Multi-Pedestrian Detection in Crowded Scenes: A Global View

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

Abstract

Recent state-of-the-art algorithms have achieved good performance on normal pedestrian detection tasks. However, pedestrian detection in crowded scenes is still challenging due to the significant appearance variation caused by heavy occlusions and complex spatial interactions. In this paper we propose a unified probabilistic framework to globally describe multiple pedestrians in crowded scenes in terms of appearance and spatial interaction. We utilize a mixture model, where every pedestrian is assumed in a special subclass and described by the sub-model. Scores of pedestrian parts are used to represent appearance and quadratic kernel is used to represent relative spatial interaction. For efficient inference, multi-pedestrian detection is modeled as a MAP problem and we utilize greedy algorithm to get an approximation. For discriminative parameter learning, we formulate it as a learning to rank problem, and propose Latent Rank SVM for learning from weakly labeled data. Experiments on various databases validate the effectiveness of the proposed approach.

1. Introduction

Pedestrian detection in real-world scenes is a crucial component for a variety of applications, such as video surveillance and drive-assistance systems, and robust pedestrian detection can also be used to improve the performance of other components such as pedestrian tracking. Recent pedestrian detection algorithms [4, 5] can achieve good performance on normal scenes, but their performance in crowded scenes are far from being satisfactory due to the ambiguous appearance caused by heavy occlusions and complex spatial relationship between different pedestrians.

Traditional template pedestrian detection methods [17, 3] always work as follows. Given an image, multi-scale windows slide over the whole image and a binary classifier is adopted to determine whether there is a pedestrian within the window. Since one pedestrian is usually bounded by different windows, non-maxima suppression (NMS) [14, 10] is often adopted to reconcile the overlap detections according to an overlap criterion. There are at least two shortcomings for these methods in crowded scenes. Firstly, the appearance of pedestrians varies a lot because of the severe occlusions thereby the single binary classifier is difficult to detect all the pedestrians correctly. Secondly, in crowded scenes, multiple pedestrians are usually very close with each other and they are not independent to some extent, therefore the fixed threshold used in NMS may not work.

To find the global optimal solution of multiple pedestrians layout in the entire image, we relax the independent assumption and model both the appearance and spatial interaction. We propose a probabilistic model to explore the relationship between the two clues, and multi-pedestrian detection is formulated as a maximizing a posteriori (MAP) problem. To better model the complex appearance and spatial relationship in crowded scenes, we utilize a mixture model with several subclasses which are automatically learned from the weakly labeled data. Every pedestrian is assumed in a special subclass and described by the sub-model. The appearance model is built on Felzenszwalb's part based model[10], where the scores of parts are used as a second level feature representation and every subclass has a weight vector to describe the importance of different parts. The spatial model is used to describe the interactions and in our model quadratic kernel[11] is used.

Since finding the global optimal solution in the model is impossible, a greedy approximate algorithm is used to efficiently infer the best hypothesis. For discriminative learning, we formulate the parameter estimation problem as a learning to rank problem [13] where we aim to find an optimal model that can generate higher score for the ground-truth layout than any other hypothesis. To handle the constraints of the exponential order, we use a data mining hard samples approach to efficiently solve the problem similar to [10] . In real world practice, we only have weakly labeled locations of pedestrians and don't have access to the subclass annotations, therefore, we take the subclass label

*Stan Z. Li is the corresponding author.

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

latent variable and propose Latent Rank SVM to simultaneously learn the model's parameter and latent subclass label.

The rest of the paper is organized as follows: Section 2 reviews the related work. Our probabilistic model in terms of appearance and spatial interaction is discussed in Section 3. Section 4 gives the optimization techniques for model learning and inference. Section 5 shows the experiments and finally in Section 6 we conclude the paper.

2. Related Work

The most popular pedestrian detection method is template based method, which uses a pre-trained classifiers to scan images in possible locations and scales to determine whether the window contain a pedestrian or not. The performance of the method benefits from low level features and training methods. In [17], Haar feature and boosting classifier were used, and in [3], HOG (Histogram of Gradient) feature and SVM classifier were used. Further improvements include [19, 6, 5, 10, 20, 18] and so on, where novel features or more powerful classifiers were utilized . A more detailed survey of these methods can be found in [7].

Compared with traditional pedestrian detection, there are few papers aiming on multi-pedestrian detection in crowded scenes despite its importance in real applications. [15] extended the implicit shape model(ISM)[14] by using both local and global information to segment different pedestrians. [1] avoided NMS by reproducing the ISM model and transform it into an energy maximization problem. Another related paper is [4], which explored spatial relationship between multi-class objects and formulated it as an energy maximization problem. Compared with [4], we aim at a different but more challenging problem and use different techniques in model, representation, learning and inference.

3. Model

In this part, we propose a unified probabilistic model to describe multiple pedestrians in crowded scenes in terms of appearance and pairwise interactions, and then the pedestrian detection task is formulated as an energy maximization problem. Firstly the model formulations are given, and then we discuss the feature representation.

3.1. Formulation

Given an image I and a search strategy, the collection of possible bounding boxes is denoted as $B = \{b_1, b_2, \cdots, b_N\}$, where N is the total number, b_i is the ith bounding box with center (x_i, y_i) and scale s_i. The label of b_i is written as t_i, where $t_i = 1$ if b_i is corresponding to a pedestrian and otherwise $t_i = 0$. The model of pedestrians in crowded scenes is $\mathcal{F}_\theta$ parameter $\theta = (\theta_a, \theta_s)$, where θ_a and θ_s are used to model appearance and spatial relationship respectively. The detection task is to infer labels of b_i in B, and we call the detection result a hypothesis as $T = \{t_1, t_2, \cdots, t_N\}$. Given an image I and model $\mathcal{F}_\theta$, the pedestrian detection task is equal to finding a hypothesis T that maximizes the probability $P(T|I, \theta)$. Using the Bayes rule, we can get:

$$\arg\max_T P(T|I,\theta) \propto \arg\max_T P(I|T,\theta)P(T|\theta) \quad (1)$$

where $p(I|T, \theta)$ describes the appearance model, and it measures the likelihood of seeing a particular image given hypothesis T and model $\mathcal{F}_\theta$; $P(T|\theta)$ is the prior term. Similar to [12], where the prior is used to model the relative position of body parts, here we use the prior term to model the relative spatial interaction between different pedestrians in the hypothesis T.

In the appearance model, the appearance likelihood of b_i in B are assumed to be independently and identically distributed, so the likelihood term $P(I|T, \theta)$ can be decomposed as:

$$P(I|T,\theta) \approx \prod_{i=1}^{N} p(I_{b_i}|T,\theta_a) = \prod_{i=1}^{N} p(I_{b_i}|t_i,\theta_a) \quad (2)$$

where I_{b_i} is the image region located at b_i. The right equation is satisfied since that $p(I_{b_i}|T, \theta)$ is only dependent with t_i and is independent with t_j, where $j \neq i$.

In the spatial interaction model, for simplicity we take pairwise interactions as our basic element and high order interaction are built indirectly through these pairwise interactions. The pairwise interaction $p(t_i, t_j|\theta_s)$ means the joint probability that b_i has the label t_i and b_j has the label t_j, given a spatial model parameterized by θ_s. For a hypothesis with N boxes, the spatial model can be decomposed as:

$$P(T|\theta) = P(T|\theta_s) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(t_i, t_j|\theta_s) \quad (3)$$

By substituting Eq. 2 and Eq. 3 into Eq. 1 and taking the logarithm operation, the MAP problem equals the following energy maximization problem:

$$\arg\max_{T, t_i \in [0,1]} \sum_{i=1}^{N} f_{\theta_a}(I_{b_i}, t_i) + \sum_{i=1}^{N}\sum_{j=1}^{N} f_{\theta_s}(t_i, t_j) \quad (4)$$

where $f_{\theta_a}(I_{b_i}, t_i) = \log(p(I_{b_i}|t_i, \theta_a))$ and $f_{\theta_s}(t_i, t_j) = \log(p(t_i, t_j|\theta_s))$.

Obviously, the traditional sliding window based detection method can be derived by setting $f_{\theta_s}(t_i, t_j) = 0$, $f_{\theta_a}(I_{b_i}, 1)$ to be the appearance score and $f_{\theta_a}(I_{b_i}, 0)$ to be a constant below pre-defined threshold. Since sliding window based algorithms often generate some overlapped detections of the same pedestrian, NMS is often used as a post-processing method which can also be shown to be a special case by setting $f_{\theta_s}(t_i, t_j) = -\infty$ if b_i, b_j have an overlap larger than a criterion and $t_i = t_j = 1$.

3.2. Representation

In this part we give the concrete form of $f_{\theta_a}(I_{b_i}, t_i)$ and $f_{\theta_s}(t_i, t_j)$ defined above, and we will first introduce the mixture model which can simplify the complexity of $f_{\theta_a}(I_{b_i}, t_i)$ and $f_{\theta_s}(t_i, t_j)$.

3.2.1 Mixture Model

We propose to use simple linear model to describe the pedestrians in crowded scenes considering the efficiency. However, a single linear model may be too simple to describe the complex appearance and spatial variations of pedestrians in real world. To better describe the complex scenes while still use linear model, we introduce a mixture model composed of K sub-classes where every pedestrian is represented by its subclass. The label of b_i is extended to be $l_i \in \{0, 1, \cdots, K\}$ from $t_i \in \{0, 1\}$, where $l_i = 0$ stands for background and $l_i = k$ stands for the kth subclass. Since in real applications only the binary label t_i is given, we will take the subclass label l_i a latent variable. Techniques used to infer l_i in training step will be discussed in the next part, here we assume subclass labels in the training data are given thus we can use them to learn the subclass model. By using the mixture model, the detection task turns out to be finding the optimal hypothesis $L = \{l_1, \cdots, l_N\}$ by maximizing the following energy function $E(L)$:

$$\arg \max_{L, l_i \in [0,K]} \sum_{i=1}^{N} f_{\theta_a}(I_{b_i}, l_i) + \sum_{i=1}^{N} \sum_{j=1}^{N} f_{\theta_s}(l_i, l_j) \quad (5)$$

3.2.2 Appearance Model

Pedestrian is usually represented by some low-level features such as HOG, SIFT, LBP which are then fed into a classifier to train the model. Although effective, it may not be robust enough for pedestrian detection with heavy occlusions. To handle the occlusions, we use a two-layer representation. The first layer is Felzenszwalb's part based model[10] which uses a deformable template to represent the part and the global appearance. In [10] part scores are summarized to give the final score. This pooling method is suitable for normal pedestrian detection task, but may have problems when a pedestrian is partially visible, because the scores in overlapped parts will be low which may result in an omission. In order to improve it, we use a second layer, where the features are the scores of each part and the appearance parameter θ_a represents the weight of each part. For example, if one subclass has only upper body, the weight corresponding to upper body part can be high while other weights are low. By using mixture model and adding part weight to each subclass, we can model the pedestrians under different occlusion conditions. Another advantage of the representation is that the dimension is relative low compared with the low level feature so that it is more efficient for parameter learning.

Given the subclass label l_i, we can simplify the appearance model as a linear function:

$$f_{\theta_a}(I_{b_i}, l_i) = \theta_{al_i}^T f(I_{b_i}) \quad (6)$$

where $f(I_{b_i})$ is the part score vector of image I in the region b_i. θ_{al_i} is the parameter vector to model the appearance of the l_ith subclass, and in our paper it stands for the weight of different parts. For the background class θ_{al_0} is directly set to be a zero vector.

3.2.3 Interaction Model

The interaction model is used to describe the spatial co-occurrence of different pedestrians. For example, a bounding box with only head part has high score may be detected as a pedestrian only when there is a pedestrian around it. Given the subclass labels l_i and l_j, the interaction model is also simplified to be a linear function as:

$$f_{\theta_s}(l_i, l_j) = \theta_{sl_il_j}^T f(b_i, b_j) \quad (7)$$

where $\theta_{sl_il_j}$ parameterizes the interaction model between the subclass l_i and l_j, $f(b_i, b_j)$ describes the relative spatial relationship between b_i and b_j. $f(b_i, b_j)$ should be flexible enough to suppress reduplicate detections and enhance true detections with low appearance scores according to the relative spatial interaction model. We use (x_i, y_i),(x_j, y_j) to denote the center of b_i and b_j, and s_i, s_j to denote the scale of b_i and b_j, then we define

$$f(b_i, b_j) = [1, dx, dy, ds, dx^2, dy^2, ds^2]^T \quad (8)$$

where 1 is used as a bias term, and the left ones are the quadratic kernel. dx, dy, ds represent $x_i - x_j$, $y_i - y_j$ and $s_i - s_j$ respectively. In our implementation, $f(b_i, b_j)$ is meaningful only when there is overlap between b_i and b_j, otherwise $f(b_i, b_j)$ is set to be a 0 vector. Similar to the appearance model, the interactions involves background are also set to be 0.

4. Optimization

The optimization includes two aspects: for model learning, we need to estimate the optimal parameters of appearance model θ_a and spatial interaction model θ_s from weakly labeled data; for inference, we need to maximize the energy function to find the best hypothesis. Firstly we give the greedy algorithm for efficient inference, and then we present our Latent Rank-SVM algorithm for discriminant parameter learning.

4.1. Inference

Given an image I, the number of the possible hypothesis is of exponential order. For example, there are about 10^5 bounding boxes in a single image and then the number of possible hypotheses is 2^{10^5}. Since it's impossible to test all hypotheses, we aim to get an approximate solution.

For efficient inference, we define $k-j$ expansion operation, which is used to select a box $b_j(l_j \neq k)$ and set its label to be k while keeps labels of other boxes unchanged. We set the initial label of all bounding box to be 0, thus the initial score of hypothesis $E(L)$ is 0. At every loop, we find a best $k-j$ expansion operation that generates a new hypothesis with the highest score. When the score of hypothesis can't be increased any more, the algorithm is terminated. Since $E(L)$ has an upper bound and keeps increasing in every loop, the algorithm can be terminated in finite steps.

To speed up inference, we set the appearance score and interaction score with background class to be 0. Since a large proportion of b_i are background, a lot of calculation can be avoided. Another technique is that we can reject a lot of bounding boxes with low appearance scores, where the threshold is selected according to the probably approximately admissible (PAA) threshold selection method used in [9].

To give the confidence of each detected bounding box , we define the score of b_i as

$$s(b_i) = E(L) - E(\widetilde{L}) \tag{9}$$

where the $\widetilde{L}$ is the hypothesis the same as L except that $\widetilde{l_i} = 0$. The score measures the contribution of b_i to the global energy maximization problem.

4.2. Learning

The variables to be determined consist of three parts: appearance model parameters $\theta_a = \{\theta_{ai} | i = 1, \cdots, K\}$, spatial interaction model parameters $\theta_s = \{\theta_{sij} | i, j = 1, \cdots, K\}$ and latent subclass labels $L_m = \{l_i | i = 1, \cdots, N_m\}$ of labeled pedestrians in training image set $\{I_1, \cdots, I_M\}$. Different from common learning problems, we have two difficulties: (1) the number of constraint is too large; (2)we don't know the subclass label L, instead we only know binary label T. First we assume that the subclass labels of training data are known, and adopt a hard samples mining approach to deal with numerous constraints, then we solve the total problem using a EM-like approach.

To formulate the Eq. 5 into a linear form, we add additional dimension to the feature representation. The original appearance feature $f(I_{b_i})$ is extended to be:

$$f'(I_{b_i}) = [\underbrace{0, \cdots, 0}_{\sum_{t=1}^{l_i-1} F}, f_1(I_{b_i}), \cdots, f_F(I_{b_i}), \underbrace{0, \cdots, 0}_{\sum_{t=l_i+1}^{K} F}]^T \tag{10}$$

where $f_j(I_{b_i})$ is the score of the jth part on I_{b_i}, and F is the length of appearance feature, which is the number of parts in this paper. Similarly the spatial interaction feature f_{b_i,b_j} is extended to be:

$$f'(b_i, b_j) = [\underbrace{0, \cdots, 0}_{\sum_{t=1}^{t1} D}, f_1(b_i, b_j), \cdots, f_D(b_i, b_j), \underbrace{0, \cdots, 0}_{\sum_{t=t2}^{K^2} D}]^T; \tag{11}$$

we use D to denote the length of original spatial interaction feature and $t1 = \sum_{t=1}^{(l_i-1)K+l_j-1} D$, $t2 = \sum_{(l_i-1)K+l_j+1}^{K^2} D$. We concatenate the appearance parameter into a vector as $w_a = [\theta_{a_1}^T, \theta_{a_2}^T, \cdots, \theta_{a_K}^T]^T$ and spatial parameter into a vector as $w_s = [\theta_{s_{11}}^T, \cdots, \theta_{s_{ij}}^T, \cdots, \theta_{s_{KK}}^T]^T$. Then the energy function in Eq. 5 equals:

$$\arg\max_L \sum_{i=1}^{N} w_a^T f'(I_{b_i}) + \sum_{i=1}^{N}\sum_{j=1}^{N} w_s^T f'(b_i, b_j) \tag{12}$$

$$= [w_a^T, w_s^T] \begin{bmatrix} \sum_{i=1}^{N} f'(I_{b_i}) \\ \sum_{i=1}^{N}\sum_{j=1}^{N} f'(b_i, b_j) \end{bmatrix} \tag{13}$$

Here we use w to denote $[w_a^T, w_s^T]^T$, and $\Phi(I, L)$ to denote the feature map $[\sum_{i=1}^{N} f'(I_{b_i})^T, \sum_{i=1}^{N}\sum_{j=1}^{N} f'(b_i, b_j)^T]^T$. Here we assume that we have known subclass label L, and the algorithm to simultaneously infer L will be discussed later. Given training image I with the ground-truth hypothesis L, we aim to find w that ensures that $E(L) = w^T\Phi(I, L)$ is larger than energy of any other hypothesis $E(H_i) = w^T\Phi(I, H_i)$. Suppose we have M training images, the objective function can be defined as:

$$\arg\min_{w, \xi_{mi} \geq 0} w^T w + C \sum_{m=1}^{M}\sum_{i=1}^{N_m} \xi_{mi} \tag{14}$$

$$s.t. \forall m \in [1, M], \forall i \in [1, N_m], \tag{15}$$

$$w^T\Phi(I_m, L_m) - w^T\Phi(I_m, H_{mi}) \geq l(L_m, H_{mi}) - \xi_{mi}$$

where $w^T w$ is the regularization, and ξ_{mi} is used to punish the incorrect hypotheses. $l(L_m, H_{mi})$ is the loss function to measure the difference between L_m and H_{mi}. The loss function is defined according to criterion of PASCAL VOC[8] where $l(L_m, H_{mi}) = 1$ if there has false positives and true negatives, otherwise $l(L_m, H_{mi}) = 0$.

The above optimization problem is a learning to rank problem and can be transformed into a linear SVM since that we can take $\Delta(I_m, L_m, H_{mi}) = \Phi(I_m, L_m) - \Phi(I_m, H_{mi})$ as the positive sample and $-\Delta(I_m, L_m, H_{mi})$ as the negative sample when $l(L_m, H_{mi}) = 1$. The difficulty is that we often have too many constraints since that the number of H_{mi} is of exponential order. To reduce the

number of constraints, here we use a data mining approach similar to [10] to find the optimal solution.

At the training procedure, only a subset of total training samples are used and kept in the cache P. Samples were divided into easy samples and hard samples, where $\Delta(I_m, L_m, H_m)$ is an easy sample if $\Delta(I_m, L_m, H_{mi}) > l(L_m, H_{mi})$, otherwise a hard one. We generate random initial negative hypotheses by disturbing the ground-truth hypotheses, and push them into the cache as well as the positive hypotheses. The model at every loop is updated based on current cache. After the model trained, we can use it to shrink cache by removing easy samples and grow the cache by finding hard samples of the current model based on the inference model discussed above. If all hard samples have been included in P, the algorithm is terminated.

In real applications, we often only have the binary label t_i instead of subclass label l_i, so that we have to infer the latent subclass with the model parameter simultaneously. The original optimization problem is extended as the following problem:

$$\arg \min_{l_i, w, \xi_{m_i} \geq 0} w^T w + C \sum_{m=1}^{M} \sum_{i=1}^{N_m} \xi_{mi} \tag{16}$$

$$s.t. \forall m, \forall i, \tag{17}$$
$$w^T \Phi(I_m, L_m) - w^T \Phi(I_m, H_{mi}) \geq l(L_m, H_{mi}) - \xi_{mi}$$

Compared with Eq. 14, above problem has latent variable l_i. To simultaneously learn the latent variable with the model parameter, we use a EM-like approach similar to latent-SVM used in [10]. Firstly we use k-means on annotated pedestrians to generate the initial hypotheses, and then we use it to train initial model $\mathcal{F}$ by Rank-SVM training algorithm discussed above. After the model trained, new subclass labels of annotated pedestrians in the training set can be inferred and then can be used to update the model again until it converged. The algorithm can't be guaranteed to get the global optimal solution but we can get a relative good solution in general.

5. Experiments

In this part we compare proposed method with MDPM[10] and FPDW[6], which are two state-of-the-art algorithms according to the experiments by [7] on general pedestrian detection benchmarks. The number of subclass in our model is set to be 3 for all experiments.

We select two challenging databases in crowded scenes: PETS2009[1] and UCSD[16]. The PETS2009 crowd database is a well-known benchmark for pedestrian counting and pedestrian tracking. In our experiment we select S1_L2 with high density crowd for training, S2_L2 with medium density crowd and S2_L3 with high density crowd for test. S1_L2 contains 201 frame and 5061 pedestrians; S2_L2 contains 436 frames and 8927 pedestrians; and S2_L3 contains 240 frames and 4509 pedestrians. All images in PETS2009 are with the resolution of 768×576, and every image is enlarged twice in detection. We annotate all the pedestrians even when they are partially visible or totally occluded in this database. The UCSD crowd database [16] were originally used for pedestrian counting and in order to evaluate detection the performance, we annotate all the pedestrians in the database. In our experiment S1 with 2997 frames are used for model learning and S2 with 1096 frames are used for test.

To give a fair comparison, we ignore information in video and detect pedestrians in every frame independently. We use the three detectors to infer multiple pedestrians layout and use the PASCAL VOC criterion[8] to generate recall-precision curve and use AP(Average-Precision) to compare different algorithms. The recall-precision curves are shown in Fig. 1 and some detection results are shown in Fig. 2.

From Fig. 1 we can find that proposed method achieves the highest AP on all the three databases. However the improvements are different according to the density level. Compared with MDPM, our proposed method can obtain about 5% to 10% AP improvement on all the three experiments. On medium density crowd, the improvement is about 5%, and on high density, the improvement is about 10%. Compared with FPDW, proposed method have similar AP on S2_L3 and improve the FPDW about 10% on S2_L3 and UCSD. The reason that FPDW has good performance on S2_L2 is that multiple feature fusion perform robust to the illumination variation in S2_L2. This motivates us to use multiple features to further improve the performance. Fig. 2 shows some detection results, we can find that some pedestrians with heavy occlusions can be successfully detected by proposed method. We also compare the three detectors on PETS2009_S2_L1, which is an easier database with pedestrians sparsely distributed in the scene. On this database, all the three detectors get similar 93% AP.

In these experiments we can prove that proposed method can improve the pedestrian detection performance over state-of-the-art static image pedestrian detectors on crowded scenes. In general, the improvement is significant on high density crowed scenes.

6. Conclusion

In this paper, we present a probabilistic model to globally model multiple pedestrians in crowded scenes in terms of appearance and spatial interactions. The pedestrian detection is formulated as a MAP problem. Mixture model is utilized to simplify the appearance and spatial interaction models. We propose Latent Rank SVM for discriminant parameter learning. Extensive experimental results show the

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

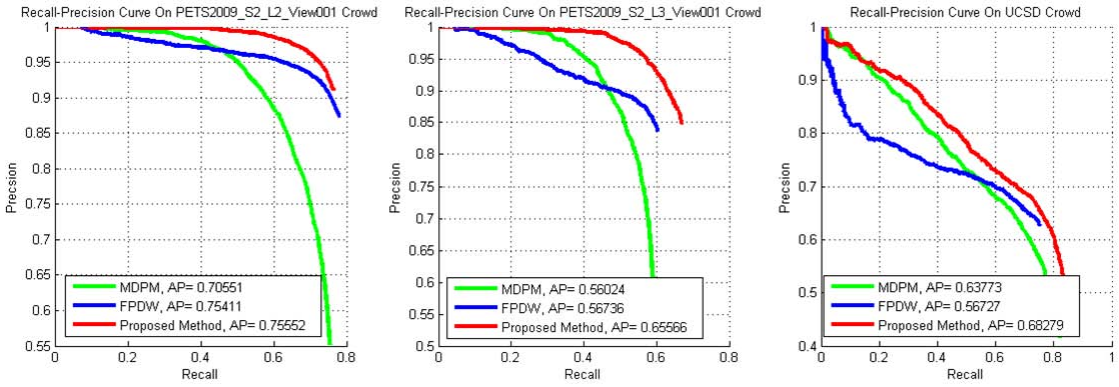

Figure 1. From left to right Recall-Precision curves of MDPM[10], FPDW[5] and proposed method on PETS2009_S2_L2_View001, PETS2009_S2_L3_View001 and UCSD respectively.

Figure 2. Detection result of proposed method on PETS2009_S2_L2_View001, PETS2009_S2_L3_View001 and UCSD.

proposed method achieve state-of-the-art pedestrian performance in challenging crowded scenes. In our future work, we will incorporate the video information into our method to improve the performance further.

Acknowledgement

This work was supported by the Chinese National Natural Science Foundation Project #61070146, #61105023, #61103156, #61105037, National IoT R&D Project #2150510, European Union FP7 Project #257289 (TABULA RASA http://www.tabularasa-euproject.org), and AuthenMetric R&D Funds.

References

[1] O. Barinova, V. Lempitsky, and P. Kohli. On detection of multiple object instances using hough transforms. In *CVPR*. IEEE, 2010.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, pages 1222–1239, 2001.

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

[4] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, pages 229–236. IEEE, 2009.

[5] P. Dollár, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *BMVC*, 2010.

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

[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 2011.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, June 2010.

[9] P. Felzenszwalb, R. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *CVPR*, pages 2241–2248. IEEE, 2010.

[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, pages 1627–1645, 2009.

[11] P. Felzenszwalb and D. Huttenlocher. Distance transforms of sampled functions. Technical report, Cornell Univ.CIS, 2004.

[12] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[13] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD*, pages 133–142. ACM, 2002.

[14] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1):259–289, 2008.

[15] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, volume 1, pages 878–885. IEEE, 2005.

[16] V. Rabaud and S. Belongie. Counting crowded moving objects. In *CVPR*, volume 1, pages 705–711. IEEE, 2006.

[17] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 63(2):153–161, 2005.

[18] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. In *CVPR*, pages 1030–1037. IEEE, 2010.

[19] X. Wang, T. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *ICCV*, pages 32–39. IEEE, 2009.

[20] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *CVPR*, pages 794–801. IEEE, 2009.