Noname manuscript No.
(will be inserted by the editor)

Learning Mutual Visibility Relationship for Pedestrian Detection with a Deep Model

Wanli Ouyang · Xingyu Zeng · Xiaogang Wang

Received: date / Accepted: date

Abstract Detecting pedestrians in cluttered scenes is a challenging problem in computer vision. The difficulty is added when several pedestrians overlap in images and occlude each other. We observe, however, that the occlusion/visibility statuses of overlapping pedestrians provide useful mutual relationship for visibility estimation - the visibility estimation of one pedestrian facilitates the visibility estimation of another. In this paper, we propose a mutual visibility deep model that jointly estimates the visibility statuses of overlapping pedestrians. The visibility relationship among pedestrians is learned from the deep model for recognizing co-existing pedestrians. Then the evidence of co-existing pedestrians is used for improving the single pedestrian detection results. Compared with existing image-based pedestrian detection approaches, our approach has the lowest average miss rate on the Caltech-Train dataset and the ETH dataset. Experimental results show that the mutual visibility deep model effectively improves the pedestrian detection results. The mutual visibility deep model leads to $6\% - 15\%$ improvements on multiple benchmark datasets.

Keywords Deep model, deep learning, pedestrian detection, object detection

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).

Wanli Ouyang, Xingyu Zeng, and Xiaogang Wang
Department of Electronic Engineering, The Chinese University of Hong Kong,
Tel.: +852-39434461
E-mail: {wlouyang, xyzeng, xgwang}@ee.cuhk.edu.hk

1 Introduction

Pedestrian detection is the task of locating pedestrians from images. It is important for applications such as video surveillance, robotics and automotive safety. Pedestrian detection results can be used as input for pedestrian tracking, person re-identification and action recognition.

Impressive progress in pedestrian detection has been achieved for better accuracy (Dai et al, 2007, Duan et al, 2010, Enzweiler et al, 2010, Wang et al, 2009, Wu and Zhu, 2011, Wu and Nevatia, 2005, Shet et al, 2007, Lin et al, 2007, Wu and Nevatia, 2009, Leibe et al, 2005, Park et al, 2010, Ding and Xiao, 2012, Yan et al, 2012, Chen et al, 2013, Mathias et al, 2013, Marın et al, 2013) and faster speed (Viola et al, 2005, Dean et al, 2013, Dollár et al, 2012a, Benenson et al, 2012, Dollár et al, 2010, 2014). However, pedestrian detection is still a challenging task due to the background clutter and the intra-class variation of pedestrians in clothing, viewpoint, lighting, and articulation.

When several pedestrians overlap in the image region, some will be occluded by others and the expected visual cues of the occluded parts are corrupted, resulting in the added difficulty in detection. The examples of overlapping pedestrians in Fig. 1 (a) often appear in real world applications.

Pedestrians with overlaps are difficult to detect, however, we observe that these pedestrians have useful mutual visibility relationship information. When pedestrians are found to overlap in the image region, there are two types of visibility relationships among their parts:

1. Compatible relationship. It means that the observation of one part is a positive indication of the other part. There are two parts, i.e. left-half part and right-half part, for each pedestrian in Fig. 1 (a). In Fig. 1 (a), given the prior knowledge that there are two pedestrians co-existing side by side, the right-half part of the *left* pedestrian Lena is compatible

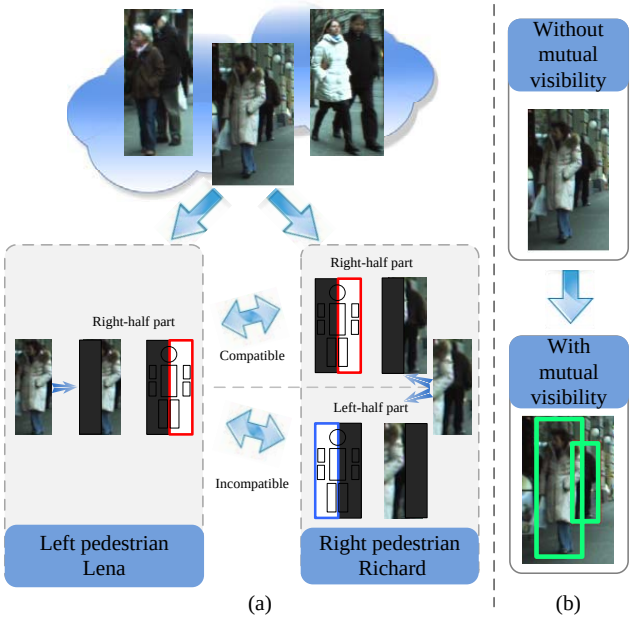


Fig. 1 **(a)** Mutual visibility relationship of parts among pedestrians and **(b)** detection results comparison of the approach without modeling mutual visibility in (Ouyang and Wang, 2012) and our approach modeling mutual visibility. The right-half of Lena is *compatible* with the *right-half* of Richard but *incompatible* with the *left-half* of Richard. With mutual visibility modeled in our approach, the missed pedestrians are found by modeling the visibility relationship among parts.

with the right-half part of the *right* pedestrian Richard[1] because these two parts often co-exist in positive training examples. The compatible relationship can be used for increasing the visibility confidence of mutually compatible pedestrian parts. Take Fig. 1 (b) as an example, if a pedestrian detector detects both Lena on the left and Richard on the right with high false positive rate, then the visibility confidence of Lena's right-half part increases when Richard's right-half part is found to be visible. And the detection confidence of Lena correspondingly increases. In this example, the compatible relationship helps to detect Lena in Fig. 1 (b).

2. Incompatible relationship. It means that the occlusion of one part indicates the visibility of the other part, and vice versa. For the example in Fig. 1 (b), Lena and Richard have so strong overlap that one occludes the other. In this case, Lena's right-half part and Richard's left-half part are incompatible because they overlap and shall not be visible simultaneously. If a pedestrian detector detects both Lena and Richard with high false positive rate in Fig. 1 (b), then the visibility confidence of Lena's right-half part increases when Richard's left-half part is found to be invisible. And Lena's detection confidence is correspondingly increased. Therefore, incompatible relationship helps to detect Lena in this example.

[1] 'Lena' and 'Richard' are used as placeholder names in this paper.

These observations motivate us to jointly estimate the occlusion status of co-existing pedestrians by modeling the mutual visibility relationship among their parts. In this paper, we propose to learn the compatible and incompatible relationship by a deep model.

The main contribution of this paper is to jointly estimate the visibility statuses of multiple pedestrians and recognize co-existing pedestrians via a mutual visibility deep model. Overlapping parts of co-existing pedestrians are placed at multiple layers in this deep model. With this deep model, 1) overlapping parts at different layers verify the visibility of each other for multiple times; 2) the complex probabilistic connections across layers are modeled with good efficiency on both learning and inference. The deep model is suitable for modeling the mutual visibility relationship because: 1) the hierarchical structure of the deep model matches with the multilayers of the parts model; 2) overlapping parts at different layers verify the visibility of each other for multiple times in the deep model; 3) the complex probabilistic connections across layers of parts are modeled with good efficiency on both learning and inference. The mutual visibility deep model effectively improves pedestrian detection performance with less than 5% extra computation in the detection process. Compared with image-based approaches, it achieves the lowest average miss rate on the Caltech-Train dataset and the ETH dataset. On the more challenging PETS dataset labeled by us, including mutual visibility leads to 10% improvement on the lowest average miss rate. Furthermore, our model takes part detection scores as input and it is complementary to many existing pedestrian approaches. It has good flexibility to integrate with other techniques, such as more discriminative features (Walk et al, 2010), scene geometric constraints (Park et al, 2010), richer part models (Zhu et al, 2010, Yang and Ramanan, 2011) and contextual multi-pedestrian detection information (Tang et al, 2012, Ouyang and Wang, 2013b, Yan et al, 2012) to further improve the performance.

2 Related Work

When pedestrians are occluded, the supposed visual cue is missing, and the performance of generic detectors in detecting them degrades severely. Since visibility estimation is the key to handle occlusions, many approaches were proposed for estimating visibility of parts (Dai et al, 2007, Duan et al, 2010, Enzweiler et al, 2010, Wang et al, 2009, Wu and Zhu, 2011, Wu and Nevatia, 2005, Shet et al, 2007, Lin et al, 2007, Wu and Nevatia, 2009, Leibe et al, 2005, Mathias et al, 2013, Marın et al, 2013). Wang *et al.* (Wang et al, 2009) used the block-wise HOG+SVM scores to estimate visibility status and combined the full-body classifier and part-based classifiers by heuristics. Enzweiler *et*

al. (Enzweiler et al, 2010) estimated the visibility of different parts using motion, depth and segmentation and then computed the classification score by summing up multiple visibility weighted cues of parts. Substructures were used in (Dai et al, 2007, Duan et al, 2010). Each substructure was composed of a set of part detectors. And the detection confidence score of an object was determined by the existence of these substructures. A set of occlusion-specic classiers were trained by a new Franken-classier in (Mathias et al, 2013). The occlusion-specic classiers, called local experts in (Marın et al, 2013) were combined by random forests in (Marın et al, 2013). The And-Or graph was used in (Wu and Zhu, 2011) to accumulate hard-thresholded part detection scores. Deformable part based model (DPM) was used in (Sadeghi and Farhadi, 2011, Li et al, 2011, Tang et al, 2012, Pepikj et al, 2013, Desai and Ramanan, 2012, Tang et al, 2013) to learn occlusion patterns and contextual multi-object detectors. The occlusion patterns learned in existing approaches (Pepikj et al, 2013, Desai and Ramanan, 2012, Ouyang and Wang, 2013a) can be used as the input of our deep model. Therefore, our approach is helpful in improving the ability of part detectors in handling occlusion using both non-occlusion patterns and occlusion patterns for pedestrian detection. As a solid example, we show in the experimental results that the occlusion patterns learned in (Ouyang and Wang, 2013a) can be used by our model for achieving a performance that is better than the performance without deep model or without occlusion patterns. Recently, the approaches in (Duan et al, 2010, Ouyang and Wang, 2012, 2013a) utilized the visibility relationship among parts for isolated pedestrian. However, the part visibility relationship among co-existing pedestrians was not explored in (Dai et al, 2007, Duan et al, 2010, Enzweiler et al, 2010, Wang et al, 2009, Wu and Zhu, 2011, Ouyang and Wang, 2013a). In order to handle inter-human occlusions, the joint part-combination of multiple humans was adopted in (Wu and Nevatia, 2005, Shet et al, 2007, Lin et al, 2007, Wu and Nevatia, 2009, Leibe et al, 2005). These approaches obtain the visibility status by occlusion reasoning using 2-D visibility scores in (Wu and Nevatia, 2005, Shet et al, 2007, Lin et al, 2007) or using segmentation results in (Wu and Nevatia, 2009, Leibe et al, 2005). They manually defined the incompatible relationship among parts of multiple pedestrians through the exclusive occupancy of segmentation region or part detection response, while our approach learns the incompatible relationship from training data. In addition, the compatible relationship was not used by these approaches.

The articulation relationship among the parts of multiple objects, parameterized by position, scale, size, rotation, was investigated as context (Yao and Fei-Fei, 2010, Yan et al, 2012, Yang et al, 2012, Desai et al, 2009). Nearby detection scores was considered as context in (Ding and Xiao, 2012, Chen et al, 2013, Zeng et al, 2013). 2-pedestrian detectors were developed in (Ouyang and Wang, 2013b, Tang et al, 2012, 2013, Pepikj et al, 2013) for capturing contextual information. But they did not consider the visibility relationship of co-existing pedestrians, which is the focus of our approach. The part visibility relationship among co-existing pedestrians has not been investigated yet and is complementary to these context-based approaches. In this paper, we develop an extended model to show that the contextual information can be used in our deep model. The deep model integrates the contextual 2-pedestrian detection information extracted using the approach in (Ouyang and Wang, 2013b) to improve the detection accuracy.

Deep learning methods aim at learning feature hierarchies, in which more abstract feature representations at higher levels are composed by lower-level features (Bengio et al, 2013). Excellent review of deep learning is provided in (Bengio et al, 2013, Bengio, 2009). Deep model has been applied for dimensionality reduction (Hinton and Salakhutdinov, 2006) (Bengio et al, 2013), hand written digit recognition (Hinton et al, 2006, Lee et al, 2009, Norouzi et al, 2009), object recognition (Jarrett et al, 2009, Lee et al, 2009, Le et al, 2012), face parsing (Luo et al, 2012), face recognition (Sun et al, 2014, Hu et al, 2014), action recognition (Ji et al, 2013, SUN et al, 2014), facial expression recognition and scene recognition (Liu et al, 2014, Ranzato et al, 2011, Farabet et al, 2013). Hinton *et al.* (Hinton et al, 2006) proved that adding a new layer, if done correctly, creates a model that has a better variational lower bound on the log probability of the training data than the previous shallower model. Krizhevsky *et al.* (Krizhevsky et al, 2012) proposed a deep model that achieved state-of-the-art performance for object recognition on the ImageNet dataset (Deng et al, 2009). Overfeat (Sermanet et al, 2013a) achieved very good object detection performance on the ImageNet Large Scale Visual Recognition Challenge 2013. Our deep model has some difference with conventional deep models in spirit. Conventional deep models assume that hidden variables had no semantic meaning and learn many layers of representation from raw data or rich feature representations; our model assigns semantic meaning to hidden nodes and uses the deep model for learning the visibility relationship from compact part detection scores. Recently, deep model was used for pedestrian detection in (Ouyang and Wang, 2012, Norouzi et al, 2009, Sermanet et al, 2013b,a, Ouyang and Wang, 2013a, Zeng et al, 2013). The approaches in (Ouyang and Wang, 2012, Norouzi et al, 2009, Sermanet et al, 2013b, Krizhevsky et al, 2012, Ouyang and Wang, 2013a, Zeng et al, 2013) focused on isolated objects or pedestrians. This paper focuses on co-existing pedestrians, which has not been considered in these works.

Algorithm 1: Overview of our pedestrian detection approach.

Input: $\mathbf{x}_1$ and $\mathbf{x}_2$, which are respectively the features of window 1 and 2.
Output: $p(y_1|\mathbf{x}_1, \mathbf{x}_2)$, which is the probability that window 1 with feature $\mathbf{x}_1$ has has a pedestrian.
1 obtain part detection scores $\mathbf{S}$ from $\mathbf{x}_1$ and $\mathbf{x}_2$ by part detector, which is the deformable part based model in our experiment;
2 estimate $p(y_1|y_2 = 0, \mathbf{x}_1)$ in (2) and $\phi(y; \mathbf{x})$ in (3) with the deep model in Section 4;
3 estimate $\phi_p(y; \mathbf{x})$ in (3) with GMM;
4 $p(y_1|\mathbf{x}_1, \mathbf{x}_2) = p(y_1, y_2 = 0|\mathbf{x}_1, \mathbf{x}_2) + p(y_1, y_2 = 1|\mathbf{x}_1, \mathbf{x}_2)$.

3 Overview of our approach

In this paper, we mainly discuss the approach for pair-wise pedestrians. The extension of the deep model to more pedestrians is discussed in Section 5. Denote the features of detection window wnd_1 by vector $\mathbf{x}_1$, containing both appearance and position information. Denote the label of wnd_1 by $y_1 = 0, 1$ for negative and positive samples respectively. Pedestrian detection with a discriminative model aims at obtaining $p(y_1|\mathbf{x}_1)$ for each window wnd_1 in a sliding window manner for all sizes of windows. We consider another detection window wnd_2 with features $\mathbf{x}_2$ and label $y_2 = 0, 1$. And we have the following by marginalizing y_2:

$$\begin{aligned} p(y_1|\mathbf{x}_1, \mathbf{x}_2) &= \sum_{y_2=0,1} p(y_1, y_2|\mathbf{x}_1, \mathbf{x}_2) \\ &= p(y_1, y_2 = 1|\mathbf{x}_1, \mathbf{x}_2) + p(y_1, y_2 = 0|\mathbf{x}_1, \mathbf{x}_2). \end{aligned} \quad (1)$$

When $y_2 = 0$, we suppose

$$p(y_1, y_2 = 0|\mathbf{x}_1, \mathbf{x}_2) = p(y_1|y_2 = 0, \mathbf{x}_1)p(y_2 = 0|\mathbf{x}_2), \quad (2)$$

where $p(y_1|y_2 = 0, \mathbf{x}_1)$ and $p(y_2 = 0|\mathbf{x}_2) = 1 - p(y_2 = 1|\mathbf{x}_2)$ are obtained from the deep model for isolated pedestrians.

When $y_2 = 1$, we have

$$p(y_1, y_2 = 1|\mathbf{x}_1, \mathbf{x}_2) \propto \phi(y; \mathbf{x})\phi_p(y; \mathbf{x}), \quad (3)$$

$\phi(y; \mathbf{x})$ in (3) is used for recognizing pair-wise co-existing pedestrians from part detection scores, where $\mathbf{x} = [\mathbf{x}_1^{\mathrm{T}}\ \mathbf{x}_2^{\mathrm{T}}]^{\mathrm{T}}$, $y = 1$ if $y_1 = 1$ and $y_2 = 1$, otherwise $y = 0$. Both $p(y_1|y_2 = 0, \mathbf{x}_1, \mathbf{x}_2)$ in (2) and $\phi(y; \mathbf{x})$ in (3) are obtained from the deep model introduced in Section 4. $\phi(y; \mathbf{x})$ is detailed in (8). $\phi_p(y; \mathbf{x})$ in (3) models probability for their relative position between wnd_1 and wnd_2. $\phi_p(y; \mathbf{x})$ is estimated from Gaussian mixture model (GMM). An overview of our approach is given in Algorithm 1.

4 The mutual visibility deep model

Since the visibility relationship of parts between pair-wise pedestrians is different when pedestrians have different relative positions, the relative positions are clustered into K mixtures using GMM. $K = 9$ in our experiments. And K deep models are trained for these K mixtures. According to our experimental results on Caltech-train, ETH and PETS, the average miss rate increases by less than 1% when $K = 3$ compared with $K = 9$. A pair of detection windows are classified into the kth mixture and then this pair are used by the kth deep model for learning and inference. The differences between the two pedestrians in horizontal location, vertical location and size, denoted by (d_x, d_y, d_s), are used as the random variables in the GMM distribution $p(d_x, d_y, d_s)$. Positive samples are used for training $p(d_x, d_y, d_s)$. $\phi_p(y; \mathbf{x})$ in (3) is obtained from $p(d_x, d_y, d_s)$ as follows:

$$p(d_x, d_y, d_s) \propto \sum_{i=1}^{9} \lambda_i e^{(d_x-\mu_{x,i})^2+(d_y-\mu_{y,i})^2+(d_s-\mu_{s,i})^2}. \quad (4)$$

where $(\mu_{x,i}, \mu_{y,i}, \mu_{s,i})$ is the ith mean of (d_x, d_y, d_s).

4.1 Preparation of part scores and overlap information

4.1.1 The parts model

Fig. 2 shows the parts model used for pedestrian 1 at window wnd_1. The parts model for pedestrian 2 at window wnd_2 is the same. As shown in Fig. 2, there are 3 layers of parts with different sizes: six small parts at layer 1, seven medium-sized parts at layer 2, and seven large parts at Layer 3. The six parts at layer 1 are left-head-shoulder, right-head-shoulder, left-torso, right-torso, left-leg and right-leg. A part at an upper layer consists of its children at the lower layer. The parts at the top layer are the possible occlusion statuses with gray color indicating occlusions. With the parts defined, we use the DPM in (Felzenszwalb et al, 2010) for jointly modeling the deformation and appearance of these parts.

4.1.2 Preparation of part detection score

With the parts model defined in Section 4.1.1, part scores are obtained for these parts. Denote $\mathbf{S} = [\mathbf{s}^{1^{\mathrm{T}}}, \ldots, \mathbf{s}^{L^{\mathrm{T}}}]^{\mathrm{T}} = \gamma(\mathbf{x})$ as the part scores of L layers, where $\gamma(\mathbf{x})$ is obtained from part detectors, $\mathbf{s}^l$ for $l = 1, \ldots, L$ denotes the scores at layer l, $L = 3$ in Fig. 2. And we have $\mathbf{s}^l = [\mathbf{s}_1^{l^{\mathrm{T}}}\ \mathbf{s}_2^{l^{\mathrm{T}}}]^{\mathrm{T}}$, where the P^l scores of the pedestrian 1 and pedestrian 2 at layer l are denoted by $\mathbf{s}_1^l = [s_{1,1}^l, \ldots, s_{1,P^l}^l]^{\mathrm{T}}$ and $\mathbf{s}_2^l =$

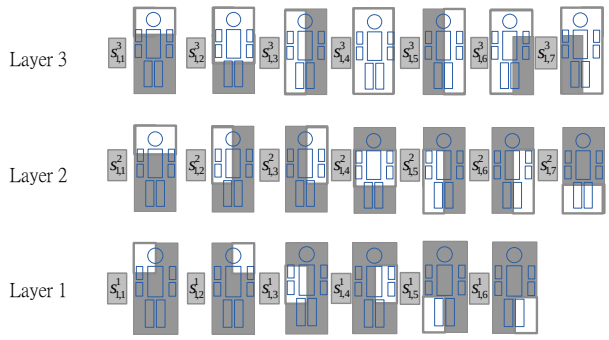


Fig. 2 The parts model for isolated pedestrians. $s^1_{1,1}$ corresponds to the score of the left-head-shoulder part at layer 1 for pedestrian 1.

$[s^1_{2,1}, \ldots, s^l_{2,P^l}]^{\mathrm{T}}$ respectively. In our implementation, HOG features and the DPM in (Felzenszwalb et al, 2010) are used for obtaining part detection scores in $\mathbf{S}$.

The deformation among parts are arranged in the star-model with full-body being the center. In this paper, it is assumed that part-based models have integrated both appearance and deformation scores into $\mathbf{S}$. And we have:

$$\tilde{s}^l_{1,i} = \max_{\mathbf{z}_{1,i}} \mathbf{F}^{\mathrm{T}}_{1,a}\psi_1(\mathbf{I}; \mathbf{z}_{1,i}) + \mathbf{F}^{\mathrm{T}}_{1,d}\psi_d(\mathbf{z}_{1,i} - \mathbf{a}_1), \quad (5)$$

where $\psi_1(\mathbf{I}; \mathbf{z}_{1,i})$ denotes the feature of the ith part at location $\mathbf{z}_{1,i}$, $\psi_d(\mathbf{z}_{1,i} - \mathbf{a}_{1,i})$ denotes the deformation term, $\mathbf{F}^{\mathrm{T}}_{1,a}$ and $\mathbf{F}^{\mathrm{T}}_{1,d}$ are parameters learned using latent SVM. $\mathbf{a}_{1,i}$ denotes the anchor position of part i. Similar definition applies for $s^l_{2,i}$.

In order to have the top layer representing occlusion status in a more direct way, $\mathbf{s}^3$ at the top layer accumulate the part detection scores that fit their possible occlusion statuses. For example,

$$\begin{aligned} s^3_{1,1} &= \tilde{s}^3_{1,1} + \tilde{s}^2_{1,1} + \sum_{i=1}^{2} \tilde{s}^1_{1,i}, \\ s^3_{1,2} &= \tilde{s}^3_{1,2} + \sum_{i=1}^{4} \tilde{s}^2_{1,i} + \sum_{i=1}^{4} \tilde{s}^1_{1,i}, \\ s^3_{1,3} &= \tilde{s}^3_{1,3} + \tilde{s}^2_{1,2} + \tilde{s}^2_{1,5} + \tilde{s}^1_{1,1} + \tilde{s}^1_{1,3} + \tilde{s}^1_{1,5}. \end{aligned} \quad (6)$$

$s^l_{n,i} = \tilde{s}^l_{n,i}$, for $n = 1, 2, l = 1, 2$. For the ith part at layer l, $l = 1, 2, 3, i = 1, \ldots, P^l$, $s^l_{1,i}$ is the accumulated score used for the deep model while $\tilde{s}^l_{1,i}$ is the part detection score obtained from DPM. Only one DPM is used for obtaining the detection scores. For example, $\tilde{s}^3_{1,2}$ is the score obtained from DPM for the head-torso part at layer 3 and $\tilde{s}^3_{1,3}$ is the score obtained from DPM for the left-half part at layer 3. The accumulated score $s^3_{1,3}$ in (6) accumulates the scores from DPM for left-head-shoulder $\tilde{s}^1_{1,1}$, left-torso $\tilde{s}^1_{1,3}$, and left-leg $\tilde{s}^1_{1,5}$ at layer 1, the scores for left-head-torso $\tilde{s}^2_{1,2}$ and left-torso-leg $\tilde{s}^2_{1,5}$ at layer 2, and the the score for left-half $\tilde{s}^3_{1,3}$ at layer 3. In our implementation of the detector, the head-shoulder part at the top layer has half of the resolution of HOG features compared with the head-shoulder part at the middle layer. Therefore, the occlusion status for the head-shoulder part at the top layer has considered visual cues of two different resolutions.

4.1.3 Preparation of overlap information

The overlap information at layer 2 in Fig. 3 is denoted by $\mathbf{o} = [\mathbf{o}_1^{\mathrm{T}}\ \mathbf{o}_2^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{o}_n = [o_{n,1}\ o_{n,2} \ldots o_{n,6}]^{\mathrm{T}}$ for $n = 1, 2$. The overlap information for six parts are left-head-shoulder $o_{n,1}$, right-head-shoulder $o_{n,2}$, left-torso $o_{n,3}$, right-torso $o_{n,4}$, left-leg $o_{n,5}$ and right-leg $o_{n,6}$. In order to obtain $\mathbf{o}$, the overlap of these six parts with the pedestrian region of the other pedestrian is computed. According to the average silhouette in Fig. 4(a), which is obtained by averaging the gradient of positive samples, two rectangles are used for approximating the pedestrian region of the other pedestrian. One rectangle is used for the head region, another rectangle is used for the torso-leg region. The union of these two rectangular regions is denoted by $A_{n'}$. Denote the region for $o_{n,i}$ by $A_{n,i}$. $o_{n,i}$ is obtained as follows:

$$o_{n,i} = \frac{area(A_{n,i} \cap A_{n'})}{area(A_{n,i})}, \quad (7)$$

where $area(\cdot)$ computes the area in this region, $\cap$ denotes intersection of region. For example, the right person in Fig. 4(b) has the left-head-shoulder, left-torso and left-leg overlapping with the pedestrian regions of the left person. The operations $area(\cdot)$ and $\cap$ in (7) can be efficiently computed using the coordinates of rectangles instead of being computed in a pixel-wise way on the rectangular regions. The overlap information $\mathbf{o}$ can also be obtained from segmentation. Compared with segmentation, the rectangular region is an approximate but faster approach for obtaining pedestrian region and computing the overlap information $\mathbf{o}$. As shown in Fig. 3, $\mathbf{o}$ is an indication on how much occlusion may be caused when the two pedestrians overlap in the image.

4.2 The inference and learning of the deep model

For the proposed deep model, this section illustrates the inference of hidden variables $\hat{\mathbf{h}}$ and learning of parameters $\mathbf{w}^l_{*,j}, \mathbf{w}_L, g^{l+1}_j$, and c^{l+1}_j. These symbols are illustrated in this section.

4.2.1 The deep model for inference

Fig. 3(a) shows the deep model used at the inference stage. The visibilities for the parts with scores $\mathbf{s}^l_1$ and $\mathbf{s}^l_2$ are denoted by $\hat{\mathbf{h}}^l_1 = [h^1_{1,1}, \ldots, h^l_{1,P^l}]^{\mathrm{T}}$ and $\hat{\mathbf{h}}^l_2 = [h^1_{2,1}, \ldots, h^l_{2,P^l}]^{\mathrm{T}}$ respectively. $\hat{\mathbf{h}}^l_1$ and $\hat{\mathbf{h}}^l_2$ are grouped into $\hat{\mathbf{h}}^l = [\hat{\mathbf{h}}^{l\,\mathrm{T}}_1\ \hat{\mathbf{h}}^{l\,\mathrm{T}}_2]^{\mathrm{T}}$. $\hat{\mathbf{h}}^l$ is considered as a vector containing

hidden variables because the visibilities of parts are not provided at the training stage or the testing stage. For the deep model in Fig. 3(a), $\hat{\mathbf{h}}_n^l$ is the vector of hidden nodes at layer l for pedestrian n. For example, $s_{1,1}^1$ is the input score for the left-head-shoulder part of pedestrian 1 and $h_{1,1}^1$ is the hidden node for its visibility inferred from the deep model.

At the inference stage, the pedestrian co-existence label y is inferred using the following model, which is shown in Fig. 3:

$$\begin{aligned}
&h_j^1 = \sigma(c_j^1 + {g_j^1}^T s_j^1),\\
&h_j^{l+1} = \sigma({\mathbf{h}^l}^{\mathrm{T}} \mathbf{w}_{*,j}^l + c_j^{l+1} + {g_j^{l+1}}^T s_j^{l+1}), l = 1, \ldots L-1,\\
&\phi(y; \mathbf{x}) = \sigma(\mathbf{w}_L^{\mathrm{T}} \mathbf{h}^L + b),\\
&\text{where } \sigma(a) = 1/(1+exp(-a)),\\
&\mathbf{h}^l = \hat{\mathbf{h}}^l \text{ if } l \neq L-1, \mathbf{h}^l = [{\hat{\mathbf{h}}^l}^{\mathrm{T}}\ \mathbf{o}^{\mathrm{T}}]^{\mathrm{T}}, \text{ if } l = L-1,
\end{aligned} \tag{8}$$

Equation (8) is illustrated as follows:

- $\mathbf{S} = [{\mathbf{s}^1}^{\mathrm{T}} \ldots {\mathbf{s}^L}^{\mathrm{T}}]^{\mathrm{T}}$ contains the part scores introduced in Section 4.1.2. s_j^l is the jth element in vector $\mathbf{s}^l$.
- g_j^{l+1} is the weight for detection score s_j^{l+1}.
- c_j^{l+1} is the bias.
- $\hat{\mathbf{h}}^l$, the vector of units at the lth layer in the deep model, contains the part visibilities for the two pedestrians at layer l.
- h_j^l, the jth element in $\hat{\mathbf{h}}^l$, is the estimated part visibility probability for the jth part at the lth layer, $l = 1, \ldots, L$. $L = 3$ in Fig. 3.
- $\phi(y; \mathbf{x})$ is the estimated value for pedestrian co-existence label y, which is used for (3).
- $\mathbf{w}_{*,j}^l$ for $l = 1, \ldots, L-1$ models the relationship between the visibilities in $\mathbf{h}^l$ at layer l and the visibility $\tilde{h}_j^{l+1}$ at layer $l+1$.
- $\mathbf{w}_L$ is the weight used for estimating the classification label y.
- $\mathbf{o}$ is the overlap information with details in Section 4.1.3.

$\mathbf{w}_{*,j}^l, \mathbf{w}_L, g_j^{l+1}$, and c_j^{l+1} are parameters to be learned. The learning of them is explained in Section 4.2.2. In this deep model, the visibility of h_j^{l+1} at the upper layer is estimated from its detection score s_j^{l+1} and the correlated visibilities $\mathbf{h}^l$ at the lower layer. In this estimation, g_j^{l+1} is the weight for detection score s_j^{l+1} and $\mathbf{w}_l$ is the weight for $\mathbf{h}^l$. And then the estimated pedestrian co-existence label $\phi(y; \mathbf{x})$ is obtained from the hidden variables in the last layer $\mathbf{h}^L$.

As shown in (8), this paper puts the input data $\mathbf{S}$ at multiple layers while existing DNN put the input data at the bottom layer. Each hidden node is connected with a single part detection score so that a hidden node explicitly corresponds to a body part in our model while a hidden node does not have explicit correspondence to a body part in conventional deep models.

4.2.2 The learning of the deep model

The following two stages are used for learning the parameters in (8).

Stage 1: Pretrain parameters $\mathbf{w}_{*,j}^l, c_j^{l+1}$, and g_j^{l+1} in (8).

Stage 2: Fine-tune all the parameters by backpropagating error derivatives. The variables are arranged as a backpropagation (BP) network as shown in Fig. 3(a).

As stated in (Erhan et al, 2010), unsupervised pretraining guides the learning of the deep model towards the basins of attraction of minima that support better generalization from the training data. Therefore, we adopt unsupervised pretraining of parameters at stage 1. The graphical model for unsupervised pretraining is shown in Fig. 3(d). The probability distribution of $p(\mathbf{h}^1, \ldots, \mathbf{h}^L, \mathbf{S})$ is modeled as follows:

$$\begin{aligned}
&p(\mathbf{h}^1, \ldots, \mathbf{h}^L, \mathbf{S})\\
&= \left(\prod_{l=1}^{L-2} p(\mathbf{s}^l|\mathbf{h}^l)\right)\left(\prod_{l=1}^{L-2} p(\mathbf{h}^l|\mathbf{h}^{l+1})\right) p(\mathbf{h}^{L-1}, \mathbf{h}^L, \mathbf{s}^L),\\
&p(s_i^l = 1|h_i^l) = \sigma(g_i^l h_i^l + b_i^l),\\
&p(h_i^l = 1|\mathbf{h}^{l+1}) = \sigma(\mathbf{w}_{i,*}^l \mathbf{h}^{l+1} + c_i^l),\\
&p(\mathbf{h}^{L-1}, \mathbf{h}^L, \mathbf{s}^L)\\
&= e^{\left[{\mathbf{h}^{L-1}}^{\mathrm{T}} \mathbf{W}^{L-1} \mathbf{h}^L + {\mathbf{c}^{\mathbf{L-1}}}^{\mathrm{T}} \mathbf{h}^{L-1} + (\mathbf{c}^{\mathbf{L}} + \mathbf{g}^{\mathbf{L}} \circ \mathbf{s}^L)^{\mathrm{T}} \mathbf{h}^L + {\mathbf{b}^L}^{\mathrm{T}} \mathbf{s}^L\right]},
\end{aligned} \tag{9}$$

where $\circ$ denotes the entrywise product, i.e. $(A \circ B)_{i,j} = A_{i,j} B_{i,j}$, $\mathbf{h}$ is defined in (8). For the model in Fig. 3(d), we have $L = 3$. $\mathbf{W}^l$, g_i,^{l} and c_i^l are the parameters to be learned. $\mathbf{W}^l$ models the correlation between $\mathbf{h}^l$ and $\mathbf{h}^{l+1}$, $\mathbf{w}_{i,*}^l$ is the ith row of $\mathbf{W}^l$, g_i^l is the weight for s_i^l, and c_i^l is the bias term. The element $w_{i,j}^l$ of $\mathbf{W}^l$ in (9) is set to zero if there is no connection between units h_i^l and h_j^{l+1} in Fig. 3(b).

Similar to the approach in (Hinton et al, 2006), the parameters in (9) are trained layer by layer and two adjacent layers are considered as a Restricted Boltzmann Machine (RBM) that has the following distributions:

$$\begin{aligned}
&p(\mathbf{h}^l, \mathbf{s}^{l+1}, \mathbf{h}^{l+1}|\mathbf{s}^l)\\
&\propto e^{\left[{\mathbf{h}^l}^{\mathrm{T}} \mathbf{W}^l \mathbf{h}^{l+1} + (\mathbf{c}^l + \mathbf{g}^{\mathbf{l}} \circ \mathbf{s}^l)^{\mathrm{T}} \mathbf{h}^l + (\mathbf{c}^{l+1} + \mathbf{g}^{l+1} \circ \mathbf{s}^{l+1})^{\mathrm{T}} \mathbf{h}^{l+1} + {\mathbf{b}^{l+1}}^{\mathrm{T}} \mathbf{s}^{l+1}\right]},\\
&p(h_i^l = 1|\mathbf{h}^{l+1}, \mathbf{s}^l) = \sigma(\mathbf{w}_{i,*}^l \mathbf{h}^{l+1} + c_i^l + {g_i^l}^{\mathrm{T}} s_i^l),\\
&p(h_j^{l+1} = 1|\mathbf{h}^l, \mathbf{s}^{l+1}) = \sigma({\mathbf{h}^l}^{\mathrm{T}} \mathbf{w}_{*,j}^l + c_j^{l+1} + {g_j^{l+1}}^{\mathrm{T}} s_j^{l+1}),\\
&p(s_i^{l+1} = 1|h_i^{l+1}) = \sigma(g_i^{l+1} h_i^{l+1} + b_i^{l+1}),
\end{aligned} \tag{10}$$

where $\mathbf{w}^l_{i,*}$ is the ith row of matrix $\mathbf{W}^l$ and $\mathbf{w}^l_{*,j}$ is the jth column of $\mathbf{W}^l$. The gradient of the log-likelihood for this RBM is computed as follows:

$$\begin{aligned}\frac{\partial L(\mathbf{S})}{\partial w^l_{i,j}} &\propto (< h^l_i h^{l+1}_j >_{data} - < h^l_i h^{l+1}_j >_{model}),\\ \frac{\partial L(\mathbf{S})}{\partial c^l_i} &\propto (< h^l_i >_{data} - < h^l_i >_{model}),\\ \frac{\partial L(\mathbf{S})}{\partial g^l_i} &\propto (< h^l_i s^l_i >_{data} - < h^l_i s^l_i >_{model}),\end{aligned} \tag{11}$$

where $w^l_{i,j}$ is the (i,j)th element in matrix $\mathbf{W}^l$, $< \cdot >_{data}$ denotes the expectation with respect to the distribution $p(\mathbf{h}^l, \mathbf{s}^{l+1})_{data}$ with $p(\mathbf{h}^l, \mathbf{s}^{l+1}|\mathbf{s}^l)_{data}$ sampled from training data, and $< \cdot >_{model}$ denotes expectation with respect to the distribution $p(\mathbf{h}^l, \mathbf{s}^{l+1}|\mathbf{s}^l)$ defined in (10). The contrastive divergence in (Hinton, 2002) is used as the fast algorithm for learning the parameters in (10). The pre-training helps to find a better initial point for learning the deep model. With pre-training, the average miss rate is reduced by about 4% on Caltech, 3% on ETH, and 3% on PETS.

To obtain the $p(y_1|y_2 = 0, \mathbf{x}_1)$ in (2) for isolated pedestrian, GMM is not used and only one deep model is trained. This deep model can be obtained by removing nodes related to the pedestrian 2 in Fig. 3(a), and then replacing y with y_1 in Fig. 3(a). The training and inference of deep model for isolated pedestrian is similar to the training and inference of the mutual visibility deep model.

As the part detection scores for wnd_1 and wnd_2 are already provided by the part-based models, the extra computations required by our approach are step 2 and step 3 in Table 1. In order to save computation, we enforce $p(y_1 = 1|\mathbf{x}_1, \mathbf{x}_2) = 0$ if the detection score of the part-base model for window wnd_1 is lower than a threshold. Similarly, we enforce $\phi(y = 1; \mathbf{x}) = 0$ if the detection score of the part-base model for window wnd_2 is lower than a threshold. Therefore, $\phi(y; \mathbf{x})$ and $\phi_p(y; \mathbf{x})$ are computed for sparse window positions. With part detection scores provided, the step 2 and step 3 in Table 1 take less than 5% the execution time of the whole detection process on a 2.27GHz CPU with multi-thread turned off on the Caltech training dataset.

4.3 Extending the deep model for using 2-pedestrian detectors

The model in Fig. 3(a) can be extended so that the visual cue from 2-pedestrian windows can be used. In the extended model, we use the 2-pedestrian detector in (Ouyang and Wang, 2013b) for obtaining part detection scores for 2-pedestrian windows. As shown in Fig. 3(b), there are 1 part covering the whole 2-pedestrian window and five parts covering local regions of the windows in (Ouyang and Wang, 2013b). Denote $\mathbf{s}^3_3$ and $\mathbf{h}^3_3$ as the part detection scores and visibilities respectively for the parts from the 2-pedestrian detector. Inference and training of the extended deep model are the same as introduced in Section 4.1 and Section 4.2.2, except for the difference that $\hat{\mathbf{h}}^3_3$ is included in $\hat{\mathbf{h}}^3$ so that $\hat{\mathbf{h}}^3 = [\hat{\mathbf{h}}^{3\mathrm{T}}_1\ \hat{\mathbf{h}}^{3\mathrm{T}}_2\ \hat{\mathbf{h}}^{3\mathrm{T}}_3]^{\mathrm{T}}$, $\mathbf{s}^3_3$ is included in $\mathbf{s}^3$ so that $\mathbf{s}^3 = [\mathbf{s}^{3\mathrm{T}}_1\ \mathbf{s}^{3\mathrm{T}}_2\ \mathbf{s}^{3\mathrm{T}}_3]^{\mathrm{T}}$.

In this extended model, the part visibilities $\hat{\mathbf{h}}^3_3$ from the 2-pedestrian detector are placed at the third layer and influenced by the visibilities of single parts in $\hat{\mathbf{h}}^2$ at the second layer. The visibility status of 2-pedestrian parts $\hat{\mathbf{h}}^3_3$ and single-pedestrian parts ($\hat{\mathbf{h}}^3_1$ and $\hat{\mathbf{h}}^3_2$) are then used for estimating the pedestrian co-existence label y.

4.4 Analysis on the deep model

In this model, the visibility of parts for one pedestrian influences the visibility of parts for another pedestrian through the $\mathbf{W}^l$ in (9). When the weight between $h^{l+1}_{1,i}$ and $h^l_{2,j}$ is positive, the ith part for pedestrian 1 at layer $l+1$ and the jth part for pedestrian 2 at layer l are considered by the deep model as compatible. On the other hand, if the weight between $h^{l+1}_{1,i}$ and $h^l_{2,j}$ is negative, they are incompatible. Fig. 5 shows examples of the weight between $\mathbf{h}^3_1$ and $\mathbf{h}^2_2$ learned from the deep model. The top example is from mixture 5 and the bottom example is from mixture 4. Denote the left pedestrian by Ped_L and denote the right pedestrian by Ped_R. For the top example in Fig. 5, the head-shoulder part of Ped_R is compatible with head-shoulder part of Ped_L but incompatible with the right-half part of Ped_L. For the bottom example in Fig. 5, the left-head-torso part of Ped_R is compatible with the left-half part of Ped_L but incompatible with the right-half part of Ped_L.

5 Discussion

This paper mainly focuses on pairwise pedestrians for simplicity. When there are $N(>2)$ pedestrians in a local image region, pair-wise relationship is still able to represent their visibility relationships. Meanwhile, our approach can be extended for considering $N(>2)$ windows simultaneously. Denote the features of N windows by $\mathbf{x}$, denote the label for the nth window by $y_n \in \{0, 1\}$, the $p(y_1|\mathbf{x}_1, \mathbf{x}_2)$ in (1) is extended to:

$$\begin{aligned}p(y_1|\mathbf{x}) &= \sum_{y_2,\cdots,y_N} p(y_1, y_2, \cdots, y_N|\mathbf{x})\\ &= \sum_u p(y_1, \sum_{n=2}^{N} y_n = u).\end{aligned} \tag{12}$$

When $\sum_{n=2}^{N} y_n = u$, a mutual visibility deep model is constructed for u pedestrians, similar to the mutual visibility deep model for pair-wise pedestrians.

6 Experimental Results

The proposed framework is evaluated on four publicly available datasets: Caltech-Train, Caltech-Test (Dollár et al, 2012b), ETH (Ess et al, 2007), and PETS2009[2]. In our implementation, the DPM in (Felzenszwalb et al, 2010) with the modified HOG feature in (Felzenszwalb et al, 2010) is used for part detection scores. The deformation among parts are arranged in the star-model with full-body being the center. Since the part detection score is considered as input of our framework, the framework keeps unchanged if other articulation models or features are used. For the experiment on the datasets Caltech-Train, ETH and PETS2009, the INRIA training dataset in (Dalal and Triggs, 2005) is used for training our parts model and deep models. For the experiment on the Caltech-Test dataset, Caltech-Train dataset is used for training parts model and deep models. In the experiments, we mainly compare with the approach D-Isol (Ouyang and Wang, 2012). It uses the same feature, the same deformable model and the same training dataset as this paper for training the parts model. D-Isol (Ouyang and Wang, 2012) only used the deep model for isolated pedestrians while both isolated pedestrians and co-existing pedestrian are considered in this paper. The FPDW in (Dollár et al, 2010) and the CrossTalk in (Dollár et al, 2012a) are also included for comparison. Using the same training dataset as this paper, FPDW and CrossTalk detect pedestrians by training cascaded classifiers on multiple features. In the experiments, D-Mut denotes the deep model in this paper with only single pedestrian detector, which was reported in (Ouyang et al, 2013). D-2PedMut denotes the results using the extended model in Section 4.3 which uses part detection scores from 2-pedestrian detector.

The per-image evaluation methodology as suggested in (Dollár et al, 2012b) is used. We use the labels and evaluation code provided by Dollár in (Dollár et al, 2012b). As in (Dollár et al, 2012b), log-average miss rate is used as the evaluation criterion.

6.1 Experimental Results on four publicly available datasets

In this section, pedestrians at least 50 pixels tall, fully visible or partial occluded are investigated in the experiments. This set of pedestrians is denoted as the subset *reasonable* in (Dollár et al, 2012b). Fig. 7 shows detection result comparison of D-Isol and D-Mut at 1 FPPI on the Caltech-Train dataset, the Caltech-Test dataset, the ETH dataset, and the PETS dataset.

Fig. 6(a) and Fig. 6(b) show the experimental results on the Caltech-Train dataset and the Caltech-Test dataset. It can be seen that our mutual visibility approach, i.e. D-Mut, has 4% and 5% miss rate improvement respectively compared with D-Isol on Caltech-Train dataset and the Caltech-Test dataset. Our extended model using 2-pedestrian detector in Section 4.3, i.e. D-2PedMut, has 2% and 11% miss rate improvement respectively compared with D-Mut on Caltech-Train dataset and the Caltech-Test dataset. LatSVM-V2 uses the same feature and deformable model as our approach, but does not use the deep model for visibility reasoning. Compared with LatSVM-V2, our D-2PedMut achieves 13% and 26% miss rate improvement respectively on the Caltech-Train dataset and the Caltech-Test dataset.

Fig. 6(c) shows the experimental results on the ETH dataset. Compared with D-Isol, D-Mut has 6% miss rate improvement on the ETH dataset. Compared with D-Mut, D-2PedMut has 4% miss rate improvement on the ETH dataset.

Fig. 6(d) shows the experimental results on the PETS2009 dataset. Compared with D-Isol, D-Mut has 8% miss rate improvement. Compared with D-Mut, D-2PedMut has 2% miss rate improvement. The PETS2009 crowd dataset is a well-known benchmark for pedestrian counting and pedestrian tracking. In our experiment, we select S2_L2 with medium density crowd and S2_L3 with high density crowd for test. S2_L2 contains 436 frames and S2_L3 contains 240 frames. There are totally 676 frames and 14385 pedestrians evaluated in this experiment. We manually labeled the pedestrians in this dataset [3]. The results for LatSVM-V2 and FPDW are obtained by running their code for this dataset. The experimental results for CrossTalk is not available on PETS2009 because the code is not available.

In order to investigate the alternative schemes in using contextual information, SVM-2PedMut and D-2PedMut-FC are included in Fig. 6. The SVM-2PedMut, D-2PedMut-FC, and D-2PedMut in Fig. 6 use the same part scores (with deformation costs included) of the two contextual single-pedestrian windows and 2-pedestrian detection results. SVM-2PedMut combines these information by SVM while D-2PedMut uses the deep model proposed in this paper. As shown in Fig. 6, the proposed deep model has $5\% - 11\%$ miss rate improvement compared with SVM on the four datasets. D-2PedMut-FC directly puts the part scores at the bottom layer and uses 3 fully connected hidden layers for recognizing co-existing pedestrians while D-2PedMut uses the proposed deep model in this paper. As shown in Fig. 6, D-2PedMut has $3\% - 11\%$ miss rate improvement compared with D-2PedMut-FC.

As a summary of the experimental results in Fig. 6, compared with the DPM model of LatSVM-V2, the deep model D-Isol for single pedestrian performs better. With mutual visibility included, D-Mut achieves better results than D-Isol. With the 2-pedestrian part detection scores in-

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

[3] http://www.ee.cuhk.edu.hk/~xgwang/2DBNped.html

cluded in our deep model, D-2PedMut has $2\% - 9\%$ miss rate improvement compared with D-Mut, which does not use 2-pedestrian detection results. D-2PedMut, which uses our deep model for combining part detection scores from 2-pedestrian window and contextual 1-pedestrian window, performs better than SVM-2PedMut and D-2PedMut-FC, which use SVM and conventional fully connected deep model respectively for combining scores.

In order to provide a context of the proposed approach compared with recent existing approaches, Fig. 8 shows the average miss rate of the approaches provided by P. Dollár online (Dollár, 2014) for the Caltech-Train, Caltech-Test and ETH dataset. The results on PETS is not provided by P. Dollár online. These approaches are VJ (Viola et al, 2005), LatSvm-V2 (Felzenszwalb et al, 2010), HOG (Dalal and Triggs, 2005), MultiFtr (Wojek and Schiele, 2008), Pls (Schwartz et al, 2009), FPDW (Dollár et al, 2010), ChnFtrs (Dollár et al, 2009), FeatSynth (Bar-Hillel et al, 2010), D-Isol (Ouyang and Wang, 2012), *+2Ped (Ouyang and Wang, 2013b), MultiResC (Park et al, 2010), MOCO (Chen et al, 2013), MT-DPM (Yan et al, 2013), ACF (Dollár et al, 2014), FisherBoost (Shen et al, 2013), ConvNet (Sermanet et al, 2013b), ACF+SDt (Park et al, 2013), MLS (Nam et al, 2011), pAUCBoost (Paisitkriangkrai et al, 2013), Roerei (Benenson et al, 2013), VeryFast (Benenson et al, 2012). Besides, we also include the most recent approaches such as RandForest (Marın et al, 2013), Franken (Mathias et al, 2013). Since it is hard to read using the ROC curve, we have chosen the average miss rate as the measure of performance in Fig. 8. LatSVM-V2+2Ped uses the same feature, deformable model, and the 2-pedestrian detector as D-2PedMut but does not use the deep model for visibility reasoning. Compared with LatSVM-V2+2Ped, our D-2PedMut has 19% and 4% miss rate improvement respectively on the Caltech-Test dataset and the ETH dataset. Since ACF+SDt (Park et al, 2013) and MultiFtr+Motion (Walk et al, 2010) use motion features, JDN (Ouyang and Wang, 2013a) learns more discriminative features, and MT-DPM+Context (Yan et al, 2013) uses the context obtained from a vehicle detector, they are excluded in the comparison. Our approach, which only uses HOG feature from static image without vehicle detection results, is complementary to the more advanced feature and contextual information of these approaches. For example, if the feature used by approach is changed to HOG+CSS, the average miss rate for D-2PedMut can be reduced from 39% to 35% on the Caltech-Test dataset. In summary, compared with still-image-based approaches reported by P. Dollár online (Dollár, 2014), our approach with only HOG feature has the lowest average miss rate on the Caltech-Train and ETH dataset. Our approach with HOG+CSS feature has the lowest average miss rate on the Caltech-Test dataset.

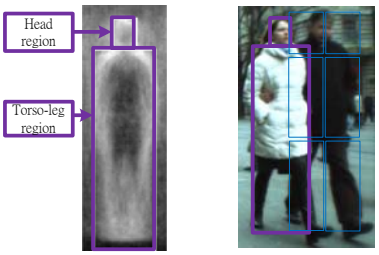


Fig. 4 **(a)** Two rectangular regions used for approximating the pedestrian region and **(b)** a pedestrian on the right with left-head-shoulder, left-torso and left-leg overlapping with the pedestrian regions of the left person.

6.2 Experimental Results on detecting occluded pedestrians

In this section, occluded pedestrians are investigated in the experiments using the occlusion label provided on the Caltech-Test dataset. Fig. 9 shows the miss rate vs. FPPI curve for the approaches HOG, FPDW, LatSvm-V2, CrossTalk, D-Isol, and D-Mut. Fig. 10 shows all the approaches evaluated in Fig. 7 on the Caltech-Test dataset. In the compared approaches, Franken and RandForest are the most recent approaches that aim at handling occluded pedestrians. The results in Fig. 7 show that our approach has lower miss rate than these approaches.

7 Conclusion

This paper proposes a mutual visibility deep model that jointly estimates the visibility statuses of multiple co-existing pedestrians. Starting from the scores of conventional part detectors, the mutual part visibility relationship among multiple pedestrians is learned by the deep model for recognizing co-existing pedestrians. Our extended deep model shows that contextual information can be used by the proposed deep model for further improving the detection accuracy. Experimental results show that the mutual visibility deep model effectively improves the pedestrian detection results. Compared with existing image-based pedestrian detection approaches evaluated in (Dollár et al, 2012b, Dollár, 2014), our approach has the lowest miss rate on the Caltech-Train dataset and the ETH dataset. Since the deep model takes the part detection scores as input, it is complementary to new investigations on features, e.g. color self similarity, local binary pattern, motion and depth, and articulation models, e.g. poselets, multi-object articulation model. Experimental results on four publicly available datasets show that the mutual visibility deep model is effective in improving pedestrian detection results.

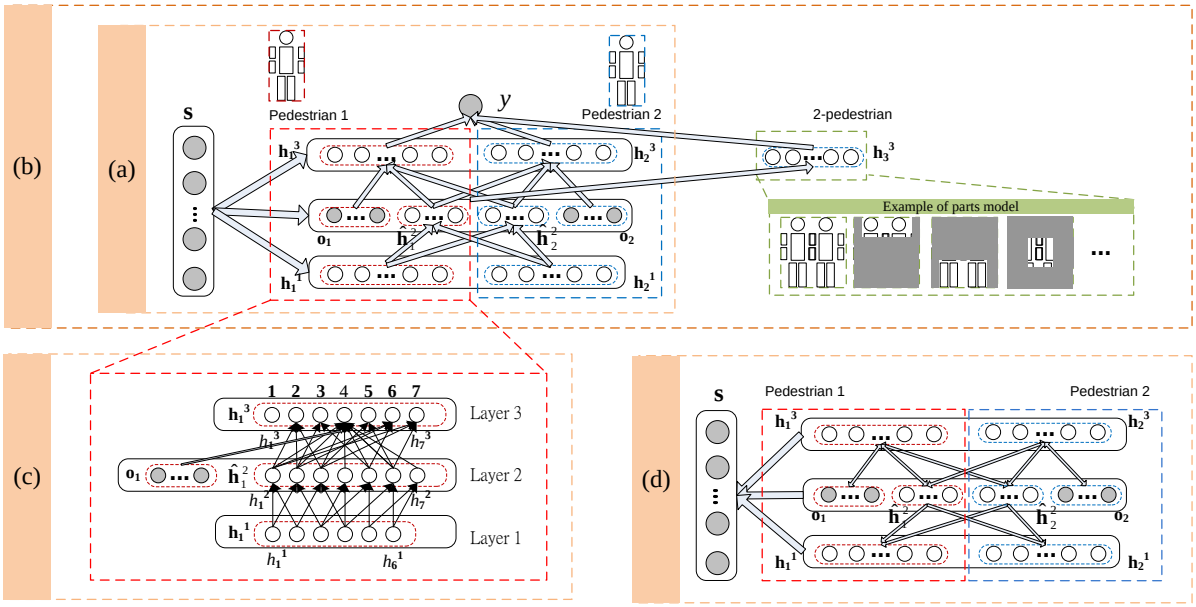


Fig. 3 The mutual visibility deep model without 2-pedestrian parts **(a)** and the extended model with 2-pedestrian parts **(b)** used for inference and fine tuning parameters, the detailed connection for pedestrian 1 **(c)**, and the model used for pretraining parameters **(d)**. In our model, hidden nodes without arrows in (c) are not connected.

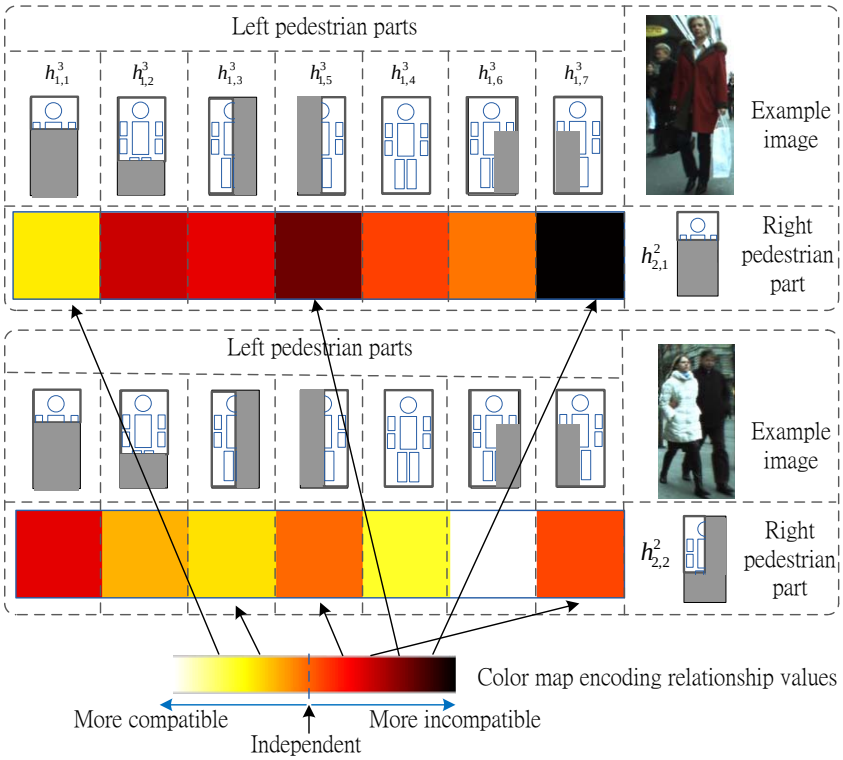
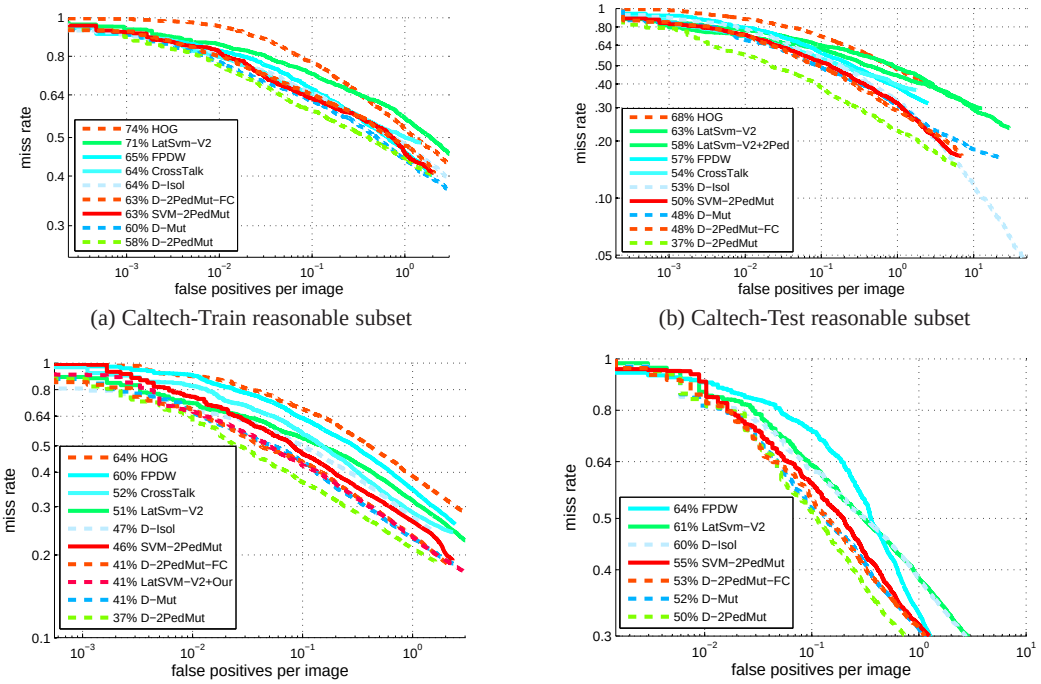


Fig. 5 Examples of correlation between $\mathbf{h}_2^2$ and $\mathbf{h}_1^3$ learned from the deep model.



(a) Caltech-Train reasonable subset

(b) Caltech-Test reasonable subset

(c) ETH reasonable subset

(d) PETS dataset

Fig. 6 Experimental results on the Caltech-Train (a), Caltech-Test (b), ETH (c), and PETS (d) dataset for HOG (Dalal and Triggs, 2005), LatSVM-V2(Felzenszwalb et al, 2010), FPDW (Dollár et al, 2010), D-Isol (Ouyang and Wang, 2012) and our mutual visibility approach, i.e. D-Mut and D-2PedMut. D-Mut denotes our approach with only single pedestrian detector. D-2PedMut denotes the result using the extended model in Section 4.3 with 2-pedestrian detector.

Fig. 7 Detection results comparison of D-Isol and D-Mut on the Caltech-Train dataset and the ETH dataset. All results are obtained at 1 FPPI.

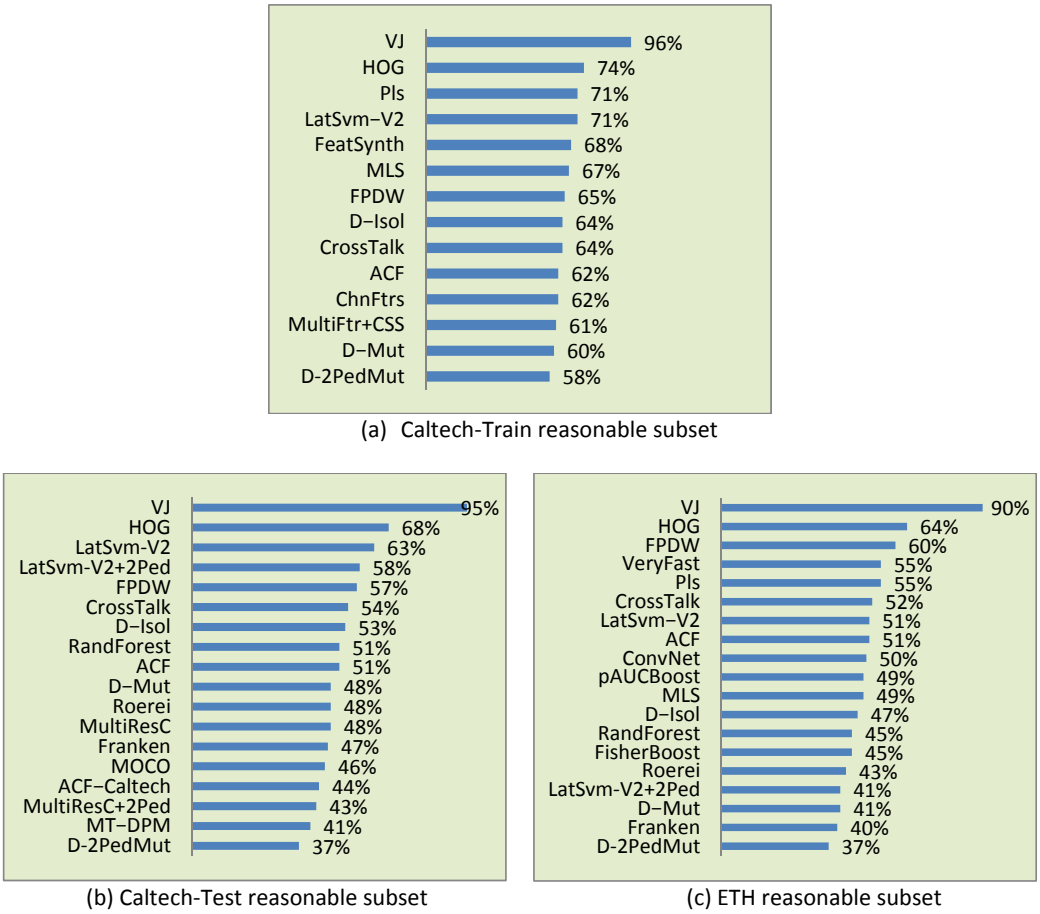


Fig. 8 Average miss rate on the Caltech-Train (a), the Caltech-Test dataset (b), and the ETH dataset (c).

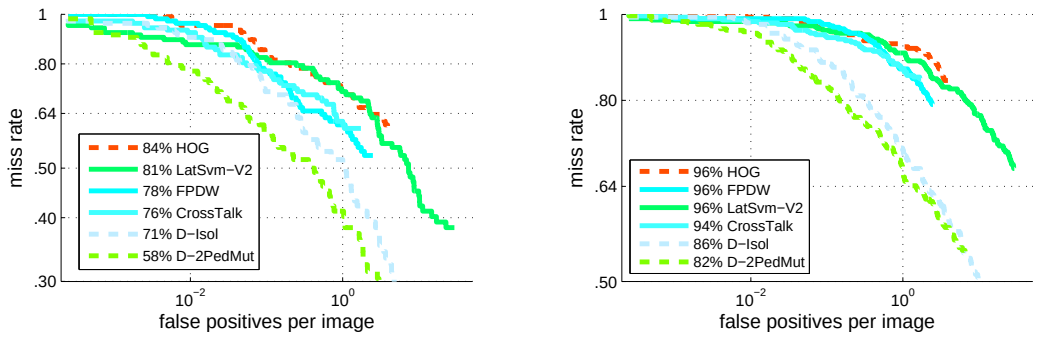


Fig. 9 Experimental results on the Caltech-Test dataset for pedestrians under *partial occlusions* (left) and *heavy occlusions* (right). The ratio of visible area is larger than 0.65 for *partial occlusions* and [0.2 0.65] for *heavy occlusions*.

References

Bar-Hillel A, Levi D, Krupka E, Goldberg C (2010) Part-based feature synthesis for human detection. In: ECCV

Benenson R, Mathias M, Timofte R, Gool LV (2012) Pedestrian detection at 100 frames per second. In: CVPR

Benenson R, Mathias M, Tuytelaars T, Van Gool L (2013) Seeking the strongest rigid detector. In: CVPR

Bengio Y (2009) Learning deep architectures for AI. Foundations and Trends in Machine Learning 2(1):1–127

Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Trans PAMI 35(8):1798–1828

Chen G, Ding Y, Xiao J, Han TX (2013) Detection evolution with multi-order contextual co-occurrence. In: CVPR

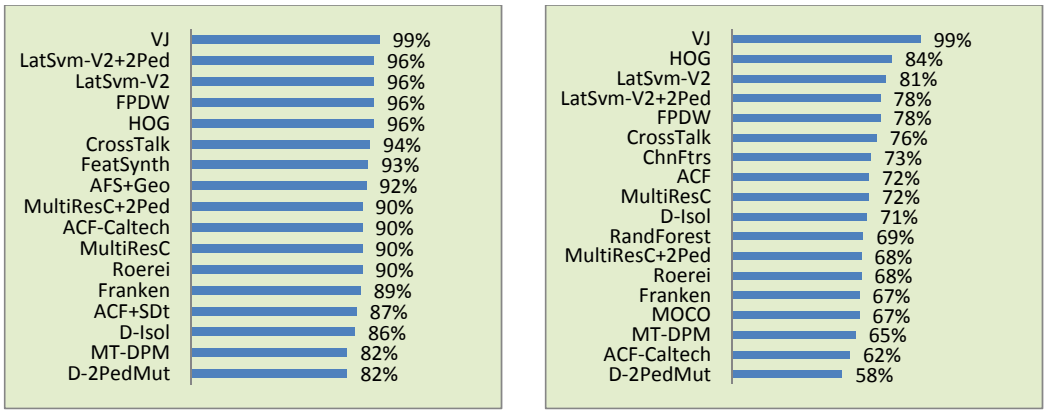


Fig. 10 Average miss rate on the Caltech Testing dataset for pedestrians under *partial occlusions* (left) and *heavy occlusions* (right).

Dai S, Yang M, Wu Y, Katsaggelos A (2007) Detector ensemble. In: CVPR

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: CVPR

Dean T, Ruzon MA, Segal M, Shlens J, Vijayanarasimhan S, Yagnik J (2013) Fast, accurate detection of 100,000 object classes on a single machine. In: CVPR

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: CVPR

Desai C, Ramanan D (2012) Detecting actions, poses, and objects with relational phraselets. In: ECCV

Desai C, Ramanan D, Fowlkes C (2009) Discriminative models for multi-class object layout. In: ICCV

Ding Y, Xiao J (2012) Contextual boost for pedestrian detection. In: CVPR

Dollár P (2014) Caltech pedestrian detection benchmark. Online (accessed on May/6/2014), URL `www.vision.caltech.edu/Image_Datasets/CaltechPedestrians`

Dollár P, Tu Z, Perona P, Belongie S (2009) Integral channel features. In: BMVC

Dollár P, Belongie S, Perona P (2010) The fastest pedestrian detector in the west. In: BMVC

Dollár P, Appel R, Kienzle W (2012a) Crosstalk cascades for frame-rate pedestrian detection. In: ECCV

Dollár P, Wojek C, Schiele B, Perona P (2012b) Pedestrian detection: an evaluation of the state of the art. IEEE Trans PAMI 34(4):743 – 761

Dollár P, Appel R, Belongie S, Perona P (2014) Fast feature pyramids for object detection. In: Accepted by IEEE Trans. PAMI

Duan G, Ai H, Lao S (2010) A structural filter approach to human detection. In: ECCV

Enzweiler M, Eigenstetter A, Schiele B, Gavrila DM (2010) Multi-cue pedestrian classification with partial occlusion handling. In: CVPR

Erhan D, Bengio Y, Courville A, Manzagol PA, Vincent P, Bengio S (2010) Why does unsupervised pretraining help deep learning? J Machine Learning Research 11:625–660

Ess A, Leibe B, Gool LV (2007) Depth and appearance for mobile scene analysis. In: ICCV

Farabet C, Couprie C, Najman L, LeCun Y (2013) Learning hierarchical features for scene labeling. IEEE Trans PAMI 30:1915–1929

Felzenszwalb P, Grishick RB, DMcAllister, Ramanan D (2010) Object detection with discriminatively trained part based models. IEEE Trans PAMI 32:1627–1645

Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Computation 14:1771–1800

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504 – 507

Hinton GE, Osindero S, Teh Y (2006) A fast learning algorithm for deep belief nets. Neural Computation 18:1527–1554

Hu J, Lu J, Tan YP (2014) Discriminative deep metric learning for face verification in the wild. In: CVPR

Jarrett K, Kavukcuoglu K, Ranzato M, LeCun Y (2009) What is the best multi-stage architecture for object recognition? In: CVPR

Ji S, Xu W, Yang M, Yu K (2013) 3d convolutional neural networks for human action recognition. IEEE Trans PAMI 35(1):221–231

Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. In: NIPS

Le QV, Ranzato M, Monga R, Devin M, Chen K, Corrado GS, Dean J, Ng AY (2012) Building high-level features using large scale unsupervised learning. In: ICML

Lee H, Grosse R, Ranganath R, Ng AY (2009) Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: ICML

Leibe B, Seemann E, Schiele B (2005) Pedestrian detection in crowded scenes. In: CVPR
Li C, Parikh D, Chen T (2011) Extracting adaptive contextual cues from unlabeled regions. In: ICCV
Lin Z, Davis LS, Doermann D, DeMenthon D (2007) Hierarchical part-template matching for human detection and segmentation. In: ICCV
Liu P, Jan S, Meng Z, Tong Y (2014) Facial expression recognition via a boosted deep belief network. In: CVPR
Luo P, Wang X, Tang X (2012) Hierarchical face parsing via deep learning. In: CVPR
Marın J, Vázquez D, López AM, Amores J, Leibe B (2013) Random forests of local experts for pedestrian detection. In: CVPR
Mathias M, Benenson R, Timofte R, Van Gool L (2013) Handling occlusions with franken-classifiers. In: CVPR
Nam W, Han B, Han JH (2011) Improving object localization using macrofeature layout selection. In: ICCV Workshop, IEEE, pp 1801–1808
Norouzi M, Ranjbar M, Mori G (2009) Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning. In: CVPR
Ouyang W, Wang X (2012) A discriminative deep model for pedestrian detection with occlusion handling. In: CVPR
Ouyang W, Wang X (2013a) Joint deep learning for pedestrian detection. In: ICCV
Ouyang W, Wang X (2013b) Single-pedestrian detection aided by multi-pedestrian detection. In: CVPR
Ouyang W, Zeng X, Wang X (2013) Modeling mutual visibility relationship in pedestrian detection. In: CVPR
Paisitkriangkrai S, Shen C, Hengel Avd (2013) Efficient pedestrian detection by directly optimize the partial area under the roc curve. In: ICCV
Park D, Ramanan D, Fowlkes C (2010) Multiresolution models for object detection. In: ECCV
Park D, Zitnick CL, Ramanan D, Dollár P (2013) Exploring weak stabilization for motion feature extraction. In: CVPR
Pepikj B, Stark M, Gehler P, Schiele B (2013) Occlusion patterns for object class detection. In: CVPR, IEEE, pp 3286–3293
Ranzato M, Susskind J, Mnih V, Hinton G (2011) On deep generative models with applications to recognition. In: CVPR
Sadeghi MA, Farhadi A (2011) Recognition using visual phrases. In: CVPR
Schwartz W, Kembhavi A, Harwood D, Davis L (2009) Human detection using partial least squares analysis. In: ICCV
Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013a) Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:13126229
Sermanet P, Kavukcuoglu K, Chintala S, Lecun Y (2013b) Pedestrian detection with unsupervised and multi-stage feature learning. In: CVPR
Shen C, Wang P, Paisitkriangkrai S, van den Hengel A (2013) Training effective node classifiers for cascade classification. IJCV 103(3):326–347
Shet VD, Neumann J, Ramesh V, Davis LS (2007) Bilattice-based logical reasoning for human detection. In: CVPR
SUN L, Jia K, Chan TH, Fang Y, Yan S (2014) Deeply-learned slow feature analysis for action recognition. In: CVPR
Sun Y, Wang X, Tang X (2014) Deep learning face representation from predicting 10,000 classes. In: CVPR
Tang S, Andriluka M, Schiele B (2012) Detection and tracking of occluded people. In: BMVC, Surrey, UK
Tang S, Andriluka M, Milan A, Schindler K, Roth S, Schiele B (2013) Learning people detectors for tracking in crowded scenes. ICCV
Viola P, Jones MJ, Snow D (2005) Detecting pedestrians using patterns of motion and appearance. IJCV 63(2):153–161
Walk S, Majer N, Schindler K, Schiele B (2010) New features and insights for pedestrian detection. In: CVPR
Wang X, Han X, Yan S (2009) An hog-lbp human detector with partial occlusion handling. In: CVPR
Wojek C, Schiele B (2008) A performance evaluation of single and multi-feature people detection. In: DAGM
Wu B, Nevatia R (2005) Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In: ICCV
Wu B, Nevatia R (2009) Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. IJCV 82(2):185–204
Wu T, Zhu S (2011) A numeric study of the bottom-up and top-down inference processes in and-or graphs. IJCV 93(2):226–252
Yan J, Lei Z, Yi D, Li SZ (2012) Multi-pedestrian detection in crowded scenes: A global view. In: CVPR
Yan J, Zhang X, Lei Z, Liao S, Li SZ (2013) Robust multi-resolution pedestrian detection in traffic scenes. In: CVPR
Yang Y, Ramanan D (2011) Articulated pose estimation with flexible mixtures-of-parts. In: CVPR
Yang Y, Baker S, Kannan A, Ramanan D (2012) Recognizing proxemics in personal photos. In: CVPR
Yao B, Fei-Fei L (2010) Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR
Zeng X, Ouyang W, Wang X (2013) Multi-stage contextual deep learning for pedestrian detection. In: ICCV
Zhu L, Chen Y, Yuille A, Freeman W (2010) Latent hierarchical structural learning for object detection. In: CVPR