Robust Multi-Resolution Pedestrian Detection in Traffic Scenes

Junjie Yan Xucong Zhang Zhen Lei Shengcai Liao Stan Z. Li*
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, China
{jjyan,xczhang,zlei,scliao,szli}@nlpr.ia.ac.cn

Abstract

The serious performance decline with decreasing resolution is the major bottleneck for current pedestrian detection techniques [14, 23]. In this paper, we take pedestrian detection in different resolutions as different but related problems, and propose a Multi-Task model to jointly consider their commonness and differences. The model contains resolution aware transformations to map pedestrians in different resolutions to a common space, where a shared detector is constructed to distinguish pedestrians from background. For model learning, we present a coordinate descent procedure to learn the resolution aware transformations and deformable part model (DPM) based detector iteratively. In traffic scenes, there are many false positives located around vehicles, therefore, we further build a context model to suppress them according to the pedestrian-vehicle relationship. The context model can be learned automatically even when the vehicle annotations are not available. Our method reduces the mean miss rate to 60% for pedestrians taller than 30 pixels on the Caltech Pedestrian Benchmark, which noticeably outperforms previous state-of-the-art (71%).

1. Introduction

Pedestrian detection has been a hot research topic in computer vision for decades, for its importance in real applications, such as driving assistance and video surveillance. In recent years, especially due to the popularity of gradient features, pedestrian detection field has achieved impressive progresses in both effectiveness [6, 31, 43, 41, 19, 33] and efficiency [25, 11, 18, 4, 10]. The leading detectors can achieve satisfactory performance on high resolution benchmarks (e.g. INRIA [6]), however, they encounter difficulties for the low resolution pedestrians (e.g. 30-80 pixels tall, Fig. 1) [14, 23]. Unfortunately, the low resolution pedestrians are often very important in real applications. For example, the driver assistance systems need detect the low resolution pedestrians to provide enough time for reaction.

Figure 1. Examples of multiple resolution pedestrian detection result of our method in the Caltech Pedestrian Benchmark [14].

Traditional pedestrian detectors usually follow the scale invariant assumption: a scale invariant feature based detector trained at a fixed resolution could be generalized to all resolutions, by resizing the detector [40, 4], image [6, 19] or both of them [11]. However, the finite sampling frequency of the sensor results in much information loss for low resolution pedestrians. The scale invariant assumption does not hold in the case of low resolution, which leads to the disastrous drop of the detection performance with the decrease of resolution. For example, the best detector achieves 21% mean miss rate for pedestrians taller than 80 pixels in Caltech Pedestrian Benchmark [14], while increases to 73% for pedestrians 30-80 pixels high.

Our philosophy is that the relationship among different resolutions should be explored for robust multi-resolution pedestrian detection. For example, the low resolution samples contain a lot of noise that may mislead the detector in the training phase, and the information contained in high resolution samples can help to regularize it. We argue that for pedestrians in different resolutions, the differences exist in the features of local patch (e.g. the gradient histogram feature of a cell in HOG), while the global spatial structure keeps the same (e.g. part configuration). To this end, we propose to conduct resolution aware transformations to map the local features from different resolutions to a common subspace, where the differences of local features are reduced, and the detector is learned on the mapped fea-

*Stan Z. Li is the corresponding author.

tures of samples from different resolutions, thus the structural commonness is preserved. Particularly, we extend the popular deformable part model (DPM) [19] to multi-task DPM (MT-DPM), which aims to find an optimal combination of DPM detector and resolution aware transformations. We prove that when the resolution aware transformations are fixed, the multi-task problems can be transformed to be a Latent-SVM optimization problem, and when the DPM detector in the mapped space is fixed, the problem equals to a standard SVM problem. We divide the complex non-convex problem into the two sub-problems, and optimize them alternatively.

In addition, we propose a new context model to improve the detection performance in traffic scenes. There is a phenomenon that quite a large number of detections (33.19% for MT-DPM in our experiments) are around vehicles. The vehicle localization is much easier than pedestrian, which motivates us to employ pedestrian-vehicle relationship as an additional cue to judge whether the detection is a false or true positive. We build an energy model to jointly encode the pedestrian-vehicle and geometry contexts, and infer the labels of detections by maximizing the energy function on the whole image. Since the vehicle annotations are often not available in pedestrian benchmark, we further present a method to learn the context model from ground truth pedestrian annotations and noisy vehicle detections.

We conduct experiments on the challenging Caltech Pedestrian Benchmark [14], and achieve significantly improvement over previous state-of-the-art methods on all the 9 sub-experiments advised in [14]. For the pedestrians taller than 30 pixels, our MT-DPM reduces 8% and our context model further reduces 3% mean miss rate over previous state-of-the-art performance.

The rest of the paper is organized as follows: Section 2 reviews the related work. The multi-task DPM detector and pedestrian-vehicle context model are discussed in Section 3 and Section 4, respectively. Section 5 shows the experiments and finally in Section 6 we conclude the paper.

2. Related work

There is a long history of research on pedestrian detection. Most of the modern detectors are based on statistical learning and sliding-window scan, popularized by [32] and [40]. Large improvements came from the robust features, such as [6, 12, 25, 3]. There are some papers fused HOG with other features [43, 7, 45, 41] to improve the performance. Some papers focused on special problems in pedestrian detection, including occlusion handling [46, 43, 38, 2], speed [25, 11, 18, 4, 10], and detector transfer in new scenes [42, 27]. We refer the detailed surveys on pedestrian detection to [21, 14].

Resolution related problems have attracted attention in recent evaluations. [16] found that the pedestrian detection performance depends on the resolution of training samples. [14] pointed that the pedestrian detection performance drops with decreasing resolution. [23] observed similar phenomenon in general object detection task. However, there are very limited works proposed to tackle this problem. The most related work is [33], which utilized root and part filters for high resolution pedestrians, while only used the rigid root filter for low resolution pedestrians. [4] proposed to use a single model per detection scale, but the paper is focused on speedup.

Our pedestrian detector is built on the popular DPM (deformable part model) [19], which combined rigid root filter and deformable part filters for detection. The DPM only performs well for high resolution objects, while our MT-DPM generalizes it to low resolution case. The coordinate descent procedure in learning is motivated by the steerable part model [35, 34], which trained the shared part bases to accelerate the detection. Note that [34] learned a shared filter bases, while our model learns a shared classifier, which result in a quite different formulation. [26] also proposed a multi-task model to handle dataset bias. The multi-task idea in this paper is motivated by works on face recognition across different domains, such as [28, 5].

Context has been used in pedestrian detection. [24, 33] captured the geometry constraint under the assumption that camera is aligned with ground plane. [9] took the appearance of nearby regions as the context. [8, 36, 29] captured the pair-wise spatial relationship in multi-class object detection. To the best of our knowledge, this is the first work to capture the pedestrian-vehicle relationship to improve pedestrian detection in traffic scenes.

3. Multi-Task Deformable Part Model

There are two intuitive strategies to handle the multi-resolution detection. One is to combine samples from different resolutions to train a single detector (Fig. 2(a)), and another is to train independent detectors for different resolutions (Fig. 2(b)). However, both of the two strategies are not prefect. The first one considers the commonness between different resolutions, while their differences are ignored. Samples from different domains would increase the complexity of the detection boundary, which probably beyond the ability of a single linear detector. On the contrary, multi-resolution model takes pedestrian detection in different resolutions as independent problems, and the relationship among them are missed. The unreliable features of low resolution pedestrians can mislead the learned detector and make it difficult to be generalized to novel test samples.

In this part, we present a multi-resolution detection method by considering the relationship of samples from different resolutions, including the commonness and the differences, which are captured by a multi-task strategy simultaneously. Considering the differences of different resolu-

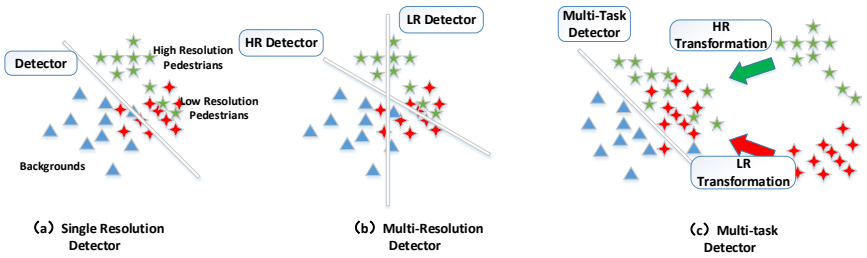


Figure 2. Different strategies for multi-resolution pedestrian detection.

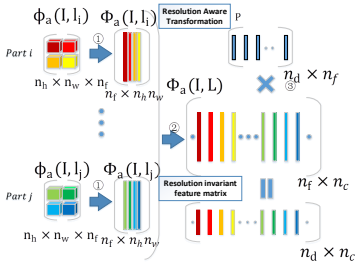


Figure 3. Demonstration of the resolution aware transformations.

tions, we use the resolution aware transformations to map features from different resolutions to a common subspace, in which they have similar distribution. A shared detector is trained in the resolution-invariant subspace by samples from all resolutions, to capture the structural commonness. It easy to see that the first two strategies are the special case of the multi-task strategy.

Particularly, we extend the the idea to popular DPM detector [19] and propose a Multi-Task form of DPM. Here we consider the partition of two resolutions (low resolution: 30-80 pixels tall, and high resolution: taller than 80 pixels, as advised in [14]). Note that extending the strategy for other local feature based linear detectors and more resolution partitions are straightforward.

3.1. Resolution Aware Detection Model

To simplify the notation, we introduce a matrix based representation for DPM. Given the image I and the collection of m part locations $L = (l_0, l_1, \cdots, l_m)$, the HOG feature $\phi_a(I, l_i)$ of the i-th part is a $n_h \times n_w \times n_f$ dimensional tensor, where n_h, n_w are the height and width of HOG cells for the part, and n_f is the dimension of gradient histogram feature vector for a cell. We reshape $\phi_a(I, l_i)$ to be a matrix $\Phi_a(I, l_i)$, where every column represents features from a cell. $\Phi_a(I, l_i)$ is further concatenated to be a large matrix $\Phi_a(I, L) = [\Phi_a(I, l_0), \Phi_a(I, l_1), \cdots \Phi_a(I, l_m)]$. The column number of $\Phi_a(I, L)$ is denoted as n_c, which is the sum number of cells in parts and root. Demonstration of the procedure is shown in Fig. 3. The appearance filters in the detector are concatenated to be a $n_f \times n_c$ matrix W_a in the same way. The spatial features of different parts are concatenated to be a vector $\phi_s(I, L)$, and the spatial prior parameter is denoted as w_s. With these notations, the detection model of DPM [19] can be written as:

$$score(I, L) = Tr(W_a^T \Phi_a(I, L)) + w_s^T \phi_s(I, L), \quad (1)$$

where $Tr(\cdot)$ is the trace operation defined as summation of the elements on the main diagonal of a matrix. Given the root location l_0, all the part locations are latent variables, and the final score is $\max_{L^*} score(I, L^*)$, where L^* is the best possible part configurations when the root location is fixed to be l_0. The problem can be solved effectively by the dynamic programming [19]. Mixture can be used to increase the flexibility, but we ignore it for simplicity in notation and adding mixture in the formulations is straightforward.

In DPM, pedestrian consists of parts, and every part consists of HOG cells. When the pedestrian resolution changes, the structure of parts and the HOG cell spatial relationship keep the same. The only difference among different resolution lies in the feature vector of evert cell, so that the resolution aware transformations P_L and P_H are defined on it. The P_L and P_H are of the dimension $n_d \times n_f$, and they map the low and high resolution samples from the original n_f dimensional feature space to the n_d dimensional subspace. The features from different resolutions are mapped into the common subspace, so that can share the same detector. We still denote the learned appearance parameters in the mapped resolution invariant subspace as W_a, which is a $n_d \times n_c$ matrix, and of the same size with $P_H \Phi_a(I, L)$. The score of a collection of part locations L in the MT-DPM is defined as:

$$\begin{cases} Tr(W_a^T P_H \Phi_a(I, L)) + w_s^T \phi_s(I, L), & \text{High Resolution} \\ Tr(W_a^T P_L \Phi_a(I, L)) + w_s^T \phi_s(I, L), & \text{Low Resolution.} \end{cases} \quad (2)$$

The model defined above provides the flexibility to describe pedestrians of different resolutions, but also brings challenges, since the W_a, w_s, P_H, P_L are all unknown. In the following part, we present the objective function of the multi-task model for learning, and show the optimization method.

3.2. Multi-Task Learning

The objective function is motivated by the original single task DPM. Its matrix form can be written as:

$$\arg \min_{W_a, w_s} \frac{1}{2}\|W_a\|_F^2 + \frac{1}{2} w_s^T w_s \quad (3)$$
$$+C \sum_N \max[0, 1 - y_n(Tr(W_a^T \Phi_a(I_n, L_n^*)) + w_s^T \phi_s(L_n^*))],$$

where $\|\cdot\|_F$ is the Frobenius Norm, and $\|W_a\|_F^2 = Tr(W_a W_a^T)$. y_n is 1 if $I_n(L_n)$ is pedestrian, and -1 for

background. The first two terms are used for regularize the detector parameters, and the last term is the hinge loss in DPM detection. The L_n^* is the optimized part configuration that maximizes the detection score of I_n. In the learning phase, the part locations are taken as latent variables, and the problem can be optimized by the Latent-SVM [19].

For multi-task learning, the relationship between different tasks should be considered. In analogy to the original DPM, MT-DPM is formulated as:

$$\arg \min_{W_a, w_s, P_H, P_L} \frac{1}{2} w_s^T w_s \tag{4}$$
$$+ f_{I_H}(W_a, w_s, P_H) + f_{I_L}(W_a, w_s, P_L),$$

where I_H and I_L denote the high and low resolution training sets, including both pedestrian and background. Since spatial term w_s is directly applied to the data from different resolutions, it can be regularized independently. f_{I_H} and f_{I_L} are used to consider the detection loss and regularize the parameters P_H, P_L and W_a. f_{I_H} and f_{I_L} are of the same form, here we take f_{I_H} as an example. It can be written as:

$$f_{I_H}(W_a, w_s, P_H) = \frac{1}{2} \|P_H^T W_a\|_F^2 \tag{5}$$
$$+ C \sum_{N_H} \max[0, 1 - y_n (Tr(W_a^T P_H \Phi_a(I_{H_n}, L_n^*)) + w_s^T \phi_s(L_n^*))],$$

where the regularization term $P_H^T W_a$ is a $n_f \times n_c$ dimensional matrix, and of the same dimension with the original feature matrix. Since P_H and W_a are applied to original appearance feature integrally in calculating the appearance score $Tr((P_H^T W_a)^T \Phi_a(I, L)$, we take them as an ensemble and regularize them together. The second term is the detection loss for resolution aware detection, corresponding to the detection model in Eq. 2. The parameters W_a and w_s are shared between f_{I_H} and f_{I_L}. Note that more partitions of resolutions can be handle naturally in Eq. 4.

In Eq. 4, we need to find an optimal combination of W_a, w_s, P_H, and P_L. However, Eq. 4 is not convex when all of them are free. Fortunately, we show that given the two transformations, the problem can be transformed into a standard DPM problem, and given the DPM detector, it can be transformed into a standard SVM problem. We conduct a coordinate descent procedure to optimize the two subproblems iteratively.

3.2.1 Optimize W_a and w_s

When P_H and P_L are fixed, we can map the features to the common space on which DPM detector can be learned. We denote $P_H P_H^T + P_L P_L^T$ as A, $A^{\frac{1}{2}} W_a$ as $\widetilde{W_a}$. For high resolution samples we denote $A^{-\frac{1}{2}} P_H \Phi_a(I_n, L_n^*)$ as $\widetilde{\Phi_a}(I_n, L_n^*)$, and for low resolution samples we denote $A^{-\frac{1}{2}} P_L \Phi_a(I_n, L_n^*)$ as $\widetilde{\Phi_a}(I_n, L_n^*)$. Eq. 4 can be reformulated as:

$$\arg \min_{\widetilde{W_a}, w_s} \frac{1}{2} \|\widetilde{W_a}\|_F^2 + \frac{1}{2} w_s^T w_s \tag{6}$$
$$+ C \sum_{N_H + N_L} \max[0, 1 - y_n (Tr(\widetilde{W_a}^T \widetilde{\Phi_a}(I_n, L_n^*)) + w_s^T \phi_s(L_n^*))],$$

which has the same form with the optimization problem in Eq. 3, and the Latent-SVM solver can be used here. Once the solution to Eq. 6 is achieved, W_a is computed by $(P_H P_H^T + P_L P_L^T)^{-\frac{1}{2}} \widetilde{W_a}$.

3.2.2 Optimize P_H and P_L

When the W_a and w_s are fixed, P_H and P_L are independent, thus the optimization problem can be divided into two subproblems: $\arg\min_{P_H} f_{I_H}(W_a, w_s, P_H)$ and $\arg\min_{P_L} f_{I_L}(W_a, w_s, P_L)$. Since they are of the same form, here we only give the details for optimizing P_H.

Given the W_a and w_s, we first infer the part location of every training samples L_n^* by finding a part configurations to maximize Eq. 2. Denoting $W_a W_a^T$ as A, $A^{\frac{1}{2}} P_H$ as $\widetilde{P_H}$, and $A^{-\frac{1}{2}} W_a \Phi_a(I_{H_n}, L_n^*)^T$ as $\widetilde{\Phi_a}(I_{H_n}, L_n^*)$, the problem of Eq. 4 equals to:

$$\arg \min_{\widetilde{P_H}} \frac{1}{2} \|\widetilde{P_H}\|_F^2 \tag{7}$$
$$+ C \sum_{N_H} \max[0, 1 - y_n (Tr(\widetilde{P_H}^T \widetilde{\Phi_a}(I_{H_n}, L_n^*)) + w_s^T \phi_s(L_n^*))].$$

The only difference between Eq. 7 and standard SVM is an additional term $w_s^T \phi_s(L_n^*)$. Since $w_s^T \phi_s(L_n^*)$ is a constant in the optimization, it can be taken as an additional dimension of $Vec(\widetilde{\Phi_a}(I_{H_n}, L_n^*))$. In this way, the Eq. 7 can be solved by a standard SVM solver. After we get $\widetilde{P_H}$, the P_H can then be computed by $(W_a W_a^T)^{-\frac{1}{2}} \widetilde{P_H}$.

3.2.3 Training Details

To start the loop of the coordinate descent procedure, one need to give initial values for either $\{W_a, w_s\}$ or $\{P_H, P_L\}$. In our implementation, we calculate the PCA of HOG features from randomly generated high and low resolution patches, and use the first n_d eigenvectors as the initial value of P_H and P_L, respectively. We use the HOG features in [19] and abandon the last truncation term, thus $n_f = 31$ in our experiment. The dimension n_d determines how much information is kept for sharing. We examine the effect of n_d in the experiments. The solver in optimizing the problem Eq. 6 and Eq. 7 are based on the [22]. The maximum number of the coordinate descent loop is set to be 8. The bin size in HOG is set to 8 for high resolution model, and 4

for low resolution. The root filter contains 8×4 HOG cells for both low and high resolution detection model.

4. Pedestrian-Vehicle Context in Traffic Scenes

A lot of detections are located around vehicles in traffic scenes (33.19% for our MT-DPM detector on Caltech Benchmark), as shown in Fig. 4. It is possible to use the pedestrian-vehicle relationship to infer whether the detection is true or false positive. For example, if we know the location of vehicles in Fig. 4, the detections above a vehicle, and detection at the wheel position of a vehicle can be safely removed. Fortunately, the vehicles are more easier to be localized than pedestrians, which has been proved in previous work (e.g. Pascal VOC [17], KITTI [20]). Since it is difficult to capture the complex relationship by handcraft rules, we build a context model and learn it automatically from data.

We split the spatial relationship between pedestrians and vehicles into five types, including: "Above", "Next-to", "Below", "Overlap" and "Far". We denote the feature of pedestrian-vehicle context as $g(p, v)$. If a pedestrian detection p and a vehicle detection[1] v have one of the first four relationships, the context features at the corresponding dimensions are defined as $(\sigma(s), \Delta c_x, \Delta c_y, \Delta h, 1)$, and other dimensions retain to be 0. If the pedestrian detection and vehicle detection are too far or there's no vehicle, all the dimensions of its pedestrian-vehicle feature is 0. Here $\Delta c_x = |c_{v_x} - c_{p_x}|$, $\Delta c_y = c_{v_y} - c_{p_y}$, and $\Delta h = h_v/h_p$, where (c_{v_x}, c_{v_y}), (c_{p_x}, c_{p_y}) are the center coordinates of vehicle detection v and pedestrian detection p, respectively. $\sigma(s) = 1/(1 + exp(-2s))$ is used to normalize the detection score to $[0, 1]$. For the left-right symmetry, the absolute operation is conducted for Δc_x. Moreover, as pointed in [33], there also has a relationship between the coordinate and the scale of pedestrians under the assumption that the cameras is aligned with ground plane. We further define this geometry context feature for pedestrian detection p as $g(p) = (\sigma(s), c_y, h, c_y^2, h^2)$, where s, c_y, h are the detection score, y-center and height of the detection respectively, and c_y and h are normalized by the height of the image.

To fully encode the context, we defined the model on the whole image. The context score is the summation of context scores of all pedestrian detections, and context score of a pedestrian is further divided to its geometry and pedestrian-vehicle scores. Suppose there are n pedestrian detections $P = \{p_1, p_2, \cdots, p_n\}$ and m vehicle detections $V = \{v_1, v_2, \cdots, v_m\}$ in an image, the context score of the image is defined as:

$$S(P, V) = \sum_{i=1}^{n} (w_p^T g(p_i) + \sum_{j=1}^{m} w_v^T g(p_i, v_j)), \quad (8)$$

[1]We use a DPM based vehicle detector trained on Pascal VOC 2012 [17] in our experiments.

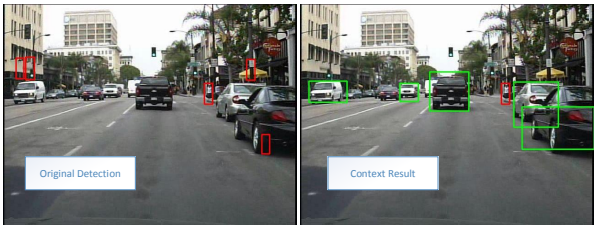


Figure 4. Examples of original detection, and the detection optimized by the context model.

where w_p and w_v are the parameters of geometry context and pedestrian-vehicle context, which ensure the truth detection (P, V) has larger context score than any other detection hypotheses.

Given the original pedestrians and vehicles detection P and V, whether each detection is a false positive or true positive is decided by maximizing the context score:

$$\arg \max_{t_{p_i}, t_{v_j}} \sum_{i=1}^{n} (t_{p_i} w_p^T g(p_i) + t_{p_i} \sum_{j=1}^{m} t_{v_j} w_v^T g(p_i, v_j)), \quad (9)$$

where t_{p_i} and t_{v_j} are the binary value, 0 means the false positive and 1 means the true positive. Eq. 9 is a integer programming problem, but becomes trivial when the label of V is fixed, since it equals to maximizing every pedestrians independently. In typical traffic scenes, the number of vehicles is limited. For example, in Caltech Pedestrian Benchmark, there are no more than 8 vehicles in an image, so that the problem can be solved by no more than 2^8 trivial sub-problems, which can be very efficient in real applications.

For the linear property, Eq. 9 is equal to:

$$\arg \max_{t_{p_i}, t_{v_j}} [w_p, w_v][\sum_{i=1}^{n} t_{p_i} g(p_i), \sum_{i=1}^{n} t_{p_i} \sum_{j=1}^{m} t_{v_j} g(p_i, v_j)]^T, \quad (10)$$

Eq. 10 provides a natural way for max-margin learning. We use w_c to denote $[w_p, w_v]$. Given the ground truth hypotheses of vehicles and pedestrians, a standard structural SVM [39] can be used here to discriminatively learn w_c by solving the following problem:

$$min_{w_c, \xi_k} \frac{1}{2} \|w_c\|_2^2 + \lambda \sum_{k}^{K} \xi_k \quad (11)$$
$$s.t. \forall P', \forall V', S(P_k, V_k) - S(P_k', V_k') \geq L(P_k, P_k') - \xi_k,$$

where P_k' and V_k' are arbitrary pedestrian and vehicle hypotheses in the kth image, and P_k and V_k are the ground truth. $L(P_k, P_k')$ is the Hamming loss of pedestrian detection hypothesis P_k' and ground truth P_k. The difficulty in pedestrian based applications is that only pedestrian ground

truth P_k is available in public pedestrian databases, and vehicle annotation V_k is unknown. To address the problem, we use the noisy vehicle detection result as the initial estimation of V_k, and jointly learn context model and infer whether the vehicle detection is true or false positive, by optimizing the following problem:

$$min_{w_c,\xi_k} \frac{1}{2}\|w_c\|_2^2 + \lambda \sum_k^K \xi_k \quad (12)$$
$$s.t. \forall P', \forall V' : \max_{\widehat{V}_k \subseteq V_k} S(P_k, \widehat{V}_k) - S(P'_k, V'_k) \geq L(P_k, P'_k) - \xi_k,$$

where $\widehat{V}_k$ is a subset of V_k, which reflects the current inference of the vehicle detections by maximizing the overall context score. Eq. 12 can be solved by optimizing the model parameters w_c and the label of vehicles $\widehat{V}_k$ iteratively. In the learning phase, the initial P'_k is the pedestrian detection result of MT-DPM.

5. Experiments

Experiments are conducted on the Caltech Pedestrian Benchmark [14][2]. Following the experimental protocol, the set00-set05 are used for training and set06-set10 are used for test. We use the ROC or the mean miss rate[3] to compare methods as advised in [14]. For more details of the benchmark, please refer to [14]. There are various sub-experiments on the benchmark to compare detectors in different conditions. Due to the space limitation, we only report the most relevant and leave results of other sub-experiments in the supplemental material. We emphasize that our method outperforms all the 17 methods evaluated in [14] on the 9 sub-experiments significantly.

In the following experiments, we examine the influence of the subspace dimension in MT-DPM, then compare it with other strategies for low resolution detection. The contribution of context model is also validated at different FPPI. Finally we compare the performance with other state-of-the-art detectors.

5.1. The Subspace Dimension in MT-DPM

The dimension of the mapped common subspace in MT-DPM reflects the tradeoff between commonness and differences among different resolutions. The high dimensional subspace can capture more differences, but may loss the generalities. We examine the parameter between 8 and 18 with a interval 2, and measure the performance on pedestrians taller than 30 pixels. We report the mean miss rate, as shown in Fig. 5. The MT-DPM achieves the lowest miss rate when the dimension is set to be 16, and tend to be stable between 14 and 18. In the following experiments, we fix it to be 16.

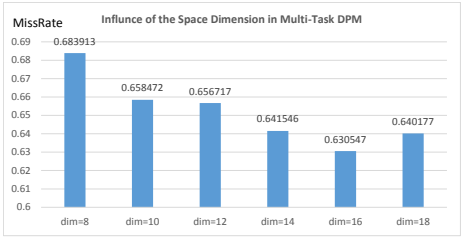


Figure 5. Influence of the subspace dimension in MT-DPM.

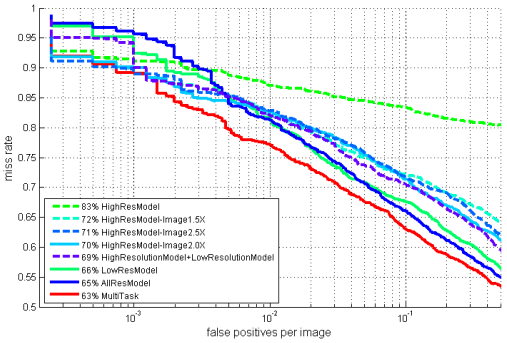


Figure 6. Results of different methods in multi-resolution pedestrian detection.

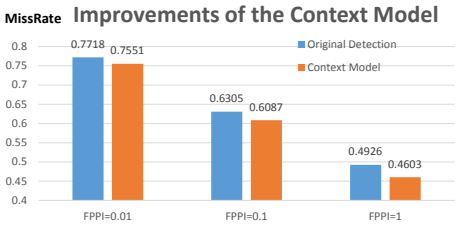


Figure 7. Contributions of the context cues in multi-resolution pedestrian detection.

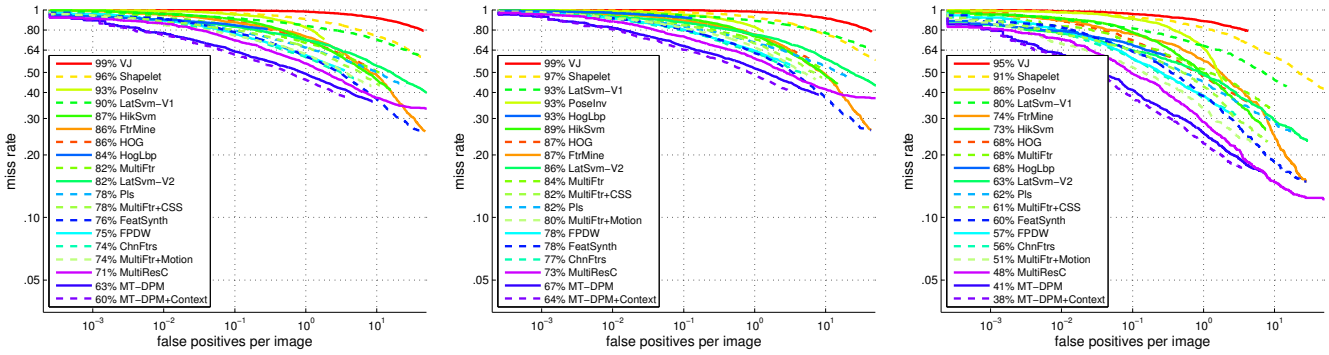
5.2. Comparisons with Other Detection Strategies

We compare the proposed MT-DPM with other strategies for multi-resolution pedestrian detection. All the detectors are based on DPM and applied on original images except for specially mentioned. The compared methods including: (1) DPM trained on the high resolution pedestrians; (2) DPM trained on the high resolution pedestrians and tested by resizing images 1.5, 2.0, 2.5 times, respectively; (3) DPM trained on low resolution pedestrians; (4) DPM trained on both high and low resolution pedestrians data (Fig. 2(a)); (5) Multi-resolution DPMs trained on high resolution and low resolution independently, and their detection results are fused (Fig. 2(b)).

ROCs of pedestrians taller than 30 pixels are reported

[2]http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

[3]We used the mean miss rate defined in P. Dollár's toolbox, which is the mean 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.



(a) Multi-resolution (taller than 30 pixels) (b) Low resolution (30-80 pixels high) (c) Reasonable (taller than 50 pixels)

Figure 8. Quantitative result of MT-DPM, MT-DPM+ Context and other methods on the Caltech Pedestrian Benchmark.

in Fig. 6. High resolution model can not detect the low resolution pedestrians directly, but some of the low resolution pedestrians can be detected by resizing images. However, the number of false positives also increases, which may hurt the performance (see HighResModel-Image1.5X, HighResModel-Image2.0X, HighResModel-Image2.5X in Fig. 6). The low resolution DPM outperforms high resolution DPM, since the low resolution pedestrians is more than high resolution pedestrians. Combining low and high resolution would always help, but the improvement depends on the strategy. Fusing low and high resolution data to train a single detector is better than training two independent detectors. By exploring the relationship of samples from different resolutions, our MT-DPM outperforms all other methods.

5.3. Improvements of Context Model

We apply the context model on the detections of MT-DPM, and optimize every image independently. The miss rate at 0.01, 0.1 and 1 FPPI for pedestrians taller than 30 pixels are shown in Fig. 7. The context model reduces the miss rate from 63.05% to 60.87% at 0.1 FPPI. The improvement of context is more remarkable when more false positives are allowed, for example, there is a 3.2% reduction of miss rate at 1 FPPI.

5.4. Comparisons with State-of-the-art Methods

In this part, we compare the proposed method with other state-of-the-art methods evaluated in [14], including: Viola-Jones [40], Shapelet [44], LatSVM-V1, LatSVM-V2 [19], PoseInv [30], HOGLbp [43], HikSVM [31], HOG[6], FtrMine [13], MultFtr [44], MultiFtr+CSS [44], Pls [37], MultiFtr+Motion [44], FPDW [11], FeatSynth [1], ChnFtrs [12], MultiResC [33]. The results of the proposed methods are denoted as MT-DPM, and MT-DPM+ Context. For the space limitation here, we only show results of multi-resolution pedestrians (Fig. 8(a), taller than 30 pixels), low resolution (Fig. 8(b), 30-80 pixels high), reasonable (Fig. 8(c), taller than 50 pixels)[4]. Our MT-DPM significantly outperforms previous state-of-the-art, at least a 6% margin mean miss rate on all the three experiments. The proposed Context model further improves the performance with about 3%. Because the ROC of [9] is not available, its performance is not shown here. But as reported in [9], it got 48% mean miss rate on the reasonable condition, while our method reduces it to 41%. The most related method is MultiResC [33], where multi-resolution model is also used. Our method outperforms it with a 11% margin for multi-resolution detection, which can prove the advantage of the proposed method.

5.5. Implementation Details

The learned MT-DPM detector can benefit from a lot of speed up methods for DPM. Specially for our implementation, we modified the code of the FFT based implementation [15] for the fast convolution computation. The time for processing one frame is less than 1s on a standard PC, including high resolution and low resolution pedestrian detection, vehicle detection and context model. More speed-up can be achieved by parallel computing or pruning the search space by the temporal information.

6. Conclusion

In this paper, we propose a Multi-Task DPM detector to jointly encode the commonness and differences between pedestrians from different resolutions, and achieve robust performance for multi-resolution pedestrian detection. The pedestrian-vehicle relationship is modeled to infer the true or false positives in traffic scenes, and we show how to learn it automatically from the data. Experiments on challenging Caltech Pedestrian Benchmark show the significant improvement over state-of-the-art performance. Our future work is to explore the spatial-temporal information and extend the proposed models to general object detection task.

[4]Results of other sub-experiments are in the supplemental material.

Figure 9. Qualitative results of the proposed method on Caltech Pedestrian Benchmark (the threshold corresponds to 0.1 FPPI).

Acknowledgement

We thank the anonymous reviewers for their valuable feedbacks. This work was supported by the Chinese National Natural Science Foundation Project #61070146, #61105023, #61103156, #61105037, #61203267, National IoT R&D Project #2150510, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, European Union FP7 Project #257289 (TABULA RASA), and AuthenMetric R&D Funds.

References

[1] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg. Part-based feature synthesis for human detection. *ECCV*, 2010. 7

[2] O. Barinova, V. Lempitsky, and P. Kholi. On detection of multiple object instances using hough transforms. *PAMI*, 2012. 2

[3] C. Beleznai and H. Bischof. Fast human detection in crowded scenes by contour integration and local shape estimation. In *CVPR*. IEEE, 2009. 2

[4] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *CVPR*. IEEE, 2012. 1, 2

[5] S. Biswas, K. W. Bowyer, and P. J. Flynn. Multidimensional scaling for matching low-resolution face images. *PAMI*, 2012. 2

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*. IEEE, 2005. 1, 2, 7

[7] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *ECCV*, 2006. 2

[8] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 2011. 2

[9] Y. Ding and J. Xiao. Contextual boost for pedestrian detection. In *CVPR*. IEEE, 2012. 2, 7

[10] P. Dollár, R. Appel, and W. Kienzle. Crosstalk cascades for frame-rate pedestrian detection. In *ECCV*. Springer, 2012. 1, 2

[11] P. Dollár, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. *BMVC 2010*, 2010. 1, 2, 7

[12] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009. 2, 7

[13] P. Dollár, Z. Tu, H. Tao, and S. Belongie. Feature mining for image classification. In *CVPR*. IEEE, 2007. 7

[14] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 2012. 1, 2, 3, 6, 7

[15] C. Dubout and F. Fleuret. Exact acceleration of linear object detectors. *ECCV*, 2012. 7

[16] M. Enzweiler and D. Gavrila. Monocular pedestrian detection: Survey and experiments. *TPAMI*, 2009. 2

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal voc 2012 results. 5

[18] P. Felzenszwalb, R. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *CVPR*. IEEE, 2010. 1, 2

[19] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010. 1, 2, 3, 4, 7

[20] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*. IEEE, 2012. 5

[21] D. Geronimo, A. Lopez, A. Sappa, and T. Graf. Survey of pedestrian detection for advanced driver assistance systems. *PAMI*, 2010. 2

[22] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/. 4

[23] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. *ECCV*, 2012. 1, 2

[24] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 2008. 2

[25] C. Huang and R. Nevatia. High performance object detection by collaborative learning of joint ranking of granules features. In *CVPR*. IEEE, 2010. 1, 2

[26] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *ECCV*. Springer, 2012. 2

[27] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*. IEEE, 2011. 2

[28] Z. Lei and S. Z. Li. Coupled spectral regression for matching heterogeneous faces. In *CVPR*. IEEE, 2009. 2

[29] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for scene understanding. In *CVPR*. IEEE, 2012. 2

[30] Z. Lin and L. Davis. A pose-invariant descriptor for human detection and segmentation. *ECCV*, 2008. 7

[31] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*. IEEE, 2008. 1, 7

[32] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 2000. 2

[33] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. *ECCV*, 2010. 1, 2, 5, 7

[34] H. Pirsiavash and D. Ramanan. Steerable part models. In *CVPR*. IEEE, 2012. 2

[35] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Bilinear classifiers for visual recognition. In *NIPS*, 2009. 2

[36] M. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 2

[37] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis. Human detection using partial least squares analysis. In *ICCV*. IEEE, 2009. 7

[38] S. Tang, M. Andriluka, and B. Schiele. Detection and tracking of occluded people. In *BMVC*, 2012. 2

[39] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 2006. 5

[40] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 2005. 1, 2, 7

[41] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. In *CVPR*. IEEE, 2010. 1, 2

[42] M. Wang, W. Li, and X. Wang. Transferring a generic pedestrian detector towards specific scenes. In *CVPR*. IEEE, 2012. 2

[43] X. Wang, T. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *ICCV*. IEEE, 2009. 1, 2, 7

[44] C. Wojek and B. Schiele. A performance evaluation of single and multi-feature people detection. *DAGM*, 2008. 7

[45] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *CVPR*. IEEE, 2009. 2

[46] J. Yan, Z. Lei, D. Yi, and S. Z. Li. Multi-pedestrian detection in crowded scenes: A global view. In *CVPR*. IEEE, 2012. 2