Modeling Mutual Visibility Relationship in Pedestrian Detection

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

Abstract

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. Experimental results show that the mutual visibility deep model effectively improves the pedestrian detection results. Compared with existing image-based pedestrian detection approaches, our approach has the lowest average miss rate on the Caltech-Train dataset, the Caltech-Test dataset and the ETH dataset. Including mutual visibility leads to $4\% - 8\%$ improvements on multiple benchmark datasets.

1. Introduction

Pedestrian detection is a challenging task due to the intra-class variation of pedestrians in clothing 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 mutual 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 pedestrian co-existing side by side, the right-half part of the *left* pedestrian is compatible with the right-half part of the *right* pedestrian 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 Alice[1] on the left and Bob on the right with high false positive rate, then the visibility confidence of Alice's right-

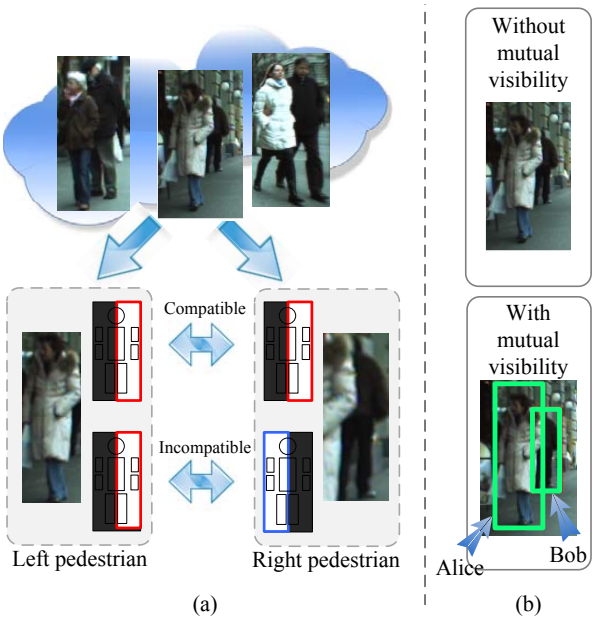

Figure 1. **(a)** Mutual visibility relationship of parts among pedestrians and **(b)** detection results comparison of the approach without modeling mutual visibility in [25] and our approach modeling mutual visibility. With mutual visibility modeled in our approach, the false positive window on the left leg is suppressed and missed pedestrian on the left is found by modeling the visibility relationship among parts.

[1] 'Alice' and 'Bob' are used as placeholder names in this paper.

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

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 discriminative 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 comoputation in the detection process. 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 8% 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 [31], scene geometric constraints [27], richer part models [40, 38] and contextual multi-pedestrian detection information [30, 26, 36] to further improve the performance.

2. Related Work

Since visibility estimation is the key to handle occlusions, many approaches were proposed for estimating visibility of parts [2, 10, 11, 32, 35, 33, 29, 22, 34, 21]. Wang *et al.* [32] 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.* [11] 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 [2, 10]. 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. The And-Or graph was used in [35] to accumulate hard-thresholded part detection scores. Recently, the approaches in [10, 25] utilized the visibility relationship among parts for isolated pedestrian. However, the part visibility relationship among co-existing pedestrians was not explored in [2, 10, 11, 32, 35]. In order to handle inter-human occlusions, the joint part-combination of multiple humans was adopted in [33, 29, 22, 34, 21]. These approaches obtain the visibility status by occlusion reasoning using 2-D visibility scores in [33, 29, 22] or using segmentation results in [34, 21]. 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 [39, 36, 37, 5]. Nearby detection scores was considered as context in [6]. But it 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.

Deep model has been applied for dimensionality reduction [17], hand written digit recognition [16, 20, 24], object recognition [18, 20], face parsing [23], facial expression recognition and scene recognition [28]. Hinton *et al.* [16] 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 . Bengio [1] proved that an architecture with insufficient depth can require many more computational elements, potentially exponentially more (with respect to input size), than architectures whose depth is matched to the task. Krizhevsky *et al.* [19] proposed a deep model that achieved state-of-the-art performance for object detection and recognition on the ImageNet dataset [4]. Recently, deep

Table 1. Overview of our pedestrian detection approach.

1. obtain part detection scores by part detector;
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, estimate $\phi_p(y; \mathbf{x})$ in (3) with GMM;
3. $p(y_1|\mathbf{x}_1, \mathbf{x}_2) = p(y_1, y_2 = 0|\mathbf{x}_1, \mathbf{x}_2) + p(y_1, y_2 = 1|\mathbf{x}_1, \mathbf{x}_2)$.

model was used for pedestrian detection in [25, 24]. The approaches in [25, 24, 19] focused on isolated objects or pedestrians. This paper focuses on co-existing pedestrians, which has not been considered in [25, 19, 24].

3. Overview of our approach

In this paper, we mainly discuss the approach for pair-wise pedestrians and extend it to more pedestrians in Section 4.3. 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 $y_1 \in \{0, 1\}$. 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 $y_2 \in \{0, 1\}$. And we have the following by marginalizing y_2:

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

When $y_2 = 0$, we have

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

where $p(y_1|y_2 = 0, \mathbf{x}_1)$ is obtained from the deep model for isolated pedestrians. $p(y_2 = 0)$ is a constant prior on wnd_2 being a background, which is obtained from cross-validation. When $y_2 = 1$, we have

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

$\phi(y; \mathbf{x})$ 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 $p(y_1|y_2 = 0, \mathbf{x}_1, \mathbf{x}_2)$ and $\phi(y; \mathbf{x})$ are obtained from the deep model introduced in Section 4. $\phi_p(y; \mathbf{x})$ in (3) models probability for the 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 Table 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. And K deep models are trained for these K mixtures. 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 $p(d_x, d_y, d_s)$.

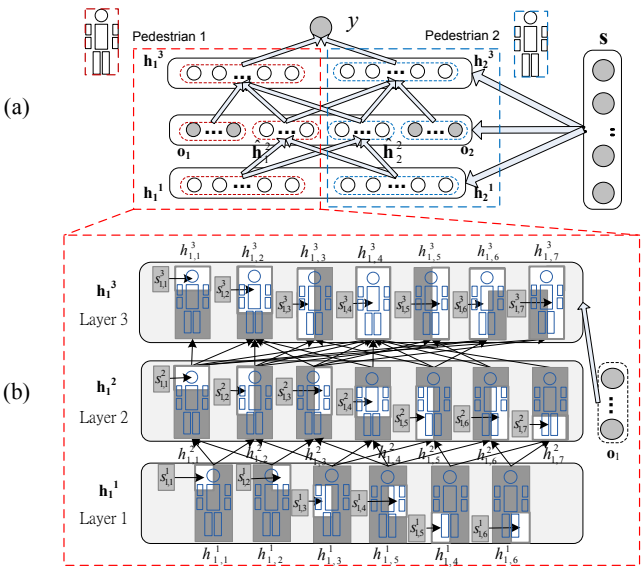

Figure 2. **(a)** The mutual visibility deep model used for inference and fine tuning parameters and **(b)** the detailed connection and parts model for pedestrian 1.

4.1. The deep model at the inference stage

Fig. 2(a) shows the deep model used at the inference stage. Fig. 2(b) 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(b), there are 3 layers of parts with different sizes. For each pedestrian, there are 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.

The detection scores for L layers are denoted by $\mathbf{s} = [\mathbf{s}^{1\mathrm{T}} \dots \mathbf{s}^{L\mathrm{T}}]^{\mathrm{T}} = \gamma(\mathbf{x})$, where $\gamma(\mathbf{x})$ is obtained from part detectors, $\mathbf{s}^l$ for $l = 1, \dots, L$ denotes the scores at layer l. For the model in Fig. 2, $L = 3$. 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 two pedestrians at layer l are denoted by $\mathbf{s}_1^l = [s_{1,1}^l, \dots, s_{1,P^l}^l]^{\mathrm{T}}$ and $\mathbf{s}_2^l = [s_{2,1}^l, \dots, s_{2,P^l}^l]^{\mathrm{T}}$. The visibilities of P^l parts are denoted by $\hat{\mathbf{h}}_1^l = [h_{1,1}^l, \dots, h_{1,P^l}^l]^{\mathrm{T}}$ and $\hat{\mathbf{h}}_2^l = [h_{2,1}^l, \dots, h_{2,P^l}^l]^{\mathrm{T}}$ respectively. The hidden variables at layer l are denoted by $\hat{\mathbf{h}}^l = [\hat{\mathbf{h}}_1^{l\mathrm{T}}\ \hat{\mathbf{h}}_2^{l\mathrm{T}}]^{\mathrm{T}}$. Since $\hat{\mathbf{h}}^l$ is not provided at training stage

or testing stage, it is considered as a hidden random vector. In our implementation, DPM in [14] is 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 $\mathbf{s}$. In order to have the top layer representing occlusion status in a more direct way, $\mathbf{s}^3$ accumulate the detection scores that fit their possible occlusion statuses. For example,

$$\begin{aligned} s^3_{1,1} &= \tilde{s}^3_{1,1} + s^2_{1,1} + s^1_{1,1} + s^1_{1,2}, \\ s^3_{1,2} &= \tilde{s}^3_{1,2} + \sum_{i=1}^{4} s^2_{1,i} + \sum_{i=1}^{4} s^1_{1,i}, \end{aligned} \tag{4}$$

where $s^l_{1,i}$ for $l = 1, 2, i = 1, \ldots, P^l$ is the detection score for the ith part at layer l, $\tilde{s}^3_{1,1}$ is the detection score for the head-shoulder part at layer 3 and $\tilde{s}^3_{1,2}$ is the detection score for the head-torso part 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.

The overlap information at layer 2 in Fig. 2 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. 3(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, denoted by A_h, another rectangle is used for the torso-leg region, denoted by A_t. Denote the region $o_{n,i}$ by $A_{n,i}$. $o_{n,i}$ is obtained as follows:

$$o_{n,i} = \frac{area(A_{n,i} \cap A_h) + area(A_{n,i} \cap A_t)}{area(A_{n,i})}, \tag{5}$$

where $area(\cdot)$ computes the area in this region, $\cap$ denotes intersection of region. For example, the right person in Fig. 3(b) has the left-head-shoulder, left-torso and left-leg overlapping with the pedestrian regions of the left person. Since $A_{n,i}$, A_h and A_t are rectangular regions, the operations $area(\cdot)$ and $\cap$ in (5) 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 $\mathbf{o}$.

At the inference stage, the pedestrian co-existence label y is inferred from features $\mathbf{x}$. The part visibility probability

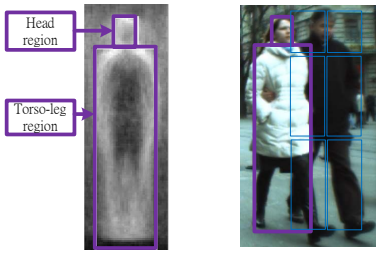

Figure 3. **(a)** Two rectangular regions used for approximating the pedestrian region and **(b)** an example with left-head-shoulder, left-torso and left-leg overlapping with the pedestrian regions of the left person.

$\tilde{h}^{l+1}_j$ is obtained using the model in Fig. 2, i.e.

$$\begin{aligned} \tilde{h}^{l+1}_j &= p(h^{l+1}_j = 1 | \mathbf{h}^l, \mathbf{x}) \\ &= \sigma(\mathbf{h}^{l\mathrm{T}} \mathbf{w}^l_{*,j} + c^{l+1}_j + g^{l+1^T}_j s^{l+1}_j), \\ \mathbf{h}^l &= \hat{\mathbf{h}}^l \text{ if } l \neq L-1, \mathbf{h}^l = [\hat{\mathbf{h}}^{l\mathrm{T}}\ \mathbf{o}^{\mathrm{T}}]^{\mathrm{T}}, \text{ if } l = L-1, \end{aligned} \tag{6}$$

where $\sigma(t) = (1 + exp(-t))^{-1}$ is the logistic function. The estimated output $\phi(y; \mathbf{x})$ is obtained as follows:

$$\phi(y; \mathbf{x}) = e^{y(\mathbf{w}_L^{\mathrm{T}} \tilde{\mathbf{h}}^L + b)} / Z, \tag{7}$$

where $Z = \sum_{y=0,1} e^{y(\mathbf{w}_L^{\mathrm{T}} \tilde{\mathbf{h}}^L + b)}$. For the model in Fig. 2, we have $L = 3$. The learning of parameters $\mathbf{w}^l_{*,j}, \mathbf{w}_L, c^{l+1}_j$ and g^{l+1}_j in (6) and (7) are explained in Section 4.2.

4.2. The learning of the deep model

The following two stages are used for learning the parameters in (6) and (7).

Stage 1: Pretrain parameters $\mathbf{w}^l_{*,j}, c^{l+1}_j$ and g^{l+1}_j in (6).

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

As stated in [12], 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. 4. The probability distribution of $p(\mathbf{h}^1, \ldots, \mathbf{h}^L | \mathbf{x})$ is modeled as follows:

$$\begin{aligned} p(\mathbf{h}^1, \ldots, \mathbf{h}^L | \mathbf{x}) &= \left(\prod_{l=1}^{L-2} p(\mathbf{h}^l | \mathbf{h}^{l+1}, \mathbf{x}) \right) p(\mathbf{h}^{L-1}, \mathbf{h}^L | \mathbf{x}), \\ p(h^l_i = 1 | \mathbf{h}^{l+1}, \mathbf{x}) &= \sigma(\mathbf{w}^l_{i,*} \mathbf{h}^{l+1} + g^l_i s^l_i + c^l_i), \\ p(\mathbf{h}^{L-1}, \mathbf{h}^L | \mathbf{x}) &= p(\mathbf{h}^{L-1}, \mathbf{h}^L | \mathbf{s}) \\ &= e^{\left[\mathbf{h}^{L-1\mathrm{T}} \mathbf{W}^{L-1} \mathbf{h}^L + (\mathbf{c}^{L-1} + \mathbf{g}^{L-1} \circ \mathbf{s}^{L-1})^{\mathrm{T}} \mathbf{h}^{L-1} + (\mathbf{c}^L + \mathbf{g}^L \circ \mathbf{s}^L)^{\mathrm{T}} \mathbf{h}^L \right]}, \end{aligned} \tag{8}$$

where $\circ$ denotes the entrywise product, i.e. $(A \circ B)_{i,j} = A_{i,j} B_{i,j}$, $\mathbf{h}$ is defined in (6). For the model in Fig. 4,

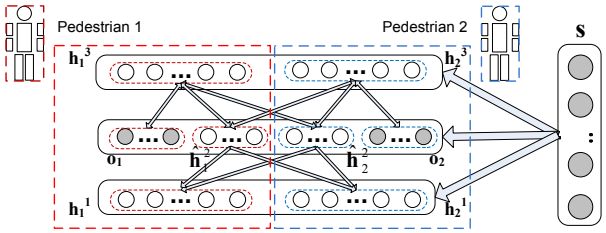

Figure 4. The mutual visibility deep model for pretraining.

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 (8) is set to zero if there is no connection between units h_i^l and h_j^{l+1} in Fig. 2(b). Since $\mathbf{s}$ is obtained from the part-based model at both the training and the testing stages, we consider them as the observed input variables and need not model $p(\mathbf{s})$. This deep model is a discriminative model.

Similar to the approach in [16], the parameters in (8) are trained layer by layer and two adjacent layers are considered as a Restricted Boltzmann Machine (RBM) that has the following distributions:

$$\begin{aligned} &p(\mathbf{h}^l, \mathbf{h}^{l+1}|\mathbf{x}) \\ &\propto e^{\left[\mathbf{h}^{l\mathrm{T}}\mathbf{W}^l\mathbf{h}^{l+1}+(\mathbf{c}^l+\mathbf{g}^l\circ\mathbf{s}^l)^{\mathrm{T}}\mathbf{h}^l+(\mathbf{c}^{l+1}+\mathbf{g}^{l+1}\circ\mathbf{s}^{l+1})^{\mathrm{T}}\mathbf{h}^{l+1}\right]}, \\ &p(h_i^l = 1|\mathbf{h}^{l+1}, \mathbf{x}) = \sigma(\mathbf{w}_{i,*}^l\mathbf{h}^{l+1} + c_i^l + g_i^{l\mathrm{T}}s_i^l), \\ &p(h_j^{l+1} = 1|\mathbf{h}^l, \mathbf{x}) = \sigma(\mathbf{h}^{l\mathrm{T}}\mathbf{w}_{*,j}^l + c_j^{l+1} + g_j^{l+1\mathrm{T}}s_j^{l+1}), \end{aligned} \tag{9}$$

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

$$\begin{aligned} \frac{\partial L(\mathbf{h}^l)}{\partial w_{i,j}^l} &\propto (< h_i^l h_j^{l+1} >_{data} - < h_i^l h_j^{l+1} >_{model}), \\ \frac{\partial L(\mathbf{h}^l)}{\partial c_i^l} &\propto (< h_i^l >_{data} - < h_i^l >_{model}), \\ \frac{\partial L(\mathbf{h}^l)}{\partial g_i^l} &\propto (< h_i^l s_i^l >_{data} - < h_i^l s_i^l >_{model}), \end{aligned} \tag{10}$$

where $w_{i,j}^l$ is the (i,j)th element in matrices $\mathbf{W}^l$, $< \cdot >_{data}$ denotes the expectation with respect to the distribution $p(\mathbf{h}^{l+1}|\mathbf{h}^l)p(\mathbf{h}^l)_{data}$ with $p(\mathbf{h}^l)_{data}$ sampled from training data, and $< \cdot >_{model}$ denotes expectation with respect to the distribution $p(\mathbf{h}^{l+1}, \mathbf{h}^l)$ defined in (9). The contrastive divergence in [15] is used as the fast algorithm for learning the parameters in (9).

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. 2 and Fig. 4, and then replacing y with y_1 in Fig. 2. The training and inference of deep model for isolated pedestrian is similar to the training and inference of the mutual visibility deep model.

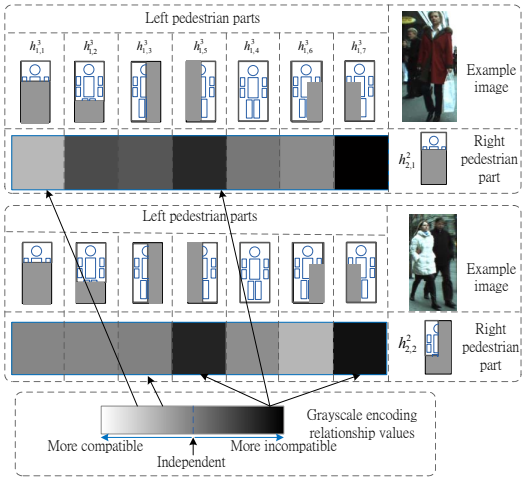

Figure 5. Examples of correlation between $\mathbf{h}_2^2$ and $\mathbf{h}_1^3$ learned from the deep model.

4.3. 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 (8). When the weight between $h_{1,i}^{l+1}$ and $h_{2,j}^l$ 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 $h_{1,i}^{l+1}$ and $h_{2,j}^l$ is negative, they are incompatible. Fig. 5 shows examples of the weight between $\mathbf{h}_1^3$ 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, 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, 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.

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 $\phi_p(y = 1; \mathbf{x}) = 0$ 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.

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:

$$p(y_1|\mathbf{x}) = \sum_{y_2,\cdots,y_N} p(y_1,y_2,\cdots,y_N|\mathbf{x}) = \sum_{u} p(y_1, \sum_{n=2}^{N} y_n = u). \tag{11}$$

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.

5. Experimental Results

The proposed framework is evaluated on four publicly available datasets: Caltech-Train, Caltech-Test [9] , ETH [13] and PETS2009[2]. In our implementation, the DPM in [14] with the modified HOG feature in [14] 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 [3] 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 [25]. It uses the same feature, the same deformable model and the same training dataset as ours for training the parts model. D-Isol [25] only used the deep model for isolated pedestrians while both isolated pedestrians and co-existing pedestrian are considered in this paper. The FPDW in [8] and the CrossTalk in [7] are also included for comparison. FPDW and CrossTalk are trained on INRIA training dataset using boosting classifier on multiple features.

The per-image evaluation methodology as suggested in [9] is used. We use the labels and evaluation code provided by Dollár in [9]. As in [9], log-average miss rate is used as the evaluation criterion.

5.1. 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 [9].

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

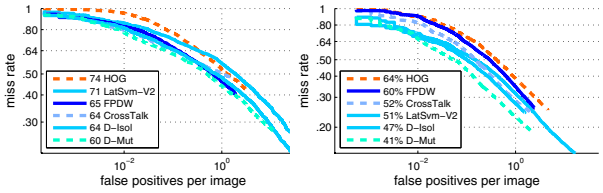

Figure 6. Experimental results on the Caltech-Train dataset (left) and the ETH dataset (right) for HOG [3], LatSVM-V2[14], FPDW [8], D-Isol [25] and our mutual visibility approach, i.e. D-Mut.

Fig. 6 shows the experimental results on the Caltech-Train dataset and the ETH dataset. Fig. 8 shows detection result comparison of D-Isol and D-Mut at 1 FPPI on the Caltech-Train dataset and the ETH dataset. It can be seen that our mutual visibility approach, i.e. D-Mut, has 4% and 6% miss rate improvement compared with D-Isol on Caltech-Train dataset and the ETH dataset respectively. Compared with LatSVM-V2, our approach achieves 11% and 10% miss rate improvement respectively on the Caltech-Train dataset and the ETH dataset. Compared with FPDW, our approach achieves 5% and 19% miss rate improvement respectively on the Caltech-Train dataset and the ETH dataset. Compared with the image-based approaches evaluated in [9], our approach has the lowest miss rate on the ETH dataset and the Caltech-Train dataset.

Fig. 7 show the experimental results on the Caltech-Test dataset and the PETS2009 dataset. Compared with D-Isol, D-Mut has 5% miss rate improvement on the Caltech-Test dataset and 8% miss rate improvement on the PETS2009 dataset. Our approach has the same miss rate as the best performing approaches [27, 6] on the Caltech-Test dataset, both of which have 48% average miss rate. More discriminative features [31] (pyramid HOG and color-self-similarity features) and scene geometric constraints [27] have been used in these approaches. These features and constraints can also be used for further improving our results. For example, with geometric constraints used in an unsupervised way, the miss rate of D-Mut can be reduced from 48% to 44%. 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.

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

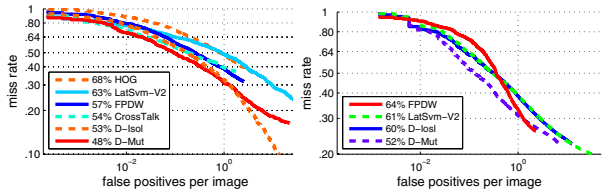

Figure 7. Experimental results on the Caltech-Test dataset (left) and the PETS2009 dataset (right).

Figure 8. 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.

5.2. Experimental Results on the effect of mutual visibility deep model

The pedestrians evaluated in Section 5.1 are divided into two subsets: *Isolated* and *Overlapped* in this section. A ground truth pedestrian is considered as overlapped if the overlapping area ratio of his/her bounding box with other pedestrian is larger than 1%. Pedestrians overlapping with other pedestrians are called O-pedestrians in this paper. The O-pedestrians are in subset *Overlapped*. Isolated Pedestrians are in subset *Isolated* and called I-pedestrians. Our approach focuses on O-pedestrians. Note that an O-pedestrian is still a pedestrian. Fig. 9 shows examples on I-pedestrian and O-pedestrian.

In this section, we specifically evaluate the detection performance on detecting I-pedestrians and O-pedestrians. Fig. 10 shows the experimental results. The improvement of our approach over D-Isol is 2% on detecting I-pedestrians for both datasets. On the other hand, the improvement on detecting O-pedestrians is 5% for the Caltech-Train dataset and 7% for the ETH dataset. Thus our approach outperforms D-Isol on pedestrian detection mainly because of its ability in detecting co-existing pedestrians, i.e. O-pedestrians. This experimental result validates the focus of this paper in improving the performance on detecting pedestrians with overlaps.

Figure 9. I-pedestrian versus O-pedestrian.

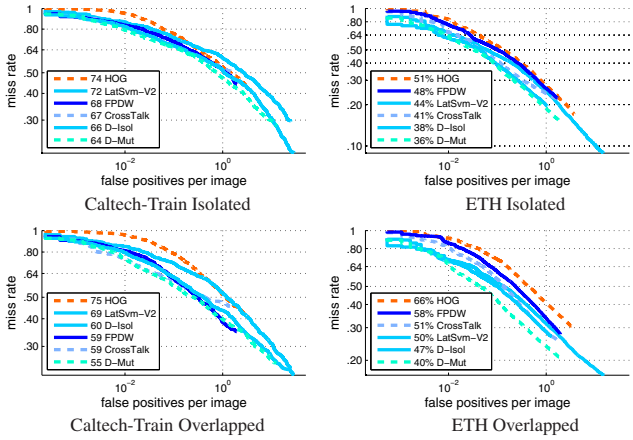

Figure 10. Experimental results on detecting Isolated pedestrians (top) and Overlapped pedestrians (bottom) on the Caltech-Train dataset (left) and the ETH dataset (right).

6. 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. 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 [9], our approach achieves the lowest miss rate on the Caltech-Train dataset, the Caltech-Test 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.

Acknowledgment: This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project No. CUHK 417110 and CUHK 417011), National Natural Science Foundation of

China (Project No. 61005057), and Guangdong Innovative Research Team Program (No.201001D0104648280)..

References

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

[2] S. Dai, M. Yang, Y. Wu, and A. Katsaggelos. Detector ensemble. In *CVPR*, 2007. 2

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

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: a large-scale hierarchical image database. In *CVPR*, 2009. 2

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

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

[7] P. Dollár, R. Appel, and W. Kienzle. Crosstalk cascades for frame-rate pedestrian detection. In *ECCV*, 2012. 6

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

[9] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: an evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743 – 761, 2012. 6, 7

[10] G. Duan, H. Ai, and S. Lao. A structural filter approach to human detection. In *ECCV*, 2010. 2

[11] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *CVPR*, 2010. 2

[12] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? *J. Machine Learning Research*, 11:625–660, 2010. 4

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

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

[15] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002. 5

[16] G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006. 2, 5

[17] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507, July 2006. 2

[18] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *CVPR*, 2009. 2

[19] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 3

[20] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML*, 2009. 2

[21] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, 2005. 2

[22] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon. Hierarchical part-template matching for human detection and segmentation. In *ICCV*, 2007. 2

[23] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *CVPR*, 2012. 2

[24] M. Norouzi, M. Ranjbar, and G. Mori. Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning. In *CVPR*, 2009. 2, 3

[25] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, 2012. 1, 2, 3, 6

[26] W. Ouyang and X. Wang. Single-pedestrian detection aided by multi-pedestrian detection. In *CVPR*, 2013. 2

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

[28] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton. On deep generative models with applications to recognition. In *CVPR*, 2011. 2

[29] V. D. Shet, J. Neumann, V. Ramesh, and L. S. Davis. Bilattice-based logical reasoning for human detection. In *CVPR*, 2007. 2

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

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

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

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

[34] B. Wu and R. Nevatia. Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. *Int'l J. Computer Vision*, 82(2):185–204, 2009. 2

[35] T. Wu and S. Zhu. A numeric study of the bottom-up and top-down inference processes in and-or graphs. *Int'l J. Computer Vision*, 93(2):226–252, Jun. 2011. 2

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

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

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

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

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