This CVPR2013 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Single-Pedestrian Detection aided by Multi-pedestrian Detection

Wanli Ouyang[1,2] and Xiaogang Wang [1,2]
[1] Shenzhen key lab of Comp. Vis. & Pat. Rec.,
Shenzhen Institutes of Advanced Technology, CAS, China
[2] Department of Electronic Engineering, The Chinese University of Hong Kong
wlouyang@ee.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk

Abstract

In this paper, we address the challenging problem of detecting pedestrians who appear in groups and have interaction. A new approach is proposed for single-pedestrian detection aided by multi-pedestrian detection. A mixture model of multi-pedestrian detectors is designed to capture the unique visual cues which are formed by nearby multiple pedestrians but cannot be captured by single-pedestrian detectors. A probabilistic framework is proposed to model the relationship between the configurations estimated by single- and multi-pedestrian detectors, and to refine the single-pedestrian detection result with multi-pedestrian detection. It can integrate with any single-pedestrian detector without significantly increasing the computation load. 15 state-of-the-art single-pedestrian detection approaches are investigated on three widely used public datasets: Caltech, TUD-Brussels and ETH. Experimental results show that our framework significantly improves all these approaches. The average improvement is 9% on the Caltech-Test dataset, 11% on the TUD-Brussels dataset and 17% on the ETH dataset in terms of average miss rate. The lowest average miss rate is reduced from 48% to 43% on the Caltech-Test dataset, from 55% to 50% on the TUD-Brussels dataset and from 51% to 41% on the ETH dataset.

1. Introduction

Pedestrian detection is one of the most important topics in object detection and has attracted a lot of attention [2, 4, 10, 31, 34]. It has been widely applied to automotive safety, robotics and intelligent video surveillance.

Pedestrian detection is challenging when multiple pedestrians are close in space. Firstly, a single-pedestrian detector tends to combine the visual cues from different pedestrians as the evidence of seeing a pedestrian and thus the detection result will drift. As a result, nearby pedestrian-existing windows with lower detection scores will be eliminated by non-maximum suppression (NMS). For the examples in Fig. 1, single bounding boxes cover multiple pedestrians, which results in inaccurate bounding boxes and missed detections. Secondly, when a pedestrian is occluded by another nearby pedestrian, its detection score may be too low to be detected. Examples are shown in Fig. 1.

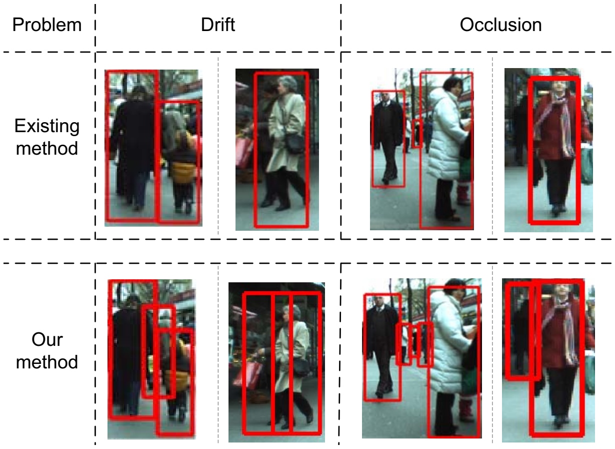


Figure 1. Examples of missed detections caused by drift and occlusion with the state-of-the-art detector in [13]. Aided by a multi-pedestrian detector, the missed pedestrians are detected. The thresholds of both approaches are fixed as 1 False Positive Per Image (FPPI). Best viewed in color.

On the other hand, the existence of multiple nearby pedestrians forms some unique patterns (as shown in Figure 2) which do not appear on isolated pedestrians. They can be used as extra visual cues to refine the detection result of single pedestrians. However, such valuable information was not explored in existing works. The motivations of this paper are two-folds:

1) It is recognized by sociologists that nearby pedestrians walk in groups and show particular spatial patterns [17, 22].

2) From the viewpoint of computer vision, these 3D spatial patterns of nearby pedestrians can be translated into unique 2D visual patterns resulting from the perspective projection of 3D pedestrians to 2D image. These unique 2D visual patterns are easy to detect and are helpful for estimating the configuration of multiple pedestrians.

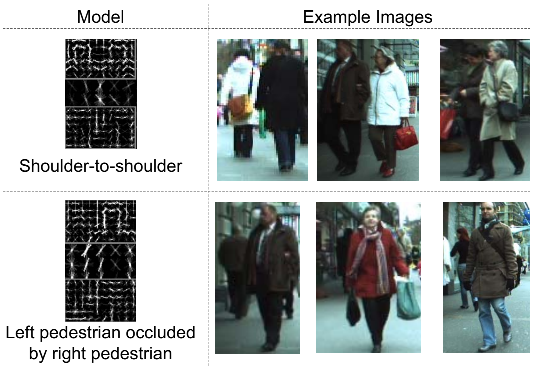


Figure 2. Visual patterns learned from training data with the HOG feature (first column) and examples detected from testing data (remaining columns). In the first row, pedestrians walk side by side. In the second row, pedestrians on the left are occluded by pedestrians on the right. Our 2-pedestrian detector captures visual cues which cannot be learned with a 1-pedestrian detector.

They inspire us to design a multi-pedestrian detector to capture these unique visual patterns. And a multi-pedestrian window found by a multi-pedestrian detector can guide the detection of each pedestrian in this window. Taking the first row in Fig. 2 as an example, when pedestrians walk side by side, they form the shoulder-to-shoulder visual pattern. Taking pedestrians in the second row as another example, the right torso of pedestrians on the left are occluded by the pedestrians on the right. 1-pedestrian detectors are not able to learn these two types of visual patterns. Instead, these visual patterns can be employed by the 2-pedestrian detector. Then the 2-pedestrian detection results are used to reinforce the evidence of detecting each of the two pedestrians.

The contribution of this paper can be summarized in three-fold. 1) A multi-pedestrian detector is learned with a mixture of deformable part-based models to effectively capture the unique visual patterns appearing in multiple nearby pedestrians. The training data is labeled as usual, i.e. a bounding box for each pedestrian. The spatial configuration patterns of multiple nearby pedestrians are learned and clustered into mixture component. 2) In the multi-pedestrian detector, each single pedestrian is specifically designed as a part, called pedestrian-part. As shown in Fig. 4(b), the filter of a pedestrian-part is different from and complementary to a 1-pedestrian detector, since it is learned under a specific multi-pedestrian configuration and under the guidance of the multi-pedestrian detector as contextual constraints. 3) A new probabilistic framework is proposed to model the configuration relationship between results of multi-pedestrian detection and 1-pedestrian detection. With this framework, multi-pedestrian detection results are used to refine 1-pedestrian detection results.

The new framework can easily integrate with any existing 1-pedestrian detector. With a fast computation approach, it only adds small computing load on the top of 1-pedestrian detectors. 15 state-of-the-art 1-pedestrian detectors are evaluated on three widely used public datasets: Caltech, TUD-Brussels and ETH. They all achieve significant improvements by integrating with our framework. The lowest miss rate is improved from 48% to 43% on the Caltech-Test dataset, from 55% to 50% on the TUD-Brussels dataset and from 51% to 41% on the ETH dataset.

2. Related Work

The progress on object detection has been achieved by the investigation on classification approaches, features and articulation handling approaches. 1) Classification approaches used include various boosting classifiers [32, 9, 37], SVM classifiers [4, 21, 13, 43], and grammar models [16] and deep model [23]. 3) Features under investigation include Haar-like features [32], edgelets [37], shapelets [27], histogram of gradients (HOG) [4], bag-of-words [18], integral histograms [26], color histograms [33], covariance descriptors [31], co-occurrence features [28], local binary patterns [34], color-self-similarity [33], depth [12], segmentation [11], features learned from training data [1] and their combinations [34, 9, 33, 28, 11]. 3) Articulation handling approaches under investigation include Deformable part-based models (DPM) [13, 43, 27], pictorial structures [14], poselet [3] and mixture of parts [41].

Context is gaining more and more attention in object detection. The context investigated in previous works includes regions surrounding objects [4, 6, 15], object-scene interaction [7], and the presence, location, orientation and size relationship among objects [2, 38, 39, 5, 25, 15, 29, 7, 42, 6, 40, 24]. They usually employ context cues in two steps: 1) single-object detection results are obtained separately; and 2) the relationship between an object and its context is modeled to refine the detection result. Therefore, the visual cues of seeing multiple objects are from single-object detectors instead of a multi-object detector. The unique visual patterns of multiple nearby pedestrians caused by inter-occlusion and spatial constraint were not explored. In [4, 15, 6], features were extracted from context regions for single-object detection but not multi-object detection. DPM is used [19, 30] to learn contextual cues. The approach in [19] only considers one contextual region with the largest score in an image, even if that image contains multiple people. So it cannot model multiple pairs of pedestrians in an image. In [19] the context cues are used to improve the centered object, but in our work the detections of two pedestrians are jointly estimated under a probabilistic model. A 2-pedestrian detector is also proposed [30]. Our paper is different from [30] in two aspects: 1) the segmentation results of pedestrian is required from the training data in [30] while our paper only requires the bounding box information of pedestrians. 2) the approach in [30] uses NMS to reject the strong overlap between the 2-pedestrian detection results and the 1-pedestrian detection results (in-compatible relationship) while this paper uses a probabilistic framework that favors the strong overlap (compatible relationship).

3. Framework of Single-Object Detection Aided by Multi-Object Detection

Denote an image with $\mathbf{I}$, and let z_1 be the configuration of an object obj_1. $p(\mathbf{I}|z_1)$ is the likelihood of seeing $\mathbf{I}$ given obj_1 with configuration $z_1 = (\mathbf{l}_1, w_1)$. z_1 contains the locations and sizes of the whole object and its parts. $w_1 = (x_1, y_1, s_1)$ is the detection window at location (x_1, y_1) with size s_1. $\mathbf{l}_1$ represents the locations and sizes of parts if the single-object detector is DPM. Object detection needs to compute the posterior distribution, $p(z_1|\mathbf{I})$. Since $p(\mathbf{I})$ is constant, the posterior is represented as $p(z_1|\mathbf{I}) = \frac{p(z_1,\mathbf{I})}{p(\mathbf{I})} \propto p(z_1, \mathbf{I})$ under the Bayes' rule. Our framework assumes multiple nearby pedestrians, and has

$$p(z_1, \mathbf{I}) = p(\mathbf{I}, z_1|c = 1)p(c = 1) + \sum_{c=2}^{C} \sum_{z_c} p(\mathbf{I}, z_1, z_c|c)p(c) \tag{1}$$

where $p(c)$ is the prior of the case when there are c nearby objects. We jointly detect c ($c = 1, \ldots, C$) nearby objects with configuration z_c and capture the visual cues of z_c as the context to assist the estimation of z_1.

3.1. Implementation for pedestrian detection

The framework in (1) is implemented as follows:

$$\begin{aligned} p(z_1, \mathbf{I}) =& p(\mathbf{I}, z_1|c = 1)p(c = 1)+ \\ & \sum_{c=2}^{C} \sum_{z_c} p(\mathbf{I}|z_1, z_c, c)p(z_1, z_c|c)p(c). \end{aligned} \tag{2}$$

$p(\mathbf{I}, z_1|c = 1)$ is estimated from a 1-pedestrian detector. $p(\mathbf{I}|z_1, z_c, c)$ is the likelihood of seeing $\mathbf{I}$ given configurations z_1, z_c and c, and calculated by a c-pedestrian detector introduced in Section 4.2. $p(z_1, z_c|c)$ models the relationship between 1-pedestrian configuration z_1 and c-pedestrian configuration z_c, and is introduced in Section 4.3.

4. Design of the multi-pedestrian detector

The location and size variation of nearby pedestrians results in the appearance variation of these pedestrians. On the other hand, sociologists have found that pedestrians walking together show a few particular spatial patterns [22]. Therefore, we address this problem with a mixture of DPM. We empirically show that such approximation can improve pedestrian detection performance (Section 6).

4.1. Considering at most two pedestrians

This paper focuses on the case when $c = 1$ and $c = 2$ because of several considerations. 1) According to sociological studies [22], the frequency of seeing two pedestrians walking together ($28\% - 42\%$) is much more than that of seeing more than two pedestrians ($< 10\%$). 2) Our approach for 2-pedestrian detector can be naturally extended for c-pedestrian detector. 3) Pair-wise relationship is a concise representation of the relationship among $c(> 2)$ pedestrians. 4) It is computationally expensive when $c > 2$.

When $c = 1$, the $p(\mathbf{I}, z_1|c = 1)$ in (2) is obtained from 1-pedestrian detector. When $c = 2$, we have

$$\sum_{z_2} p(\mathbf{I}, z_1, z_2|c = 2)p(c = 2). \tag{3}$$

The evidence from a 2-pedestrian detection in (3) is used as the extra information to refine the 1-pedestrian detection result in (2). The priors $p(c = 1)$ and $p(c = 2)$ are used as the weights to balance the 1-pedestrian detection result and the evidence from 2-pedestrian detection. These weights are obtained by cross-validation. In our implementation, we have $z_2 = (\mathbf{l}_2, w_2, m_2)$. Since the configurations of two pedestrians are complex, we assume that they are sampled from a mixture model and m_2 is the configuration mixture type. $w_2 = (x, y, s)$ represents the 2-pedestrian detection window at location (x, y) with size s, and $\mathbf{l}_2$ represents the locations and sizes of parts in w_2. In the remaining of this paper, we drop the conditional term $c = 2$ to simplify notations because it is implicitly assumed by $\mathbf{l}_2$, m_2 and w_2. We have the following for (3) by replacing z_2 with $(\mathbf{l}_2, w_2, m_2)$ and then using the sum-product rule:

$$\begin{aligned} &\sum_{z_2} p(\mathbf{I}, z_1, z_2|c = 2)p(c = 2) \\ &= \sum_{\mathbf{l}_2, w_2, m_2} p(\mathbf{I}, z_1, \mathbf{l}_2, w_2, m_2)p(c = 2) \\ &= \sum_{\mathbf{l}_2, w_2, m_2} p(\mathbf{I}, z_1, \mathbf{l}_2|w_2, m_2)p(w_2|m_2)p(m_2)p(c = 2) \\ &= p(c = 2) \sum_{m_2} p(m_2) \sum_{w_2} p(w_2|m_2) \sum_{\mathbf{l}_2} p(\mathbf{I}, z_1, \mathbf{l}_2|w_2, m_2). \end{aligned} \tag{4}$$

Details on the mixture model m_2 and its detection window w_2 are provided in Section 4.2. $p(\mathbf{I}, z_1, \mathbf{l}_2|w_2, m_2)$ in (4) is the joint distribution of image $\mathbf{I}$, configurations z_1 and $\mathbf{l}_2$ given mixture m_2 and window w_2. An overview of this implementation is shown in Fig. 3. The 1-Pedestrian, 2-pedestrian and pedestrian-part detection scores in Fig. 3 are integrated into $p(\mathbf{I}, z_1, \mathbf{l}_2|w_2, m_2)$, which is detailed in Section 4.2. The evidence to 1-pedestrian in Fig. 3 is obtained using (4) and is then added to 1-pedestrian detection results using (2) to obtain the refined detection result in Fig. 3.

4.2. Mixture of DPM for 2-pedestrian detection

In order to learn the mixture type $m_2 = 1, \ldots, M$, the configuration space of z_2 is divided into $M = S \cdot A$ clusters with the following two steps.

1) The two pedestrians form a 2-pedestrian bounding box. The positive training samples are divided into A groups according to their aspect ratios.

2) Each aspect ratio group is further divided into S clusters. The relative location and size between the two pedestrians are used as features for clustering. Many clustering

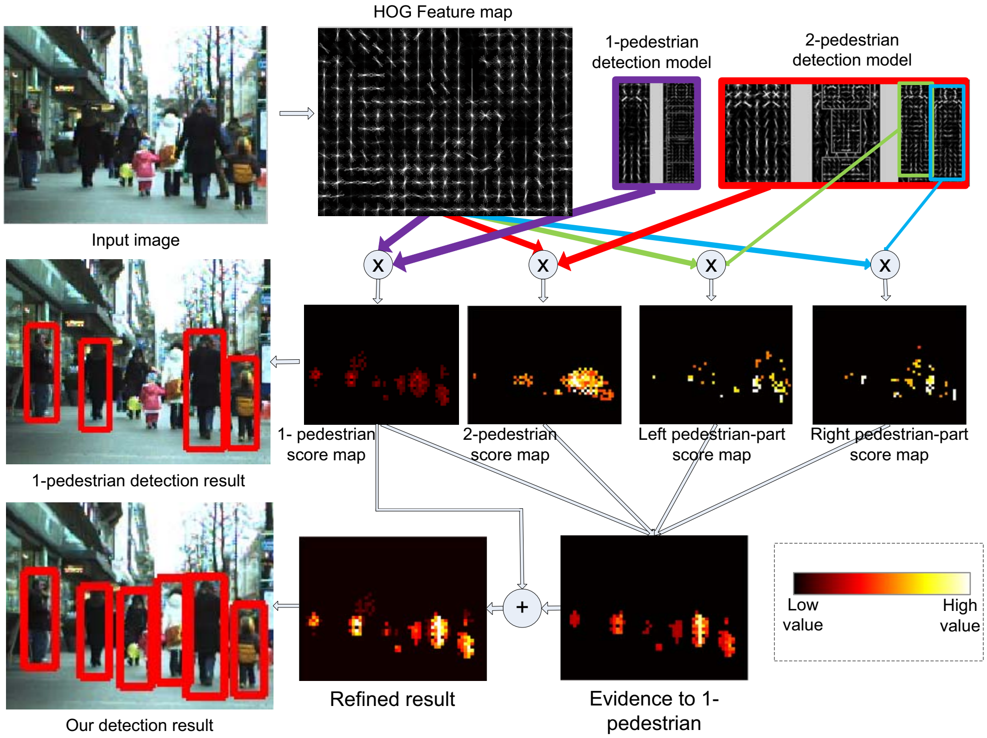


Figure 3. Use 2-pedestrian detection result to refine 1-pedestrian detection. The detection scores of 1-pedestrian, 2-pedestrians and pedestrian-parts are integrated as the evidence to 1-pedestrian configuration z_1. This evidence is added to the result obtained with the 1-pedestrian detector. Examples in the left column are obtained at 1FPPI on the ETH dataset. **This figure is best viewed in color.**

approaches can be adopted. We empirically evaluate the mixture of Gaussian (MoG) in the experiments. Fig. 4(a) shows three examples of the detectors learned for the 9 clusters. It can be seen that each detector captures a specific configuration relationship between the two pedestrians.

After the clustering step, the positive training samples in a cluster and all the negative samples are used to train a DPM [13]. Each cluster corresponds to a mixture type m_2. The 2-pedestrian model for a mixture type m_2 consists of one root filter and five deformable part filters with deformation under the star model learned with the Latent SVM in [13]. The 2-pedestrian bounding box is used to train the root filter. Three parts are greedily selected and initialized from the root filter using the approach in [13]. Besides, we add two extra parts that correspond to the two pedestrians in a 2-pedestrian training sample. They are called pedestrian-parts. The anchor locations and sizes of the two pedestrian-parts are obtained from the mean of the training samples in this cluster. In order to transfer the knowledge of the 1-pedestrian detector to the 2-pedestrian detector, the initial filters for the two pedestrian-parts are obtained from the root filter of the 1-pedestrian detector. With the positive samples and initial part filters defined, the DPM with Latent SVM in [13] is then used to train the 2-pedestrian detector. Examples of the learned model are shown in Fig. 4. The configuration $\mathbf{l}_2$ contains the sizes and locations of parts. Since the pedestrian-parts are explicitly modeled as parts in the 2-pedestrian model, the size and location of each pedestrian in the 2-pedestrian window are also inferred with DPM at the detection stage. This is the key to build the relationship between the 2-pedestrian detection result and the 1-pedestrian detection result.

$p(m_2)$ in (4) could be estimated from the training set. But it could be biased because of insufficient training data. It is assumed to be uniform in our implementation. Given the mixture model m_2, $p(w_2|m_2)$ in (4) can be densely sampled from the image in a sliding window manner with varying window sizes.

To represent the relationship between the pedestrian-part and the single-pedestrian detection result, we introduce a hidden variable h. $h = 0$ when the left pedestrian-part in $\mathbf{l}_2$ is considered to match the single pedestrian with configuration z_1, and $h = 1$ when the right pedestrian-part matches the single pedestrian. With h included, we have the following for the $p(\mathbf{I}, z_1, \mathbf{l}_2|w_2, m_2)$ in (4):

$$
\begin{aligned}
&p(\mathbf{I}, z_1, \mathbf{l}_2|w_2, m_2) = \sum_h p(\mathbf{I}, z_1, \mathbf{l}_2, h|w_2, m_2)\\
&= \sum_h p(\mathbf{I}, z_1, \mathbf{l}_2|h, w_2, m_2)p(h|w_2, m_2),\\
&\text{where} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5)\\
&p(\mathbf{I}, z_1, \mathbf{l}_2|h, w_2, m_2)\\
&= p(\mathbf{I}, \mathbf{l}_1|w_1, \mathbf{l}_2, h, w_2, m_2)p(w_1|\mathbf{l}_2, h, w_2, m_2)p(\mathbf{l}_2|w_2, m_2),
\end{aligned}
$$

$p(w_1, |\mathbf{l}_2, w_2, m_2, h)$ models the relationship between z_1 and z_c and will be detailed in Section 4.3. $z_1 =$

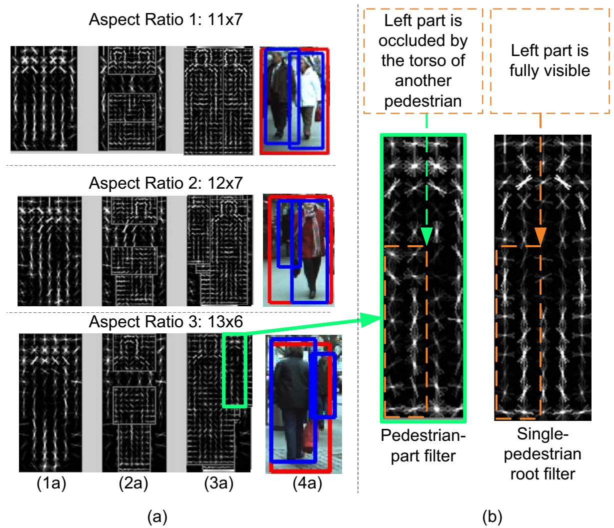

Aspect Ratio 1: 11x7
Aspect Ratio 2: 12x7
Aspect Ratio 3: 13x6
(1a)
(2a)
(3a)
(4a)
(a)
Left part is occluded by the torso of another pedestrian
Left part is fully visible
Pedestrian-part filter
Single-pedestrian root filter
(b)

Figure 4. (a) Examples of 2-Pedestrian detectors learned for different clusters, (b) pedestrian-part filter and the single-pedestrian root filter in [13]. (1a): root filter; (2a): three part filters found from root filter; (3a): pedestrian-part filters; (4a): examples detected by the detectors in the same rows. Red rectangles are 2-pedestrian detection results. Blue rectangles indicate pedestrian-part locations. Best viewed in color.

$(\mathbf{l}_1, w_1)$ and $p(h|w_2, m_2) = 0.5$. The $p(\mathbf{l}_2|w_2, m_2)$ and $p(\mathbf{I}, \mathbf{l}_1|w_1, \mathbf{l}_2, h, w_2, m_2)$ in (5) are implemented as:

$$\begin{aligned}&p(\mathbf{I}, \mathbf{l}_1|w_1, \mathbf{l}_2, h, w_2, m_2)p(\mathbf{l}_2|w_2, m_2)\\ \propto&\phi_a(\mathbf{I}, \mathbf{l}_1; w_1)\phi_b(\mathbf{I}; \mathbf{l}_2, w_2, m_2, h)\phi_c(\mathbf{I}, \mathbf{l}_2; w_2, m_2)\\ =&\lambda_1\lambda_p\lambda_2,\end{aligned} \tag{6}$$

where $\phi_a(\mathbf{I}, \mathbf{l}_1; w_1) = \lambda_1$ is from 1-pedestrian detection score. $\lambda_2 = \phi_c(\mathbf{I}, \mathbf{l}_2; w_2, m_2) = \phi_{c,1}(\mathbf{I}; \mathbf{l}_2, w_2, m_2) \cdot \phi_{c,2}(\mathbf{l}_2; w_2, m_2)$ in (6) is from the 2-pedestrian detection score obtained by DPM in our implementation. $\phi_{c,1}(\mathbf{I}; \mathbf{l}_2, w_2, m_2)$ is the appearance score and $\phi_{c,2}(\mathbf{l}_2; w_2, m_2)$ is the deformation score. The $\phi_b(\mathbf{I}; \mathbf{l}_2, w_2, m_2, h) = \lambda_p$ in (6) is obtained from the pedestrian-part score, which is used as extra information to refine 1-pedestrian detection result. In Fig. 3, the 1-pedestrian score map is from λ_1, the 2-pedestrian score map is from λ_2, and the pedestrian-part score maps are from λ_p.

4.3. Modeling the relationship between 2- and 1-pedestrian detection results

With the pedestrian-parts designed in the 2-pedestrian detector, this relationship becomes matching the pedestrian-part in the 2-pedestrian detector with the 1-pedestrian detection result. It is modeled with $p(w_1|\mathbf{l}_2, w_2, m_2, h)$ in (5), which is a Gaussian distribution:

$$p(w_1|\mathbf{l}_2, w_2, m_2, h) = (2\pi)^{-\frac{3}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(w_1-u)^T\Sigma^{-1}(w_1-u)}, \tag{7}$$

where Σ is the covariance matrix estimated from training samples for each mixture m_2, $w_1 = (x_1, y_1, s_1)$ is the location and size of z_1, $u = (x_{2,h}, y_{2,h}, s_{2,h})$ is the location and size of the pedestrian-part h in $\mathbf{l}_2$. $p(w_1|\mathbf{l}_2, w_2, m_2, h)$ is the largest if the 1-pedestrian detection window w_1 perfectly matches the pedestrian-part.

5. Reduction of computational complexity

Suppose the number of possible configurations for w_1 in $z_1 = (w_1, \mathbf{l}_1)$ is L_c. The number of possible configurations for the 5 parts in $\mathbf{l}_2$ is $O(L_c^5)$ and the number of possible configurations for w_2 is $O(L_c)$. The number of possible configurations for m_2 is M. Overall, the computational complexity of (4) is $O(ML_c^7)$, which is unaffordable and a fast approach is required.

In order to reduce the computational complexity, we have the following approximation for (4):

$$\begin{aligned}&\sum_{\mathbf{l}_2, w_2, m_2} p(\mathbf{I}, z_1, \mathbf{l}_2, w_2, m_2)\\ \approx&\sum_{h, w_2, m_2} p(w_2, m_2)\phi_a(\mathbf{I}, \mathbf{l}_1; w_1)\phi_b(\mathbf{I}; \tilde{\mathbf{l}}_2, w_2, m_2, h)\\ &\qquad\phi_c(\mathbf{I}, \tilde{\mathbf{l}}_2; w_2, m_2)p(w_1|\tilde{\mathbf{l}}_2, w_2, m_2, h)p(h)\\ =&\sum_{h, w_2, m_2} p(w_2, m_2)\lambda_1\tilde{\lambda}_p\tilde{\lambda}_2 p(w_1|\tilde{\mathbf{l}}_2, w_2, m_2, h)p(h),\end{aligned} \tag{8}$$

where $\tilde{\mathbf{l}}_2 = \operatorname*{argmax}_{\mathbf{l}_2} \phi_c(\mathbf{I}, \mathbf{l}_2; w_2, m_2)$.

In this way, the summation with regard to $\mathbf{l}_2$ in (8) is approximated by maximization, which can be efficiently computed with the distance transform in [13]. Denote the number of candidates for z_1 by $Cand_1$, and the number of candidate windows w_2 for M mixtures by $Cand_2$. The procedure and computational complexity of computing (8) is as follows:

Step 1. Obtain the 1-pedestrian detection result, which is used for $p(\mathbf{I}, z_1|c = 1)$ in (2) and λ_1 in (8). Only $Cand_1$ candidate windows, which are detected by the single-pedestrian detector, are used for the next steps. $O(L_c)$ operations are required.

Step 2. Obtain the 2-pedestrian detection results, which is used for $\tilde{\lambda}_p$ and $\tilde{\lambda}_2$. Since there should not be any 2-pedestrian window in which no pedestrian is found, the 2-pedestrian detector can be evaluated only around $Cand_1$ 1-pedestrian candidate windows to save computation (i.e. we assume that if two nearby pedestrians exist, at least one pedestrian will be detected by the single-pedestrian detector around this region). $O(Cand_1)$ operations are required.

Step 3. For each 1-pedestrian candidate z_1, compute (8) for $Cand_2$ 2-pedestrian candidate windows using the results obtained in Step 1 and Step 2. In practice, most λ_1 and $\tilde{\lambda}_2$ are very close to 0, i.e. $Cand_1, Cand_2 \ll L_c$. This allows us to compute $p(w_1|\tilde{\mathbf{l}}_2, w_2, m_2, h)$ only for $Cand_1Cand_2$ non-zero λ_1 and $\tilde{\lambda}_2$. With the terms computed, the computational complexity for summing up them w.r.t. h, w_2 and m_2 in (8) is $O(Cand_1Cand_2)$ by enforcing sparsity on 1-pedestrian and 2-pedestrian candidate windows. $O(Cand_1 Cand_2)$ operations are required.

Take our experiment on the Caltech dataset [10] as an example, we have $L_c > 40,000$, $Cand_2 = 20$, $Cand_1 = 140$ and $Cand_1 Cand_2 = 2,800$ per image on average. Therefore, the computation required for Step 2 and Step 3, i.e. $O(Cand_1) + O(Cand_1 Cand_2)$, is relatively small compared with the computation required for singe-pedestrian detection in Step 1, i.e. $O(L_c)$.

6. Experimental Results

The proposed framework is evaluated on three public datasets: Caltech [10], TUD-Brussels [36] and ETH [12]. We use the modified HOG [13] as feature and the DPM in [13] to learn the 2-pedestrian detector. HOG+DPM is used because it is off-the-shelf, open-source, and widely used. Since the detection scores of multi-pedestrian detector and 1-pedestrian detector are considered as input, the framework keeps unchanged if other detection models or features are used for 1-pedestrian detector or 2-pedestrian detector. Existing pedestrian detection results can be directly used as the input of our framework.

The 1-pedestrian detection approach in [13] used the same feature and DPM as our 2-pedestrian detector. It is denoted as LatSVM-V2 in the experimental results. Our framework using LatSVM-V2 as the 1-pedestrian detector is denoted as LatSVM-V2+Our in the experimental results. Other single-pedestrian detectors trained with different models, features and datasets are also integrated with our 2-pedestrian detector and compared in Section 6.2.

The labels and evaluation code provided by Dollár *et al.* online are used for evaluation following the criteria proposed in [10]. As in [10], the *log-average miss rate* is used to summarize the detector performance, and is computed by averaging the miss rate at nine FPPI rates evenly spaced in the log-space in the range from 10^{-2} to 10^{0}. In the experiments, we evaluate the performance on the *reasonable* subset of the evaluated datasets, which is the most popular portion of the datasets. It consists of pedestrians of ≥ 50 pixels in height, who are fully visible or less than 35% occluded.

6.1. Preparation of 2-Pedestrian Training Data

Since there is no 2-pedestrian detection training dataset, we construct it based on the INRIA training dataset [4] as follows:

1) All the negative images are used for negative samples.

2) Because most pedestrians labeled in INRIA are isolated pedestrians, this results in a very small number of 2-pedestrian positive samples (656). We labeled more pedestrians in the positive images. The number of positive 2-pedestrians increases from the original 656 to 4398. [1].

3) If the bounding boxes of two pedestrians have overlap, the bounding box that exactly covers the two pedestrians is considered as the label of the 2-pedestrian positive sample.

Once the 2-pedestrian detection model is learned from this training set, it is fixed and tested on other datasets.

6.2. Experimental Results on Caltech, TUD-Brussels and ETH

All the state-of-the-art approaches evaluated on the TUD-Brussels and EHTZ dataset in [10] are evaluated in this experiment. First of all, we compare with the approach in [13] which used the same feature and learning model as our 2-pedestrian detector. Compared with LatSVM-V2, our approach has 10%, 7% and 5% log-average miss rate improvement on the datasets ETH, TUD-Brussels and Caltech-Test respectively. In order to exclude the factor of using a larger training set, we also train the 1-pedestrian detector with DPM 6.1 on our extended INRIA dataset described in Section 6.1. It is denoted by LatSvm-V2-E. By combining with LatSVM-V2-E, our approach (LatSvm-V2-E+our) has 9%, 7% and 5% log-average miss rate improvement over LatSVM-V2-E on the datasets ETH, TUD-Brussels and Caltech-Test respectively.

We also investigate other 1-pedestrian detectors and integrate them with our 2-pedestrian detector in this experiment. The evaluated 1-pedestrian detectors are VJ [32], Shapelet [27], PoseInv [20], LatSVM-V1 [13], HikSVM [21], HOG [4], MultiFtr [35], HogLbp [34], Pls [28], MultiFtr+CCS, MultiFtr+Motion [33], FPDW [8], ChnFtrs [9], and MultiResC [25]. MultiResC is only evaluated on the Caltech-Test dataset, since its results on ETH and TUD-Brussels is not available. For 1-pedestrian detection results, the range of detection score s has large variation for different approaches. s is normalized to s_{norm} as following:

$$s_{norm} = \sigma(a * s + b), a = 6/s_{max}, b = -0.6a. \quad (9)$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function, s_{max} is the maximum detection score of the first 100 images for each approach. s_{norm} is used as $p(\mathbf{I}, z_1|c = 1)$ in (2). Fig. 5 shows the results on the three datasets. Fig. 6 shows the improvement of our framework for each of these approaches on the two datasets. Our framework significantly improves all the state-of-the-art pedestrian detectors by integrating with them. The average improvement is about 9% on the Caltech-Test dataset, 11% on the TUD-Brussels dataset and 17% on the ETH dataset. It is reported in [10] that LatSvm-V2 has the best performance among the 14 state-of-the-art approaches evaluated on the ETH dataset. The averaeg miss rate for LatSvm-V2 is 51%. By integrating with our framework, 10 algorithms outperform LatSVM-V2 and

[1] http://www.ee.cuhk.edu.hk/~xgwang/2ped.html

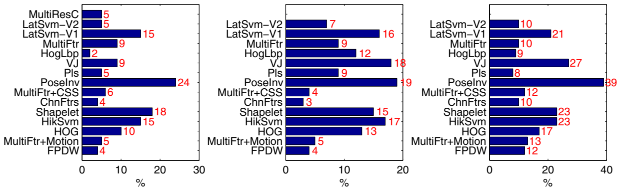


Figure 6. Miss rate improvement of the framework for each of the state-of-the-art 1-pedestrian detectors on Caltech-Test (left), TUD-Brussels (middle) and ETH (right). X-axis denotes the miss rate improvement.

the best performing one (LatSVM-V2+Our) reaches the average miss rate of 41%. The current best performing approaches on the Caltech-Test dataset is the MultiResC and the contextual boost in [6], both of which use context information and have an average miss rate 48%. With our framework, MultiResC+Our is improved to 43%. With our framework, the current best performing approach on the TUD-Brussels dataset, i.e. MultiFtr+Motion, is improved from 55% to 50%. This experiment shows that the multi-pedestrian detector provides rich complementary information to current state-of-the-art 1-pedestrian detection approaches even when context [25] or motion [33] is used by these approaches.

7. Conclusion

In this paper, we propose a new probabilistic framework for single pedestrian detection aided by multi-pedestrian detection. DPM is used to learn the multi-pedestrian detector which effectively captures the unique visual patterns appearing in multiple nearby pedestrians. Detection performance is improved by modeling the relationship between the configurations of single-pedestrian detection results and those of multi-pedestrian detection results. It is very flexible to incorporate with new features (e.g. color self-similarity, local binary pattern, motion and depth), other deformable part-based models (e.g. the tree and loopy models), and learning methods (e.g. boosting). Existing pedestrian detection results can be directly used as the input of our framework. Extensive experimental evaluation shows that the proposed framework can significantly improve all the state-of-the-art single-pedestrian detection approaches, and that the multi-pedestrian detector provides rich complementary information to current state-of-the-art single-pedestrian detection approaches, even if motion or context is used by these approaches. Over the 15 state-of-the-art approaches under investigation, the average improvement is 9% on the Caltech-Test dataset, 11% on the TUD-Brussels dataset and 17% on the ETH dataset. The lowest miss rate is reduced from 48% to 43% on the Caltech-Test dataset, from 55% to 50% on the TUD-Brussels dataset and from 51% to 41% on the ETH dataset.

Acknowledgment: This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project No. CUHK 417110 and CUHK 417011), National Natural Science Foundation of China (Project No. 61005057), and Guangdong Innovative Research Team Program (No.201001D0104648280).

References

[1] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg. Part-based feature synthesis for human detection. In *ECCV*, 2010. 2

[2] O. Barinova, V. Lempitsky, and P. Kohli. On detection of multiple object instances using hough transforms. In *CVPR*, 2010. 1, 2

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

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 2, 6, 8

[5] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009. 2

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

[7] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009. 2

[8] P. Dollár, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *BMVC*, 2010. 6

[9] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009. 2, 6

[10] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: an evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743 – 761, 2012. 1, 6

[11] M. Enzweiler and D. M. Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. *IEEE Trans. Image Process.*, 20(10):2967–2979, 2011. 2

[12] A. Ess, B. Leibe, and L. V. Gool. Depth and appearance for mobile scene analysis. In *ICCV*, 2007. 2, 6

[13] P. Felzenszwalb, R. B. Grishick, D.McAllister, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:1627–1645, 2010. 1, 2, 4, 5, 6

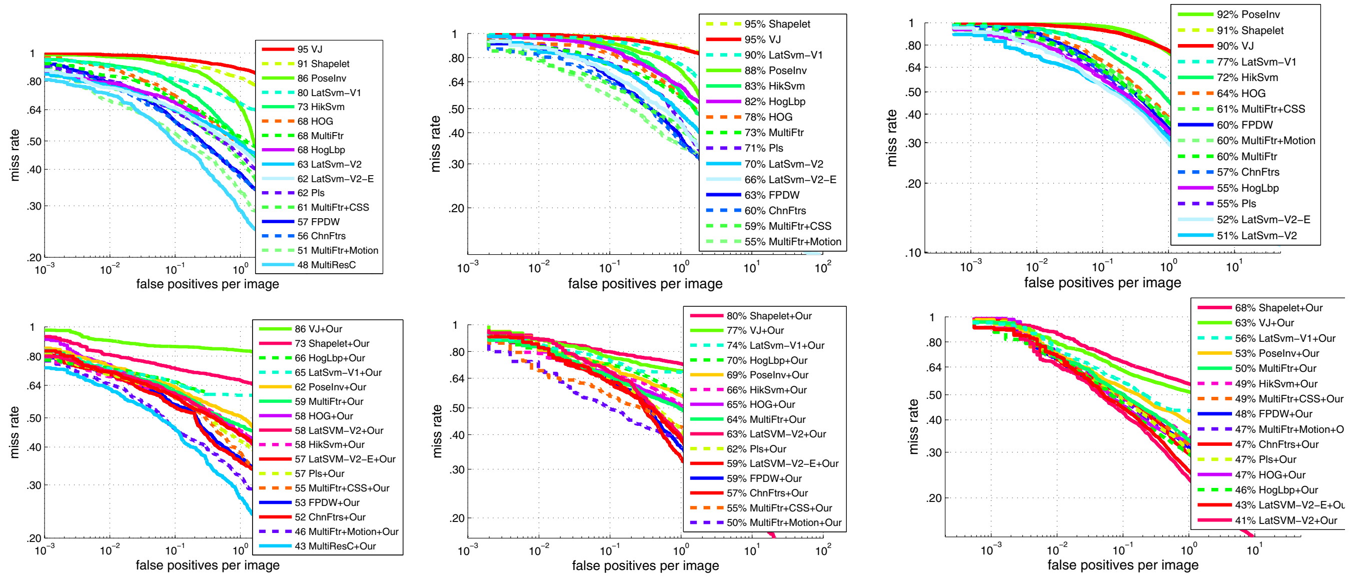
[14] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *Int'l J. Computer Vision*, 61:55–79, 2005. 2

[15] C. Galleguillosy, B. McFeey, S. Belongiey, and G. Lanckriet. Multi-class object localization by combining local contextual interactions. In *CVPR*, 2010. 2

[16] R. Girshick, P. Felzenszwalb, and D. McAllester. Object detection with grammar models. In *NIPS*, 2011. 2

[17] A. Hare. *Handbook of small group research*. Macmillan, 1962. 1

[18] C. Lampert, M. Blaschko, and T. Hofmann. Beyond sliding windows: object localization by efficient subwindow search. In *CVPR*, 2008. 2



(a) Caltech-Test (b) TUD-Brussels (c) ETH

Figure 5. Detection results of existing approaches (top) and integrating them with our framework (bottom) on the datasets Caltech-Test (a), TUD-Brussels (b) and ETH (c). The results of integrating existing approaches with our framework are denoted by '+Our'. For example, the result of integrating HOG [4] with our framework is denoted by HOG+Our.

[19] C. Li, D. Parikh, and T. Chen. Extracting adaptive contextual cues from unlabeled regions. In *ICCV*, pages 511–518. IEEE, 2011. 2

[20] Z. Lin and L. Davis. A pose-invariant descriptor for human detection and segmentation. In *ECCV*, 2008. 6

[21] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008. 2, 6

[22] M. Moussaid, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PLoS ONE*, 5(4):e10047, 2010. 1, 3

[23] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, 2012. 2

[24] W. Ouyang, X. Zeng, and X. Wang. Modeling mutual visibility relationship in pedestrian detection. In *CVPR*, 2013. 2

[25] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *ECCV*, 2010. 2, 6, 7

[26] F. Porikli. Integral histogram: a fast way to extract histograms in cartesian spaces. In *CVPR*, 2005. 2

[27] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. In *CVPR*, 2007. 2, 6

[28] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis. Human detection using partial least squares analysis. In *ICCV*, 2009. 2, 6

[29] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *CVPR*, 2011. 2

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

[31] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(10):1713–1727, Oct. 2008. 1, 2

[32] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *Int'l J. Computer Vision*, 63(2):153–161, 2005. 2, 6

[33] S. Walk, N. Majer, K. Schindler, and B. Schiele. New features and insights for pedestrian detection. In *CVPR*, 2010. 2, 6, 7

[34] X. Wang, X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *CVPR*, 2009. 1, 2, 6

[35] C. Wojek and B. Schiele. A performance evaluation of single and multi-feature people detection. In *DAGM*, 2008. 6

[36] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *CVPR*, 2009. 6

[37] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, 2005. 2

[38] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *Int'l J. Computer Vision*, 75(2):247–266, 2007. 2

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

[40] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *CVPR*, 2012. 2

[41] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 2

[42] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 2

[43] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, 2010. 2