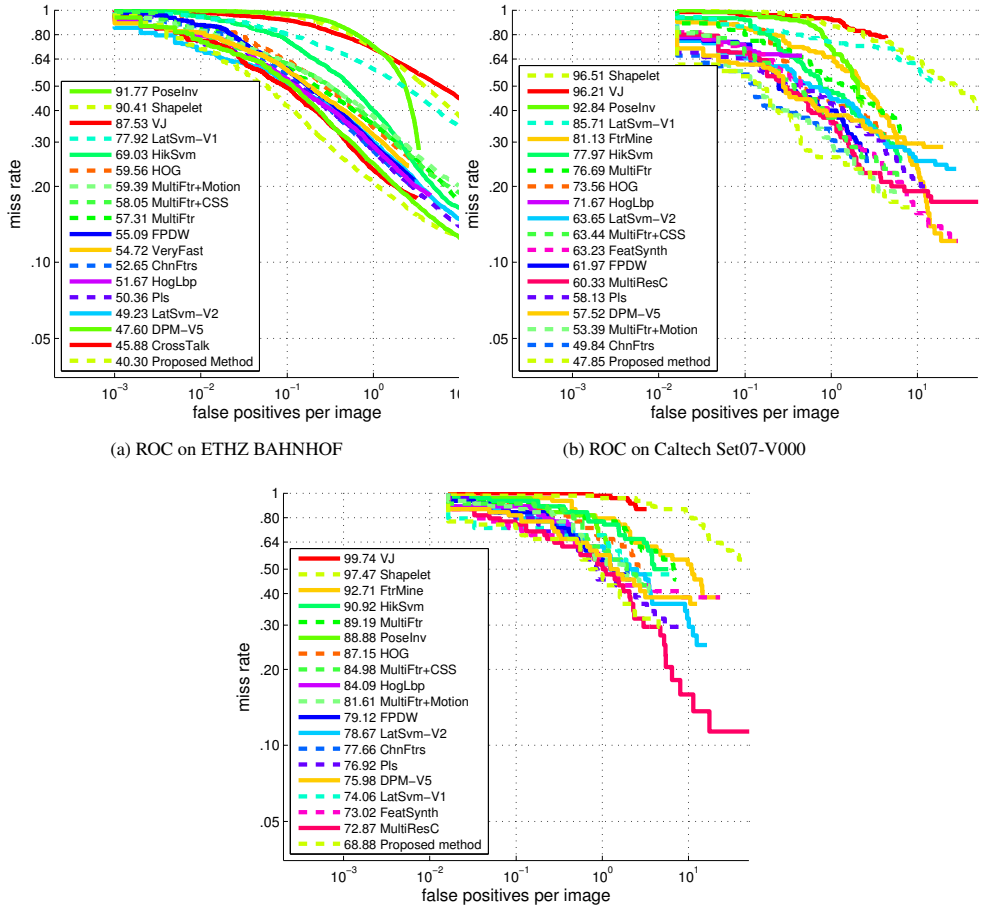

**Fig. 3.** The convergence illustration of the proposed optimization method in video based pedestrian detection.

From Fig. 3, we can find the noticeable performance improvement and the fast convergence rate of our approach. On the 8 videos, the proposed method has an average of 16% reduction in mean miss rate and the performance is close to convergence after 3 loops. Since the first loop can mine most of the instances, it contributes most to the performance.

### 5.3 Comparisons with State-of-the-art Methods

In this part, we compare the proposed method with other state-of-the-art algorithms, collected in [14], including Viola-Jones [42], Shapelet [43], LatSVM-V1, LatSVM-V2

[6], PoseInv [44], HOGLbp [4], HikSVM [3], HOG[1], FtrMine [45], MultFtr [43], MultiFtr+CSS [43], Pls [46], MultiFtr+Motion [43], FPDW [10], FeatSynth [47], ChnFtrs [48], MultiResC [7], Veryfast [11], and CrossTalk [12]. We show the results of the video Bahnhof in ETHZ, the set07-V000 and set07-V011 in Caltech [3].




(a) ROC on ETHZ BAHNHOF

(b) ROC on Caltech Set07-V000

(c) ROC on Caltech Set07-V011

**Fig. 4.** Quantitative evaluations on ETHZ and Caltech.

Fig. 4 illustrates the quantitative results of different methods. On all the three videos, the proposed method outperforms the baseline DPM (version 5) by more than 7%, and outperforms the published state-of-the-art results. It improves about 5% on the ETHZ

[3] The two videos are selected as they contain more people than other videos.

Bahnhof, 2% on Caltech Set07-V000, and 4% on Caltech Set07-V011 than the best published results. Some qualitative examples are shown in Fig. 5.

While the structural model optimization step is very efficient, most of the calculation is spent on frame-wise detection. In our implementation, we modify the code of the FFT based implementation [49] for fast convolution computation. Some techniques can be used to further accelerate the loop, such as the cascade detection or only detecting a subset in the early steps of the outer loop in Algorithm 1, and we leave it in future work.

**Fig. 5.** Qualitative results of the proposed video base pedestrian detection on the ETHZ, Caltech and PETS2009.

## 6 Conclusion

In this paper, we propose a joint structural model to adapt the generic pedestrian detector for video based pedestrian detection. The instance mining and detector adaptation are formulated in one objective function, and an alternating minimization procedure is adopted to optimize it. The DPM is extended to be adaptive-DPM, where a feature transformation defined on low-level HOG cell is used to reduce the domain shift. We demonstrate noticeable improvement over the methods that treat the two tasks independently, and other state-of-the-art detectors on challenging videos from Caltech, ETHZ and PETS2009.

**Acknowledgement**. This work was supported by the Chinese National Natural Science Foundation Projects #61105023, #61103156, #61105037, #61203267, #61375037, #61473291, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds.

## References


1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, IEEE (2005)
2. Yan, J., Lei, Z., Yi, D., Li, S.Z.: Multi-pedestrian detection in crowded scenes: A global view. In: CVPR, IEEE (2012)
3. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR, IEEE (2008)
4. Wang, X., Han, T., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: ICCV, IEEE (2009)
5. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. In: CVPR, IEEE (2010)
6. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI (2010)
7. Park, D., Ramanan, D., Fowlkes, C.: Multiresolution models for object detection. ECCV (2010)
8. Yan, J., Zhang, X., Lei, Z., Liao, S., Li, S.Z.: Robust multi-resolution pedestrian detection in traffic scenes. In: CVPR, IEEE (2013)
9. Huang, C., Nevatia, R.: High performance object detection by collaborative learning of joint ranking of granules features. In: CVPR, IEEE (2010)
10. Dollár, P., Belongie, S., Perona, P.: The fastest pedestrian detector in the west. BMVC 2010 (2010)
11. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. In: CVPR, IEEE (2012)
12. Dollár, P., Appel, R., Kienzle, W.: Crosstalk cascades for frame-rate pedestrian detection. In: ECCV, Springer (2012)
13. Yan, J., Lei, Z., Wen, L., Li, S.Z.: The fastest deformable part model for object detection. In: CVPR. (2014)
14. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. TPAMI (2012)
15. Yang, M., Zhu, S., Lv, F., Yu, K.: Correspondence driven adaptation for human profile recognition. In: CVPR, IEEE (2011)
16. Sharma, P., Huang, C., Nevatia, R.: Unsupervised incremental learning for improved object detection in a video. In: CVPR, IEEE (2012)
17. Wang, X., Hua, G., Han, T.X.: Detection by detections: Non-parametric detector adaptation for a video. In: CVPR, IEEE (2012)
18. Tang, K., Ramanathan, V., Fei-Fei, L., Koller, D.: Shifting weights: Adapting object detectors from image to video. In: NIPS. (2012)
19. Wang, M., Wang, X.: Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In: CVPR, IEEE (2011)
20. Wang, M., Li, W., Wang, X.: Transferring a generic pedestrian detector towards specific scenes. In: CVPR, IEEE (2012)
21. Sharma, P., Nevatia, R.: Efficient detector adaptation for improved object detection in a video. In: CVPR, IEEE (2013)
22. Yang, Y., Shu, G., Shah, M.: Semi-supervised learning of feature hierarchies for object detection in a video. In: CVPR, IEEE (2013)
23. Enzweiler, M., Gavrila, D.: Monocular pedestrian detection: Survey and experiments. TPAMI (2009)
24. Geronimo, D., Lopez, A., Sappa, A., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. PAMI (2010)

25. Roth, P.M., Sternig, S., Grabner, H., Bischof, H.: Classifier grids for robust adaptive object detection. In: CVPR, IEEE (2009)
26. Pang, J., Huang, Q., Yan, S., Jiang, S., Qin, L.: Transferring boosted detectors towards viewpoint and scene adaptiveness. TIP (2011)
27. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: ECCV. Springer (2010)
28. Kulis, B., Saenko, K., Darrell, T.: What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In: CVPR, IEEE (2011)
29. Gao, T., Stark, M., Koller, D.: What makes a good detector?–structured priors for learning from few examples. In: ECCV. Springer (2012)
30. Gopalan, R., Li, R., Chellappa, R.: Domain adaptation for object recognition: An unsupervised approach. In: ICCV, IEEE (2011)
31. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: CVPR, IEEE (2012)
32. Pirsiavash, H., Ramanan, D.: Steerable part models. In: CVPR, IEEE (2012)
33. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR, IEEE (2008)
34. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR, IEEE (2011)
35. Berclaz, J., Fleuret, F., Fua, P.: Multiple object tracking using flow linear programming. In: PETS-Winter, IEEE (2009)
36. Jiang, H., Fels, S., Little, J.J.: A linear programming approach for multiple object tracking. In: CVPR, IEEE (2007)
37. Yang, B., Huang, C., Nevatia, R.: Learning affinities and dependencies for multi-target tracking using a crf model. In: CVPR, IEEE (2011)
38. Andriyenko, A., Schindler, K.: Globally optimal multi-target tracking on a hexagonal lattice. In: ECCV. Springer (2010)
39. Wen, L., Li, W., Yan, J., Lei, Z., Yi, D., Li, S.Z.: Multiple target tracking based on undirected hierarchical relation hypergraph. (2014)
40. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge university press (2009)
41. Ess, A., Leibe, B., Schindler, K., , van Gool, L.: A mobile vision system for robust multi-person tracking. In: CVPR, IEEE (2008)
42. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. IJCV (2005)
43. Wojek, C., Schiele, B.: A performance evaluation of single and multi-feature people detection. DAGM (2008)
44. Lin, Z., Davis, L.: A pose-invariant descriptor for human detection and segmentation. ECCV (2008)
45. Dollár, P., Tu, Z., Tao, H., Belongie, S.: Feature mining for image classification. In: CVPR, IEEE (2007)
46. Schwartz, W., Kembhavi, A., Harwood, D., Davis, L.: Human detection using partial least squares analysis. In: ICCV, IEEE (2009)
47. Bar-Hillel, A., Levi, D., Krupka, E., Goldberg, C.: Part-based feature synthesis for human detection. ECCV (2010)
48. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: BMVC. (2009)
49. Dubout, C., Fleuret, F.: Exact acceleration of linear object detectors. ECCV (2012)