Single-Pedestrian Detection Aided by Two-Pedestrian Detection

Wanli Ouyang, *Member, IEEE*, Xingyu Zeng, and Xiaogang Wang, *Member, IEEE*

Abstract—In this paper, we address the challenging problem of detecting pedestrians who appear in groups. A new approach is proposed for single-pedestrian detection aided by two-pedestrian detection. A mixture model of two-pedestrian detectors is designed to capture the unique visual cues which are formed by nearby 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 two-pedestrian detectors, and to refine the single-pedestrian detection result using two-pedestrian detection. The two-pedestrian detector can integrate with any single-pedestrian detector. Twenty-five state-of-the-art single-pedestrian detection approaches are combined with the two-pedestrian detector on three widely used public datasets: Caltech, TUD-Brussels, and ETH. Experimental results show that our framework improves all these approaches. The average improvement is 9 percent on the Caltech-Test dataset, 11 percent on the TUD-Brussels dataset and 17 percent on the ETH dataset in terms of average miss rate. The lowest average miss rate is reduced from 37 to 32 percent on the Caltech-Test dataset, from 55 to 50 percent on the TUD-Brussels dataset and from 43 to 38 percent on the ETH dataset.

Index Terms—Part based model, discriminative model, pedestrian detection, object detection, human detection, contextual information

1 Introduction

Object detection is one of the central problems in computer vision. Pedestrian detection is one of the most important topics in object detection and has attracted much attention [3], [10], [22], [76], [83]. For research, pedestrian detection incorporates most of the challenges characterizing object detection—illumination change, nonrigid deformation, viewpoint change, occlusion, and intraclass appearance variability. For application, pedestrian detection has many practical applications such as automotive safety, robotics, content based image retrieval, assistive technology for the visually impaired, advanced human-computer interface, and video surveillance. Therefore, there is considerable interest in building automated vision systems for detecting pedestrians [35].

Early approaches used Haar wavelets with polynomial SVM [61], hierarchical Chamfer matching [33] and Haar-like features with AdaBoost [79]. In the recent years, there has been a surge of interest in pedestrian detection [10], [16], [20], [21], [23], [30], [46], [49], [62], [67], [69], [76], [80], [83], [85], [91]. The spectacular progress in object detection and pedestrian detection has been achieved by new classification approaches, features, deformation models, fast algorithms and datasets.

- The investigated classification approaches include various boosting classifiers [21], [78], [87], linear SVM [10], [25], [82], histogram intersection kernel SVM [49], latent SVM [30], confidence-encoded SVM [81], probabilistic models [3], [50], multiple kernel SVM [77], structural SVM [98], grammar models [36] and deep models [48], [56], [57], [58], [59], [60], [70], [96], [97].
- Features under investigation include Haar-like features [79], edgelets [87], shapelet [67], histogram of gradients (HOG) [10], dense SIFT [77], bag-of-words [43], integral histogram [66], color histogram [80], gradient histogram [99], covariance descriptor [76], co-occurrence features [69], local binary pattern [83], color-self-similarity [80], depth [27], [28], segmentation [23], [27], motion [11], [27], features learned from training data [2], [54] and their combinations [21], [23], [24], [27], [40], [43], [69], [77], [80], [83].
- In recent years, models handling deformation [30], [31], [45], [94], [98] and appearance variation of parts [7], [8], [94] achieved great success on object detection.
- Fast object detection algorithms often gain high interests [4], [12], [19], [40], [90]. Fast pedestrian detection algorithms can achieve frame-rate detection speed [19] and 100 frames per second with GPU acceleration and geometric constraint [4].
- For object detection, the datasets in PASCAL Visual Object Classes (VOC) Challenge [29] and the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) are the most relevant ones [13], [39]. For pedestrian detection, datasets such as MIT [61], INRIA [10], ETH [28], TUD-Brussels [86], Caltech [22], and KITTI [34] are designed with increasing difficulty and pedestrian information such as motion, stereo, and occluded region. Surveys and performance evaluations on recent pedestrian detection approaches are provided in [22], [26], [35], [52], [85].

- *The authors are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong. E-mail: {wlouyang, xyzeng, xgwang}@ee.cuhk.edu.hk.*

Manuscript received 29 Nov. 2013; revised 4 Oct. 2014; accepted 24 Nov. 2014. Date of publication 18 Dec. 2014; date of current version 7 Aug. 2015. Recommended for acceptance by D. Ramanan.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2014.2377734

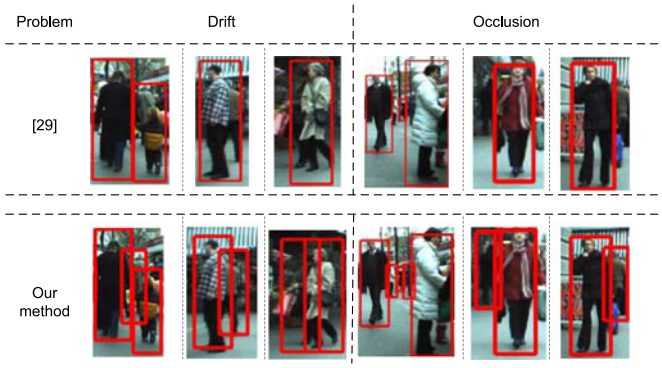


Fig. 1. Examples of missed detections caused by drift and occlusion with the state-of-the-art detector in [30]. Aided by a two-pedestrian detector, the missed pedestrians are detected. The thresholds of both approaches are fixed at one False Positive Per Image (FPPI). Best viewed in color.

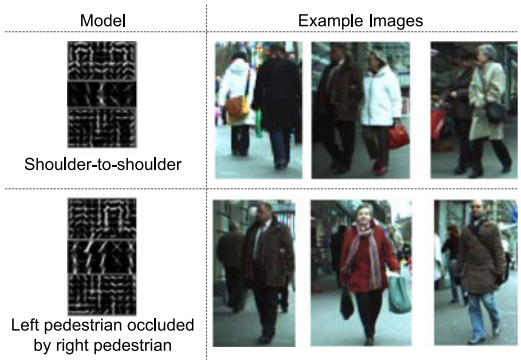


Fig. 2. Visual patterns learned from training data with the HOG feature (first column) and examples detected from testing data (remaining columns). In the first row, pedestrians walk side by side. In the second row, pedestrians on the left are occluded by pedestrians on the right. Our two-pedestrian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

Pedestrian detection is challenging when multiple pedestrians are close in space. First, 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. Second, 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.

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

1) Sociologists find that nearby pedestrians walk in groups and show particular spatial patterns [37], [51].
2) From a computer vision viewpoint, 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 images. These unique 2D visual patterns are helpful for estimating the configuration of multiple pedestrians.

They inspire us to design a two-pedestrian detector to capture these unique visual patterns. A two-pedestrian window found by a two-pedestrian detector can then 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. One-pedestrian detectors are not able to learn these two types of visual patterns. Instead, these visual patterns can be employed by the two-pedestrian detector. Then the two-pedestrian detection results are used to reinforce the detection probabilities for the two pedestrians. In Fig. 2, when the probabilities of the two pedestrians increase, they will not be suppressed by NMS in the drift example and will be found in the occlusion example.

The contributions of this paper are as follows:

1) A two-pedestrian detector is learned to effectively capture the unique visual patterns appearing in nearby pedestrians. The training data is labeled as usual, i.e., a bounding box for each pedestrian. The spatial configuration patterns of nearby pedestrians are learned and clustered into different appearance patterns. In the two-pedestrian detector, each single pedestrian is specifically designed as a part, called pedestrian-part. As shown in Fig. 9, the filter of a pedestrian-part is different from and complementary to a one-pedestrian detector, since it is learned under a specific two-pedestrian configuration and under the guidance of the two-pedestrian detector as contextual constraints.
2) A new probabilistic framework is proposed to model the configuration relationship between results of two-pedestrian detection and one-pedestrian detection. With this framework, two-pedestrian detection results are used to refine one-pedestrian detection results.
3) A way of reducing the unaffordable computational complexity of the probabilistic framework to acceptable computational complexity.

The new framework can easily integrate with any existing one-pedestrian detector. With a fast computation approach, it only adds small computing load on the top of one-pedestrian detectors. Twenty-five state-of-the-art one-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 37 to 32 percent on the Caltech-Test dataset, from 55 to 50 percent on the TUD-Brussels dataset and from 43 to 38 percent on the ETH dataset.

2 Related Work

Context is gaining more and more attention in object detection. Researches on visual cognition, computer vision and cognitive neuroscience have shown that the ability in recognizing objects is affected by the contextual information like

non-target objects, object size and location, consistency in object and contextual scenes. A review of context in object recognition is provided in [55]. The context investigated in previous works includes regions surrounding objects [10], [16], [32], object-scene interaction [17], and the presence, location, orientation and size relationship among objects [3], [14], [15], [16], [17], [32], [60], [62], [68], [72], [74], [88], [91], [93], [95]. 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. These approaches can be considered as two categories:

1. Assemble detected parts into detected objects. Multiple objects may inter-occlude one another. By considering objects as consisting of parts, many approaches consider multiple objects detection as an assembly problem [3], [41], [42], [88], [89], [91]. These approaches are based on the observation that one part, e.g. head-shoulder, only belongs to one instance, e.g., human. The joint part-combination of multiple objects is adopted in [3], [41], [42], [43], [47], [71], [87], [89]. In these approaches, the existence of an object in a image suppresses the existence of an object nearby. Therefore, these approaches are not able to model the following observation: the evidence of seeing an object should be able to amplify the confidence of seeing another object nearby. To take this observation into account, recently Yan et al. [91] include the context of objects in assembling multiple pedestrian detection results. In their approach, the context is the location and size difference between two pedestrians.
2. Refine the detection of one object by the detection of other objects. In this category, there are two subcategories.

In the first sub-category, the descriptions of other objects, e.g., cars or bicycles, in other windows are used as context features to help classify the existence of specific object class, e.g. person, in a specific window [14], [15], [16], [17], [32], [62], [72], [95]. Divvala et al. [17] did an empirical study of contexts such as object presence, location and spatial support. These contexts are used as features and concatenated with appearance features for training a logistic regression classifier. These contexts improves the average precision from 22.4 percent to 23.9 on Pascal VOC 2008. Desai et al. [15] grouped the spatial layout of other objects into a 7-dimensional features for each class. The seven-dimensional features are near, far, above, on-top, next-to, overlap, below. The spatial layout contexts improves the average precision from 26 to 27.2 percent on Pascal VOC 2007. Desai and Ramanan [14] models the spatial relationship between human pose and interacting objects and use them for detecting actions, poses, and objects. Yao and Fei-fei [95] studied the strong contextual information between human and the objects in sports, e.g., cricket bat, cricket ball. They consider the location, size and orientation as the context for detecting objects and estimating human pose. In their approach, objects are involved in human-object interaction and human activity class is further used for object detection and human pose estimation. Song et al. [72] considered the context as the existence probability of certain object class in each image. For each class, the existence probability is measured by classification score for object classification and measured by highest detection score for object detection in [72]. The existence probability contexts improves the average detection precision from 34.5 to 36.8 percent on Pascal VOC 2010. Galleguillos et al. [32] considered different levels of context such as pixel, region and object. In this approach, the pixel and region context is combined with appearance features for single object detection and then object level interaction is combined using a conditional random field. With the segmentation information and multi-kernel large margin nearest neighbor approach, the contextual information improves the average precision from 26 to 33 percent on Pascal VOC 2007. Ding and Xiao considered the detection score of windows surrounding current window as the context feature in [16].

In the second sub-category, geometric constraints are used for pedestrians and cars [38], [62]. Large improvement is observed using geometric constraints. Objects of the same class are assumed to have similar height. The context is considered as the difference between the current window size and the window size estimated from scene geometry in [38], [62]. The estimated window size is obtained by putting objects in perspective view and assuming that all objects of interest rest on the ground plane.

In these approaches, the visual cue of multiple overlapping objects is captured by separate single object detectors. In these approaches, the visual cue of seeing multiple objects is from the single object detectors. The global visual cue of multiple nearby objects caused by inter-occlusion and spatial constraint is not explored. In this paper, we will explore the unique visual cue of multiple pedestrians by designing a multi-pedestrian detector.

Deformable part based model (DPM) is used in [44], [65], [68], [74], [75] to learn contextual cues. The approach in [44] only considers one contextual region with the largest score in an image, even if that image contains multiple persons. So it cannot model multiple pairs of pedestrians in an image. Similarly, the visual phrases in [68] cannot model multiple pairs of pedestrians in a spatial region.

The most closely related works are the approaches in [65], [74], [75], which are the only other two-pedestrian detectors. The approach in [65] learns both single-object occlusion patterns and double-object occlusion patterns using the 3D object information and labeled occlusion patterns in the training set. Different from our approach, the approach in [65] directly used a double-object detector to detect a pair of neighboring objects or a single object, but did not integrate its result with a single-object detector, which is a key contribution of our paper. 3D information is required for learning the double-object detector in [65] while only 2D bounding box information is required by our approach. Two-pedestrian detectors trained by DPM are proposed in [74], [75] for detecting inter-occluded pedestrians. Our work is different from [74], [75] in five aspects: 1) the manually labeled segmentation maps of pedestrians are required in the training data in [74], [75] in order to infer occlusion patterns, while our

work only requires the bounding box information of pedestrians. 2) The pair of pedestrians in the dataset investigated in [74], [75] have small variation in window sizes and 3D spatial locations while we consider more general and natural cases in which the pair of pedestrians have large variation in window sizes and 3D spatial locations. 3) The approach in [74], [75] use NMS to reject the strong overlap between the two-pedestrian detection results and the one-pedestrian detection results (incompatible relationship) while this paper uses a probabilistic framework that favors the strong overlap (compatible relationship). 4) The experiment is done on more specific datasets in [74], [75] while our experimental results are done on more general pedestrian detection datasets that complement the ones in [74], [75]. 5) The experiment in [74], [75] only shows the effectiveness for a one-pedestrian detection approach while our experiment shows the effectiveness for 25 one-pedestrian detection approaches.

3 Framework Overview

Denote an image by $\mathbf{I}$, and let $\mathbf{z}_1$ be the configuration of an object denoted obj_1. $p(\mathbf{I}|\mathbf{z}_1)$ is the likelihood of seeing $\mathbf{I}$ given obj_1 with configuration $\mathbf{z}_1 = (\mathbf{l}_1, w_1)$. $\mathbf{z}_1$ is a vector that contains the locations, orientations and scales 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. $\mathbf{l}_1$ is a vector that represents the locations and sizes of parts if the single-object detector is the DPM in [30]. Object detection needs to compute the posterior distribution, $p(\mathbf{z}_1|\mathbf{I})$. Since $p(\mathbf{I})$ is assumed to be constant, the posterior is represented as $p(\mathbf{z}_1|\mathbf{I}) = \frac{p(\mathbf{z}_1,\mathbf{I})}{p(\mathbf{I})} \propto p(\mathbf{z}_1, \mathbf{I})$ under the Bayes' rule. Considering multiple nearby pedestrians, we have

$$p(\mathbf{z}_1, \mathbf{I}) = p(\mathbf{I}, \mathbf{z}_1|c=1)p(c=1) + \sum_{c=2}^{C} \sum_{\mathbf{z}_c \in \mathbb{Z}^c} p(\mathbf{I}, \mathbf{z}_1, \mathbf{z}_c|c)p(c), \quad (1)$$

where $p(c)$ is the prior of the case when there are c nearby objects, $\mathbf{z}_c$ denotes the configuration of c nearby objects, $\mathbb{Z}^c$ denotes the set of all configurations for $\mathbf{z}_c$. c $(c = 1, \ldots, C)$ nearby objects are considered and the visual cues of $\mathbf{z}_c$ are captured as the context to assist the estimation of $\mathbf{z}_1$.

An overview of our implementation is shown in Fig. 3. In our implementation, $p(\mathbf{I}, \mathbf{z}_1|c=1)$ is estimated from a one-pedestrian detector. $p(\mathbf{I}|\mathbf{z}_1, \mathbf{z}_c, c)$, which is detailed in Section 4.2, is the likelihood of seeing $\mathbf{I}$ given the configurations $\mathbf{z}_1, \mathbf{z}_c$, and c. $p(\mathbf{z}_1, \mathbf{z}_c|c)$, which is introduced in Section 4.3, models the joint probability of the one-pedestrian configuration $\mathbf{z}_1$ and the c-pedestrian configuration $\mathbf{z}_c$.

4 Design of the Two-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 [51]. Therefore, we handle the appearance variation using a mixture of DPM. We empirically show that such approximation is reasonable (Fig. 4) and can improve pedestrian detection performance (Section 6).

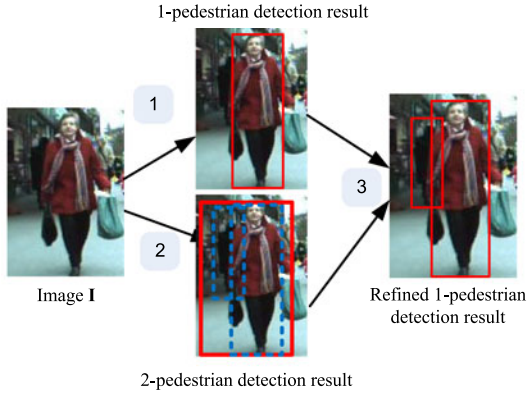


Fig. 3. Overview of our implementation of the framework introduced in Eq. (1).

4.1 Considering at Most Two Pedestrians

This paper focuses on the case when $c = 1$ and $c = 2$ because of several considerations. 1) The frequency of two pedestrians overlapping with each other (about 60 percent on the ETHZ dataset, 50 percent on the TUD-Brussels dataset, and 30 percent on the Caltech Testing dataset) is much larger than the frequency of more than two pedestrians (<6% on these three datasets). 2) Our approach with two-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 $C = 2$, the $p(\mathbf{z}_1, \mathbf{I})$ in (1) is as follows:

$$p(\mathbf{I}, \mathbf{z}_1|c=1)p(c=1) + \sum_{\mathbf{z}_2 \in \mathbb{Z}^2} p(\mathbf{I}, \mathbf{z}_1, \mathbf{z}_2|c=2)p(c=2), \quad (2)$$

where $\mathbb{Z}^2$ denotes the set of all configurations for $\mathbf{z}_2$. The $p(\mathbf{I}, \mathbf{z}_1|c=1)$ in (2) is obtained from a 1-pedestrian detector. The second term in (2) is the evidence from a two-pedestrian detector, which is used as the extra information to refine the one-pedestrian detection result. The priors $p(c = 1)$ and $p(c = 2)$ in (2) are used as the weights to balance the one-pedestrian detection result and the evidence from two-pedestrian detection. These weights are obtained by cross-validation. In our implementation, we have $\mathbf{z}_2 = (\mathbf{l}_2, w_2, m_2)$. Since the configurations of two pedestrians are complex, we assume that they are sampled from a mixture model and m_2 is the mixture type of configuration $\mathbf{z}_2$. Details on the mixture model m_2 and its detection window w_2 are provided in Section 4.2. $w_2 = (x, y, s)$ represents the two-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 the second term in (2) by taking out $p(c = 2)$, replacing $\mathbf{z}_2$ with $(\mathbf{l}_2, w_2, m_2)$,

Fig. 4. The configurations of two-pedestrian samples from the INRIA dataset together with four sample images. In each sample, the left pedestrian is considered as the anchor. *X-axis*: the horizontal distance between the two pedestrians divided by the width of the left bounding box. *Y-axis*: the size of the right pedestrian divided by the size of the left pedestrian in log scale. Samples are not uniformly distributed in the configuration space. A single detector cannot handle the large appearance variation. It is reasonable to cluster these samples to train a mixture model.

and then using the sum-product rule:

$$\begin{aligned}
&\sum_{\mathbf{z}_2 \in \mathbb{Z}^2} p(\mathbf{I}, \mathbf{z}_1, \mathbf{z}_2 | c = 2) \\
&= \sum_{\mathbf{l}_2, w_2, m_2} p(\mathbf{I}, \mathbf{z}_1, \mathbf{l}_2, w_2, m_2) \\
&= \sum_{\mathbf{l}_2, w_2, m_2} p(\mathbf{I}, \mathbf{z}_1, \mathbf{l}_2 | w_2, m_2) p(w_2 | m_2) p(m_2) \\
&= \sum_{m_2} p(m_2) \sum_{w_2} p(w_2 | m_2) \sum_{\mathbf{l}_2} p(\mathbf{I}, \mathbf{z}_1, \mathbf{l}_2 | w_2, m_2).
\end{aligned} \tag{3}$$

The $p(\mathbf{I}, \mathbf{z}_1, \mathbf{l}_2 | w_2, m_2)$ in (3) is the joint distribution of image $\mathbf{I}$, configurations $\mathbf{z}_1$ and $\mathbf{l}_2$ given mixture m_2 and window w_2. An overview of this implementation is shown in Fig. 5. The one-Pedestrian, two-pedestrian and pedestrian-part detection scores in Fig. 5 are integrated into the evidence to one-pedestrian configuration $p(\mathbf{I}, \mathbf{z}_1, \mathbf{l}_2 | w_2, m_2)$, which is detailed in Section 4.2. The evidence to one-pedestrian configuration in Fig. 5 is then added to one-pedestrian detection results using (1) to obtain the refined detection result in Fig. 5.

4.2 Mixture of DPM for Two-Pedestrian Detection

In order to learn the mixture type $m_2 = 1, \ldots, M$, the configuration space of $\mathbf{z}_2$ is divided into $M = S \cdot A$ clusters using the following two steps.

1) The two pedestrians form a two-pedestrian bounding box. The positive training samples are divided into A groups according to their aspect ratios, i.e., the height divided by the width of the bounding box. Fig. 6 shows the results of dividing the INRIA training samples into $A = 3$ groups.
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 approaches can be adopted. We empirically evaluate the mixture of Gaussian (MoG), spectral clustering and K-means in the experiments. Fig. 7 shows the clustering results for $A = 3$ and $S = 3$. Fig. 8 shows the detectors learned for the nine clusters using MoG.

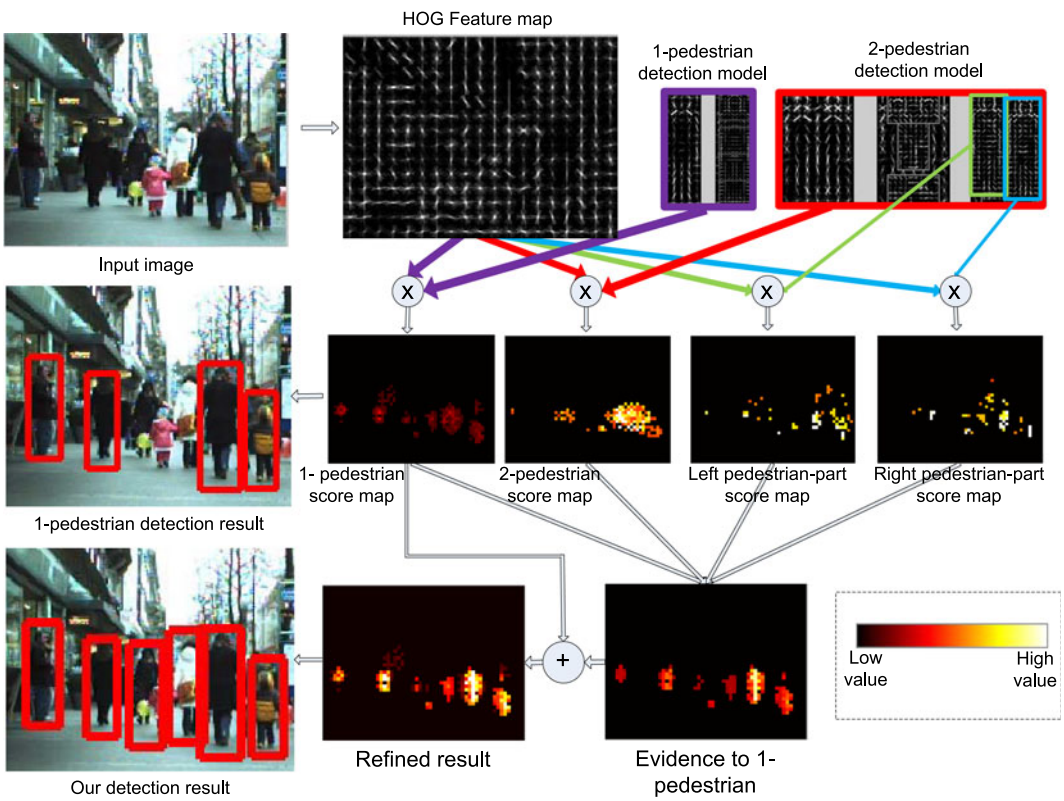


Fig. 5. Use two-pedestrian detection result to refine one-pedestrian detection. The detection scores of one-pedestrian λ_1, two-pedestrians λ_2 and pedestrian-parts λ_p are integrated as the evidence to one-pedestrian configuration $\mathbf{z}_1$. This evidence is added to the result obtained with the one-pedestrian detector. Examples in the left column are obtained at 1 FPPI on the ETH dataset. Although only activations in one scale are shown, two-pedestrian activations at one scale are able to affect a one-pedestrian activation of a closely related scale. This figure is best viewed in color.

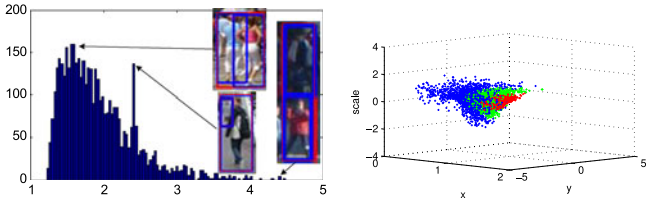


Fig. 6. The number of samples (*Y*-axis) with respect to the aspect ratio (*X*-axis) measured by height divided width (left) and division of the INRIA training samples in Fig. 4 into $A = 3$ groups according the aspect ratio of two-pedestrian bounding box (right). Best viewed in color.

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 [30]. Each cluster corresponds to a mixture type m_2 in (3). The two-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 [30]. The two-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 [30]. Besides, we add two extra parts that correspond to the two pedestrians in a two-pedestrian training sample. They are called pedestrian-parts. The anchor locations and sizes of the two pedestrian-parts are obtained from the average locations and sizes of the training samples in this cluster. In order to transfer the knowledge of the one-pedestrian detector to the two-pedestrian detector, the initial filters for the two pedestrian-parts are obtained from the root filter of the one-pedestrian detector. With the positive samples and initial part filters defined, the DPM with Latent SVM and HOG feature in [30] is then used to train the two-pedestrian detector. The learned models using MoG are shown in Fig. 8. The configuration $\mathbf{l}_2$ contains the sizes and locations of parts. Since the pedestrian-parts are explicitly modeled as parts in the two-pedestrian model, the size and location

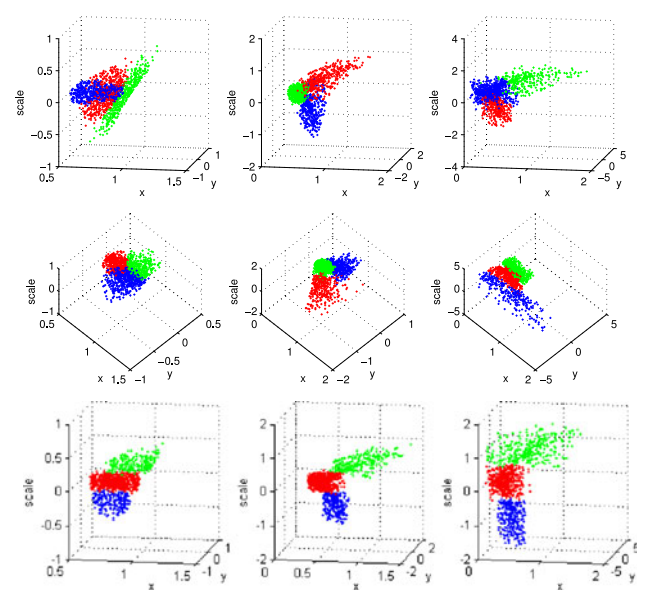


Fig. 7. Division of the INRIA training samples in each aspect ratio group into $S = 3$ clusters (best viewed in color). Each column corresponds to an aspect ratio. First row: result of Gaussian mixture model. Second: result of spectral clustering in [53]. Third row: result of K-means.

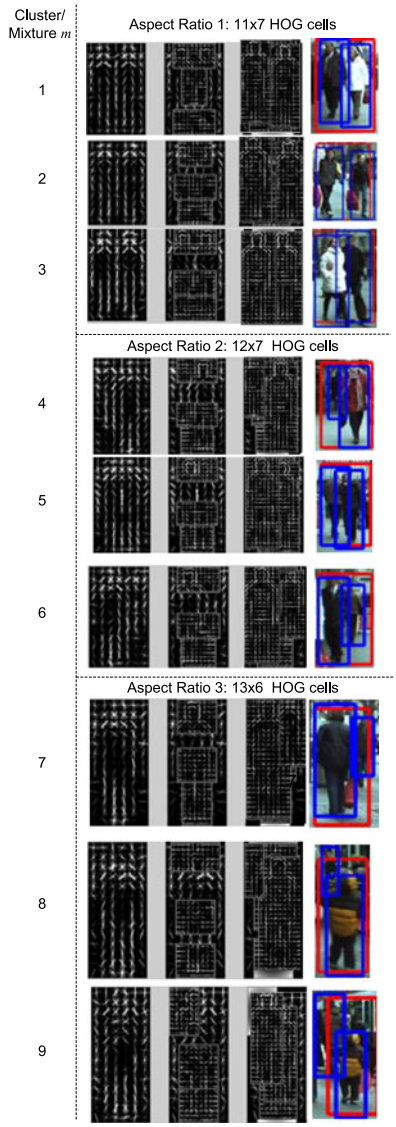


Fig. 8. Two-Pedestrian detectors learned for different clusters. Column 1: root filter; Column 2: three part filters found from root filter; Column 3: two pedestrian-part filters; Column 4: examples detected by the detectors in the same rows. Red rectangles are two-pedestrian detection results. Blue rectangles indicate pedestrian-part locations. Best viewed in color.

of each pedestrian in the two-pedestrian window are also inferred by DPM at the detection stage. This is the key to build the relationship between the two-pedestrian detection result and the one-pedestrian detection result.

$p(m_2)$ in (3) 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. We tried estimating $p(m_2)$ from training data but did not observe improvement. Given the mixture model m_2, $p(w_2|m_2)$ in (3) 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 $h \in \{0, 1\}$. $h = 0$ when the left pedestrian-part in $\mathbf{l}_2$ is considered to match the single pedestrian with configuration $\mathbf{z}_1$, and $h = 1$ when the right pedestrian-part matches the single pedestrian. With h included, we have the following for the $p(\mathbf{I}, \mathbf{z}_1, \mathbf{l}_2|w_2, m_2)$ in (3):

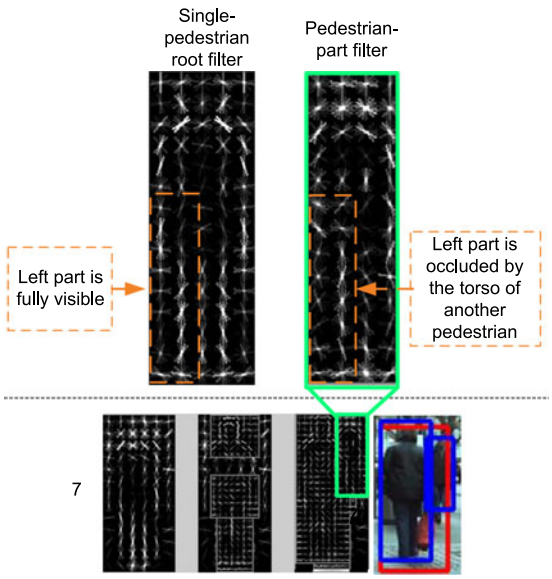


Fig. 9. The single-pedestrian root filter in [30] and our pedestrian-part filter of mixture type 7.

$$p(\mathbf{I},\mathbf{z}_1,\mathbf{l}_2|w_2,m_2)=\sum_h p(\mathbf{I},\mathbf{z}_1,\mathbf{l}_2,h|w_2,m_2)$$
$$=\sum_h p(\mathbf{I},\mathbf{z}_1,\mathbf{l}_2|h,w_2,m_2)p(h|w_2,m_2), \quad (4)$$

where $p(h|w_2,m_2)=0.5, \mathbf{z}_1=(\mathbf{l}_1,w_1)$,

$$p(\mathbf{I},\mathbf{z}_1,\mathbf{l}_2|h,w_2,m_2)$$
$$=p(\mathbf{I},\mathbf{l}_1|w_1,\mathbf{l}_2,h,w_2,m_2)p(w_1|\mathbf{l}_2,h,w_2,m_2)p(\mathbf{l}_2|w_2,m_2) \quad (5)$$

and we suppose $p(\mathbf{l}_2|w_2,m_2,h)=p(\mathbf{l}_2|w_2,m_2)$. The right-hand-side terms in (5) are enumerated as follows:

- $p(w_1,|\mathbf{l}_2,h,w_2,m_2)$ models the relationship between the one-pedestrian configuration $\mathbf{z}_1$ and two-pedestrian configuration $\mathbf{z}_2$ and will be detailed in Section 4.3. This is implemented by matching $\mathbf{z}_1$ with $\mathbf{z}_2$.
- Consider the $p(\mathbf{I},\mathbf{l}_1|w_1,\mathbf{l}_2,h,w_2,m_2)$ and $p(\mathbf{l}_2|w_2,m_2)$ in (5) together, we have:

$$p(\mathbf{I},\mathbf{l}_1|w_1,\mathbf{l}_2,h,w_2,m_2)p(\mathbf{l}_2|w_2,m_2)$$
$$\propto \phi_1(\mathbf{I},\mathbf{l}_1;w_1)\phi_p(\mathbf{I};\mathbf{l}_2,w_2,m_2,h)$$
$$\phi_{2,1}(\mathbf{I};\mathbf{l}_2,w_2,m_2)\phi_{2,2}(\mathbf{l}_2;w_2,m_2) \quad (6)$$
$$=\phi_1(\mathbf{I},\mathbf{l}_1;w_1)\phi_p(\mathbf{I};\mathbf{l}_2,w_2,m_2,h)\phi_2(\mathbf{I},\mathbf{l}_2;w_2,m_2)$$
$$=\lambda_1\lambda_p\lambda_2,$$

where $p(\mathbf{l}_2|w_2,m_2)\propto\phi_{2,2}(\mathbf{l}_2;w_2,m_2)$, $p(\mathbf{I},\mathbf{l}_1|w_1,\mathbf{l}_2,h,w_2,m_2)$ is approximated by the products of three terms, $\phi_1(*)$, $\phi_p(*)$, and $\phi_{2,1}(*)$, subject to normalization factor. $\phi_1(*)$ is from one-pedestrian detector, $\phi_p(*)$ is from the pedestrian-part, and $\phi_{2,1}(*)$ is from the two-pedestrian detector. The λ_1, λ_p, and λ_2 in (6) are illustrated as follows:

- $\lambda_1=\phi_1(\mathbf{I},\mathbf{l}_1;w_1)$ is from one-pedestrian detection score. For example, $\phi_1(\mathbf{I},\mathbf{l}_1;w_1)$ can be implemented using the DPM in [30] as follows:

$$\lambda_1=\phi_1(\mathbf{I},\mathbf{l}_1;w_1)=e^{\mathbf{F}_a^{\mathrm{T}}\psi_1(\mathbf{I};\mathbf{l}_1)+\mathbf{F}_d^{\mathrm{T}}\psi_d(\mathbf{l}_1-\mathbf{a}_1)}, \quad (7)$$

where $\mathbf{F}_a$ and $\mathbf{F}_d$ are linear weights learned from SVM, $\psi_1(\mathbf{I}_1,\mathbf{l}_1,w,m)$ represents appearance features, $\mathbf{a}_1$ is the anchor position of parts and $\mathbf{F}_d^{\mathrm{T}}\psi(\mathbf{l}_1-\mathbf{a}_1)$ calculates the cost of deforming parts from anchor $\mathbf{a}_1$ to location $\mathbf{l}_1$.

- $\lambda_p=\phi_p(\mathbf{I};\mathbf{l}_2,w_2,m_2,h)$ is from the pedestrian-part score, which is used as the extra information to refine one-pedestrian detection results.
- $\lambda_2=\phi_2(\mathbf{I},\mathbf{l}_2;w_2,m_2)$ is from the two-pedestrian detection score obtained by DPM. And we have:

$$\lambda_2=\phi_2(\mathbf{I},\mathbf{l}_2;w_2,m_2)$$
$$=\phi_{2,1}(\mathbf{I};\mathbf{l}_2,w_2,m_2)\phi_{2,2}(\mathbf{l}_2;w_2,m_2) \quad (8)$$
$$=e^{\mathbf{F}_{2,1}^{\mathrm{T}}\psi_{2,1}(\mathbf{I};\mathbf{l}_2)}e^{\mathbf{F}_{2,2}^{\mathrm{T}}\psi_{2,2}(\mathbf{l}_2-\mathbf{a}_2)},$$

where $\mathbf{F}_{2,1}$ and $\mathbf{F}_{2,2}$ are linear weights learned from training data, $\psi_{2,1}(\mathbf{I},\mathbf{l}_2,w,m)$ represents appearance features, $\mathbf{a}_2$ is the anchor position of parts and $\mathbf{F}_{2,2}^{\mathrm{T}}\psi(\mathbf{l}_2-\mathbf{a}_2)$ calculates the cost of deforming parts from anchor $\mathbf{a}_2$ to location $\mathbf{l}_2$.

In Fig. 5, the one-pedestrian score map is from λ_1, the two-pedestrian score map is from λ_2, and the pedestrian-part score maps are from λ_p.

4.3 Modeling the Relationship between Two- and One-Pedestrian Detection Results

With the pedestrian-parts designed in the two-pedestrian detector, the relationship between two- and one-pedestrian detection results is implemented by matching the pedestrian-part in the two-pedestrian detector with the one-pedestrian detection result. For a one-pedestrian activation, two-pedestrian activations at different scales affect it through $p(w_1|\mathbf{l}_2,w_2,m_2,h)$ in (4), which is a Gaussian distribution:

$$p(w_1|\mathbf{l}_2,w_2,m_2,h)=(2\pi)^{-\frac{3}{2}}|A|^{\frac{1}{2}}e^{-\frac{1}{2s_1\cdot s_1}(w_1-u)^T A(w_1-u)}, \quad (9)$$

where A is the precision matrix estimated from training samples for each mixture m_2, $w_1=(x_1,y_1,s_1)$ is the location and scale of $\mathbf{z}_1$ normalized by the scale s_1, $u=(x_{2,h},y_{2,h},s_{2,h})$ is the location and scale of the pedestrian-part h in $\mathbf{l}_2$ normalized by the size s_1. $p(w_1|\mathbf{l}_2,w_2,m_2,h)=0$ when the overlap between the one-pedestrian window and the pedestrian-part is smaller than 0.5. $p(w_1|\mathbf{l}_2,w_2,m_2,h)$ is the largest if the one-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 $\mathbf{z}_1=(w_1,\mathbf{l}_1)$ is L_c. The number of possible configurations for the five 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 (3) is $O(ML_c^7)$. For example, a 640×480 image has $L_c>40{,}000$ considering sliding windows for different scales. The computation will be greater than $M\times10^{32}$. The computation ability of 4G Hz CPU is about 2×10^{10} operations per second. Therefore, direct computation of (3) is unaffordable and a fast approach is required.

The first step for faster speed is to assume that the configuration prior $\phi_{2,2}(\mathbf{l}_2; w_2, m_2)$ is sharply peaked around value $\tilde{\mathbf{l}}_2$ so that the summation in (3) can be approximated with maximization as follows:

$$\begin{aligned}&\sum_{\mathbf{l}_2,w_2,m_2} p(\mathbf{I}, \mathbf{z}_1, \mathbf{l}_2, w_2, m_2)\\ &\approx \sum_{h,w_2,m_2} \max_{\mathbf{l}_2} p(w_2, m_2)\lambda_1\phi_p(\mathbf{I}; \mathbf{l}_2, w_2, m_2, h)\\ &\quad \phi_2(\mathbf{I}, \mathbf{l}_2; w_2, m_2)p(w_1|\mathbf{l}_2, w_2, m_2, h)p(h),\end{aligned} \tag{10}$$

where $\lambda_1 = \phi_1(\mathbf{I}, \mathbf{l}_1; w_1)$. Although with the approximation above, the computational complexity is still not changed because the best configuration $\tilde{\mathbf{l}}_2$ is dependent on w_1. Since $\phi_{2,2}(\mathbf{l}_2; w_2, m_2)$ is supposed to be sharply peaked around $\tilde{\mathbf{l}}_2$, $p(w_1|\mathbf{l}_2, w_2, m_2, h)$ is considered as being non-zero only when $\mathbf{l}_2 = \tilde{\mathbf{l}}_2$. Then $p(w_1|\mathbf{l}_2, w_2, m_2, h)$ is moved out of the maximization operation to have

$$\begin{aligned}&\sum_{h,w_2,m_2} \max_{\mathbf{l}_2} \{p(w_2, m_2)\lambda_1\phi_p(\mathbf{I}; \mathbf{l}_2, w_2, m_2, h)\\ &\quad \cdot \phi_2(\mathbf{I}, \mathbf{l}_2; w_2, m_2)p(w_1|\mathbf{l}_2, w_2, m_2, h)p(h)\}\\ &\approx \sum_{h,w_2,m_2} \{p(w_2, m_2)p(h)\lambda_1\phi_p(\mathbf{I}; \tilde{\mathbf{l}}_2, w_2, m_2, h)\\ &\quad \cdot \phi_2(\mathbf{I}, \tilde{\mathbf{l}}_2; w_2, m_2)p(w_1|\tilde{\mathbf{l}}_2, w_2, m_2, h)\}\\ &= \sum_{h,w_2,m_2} p(w_2, m_2)p(h)\lambda_1\tilde{\lambda}_p\tilde{\lambda}_2 p(w_1|\tilde{\mathbf{l}}_2, w_2, m_2, h),\\ &\text{where } \tilde{\mathbf{l}}_2 = \operatorname*{argmax}_{\mathbf{l}_2} \phi_p(\mathbf{I}; \mathbf{l}_2, w_2, m_2, h)\phi_2(\mathbf{I}, \mathbf{l}_2; w_2, m_2),\\ &\quad \tilde{\lambda}_p = \phi_p(\mathbf{I}; \tilde{\mathbf{l}}_2, w_2, m_2, h),\\ &\quad \tilde{\lambda}_2 = \phi_2(\mathbf{I}, \tilde{\mathbf{l}}_2; w_2, m_2).\end{aligned} \tag{11}$$

The search for best configuration of $\mathbf{l}_2$ in (11) is the search of the best part location in two-pedestrian detector, which is independent of the single detector configuration $\mathbf{z}_1$. This search can be efficiently solved by the distance transform [31], which is also used for DPM in [30] with computational complexity $O(KL_c)$ for all candidate windows w, where K is the number of parts. Therefore, the $\arg\max_{\mathbf{l}_2}\phi_p(\mathbf{I}; \mathbf{l}_2, w_2, m_2, h)\phi_2(\mathbf{I}, \mathbf{l}_2; w_2, m_2)$ in (11) has computational complexity $O(L_f L_c + KL_c)$, where L_f is the length of features, $L_f L_c$ is used for obtaining the linear SVM filtering results on visual features like HOG, KL_c is used for obtaining the best configuration $\tilde{\mathbf{l}}_2$ for all windows with mixture type m. Since $K \ll L_f$, $O(L_f L_c + KL_c) \approx O(L_f L_c)$. In summary, the computational complexity in obtaining $\tilde{\mathbf{l}}_2, \tilde{\lambda}_p$ and $\tilde{\lambda}_2$ is $O(L_f L_c)$ if the two-pedestrian detector is computed in sliding window manner.

The overall implementation is as follows:

$$\begin{aligned}p(\mathbf{z}_1, \mathbf{I}) &\approx p(c=1)p(\mathbf{I}, \mathbf{z}_1|c=1)\\ &+ p(c=2)\sum_{h,w_2,m_2} p(w_2, m_2)p(h)\lambda_1\tilde{\lambda}_p\tilde{\lambda}_2 p(w_1|\tilde{\mathbf{l}}_2, w_2, m_2, h),\end{aligned} \tag{12}$$

where λ_1 and $p(\mathbf{I}, \mathbf{z}_1|c=1)$ are from the single pedestrian detection, $p(w_2, m_2)$ is sampled in sliding window for all mixture types, $p(h) = 0.5$, $\tilde{\lambda}_p$, and $\tilde{\lambda}_2$ are given in (11) and illustrated after equation (6), $p(w_1|\tilde{\mathbf{l}}_2, w_2, m_2, h)$ is given in (9). By default, $p(c=1) = 1/3$ and $p(c=2) = 2/3$. Tuning this parameter on the Caltech and TUD datasets, the average miss rate is further reduced less by than 1 percent. On ETHZ, it is reduced by 2 percent by setting $p(c=1) = 0.4$ and $p(c=2) = 0.6$. We resize the 1-pedestrian bounding box into 1:3 for width:height.

The computation related to the two-pedestrian detector can be further reduced by cascading two-pedestrian detector after the one-pedestrian detection results is obtained. Denote the number of candidates for $\mathbf{z}_1$ by $Cand_1$, and the number of candidate windows w_2 for M mixtures by $Cand_2$. The procedure and computational complexity of computing (11) for all configurations of $\mathbf{z}_1$ is as follows:

- *Step* 1. Obtain the one-pedestrian detection result, which is used for $p(\mathbf{I}, \mathbf{z}_1|c=1)$ and λ_1 in (12).
  - *Analysis*. $O(L_c)$ operations are required in this step. Only $Cand_1$ candidate windows, which are detected by the single-pedestrian detector, are used for the next steps.
- *Step* 2. Obtain the two-pedestrian detection results, which is used for $\tilde{\lambda}_p$ and $\tilde{\lambda}_2$ in (11).
  - *Analysis*. We assume that if two nearby pedestrians exist, at least one pedestrian will be detected by the single-pedestrian detector around this region. With this assumption, the two-pedestrian detector can be evaluated only around $Cand_1$ one-pedestrian candidate windows to save computation. $O(Cand_1)$ operations are required in this step.
- *Step* 3. For each one-pedestrian candidate $\mathbf{z}_1$, compute (11) for $Cand_2$ two-pedestrian candidate windows using the results obtained in Step 1 and Step 2.
  - *Analysis*. 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 $Cand_1 Cand_2$ non-zero λ_1 and $\tilde{\lambda}_2$. With the terms computed, the computational complexity for summing up them w.r.t. w_1, h, w_2 and m_2 in (11) is $O(Cand_1 Cand_2)$ by enforcing sparsity on one-pedestrian and two-pedestrian candidate windows. $O(Cand_1 Cand_2)$ operations are required in this step.

Take our experiment on the Caltech dataset [22] 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.

6 Experimental Results

The proposed framework is evaluated on three public datasets: Caltech [22], TUD-Brussels [86] and ETH [28]. We use the modified HOG [30] as feature and the DPM in [30] to learn the two-pedestrian detector. HOG+DPM is used because it is off-the-shelf, open-source, and widely used. Since the detection scores of two-pedestrian detector and one-pedestrian detector are considered as input, the framework keeps unchanged if other detection models or features are used for one-pedestrian detector or two-pedestrian detector. Existing pedestrian detection results can be directly used as the input of our framework.

The one-pedestrian detection approach in [30] used the same feature and DPM as our two-pedestrian detector. It is denoted as LatSVM-V2 in the experimental results. Our framework using LatSVM-V2 as the one-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 two-pedestrian detector and compared in Section 6.2. The detection windows of the LatSVM-V2 and LatSVM-V2-E are obtained by running the authors' code. The detection windows of other approaches are obtained from Dollar's webpage,[1] NMS is used for these approaches before they are aided by the two-pedestrian detector.

The labels and evaluation code provided by Dollár et al. online are used for evaluation following the criteria proposed in [22]. As in [22], the *log-average miss rate* is used to summarize the detector performance, which 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 datasets, which is the most popular portion of the datasets. It consists pedestrians of ≥ 50 pixels in height, and less than 35 percent occluded.

6.1 Preparation of Two-Pedestrian Training Data

Since there is no two-pedestrian detection training dataset, we construct it based on the INRIA training dataset [10] 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 two-pedestrian positive samples (656). We labeled more pedestrians in the positive images. The number of positive one-pedestrian samples increases from the original 1,237 to 2,738. The number of positive two-pedestrian samples increases from the original 656 to 4,398.[2]
3) If the bounding boxes of two pedestrians overlap, the bounding box that exactly covers the two pedestrians is considered as the label of the two-pedestrian positive sample.

Once the two-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

First of all, we compare with the approach in [30] which used the same feature and learning model as our two-pedestrian detector. Compared with LatSVM-V2, our approach has 10 percent (from 51 to 41 percent), 7 and 5 percent log-average miss rate improvement on the datasets ETH,[3] 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 on our extended INRIA dataset described in Section 6.1. It is denoted by

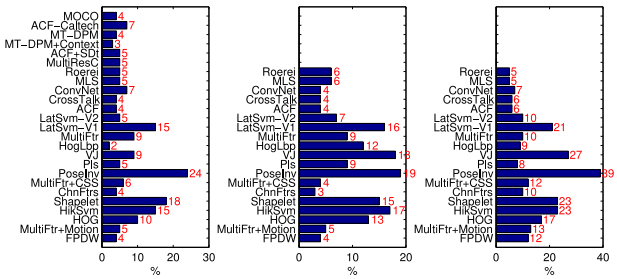


Fig. 10. Miss rate improvement of the framework for each of the state-of-the-art one-pedestrian detectors on Caltech-Test (left), TUD-Brussels (middle) and ETH (right). *X*-axis denotes the miss rate improvement.

LatSvm-V2-E. By combining with LatSVM-V2-E, our approach (LatSvm-V2-E+our) has 9, 7 and 5 percent log-average miss rate improvement over LatSVM-V2-E on the datasets ETH, TUD-Brussels and Caltech-Test respectively. Considering LatSvm-V2 and LatSvm-V2-E on the three datasets, the inclusion of larger training set does not influence the relative performance rank of DPM compared with other approaches, e.g., FPDW and Pls.

We also investigate other one-pedestrian detectors and integrate them with our two-pedestrian detector in this experiment. All the approaches evaluated on the Caltech, TUD-Brussels and EHTZ datasets in [22] are evaluated in this experiment. We resize the one-pedestrian bounding box of these approaches into 1:3 for width:height. These approaches are VJ [79], Shapelet [67], PoseInv [46], LatSvm-V1 [30], LatSvm-V2 [30], HikSVM [49], HOG [10], MultiFtr [85], HogLbp [83], Pls [69], MultiFtr+CCS, MultiFtr+Motion [80], FPDW [20], ChnFtrs [21]. We also include recent approaches such as MultiResC [62], Rorei [5], MOCO [9], Crosstalk [19], MT-DPM [92], ACF [18], and ConvNet [70]. MultiResC, MOCO, ACF-Caltech, MT-DPM, MT-DPM +Context, and ACF+SDt are only evaluated on the Caltech-Test dataset, because their results on ETH and TUD-Brussels are not available. For one-pedestrian detection results, the range of detection score s has large variation for different approaches. s is normalized to s_{norm} as follows:

$$s_{norm} = \sigma(a * s + b), a = 6/s_{max}, b = -0.6a, \qquad (13)$$

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 $p(\mathbf{I}, \mathbf{z}_1|c = 1)$ in (1). This normalization was chosen empirically and the same for all the methods. Tuning them per method can improve performance but is not investigated in the final results. Fig. 14 shows the results on the three datasets. Fig. 10 shows the improvement of our framework for each of these approaches on the three datasets. Our framework significantly improves all the state-of-the-art pedestrian detectors by integrating with them. On the ETH dataset, it is reported that LatSvm-V2 has the best performance among the 14 state-of-the-art approaches evaluated in [22]. The average miss rate for LatSvm-V2 is 51 percent. By integrating with our framework, 10 algorithms outperform LatSVM-V2 and the best performing one (LatSVM-V2+Our) reaches the average miss rate of 41 percent. The current best performing approaches on the Caltech-Test dataset is the ACF+SDt in [18] with the motion feature in [63], which has an average

1. www.vision.caltech.edu/Image_Datasets/CaltechPedestrians
2. Publicly available on www.ee.cuhk.edu.hk/~wlouyang/projects/ouyangWcvpr13MultiPed
3. Demo results on https://www.youtube.com/watch?v=0JTg93ur52A

miss rate 37 percent. With our framework, ACF+SDt+Our is improved from 37 to 32 percent. With our framework, the current best performing approach on the TUD-Brussels dataset, i.e. MultiFtr+Motion, is improved from 55 to 50 percent.

Fig. 12 shows the effect of different clustering approaches, i.e., mixture of Gaussian, K-means and spectral clustering, for obtaining the mixture type m of the two-pedestrian detector on the algorithms evaluated in [22]. The three approaches achieve similar improvement. Spectral clustering performs slightly better than the other two. When spectral clustering is used, the average improvement is about 9 percent on the Caltech-Test dataset, 11 percent on the TUD-Brussels dataset and 17 percent on the ETH dataset. This experiment shows that the two-pedestrian detector provides rich complementary information to current state-of-the-art one-pedestrian detection approaches even when context [9], [18], [62], [92] or motion [18], [80] is used by these approaches.

6.3 Investigation on the Varied Improvement of Our Framework

The improvement of LatSvm-V2-E+Our compared with LatSvm-V2-E varies from 5 percent on the Caltech Testing dataset to 9 percent on the ETH dataset. In order to investigate the reason, we show total number of pedestrians and the number of M-pedestrians in the four datasets in Fig. 13. Pedestrians overlapping with other pedestrians are called M-pedestrians in this paper. Our approach focuses on M-pedestrians. The M-pedestrians in the subset *reasonable* (explained in the paragraph before Section 6.1) are put into the subset *multiple* for evaluation. As shown in Fig. 15, the miss rate reduction of our LatSvm-V2-E+Our compared with LatSvm-V2-E is close on the three datasets. Note that an M-pedestrian is still a 1-pedestrian. As shown in Fig. 13, the number of M-pedestrians divided by the number of all pedestrians is the largest for the ETH dataset, i.e., about 2/3, but the smallest for the Caltech Testing dataset, i.e., about 1/3. When this ratio is larger, the improvement from M-pedestrians contribute more to the overall result. Therefore, the subset reasonable in Fig. 14 is similar to the subset multiple in Fig. 15 for the ETH dataset where M-pedestrians appears more frequently.

6.4 Investigation on Joint Detection Scheme

Fig. 16 shows the experimental results on different joint detection schemes. All the approaches in Fig. 16 are trained on the extended INRIA training dataset described in Section 6.1. LatSvm-V2-E denotes the result using only the one-pedestrian DPM. LatSvm-V2-E+Joint denotes the joint person detection scheme in [75], which uses two stages of NMS for combining one-pedestrian and two-pedestrian detection results. The result from [75] is obtained by using the authors source code. LatSvm-V2-E +Our denotes our scheme. Both LatSvm-V2-E+Joint and LatSvm-V2-E+Our use the same one-pedestrian and two-pedestrian detection results. They are only different in the joint detection scheme. Experimental results in Fig. 16 show that our approach performs better than the approach in [75]. The post-processing steps (bounding box regression and two-level NMS) of LatSvm-V2-E+Joint are explicitly designed and work well for side-view person/person pairs who walk very close to each other. However, it does not perform well when pedestrians have large variation in window size, 3D spatial location, and walking directions, which is the case for the ETH, TUD, and Caltech datasets. Note that we only re-implement the part of integrating one-pedestrian detection results and two-pedestrian detection results, but not the whole approach of [75]. Because integration scheme is the overlapping part of our approach and the approach in [75]. Tang et al. [75] has major contributions regarding jointly training one-pedestrian and two-pedestrian detectors with segmentation maps from the training set, synthetically generating two-people samples, and integrating the model into a tracking approach. These contributions are orthogonal to our contributions and thus are not implemented.

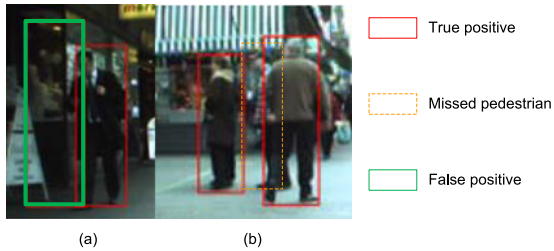


Fig. 11. Failure cases caused by NMS (a) and the imperfect two-pedestrian detector (b).

7 Discussion

In this paper, we have used two-pedestrian detection results to improve the one-pedestrian detection results. Since both two-pedestrian and one-pedestrian detection scores are considered as the input of our framework, the framework keeps unchanged if other pedestrian detection approaches are used. Therefore, the other models like the tree model in [31], [73], [98], the loopy graph model in [64], [84], the complete graph model in [6] and the pictorial structures in [1], [31] can be used for both single- and two-pedestrian detection.

Since we do NMS after the single-pedestrian detection results are refined by the two-pedestrian detector, it may miss one of the pedestrians if their overlap is larger than 0.5. An example is shown in Fig. 11a. This problem could be handled with more advanced post-processing approaches, e.g., mean-shift.

This paper shows by experiment that the two-pedestrian detector improves the detection performance. Nevertheless, investigation on K-pedestrian detector ($K > 2$) in crowded scenes is a potential way of improvement.

In this paper, the one-pedestrian and two-pedestrian detection results are assumed to be provided in order to be independent of detectors and features. However, interaction between one-pedestrian and two-pedestrian detector in the training stage is a future work for improvement. For example, one-pedestrian detector can be influenced by the two-pedestrian detector and vise versa.

Fig. 10 shows that one-pedestrian detection result using motion (MultiFtr+Motion and ACF+SDt) can be improved with the help of two-pedestrian detection results. Although

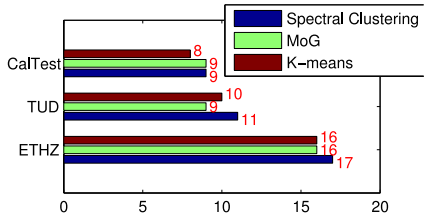


Fig. 12. Average improvement of the framework for the algorithms in [22] using different clustering approaches.

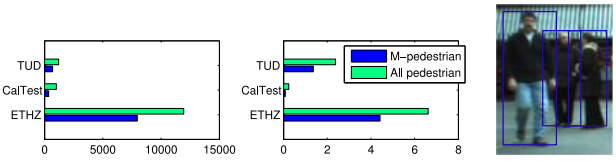


Fig. 13. The *overall* number of all pedestrians and M-pedestrians on the *x*-axis for different datasets (left), the *average* number of all pedestrians and M-pedestrians *per image* on the *x*-axis (middle) and 3 M-Pedestrians (right). An M-pedestrian is a pedestrian whose bounding box overlaps with other pedestrians.

we only use the HOG feature extracted from static images for two-pedestrian detection, the framework is naturally applicable for single- and two-pedestrian detection using multiple cues like depth, motion and segmentation.

Our approach can also be considered as a rescoring approach. Probabilistic formulation is chosen because it is the idea behind our final implementation and it is a principled tool. Hough voting is also widely used in rescoring activations [3], [8], [41], [42], [89]. As pointed out in [3], to some extent, Hough voting can be viewed as logarithms of $p(x_i = h|I_i)$, where $x_i = h$ implies that the voting element i is generated by the object h, I_i is the descriptor of the voting element i. However, rather than fusing all the votes in a principled way, Hough voting simply sums them up. Therefore, the probabilistic representation of Hough voting has the joint product form so that its logarithm can be represented by summation. However, our model in (12) cannot be represented by the joint product form because of the summation over $p(c = 1)$ and $p(c = 2)$, and the summations over h, w_2 and m_2. The summations come from the marginalization, e.g. over $c = 1$ and $c = 2$. Thus ours is not a Houg voting method.

Because of the imperfectness of the two-pedestrian detector, some false positives will have their detection scores increased, e.g., in Fig. 11b.

This paper focuses on detecting pedestrians. In the future work, we will investigate on using the model for other objects, e.g., cars and articulated persons.

7.1 Relationship Among Existing Approaches

There are many approaches that consider multiple pedestrians for detection. These approaches can be grouped into two categories:

1. Find out the best configuration of multiple pedestrians [3], [41], [42], [88], [89], [91]. In these approaches, pedestrians are represented as an assembly of several parts, and joint part combination for multiple pedestrians are sought. Denote the configuration of the nth object by $z^{(n)}$ for $n = 1, \ldots, N$. These approaches can be considered as solving one of the following problems:

$$\text{argmax}_{z^{(1)}\ldots z^{(N)}} \prod_n p(\mathbf{I}|z^{(n)}) \tag{14}$$

$$\text{or} \quad \arg\max_{z^{(1)}\ldots z^{(N)}} p(z^{(1)} \ldots z^{(N)}) \prod_n p(\mathbf{I}|z^{(n)}). \tag{15}$$

Non-maximum suppression is also an approach for solving the problem in (14). Since the search space for $z^{(1)} \ldots z^{(N)}$ is too large, greedy search is often adopted [3], [89], [91]. These approaches are useful for handling inter-object occlusion.

2. Use the existence, location, orientation and size information of the other objects for refining the configuration of the target object [16], [62], [93]. Let $\mathbf{z}_C = (\mathbf{l}_{1,1}, \ldots, \mathbf{l}_{B,Q})$, where $C = B \cdot Q$, $\mathbf{l}_{b,q}$ for $b = 1, \ldots, B, q = 1, \ldots, Q$ is the configuration information of the qth contextual object with class label b. $\mathbf{z}_C$ is the configurations of the contextual objects for the target object $\mathbf{z}_1$. These approaches can be represented as follows by setting $p(c = C) = 1$ in (1):

$$\begin{aligned} p(\mathbf{z}_1, \mathbf{I}) &= \sum_{c=C} \sum_{\mathbf{z}_c} p(\mathbf{I}, \mathbf{z}_1, \mathbf{z}_C) p(c = C) \\ &= \sum_{\mathbf{z}_C} p(\mathbf{I}, \mathbf{z}_1, \mathbf{z}_C) = \sum_{\mathbf{l}_1,\ldots,\mathbf{l}_Q} p(\mathbf{I}, \mathbf{z}_1, \mathbf{l}_{1,1}, \ldots, \mathbf{l}_{B,Q}) \\ &= \sum_{\mathbf{l}_1,\ldots,\mathbf{l}_Q} \left[p(\mathbf{z}_1, \mathbf{l}_{1,1}, \ldots, \mathbf{l}_{B,Q}) p(\mathbf{I}|\mathbf{z}_1) \prod_{b,q} p(\mathbf{I}|\mathbf{l}_{b,q}) \right]. \end{aligned} \tag{16}$$

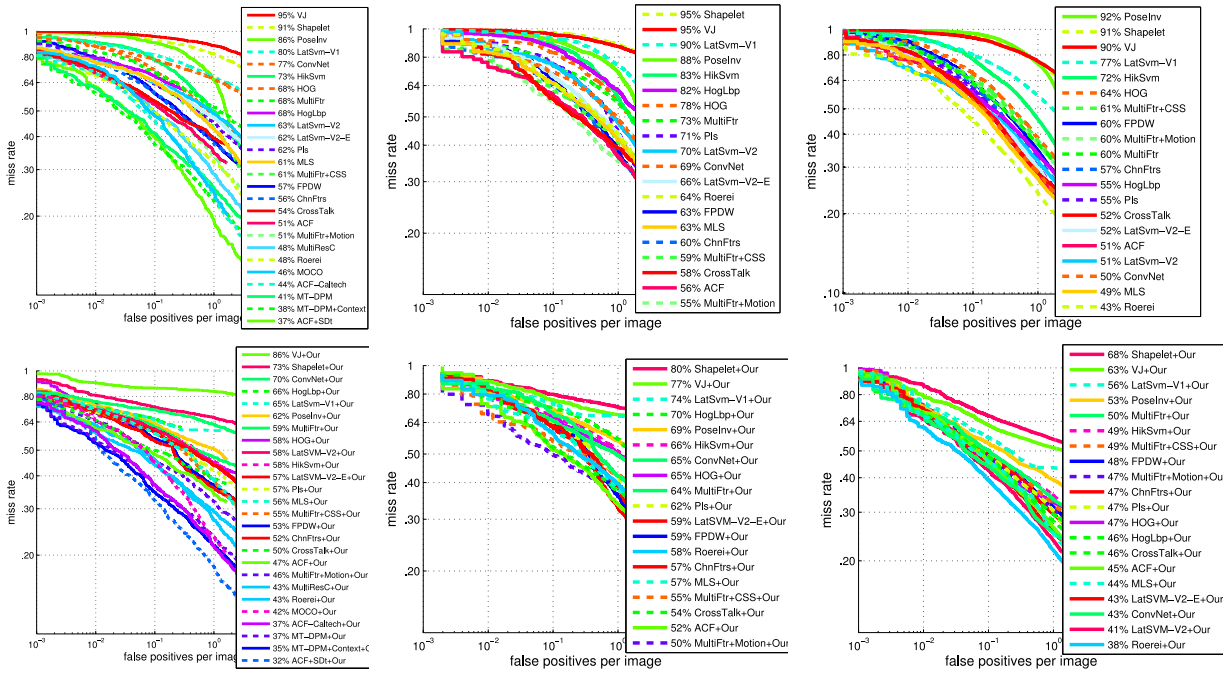
Many approaches have focused on modeling the configuration relationship $p(\mathbf{z}_1, \mathbf{l}_{1,1}, \ldots, \mathbf{l}_{B,Q})$ in (16). $\mathbf{l}_{b,q}$ and $p(\mathbf{z}_1, \mathbf{l}_{1,1}, \ldots, \mathbf{l}_{B,Q})$ are implemented in different ways by different approaches.

1) The approach in [93] jointly estimated the pose of two humans. This approach has $B = 1, Q = 1$. $\mathbf{z}_1$ is the pose of one human and $\mathbf{l}_{1,1}$ is the contextual pose of another human in [93].
2) The approach in [62] considered the estimated sizes of objects as context. This constraint of object size is obtained from the assumption of small 3D object size variation, ground plane, and perspective view. These approaches can be considered as having $B = 1, Q = 1$ in (16). $p(\mathbf{z}_1, \mathbf{l}_{1,1})$ is the probability that the size of $\mathbf{z}_1$ conforms to its size estimated from $\mathbf{l}_{1,1}$ using the geometric constraint. The term $p(\mathbf{I}|\mathbf{l}_{1,1})$ in (16) is not considered and can be represented as a constant.
3) In [16], the presence of pedestrians in the neighboring windows is used as the context. This approach can be considered as implementing (16) by setting $B = 1$, $p(\mathbf{z}_1, \mathbf{l}_{1,1}, \ldots, \mathbf{l}_{1,Q}) = p(\mathbf{z}_1) \prod_q p(\mathbf{l}_{1,q}|\mathbf{z}_1)$, where $\mathbf{l}_{1,q}$ refers to the qth window near $\mathbf{z}_1$. The $p(\mathbf{I}|\mathbf{l}_{1,q})$ in (16) is obtained from single-pedestrian detector in [16].

These approaches are useful for exploiting the contextual information of co-occurring objects.

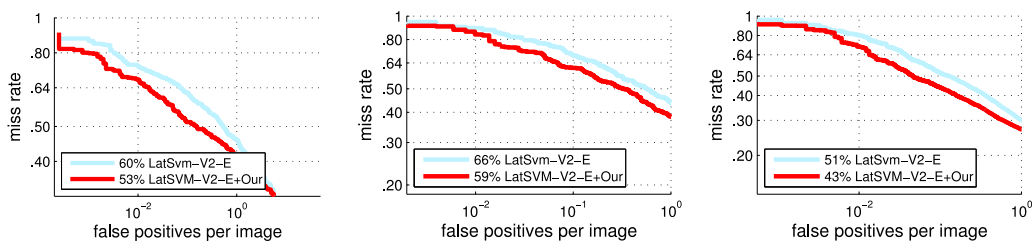
The above two categories can be used with each other. For example, both NMS, which falls into category 1, and contexts, which falls into category 2, are used in [62].

These previous approaches model the location and size correlation among pedestrians. This paper models the



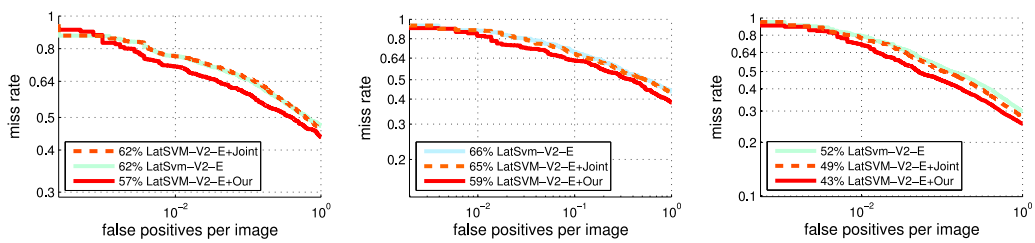
(a) Caltech-Test (b) TUD-Brussels (c) ETH

Fig. 14. 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', e.g. integration of HOG [10] with our framework is denoted by HOG+Our.



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

Fig. 15. Experimental results in detecting M-pedestrians.



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

Fig. 16. Detection results for different schemes. LatSvm-V2-E denotes single pedestrian detection results using DPM. LatSVM-V2-E+Joint denotes the joint person detector in [75]. LatSVM-V2-E+Our denotes our approach.

visual cue of two-pedestrians and models the relationship between single-pedestrian detection results and two-pedestrian detection results.

8 Conclusion

In this paper, we propose a new probabilistic framework for single pedestrian detection aided by two-pedestrian detection. DPM is used to learn the two-pedestrian detector which effectively captures the unique visual patterns appearing in nearby pedestrians. Detection performance is improved by modeling the probabilistic relationship between the configurations of single-pedestrian detection results and those of two-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 two-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. The lowest miss rate is reduced from 37 to 32 percent on the Caltech-Test dataset, from 55 to 50 percent on the TUD-Brussels dataset and from 43 to 38 percent on the ETH dataset. For the 14 state-of-the-art approaches evaluated in [22], the average improvement is 9 percent on the Caltech-Test dataset, 11 percent on the TUD-Brussels dataset and 17 percent on the ETH dataset.

Acknowledgments

The authors would like to thank anonymous reviewers for their constructive comments, Siyu Tang and Bernt Schiele from the Max Planck Institut Informatik for providing their source code and constructive comments. This work was supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project No. CUHK 417011 and CUHK 419412), Shenzhen Basic Research Program (JCYJ20130402113127496) and Guangdong Innovative Research Team Program (No. 201001D0104648280).

References

[1] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1014–1021.
[2] A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg, "Part-based feature synthesis for human detection," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 127–142.
[3] O. Barinova, V. Lempitsky, and P. Kohli, "On detection of multiple object instances using hough transforms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2233–2240.
[4] R. Benenson, M. Mathias, R. Timofte, and L. V. Gool, "Pedestrian detection at 100 frames per second," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2903–2910.
[5] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, "Seeking the strongest rigid detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3666–3673.
[6] M. Bergtholdt, J. H. Kappes, S. Schmidt, and C. Schnorr, "A study of parts-based object class detection using complete graphs," *Int. J. Comput. Vis.*, vo. 87, nos. 1/2, pp. 93–117, 2010.
[7] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 168–181.
[8] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1365–1372.
[9] G. Chen, Y. Ding, J. Xiao, and T. X. Han, "Detection evolution with multi-order contextual co-occurrence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1798–1805.
[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.
[11] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 428–441.
[12] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik, "Fast, accurate detection of 100,000 object classes on a single machine," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1814–1821.
[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[14] C. Desai and D. Ramanan, "Detecting actions, poses, and objects with relational phraselets," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 158–172.
[15] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for multi-class object layout," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 229–236.
[16] Y. Ding and J. Xiao, "Contextual boost for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2895–2902.
[17] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1271–1278.
[18] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.
[19] P. Dollár, R. Appel, and W. Kienzle, "Crosstalk cascades for frame-rate pedestrian detection," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 645–659.
[20] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *Proc. British Mach. Vis. Conf.*, 2010, pp. 1–11.
[21] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. British Mach. Vis. Conf.*, 2009, pp. 91.1–91.11.
[22] 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.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[23] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 990–997.
[24] M. Enzweiler and D. Gavrila, "A mixed generative-discriminative framework for pedestrian classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[25] M. Enzweiler and D. Gavrila, "Integrated pedestrian classification and orientation estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 982–989.
[26] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.
[27] M. Enzweiler and D. M. Gavrila, "A multilevel mixture-of-experts framework for pedestrian classification," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2967–2979, Oct. 2011.
[28] A. Ess, B. Leibe, and L. V. Gool, "Depth and appearance for mobile scene analysis," in *Proc. 11th IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[29] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[30] P. Felzenszwalb, R. B. Grishick, D. McAllister, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, pp. 1627–1645, Sep. 2010.
[31] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, pp. 55–79, 2005.
[32] C. Galleguillos, B. McFee, S. Belongie, and G. Lanckriet, "Multi-class object localization by combining local contextual interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 113–120.
[33] D. Gavrila, "Multi-feature hierarchical template matching using distance transforms," in *Proc. 14th IEEE Int. Conf. Pattern Recognit.*, 1998, pp. 439–444.
[34] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.
[35] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey on pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.
[36] R. Girshick, P. Felzenszwalb, and D. McAllester, "Object detection with grammar models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 442–450.
[37] A. Hare, *Handbook of Small Group Research*. New York, NY, USA: Macmillan, 1962.
[38] D. Hoiem, A. Efros, and M. Hebert, "Putting objects in perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 2137–2144.
[39] S. V. Lab [Online]. Available: http://image-net.org/challenges/LSVRC/2011/index, 2014.

[40] C. Lampert, M. Blaschko, and T. Hofmann, "Beyond sliding windows: object localization by efficient subwindow search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[41] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Proc. 10th Eur. Conf. Comput. Vis. Workshop Stat. Learn. Comp. Vis*, 2004, pp. 17–32.

[42] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *Int. J. Comput. Vis.*, vol. 77, nos. 1/3, pp. 259–289, 2008.

[43] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 878–885.

[44] C. Li, D. Parikh, and T. Chen, "Extracting adaptive contextual cues from unlabeled regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 511–518.

[45] H. Li, X. Huang, J. Huang, and S. Zhang, "Feature matching with affine-function transformation models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2407–2422, Dec. 2014.

[46] Z. Lin and L. Davis, "A pose-invariant descriptor for human detection and segmentation," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 423–436.

[47] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon, "Hierarchical part-template matching for human detection and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[48] P. Luo, Y. Tian, X. Wang, and X. Tang, "Switchable deep network for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 899–906.

[49] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[50] K. Mikolajczyk, B. Leibe, and B. Schiele, "Multiple object class detection with a generative model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 26–36.

[51] 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*, vol. 5, no. 4, p. e10047, 2010.

[52] S. Munder and D. M. Gavrila, "An experimental study on pedestrian classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1863–1868, Nov. 2006.

[53] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.

[54] M. Norouzi, M. Ranjbar, and G. Mori, "Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2735–2742.

[55] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends Cogn. Sci.*, vol. 11, no. 12, pp. 520–527, 2007.

[56] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2337–2344.

[57] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian, Z. Zhu, R. Wang, C.-C. Loy, X. Wang, and X. Tang, "Deepid-net: Multi-stage and deformable deep convolutional neural networks for object detection," arXiv preprint arXiv:1409.3505, 2014.

[58] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3258–3265.

[59] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2056–2063.

[60] W. Ouyang, X. Zeng, and X. Wang, "Modeling mutual visibility relationship in pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3222–3229.

[61] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Int. J. Comput. Vis.*, vol. 38, no. 1, pp. 15–33, 2000.

[62] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2010, pp. 241–254.

[63] D. Park, C. L. Zitnick, D. Ramanan, and P. Dollár, "Exploring weak stabilization for motion feature extraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2882–2889.

[64] M. Pedersoli, A. Vedaldi, and J. Gonzalez, "A coarse-to-fine approach for fast deformable object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1353–1360.

[65] B. Pepikj, M. Stark, P. Gehler, and B. Schiele, "Occlusion patterns for object class detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3286–3293.

[66] F. Porikli, "Integral histogram: A fast way to extract histograms in cartesian spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 829–836.

[67] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[68] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1745–1752.

[69] W. Schwartz, A. Kembhavi, D. Harwood, and L. Davis, "Human detection using partial least squares analysis," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, 2009, pp. 24–31.

[70] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. Lecun, "Pedestrian detection with unsupervised and multi-stage feature learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3626–3633.

[71] V. D. Shet, J. Neumann, V. Ramesh, and L. S. Davis, "Bilattice-based logical reasoning for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[72] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1585–1592.

[73] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 723–730.

[74] S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele, "Learning people detectors for tracking in crowded scenes," *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1049–1056.

[75] S. Tang, M. Andriluka, and B. Schiele, "Detection and tracking of occluded people," in *Proc. British Mach. Vis. Conf.*, Surrey, U.K., 2012, pp. 9.1–9.11.

[76] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.

[77] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 606–613.

[78] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[79] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, 2005.

[80] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1030–1037.

[81] M. Wang, W. Li, and X. Wang, "Transferring a generic pedestrian detector towards specific scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3274–3281.

[82] M. Wang and X. Wang, "Automatic adaptation of a generic pedestrian detector to a specific traffic scene," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3401–3408.

[83] X. Wang, X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 32–39.

[84] Y. Wang, D. Tran, and Z. Liao, "Learning hierarchical poselets for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1705–1712.

[85] C. Wojek and B. Schiele, "A performance evaluation of single and multi-feature people detection," in *Proc. 30th DAGM Symp. Pattern Recognit.*, 2008, pp. 82–91.

[86] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 794–801.

[87] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 90–97.

[88] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, 2007.

[89] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses," *Int. J. Comput. Vis.*, vol. 82, no. 2, pp. 185–204, 2009.
[90] J. Yan, Z. Lei, L. Wen, and S. Z. Li, "The fastest deformable part model for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2497–2504.
[91] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Multi-pedestrian detection in crowded scenes: A global view," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3124–3129.
[92] J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li, "Robust multi-resolution pedestrian detection in traffic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3033–3040.
[93] Y. Yang, S. Baker, A. Kannan, and D. Ramanan, "Recognizing proxemics in personal photos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3522–3529.
[94] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1385–1392.
[95] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 17–24.
[96] X. Zeng, W. Ouyang, M. Wang, and X. Wang, "Deep learning of scene-specific classifier for pedestrian detection," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2014, pp. 472–487.
[97] X. Zeng, W. Ouyang, and X. Wang, "Multi-stage contextual deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 121–128.
[98] L. Zhu, Y. Chen, A. Yuille, and W. Freeman, "Latent hierarchical structural learning for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1062–1069.
[99] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1491–1498.

Wanli Ouyang received the BS degree in computer science from Xiangtan University, Hunan, China, in 2003. He received the MS degree in computer science from the College of Computer Science and Technology, Beijing University of Technology, Beijing, China. He received the PhD degree in the Department of Electronic Engineering, The Chinese University of Hong Kong, where he is currently a research assistant professor. His research interests include image processing, computer vision, and pattern recognition. He is a member of the IEEE.

Xingyu Zeng received the BS degree in electronic engineering and information science from the University of Science and Technology of China in 2011. He is currently working toward the PhD degree in the Department of Electronic Engineering at the Chinese University of Hong Kong. His research interests include computer vision and deep learning.

Xiaogang Wang received the BS degree from the University of Science and Technology of China in 2001, the MS degree from the Chinese University of Hong Kong in 2003, and the PhD degree from the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology in 2009. He is currently an assistant professor in the Department of Electronic Engineering at The Chinese University of Hong Kong. His research interests include computer vision and machine learning. He is a member of the IEEE.