A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling

Wanli Ouyang and Xiaogang Wang
Department of Electronic Engineering, The Chinese University of Hong Kong
wlouyang@ee.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk

Abstract

Part-based models have demonstrated their merit in object detection. However, there is a key issue to be solved on how to integrate the inaccurate scores of part detectors when there are occlusions or large deformations. To handle the imperfectness of part detectors, this paper presents a probabilistic pedestrian detection framework. In this framework, a deformable part-based model is used to obtain the scores of part detectors and the visibilities of parts are modeled as hidden variables. Unlike previous occlusion handling approaches that assume independence among visibility probabilities of parts or manually define rules for the visibility relationship, a discriminative deep model is used in this paper for learning the visibility relationship among overlapping parts at multiple layers. Experimental results on three public datasets (Caltech, ETH and Daimler) and a new CUHK occlusion dataset[1] specially designed for the evaluation of occlusion handling approaches show the effectiveness of the proposed approach.

1. Introduction

Object detection is a fundamental problem in computer vision with wide applications such as surveillance, image retrieval, robotics and intelligent vehicles. Since pedestrian detection is one of the most important topics in object detection, it has attracted much attention in recent years.

Many classification approaches, features and deformation models have been used for achieving the progress on object detection. The classification approaches widely used include various boosting classifiers [7, 32], linear SVM [5], histogram intersection kernel SVM [20], latent SVM [12], multiple kernel SVM [26] and structural SVM [34]. The investigation on features includes Haar-like features, histogram of gradients (HOG), integral histogram, color histogram, gradient histogram, covariance descriptor, local binary pattern, features learned from deep model, depth, segmentation and motion [1, 5, 7, 10, 19, 22, 25, 27, 29, 28].

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

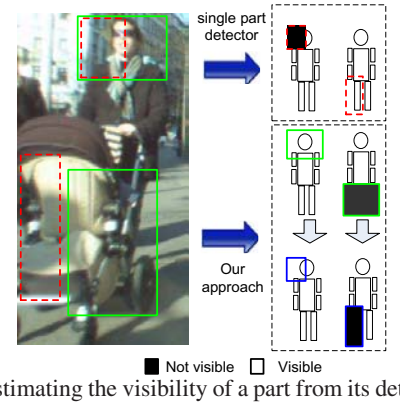


Figure 1. Estimating the visibility of a part from its detection score or from its correlated parts. Rectangles with dashed line denote wrong estimation, rectangles with solid line denote correct estimation. Parts estimated as not visible are represented by black rectangles. Single part detection score gives wrong visibility estimation. With the help of visibility correlation among parts, our approach can find the correct estimation successfully.

Recent deformable models for object detection mainly model the translational deformation of parts [12, 21, 34].

Generic detectors [5, 29, 12, 34, 22] assume that pedestrians are fully visible, and their performance degrades when pedestrians are partially occluded. For example, many part-based deformable models [12, 34] summed the scores of part detectors. A pedestrian-existing input window is considered as having high summed score. If one part is occluded, the score of its part detector could be very low and consequently the summed score will also be low. However, occlusions occur frequently, especially in crowded scenes. As pointed out in [10], the key to successful detection of partially occluded pedestrians is to utilize additional information about which body parts are occluded. For example, the additional information used in [10] was from motions, depth and segmentation results. In this paper, it is only inferred from the appearance of single images through exploring the strong correlations among the visibilities of different parts with multiple sizes. Once the occluded parts are identified, their effect should be properly removed from the final combined score.

Most previous approaches [4, 10, 29, 32] rely on the detection score of a part for estimating its visibility. However, part detectors are imperfect and such estimation is inaccu-

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

rate. Take the pedestrian in Fig. 1 as an example. Although the part of left-head-shoulder is visible, its part detection score is relatively low because its visual cue in the image does not fit the part detector well. Although the part of left-leg is invisible, its part detector finds a meaningless false-positive window on the baby carriage with a relatively high detection score. If the detection scores of parts are directly used for estimating visibility, the pedestrian will be wrongly estimated as having left-head-shoulder invisible and left-leg visible.

This paper is motivated by the fact that it is more reliable to design overlapping parts at multiple layers and verify the visibility of a part for multiple times at different layers. The detection score of one part provides valuable contextual information for the estimation on its overlapping parts. Take the pedestrian in Fig. 1 as an example. The left-head-shoulder and head-shoulder are overlapping parts at different layers. Similarly for the left-leg and the two-legs. The part of head-shoulder has a high detection score because its visual cue in the image fits the corresponding part detector well. And the part of two-legs has a low detection score because it does not find any visual cue to fit the detector. If the correlation among parts is modeled in a correct way, the detection score of the head-shoulder can be used to recommend the left-head-shoulder as visible and that of the two-legs can be used to recommend the left-leg as invisible. Therefore, the major challenges are how to model the relationship of the visibilities of different parts and how to properly combine the results of part detectors according to the estimation of part visibility.

There are two contributions of this paper.

1. A probabilistic framework for pedestrian detection which models the visibility of parts as hidden variables. It is shown that various heuristic occlusion handling approaches (such as linear combination and hard-thresholding) are considered as its special cases but did not fully explore its power in modeling the correlations of different parts.

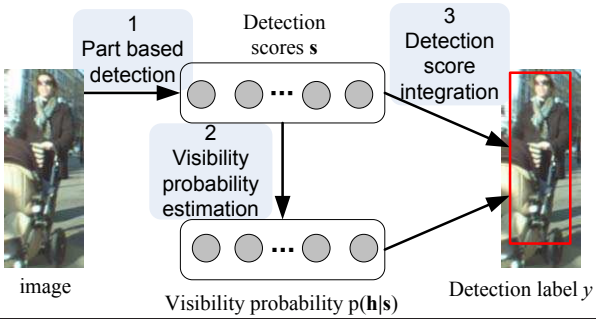
2. A discriminative deep model to learn the correlations of different parts, which is inspired by the great success of deep models [2, 15, 18] in various applications of dimension reduction [16] and recognition [15, 17, 18, 24]. The new model has some attractive features. First, the hierarchical structure of our deep model well matches with the multi-layers of the parts model. Different from the Deep Belief Networks (DBN) in [15, 16], whose hidden variables had no semantic meaning, our model consider each hidden variable as representing the visibility of a part. By including multiple layers, our deep model achieves a better variational lower bound on the training data, and in the meanwhile, achieves more reliable visibility estimation. Second, it well models the complex probabilistic connections across layers with good efficiency on both learning and inference. Third, our discriminative deep model only uses the scores of part detectors as input without requiring any occlusion information for supervision at the training stage.

Finally, although the above discussions focus on occlusions, the proposed framework is also effective to handle abnormal deformations to some extent. If some parts are abnormally deformed and cannot be detected by part detectors, they can be treated as occlusions and removed from the integration of parts.

2. Related Work

Deformation and occlusion are two major problems to be solved in object detection. To handle the deformation problem, deformable part-based models have been widely used [12, 21, 34]. In these models, the appearance of each part and the deformation among parts were considered for detection. For example, the state-of-the-art approach in [12] combined both the appearance score and the translational deformation score. To model the deformation, the star model was used in [12], the tree model was used in [21, 34] and a loopy graph model was used in [30]. Detectors using boosting to select features from a large pool of local candidate features also consider objects as being composed of parts [6, 7, 25].

Since visibility estimation plays a key role for detectors in handling occlusions, various approaches [4, 9, 10, 19, 29, 32, 33] were proposed to estimate the visibility of parts. Most existing approaches [4, 10, 19, 29, 32, 33] assumed that the visibility of a part is independent of other parts and estimated the visibility by hard-thresholding the detection scores of parts. Recently, Duan et al. [9] used manually defined rules to describe the relationship between the visibility of a part and its overlapping larger parts and smaller parts, e.g. if the head or the torso are invisible, its larger part of upper-body should also be invisible. In their approach, the binary visibility status of a part is obtained by hard-thresholding its detection score. Then rules are used to determine whether the combination of the binary visibility status of different parts is correct. If yes, the current window is detected as positive; otherwise, negative. This approach has certain drawbacks. First, hard-thresholding does not distinguish partial occlusions from full occlusions. A probabilistic model would be a more reasonable way to describe occlusions. Second, a larger part that is misclassified as being occluded by hard-thresholding its detection score cannot be corrected by the rules. Third, the rules were defined manually but not learned from training data. The relationship among the visibilities of parts systematically learned from training data may open the door to more robust methods with a wider spectrum of applications. Considering the problems faced by the approaches discussed above, we propose to use a discriminative deep model to automatically learn the probabilistic dependency of the visibilities of different parts.



1. obtain the detection scores $\mathbf{s}$ by part detector;
2. use the $\mathbf{s}$ to estimate visibility probability $p(\mathbf{h}|\mathbf{s})$;
3. combine the detection scores with the visibility probability to estimate the probability of an input window being pedestrian $p(y|\mathbf{s})$, c.f. (3).

Figure 2. Framework overview.

3. A Framework for Pedestrian Detection with Hidden Occlusion Variables

Denote the label of the current detection window by y. Denote the detection scores of the P parts by $\mathbf{s} = [s_1, \ldots, s_P]^{\mathrm{T}}$. In this paper, it is assumed that part-based models have integrated both the appearance scores and the deformation scores into $\mathbf{s}$. Denote the visibility of the P parts by $\mathbf{h} = [h_1, \ldots, h_P]^{\mathrm{T}} \in \{0, 1\}^P$, with $h_i = 1$ meaning visible and $h_i = 0$ meaning invisible. The overview of the framework is shown in Fig. 2.

Since $\mathbf{h}$ is not provided at the training or testing stages, it is a hidden random vector. $p(y|\mathbf{s})$ can be obtained by marginalizing out hidden variables $\mathbf{h}$:

$$p(y|\mathbf{s}) = \sum_{\mathbf{h}} p(y, \mathbf{h}|\mathbf{s}) = \sum_{\mathbf{h}} p(y|\mathbf{h}, \mathbf{s}) p(\mathbf{h}|\mathbf{s}). \quad (1)$$

It can be implemented by setting $p(y|\mathbf{h}, \mathbf{s}) = e^{\sum_i y h_i s_i}$:

$$p(y|\mathbf{s}) = \sum_{\mathbf{h}} e^{\sum_i y s_i h_i} p(\mathbf{h}|\mathbf{s}). \quad (2)$$

The computational complexity of (2) is exponential to the dimension of $\mathbf{h}$. A faster and approximate solution to the above is as follows:

$$p(y|\mathbf{s}) \approx e^{\sum_i y s_i \tilde{h}_i} / Z. \quad (3)$$

$Z = 1 + e^{\sum_i s_i \tilde{h}_i}$ is the partition function to have $\sum_y (e^{\sum_i y s_i \tilde{h}_i} / Z) = 1$. $\tilde{h}_i$ is sampled from $p(h_i|\mathbf{h} \backslash h_i, \mathbf{s})$, or alternatively calculated by a mean-field approximation, in which instead of the average over all $\mathbf{h}$ configurations according to $p(\mathbf{h}|\mathbf{s})$ in (2) one replaces $\mathbf{h}$ by its average configuration $\tilde{\mathbf{h}} = E[\mathbf{h}|\mathbf{s}]$. The mean-field approximation is also used in [14, 15, 16] for computing the posterior of RBM and DBN. More details are provided in [2]. Different approaches have different implementations of the $\tilde{h}_i$ in (3). $\tilde{h}_i$ is call the visibility term.

Many deformable part-based models [12, 23, 34] directly sum up part-based detection scores. They can be considered as setting $\tilde{h}_i = 1$ in (3) and have

$$p(y = 1|\mathbf{s}) \approx exp(\sum_i s_i)/Z \propto \sum_i s_i. \quad (4)$$

After obtaining $\mathbf{s}$ from the part-based model, many occlusion handling methods calculate the $p(y|\mathbf{s})$ by weighted sum of detection scores. These approaches obtain the $\tilde{h}_i$ in (3) from thresholding detection scores [29, 33], from linear SVM in [10] when only intensity information is available or from other cues like depth and motion [10]. With deformation among parts and multiple cues already integrated into s_i, these approaches assume that the $\tilde{h}_i$ in (3) is dependent on s_i, i.e. $\tilde{h}_i = f(s_i)$, where f is the mapping of s_i to $\tilde{h}_i$.

In summary, many approaches are special cases of the framework in (3) by setting $\tilde{h}_i = 1$ or by considering the visibility term $\tilde{h}_i$ as depending on s_i. The power of this framework in considering the visibility relationship among parts is not explored. In this paper, we explore this power and construct a deep model that learns the visibility relationship among parts. In our model, $\tilde{h}_i = p(h_i|\mathbf{h} \backslash h_i, \mathbf{s}) \neq p(h_i|s_i)$ and $p(h_i|\mathbf{h} \backslash h_i, \mathbf{s})$ is learned from a deep model that will be introduced in the next section [15].

4. The Discriminative Deep Model for Part Visibility Estimation

4.1. The Restricted Boltzmann Machine (RBM)

Since RBM is a building block of our deep model introduced in the next section, a brief introduction on RBM is provided. Denote the binary visible variables by vector $\mathbf{x} = [x_1, x_2, \ldots, x_N]^{\mathrm{T}}$. Denote the binary hidden variables by $\mathbf{h}$. The RBM defines a probability distribution over $\mathbf{h}$ and $\mathbf{x}$ as

$$p(\mathbf{x}, \mathbf{h}) \propto e^{\left[\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{h} + \mathbf{c}^{\mathrm{T}}\mathbf{h} + \mathbf{b}^{\mathrm{T}}\mathbf{x}\right]}. \quad (5)$$

$\mathbf{x}$ forms the visible layer and $\mathbf{h}$ forms the hidden layer. There are symmetric connections $\mathbf{W}$ between the visible layer and the hidden layer, but no connection for variables within the same layer. The graphical model of RBM is shown in Fig. 3(a). This particular configuration makes it easy to compute the conditional probability distributions:

$$\begin{aligned} p(x_n = 1|\mathbf{h}) &= \sigma(\mathbf{w}_{n,*}\mathbf{h} + b_n), \\ p(h_i = 1|\mathbf{x}) &= \sigma(\mathbf{x}^{\mathrm{T}}\mathbf{w}_{*,i} + c_i), \end{aligned} \quad (6)$$

where $\mathbf{w}_{n,*}$ is the nth row of $\mathbf{W}$, $\mathbf{w}_{*,i}$ is the ith column of $\mathbf{W}$ and $\sigma(t) = (1 + exp(-t))^{-1}$ is the logistic function. The contrastive divergence in [14] is used for learning the parameters $\mathbf{W}$, $\mathbf{c}$ and $\mathbf{b}$ in (5).

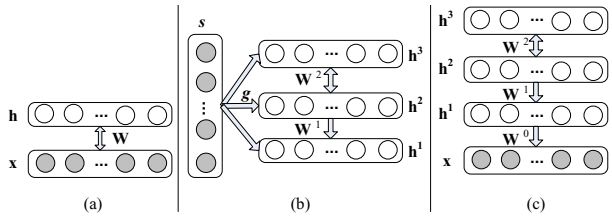


Figure 3. **(a)** RBM, **(b)** our deep model and **(c)** DBN.

4.2. The Deep Model for Visibility Estimation

Parts model. Our parts model consists of 3 layers that have different sizes of parts as shown in Fig. 4. Parts at the bottom layer have the smallest size, and those at the top layer have the largest. A part at upper layer is composed of its children in the lower layer. The top layer is the possible occlusion statuses. Gray color indicates occlusion. The other two layers are body parts. An occlusion status is obtained by combining one or several parts in the middle layer. The leftmost part, i.e. head-shoulder, appears twice (representing occlusion status at the top layer and part in the middle layer respectively) in this figure because this part itself can generate an occlusion status.

Deep model. The graphical model of the proposed deep model is shown in Fig. 3(b). Detailed information is shown in Fig. 4. Denote the visibility of P_l parts in layer l by $\mathbf{h}^l = [h^l_1 \ \ldots h^l_{P_l}]^{\mathrm{T}}$. There are connections for variables between adjacent layers and no connections for variables within the same layer. A part can have multiple parents and multiple children. In this way, the visibility of one part is correlated with the visibility of other parts at the same layer through shared parents. Given $\mathbf{s}$, the probability distribution of $\mathbf{h}^1, \ldots, \mathbf{h}^L$ is as follows:

$$
\begin{aligned}
&p(\mathbf{h}^1, \ldots, \mathbf{h}^L|\mathbf{s}) = \left(\prod_{l=1}^{L-2} p(\mathbf{h}^l|\mathbf{h}^{l+1}, \mathbf{s})\right) p(\mathbf{h}^{L-1}, \mathbf{h}^L|\mathbf{s}),\\
&p(h^l_i|\mathbf{h}^{l+1}, \mathbf{s}) = \sigma(\mathbf{w}^l_{i,*}\mathbf{h}^{l+1} + g^l_i s^l_i + c^l_i),\\
&p(\mathbf{h}^{L-1}, \mathbf{h}^L|\mathbf{s})\\
&= e^{\left[\mathbf{h}^{L-1\mathrm{T}}\mathbf{W}^{L-1}\mathbf{h}^L + \mathbf{c}^{\mathbf{L-1}\mathrm{T}}\mathbf{h}^{L-1} + \mathbf{c}^{\mathbf{L}\mathrm{T}}\mathbf{h}^L + \mathbf{g}^{\mathbf{L-1}\mathrm{T}}\mathbf{s}^{L-1} + \mathbf{g}^{\mathbf{L}\mathrm{T}}\mathbf{s}^L\right]}.
\end{aligned}
\tag{7}
$$

For the model in Fig. 4, we have $L = 3$. $\mathbf{W}^l$, g^l_i and $\mathbf{c}^l$ are the parameters to be learned. $\mathbf{W}^l$ models the correlation between $\mathbf{h}^l$ and $\mathbf{h}^{l+1}$, $\mathbf{w}^l_{i,*}$ is the ith row of $\mathbf{W}^l$, g^l_i balances the weights between the detection score s^l_i and correlation with other parts, and $\mathbf{c}^l$ is the bias term. Since the detection scores $\mathbf{s}$ are 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 $p(\mathbf{s})$. And this model can be considered as a conditional random field (CRF). Note that, given $\mathbf{s}$, h^l_i and h^l_j are not independent, i.e. $p(h^l_i, h^l_j|\mathbf{s}) \neq p(h^l_i|\mathbf{s})p(h^l_j|\mathbf{s})$, although they are independent given $\mathbf{h}^{l+1}$ and $\mathbf{s}$, i.e. $p(h^l_i, h^l_j|\mathbf{h}^{l+1}, \mathbf{s}) = p(h^l_i|\mathbf{h}^{l+1}, \mathbf{s})p(h^l_j|\mathbf{h}^{l+1}, \mathbf{s})$. In this way, the correlation among parts at the same layer is also modeled.

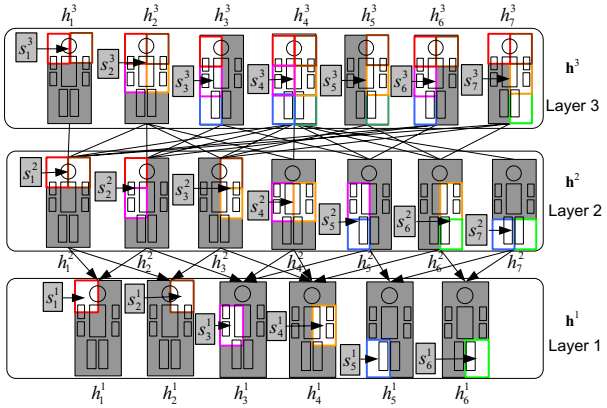


Figure 4. The parts model used. s^i_l is detection score of each part, h^l_i is the visibility of ith part in the lth layer. For example, h^1_1 indicates the visibility of the left-head-shoulder part.

Since the proposed model is a loopy graphical model, it is normally time consuming and hard to train. Hinton et al. [15, 16] proposed a fast learning algorithm for deep belief net (DBN) which has shown its success in many applications. In this work, we adopt a similar learning algorithm to train our model. DBN is a generative model and does not have semantic meanings for the hidden variables. Our model is a conditional model and has semantic meaning for each hidden variable. Because of these differences, the DBN algorithm cannot be directly used for our model. We modified the training and inference algorithms in [15] when we apply them for our model.

The training algorithm is to learn the visibility correlation $\mathbf{W}^l$, detection score weight $\mathbf{g}^l$ and bias $\mathbf{c}^l$ in (7), with two stages.

1. Stage 1: For l=1 to 2 { Train parameters for layer l and $l+1$ using RBM. }

2. Stage 2: Fine-tune all the parameters by backpropagating error derivatives.

At Stage 1, the parameters are trained layer by layer and two adjacent layers are considered as an RBM that has the following distributions:

$$
\begin{aligned}
p(h^l_i|\mathbf{h}^{l+1}, \mathbf{s}) &= \sigma(\mathbf{w}^l_{i,*}\mathbf{h}^{l+1} + c^l_i + g^l_i s^l_i),\\
p(h^{l+1}_j|\mathbf{h}^l, \mathbf{s}) &= \sigma(\mathbf{h}^{l\mathrm{T}}\mathbf{w}^l_{*,j} + c^{l+1}_j + g^{l+1}_j s^{l+1}_j).
\end{aligned}
\tag{8}
$$

The contrastive divergence in [14] is used for fast learning of the parameters in (8). In the appendix, we prove that this layer-wise training algorithm is optimizing a lower bound of the likelihood function. At Stage 2, the variables are arranged as a backpropagation (BP) network as shown in Fig. 5 for fine tuning all parameters.

The inference stage is to infer the label y from detection scores $\mathbf{s}$. At the inference stage, we use the framework in

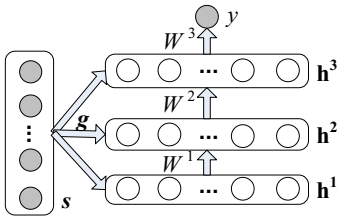


Figure 5. The BP network for fine tuning and estimating visibility.

(3) for obtaining $p(y|\mathbf{s})$. And the part visibility probability $\tilde{h}_i^{l+1}$ in (3) is obtained using the BP network in Fig. 5, i.e.

$$\begin{aligned}\tilde{h}_i^{l+1} &= p(h_i^{l+1}|\mathbf{h}\backslash h_i^{l+1}, \mathbf{s}) = p(h_i^{l+1}|\mathbf{h}^l, \mathbf{s}) \\ &= \sigma({\mathbf{h}^l}^{\mathrm{T}}\mathbf{w}_{*,j}^l + c_j^{l+1} + g_j^{l+1}s_j^{l+1}).\end{aligned} \tag{9}$$

In order to reduce bias of training data and regularize the training process, we enforce that the visibility correlation parameter W to be non-negative. Therefore, our training process have used the the prior knowledge that negative correlation among visibility of parts is unreasonable. Furthermore, the element $w_{i,j}^l$ of $\mathbf{W}^l$ in (7) is set to zero if there is no connection between units h_i^l and h_j^{l+1} in Fig. 4. In this way, we keep the most important edges based on our knowledge. There are other ways for modeling the connection among parts, e.g. the full-connected part models [3]. They could be helpful for finding more connections but will increase model complexity, reduce efficiency and need more training samples. They are not DBN, need more complex inference/learning algorithms, and may need pruning edges in training.

5. Experimental Results

The proposed framework is evaluated on four datasets: Caltech [8], ETHZ [11] and Daimler [10] datasets are publicly available; the CUHK occlusion dataset is constructed by us. The INRIA training dataset in [5] is used to train our approach. Occlusion information is not required for training. Once the model is learned from this training set, it is fixed and tested on the four datasets mentioned above.

In the experiment, we use the modified HOG in [12] as the feature for detection. In our implementation, the deformable part-based model in [12] is used for modeling the deformation among the 20 parts in Fig. 4. The parts are arranged in the star-model with the full body part being the root. Since the detection scores obtained from our parts model are considered as the input of our deep model, the deep model keeps unchanged if other deformable part-based models are used.

The approaches to be compared and our approach use the same features. They are also trained from the INRIA dataset. The evaluation criteria proposed in [8] is used. The labels and evaluation code provided by Dollár et al. online [2] is used for evaluating the Caltech dataset and the ETHZ dataset. As in [8], *log-average miss rate* is used to summarize the detector performance, computed by averaging miss rate at nine FPPI rates evenly spaced in log-space in the range from 10^{-2} to 10^0.

Table 1. The composition of the dataset.

Dataset	Images	Dataset	Images
Caltech train [8]	105	INRIA test [5]	70
TUD-Brussels [31]	110	ETHZ [11]	211
Caviar [13]	355	Our	212

5.1. Experimental Results on the CUHK Occlusion Dataset

Most existing datasets are not specifically designed for evaluating occlusion handling. For example, in the Caltech dataset, only 105 out of 4250 images for evaluation have occluded pedestrians. If such datasets are used for evaluation, it is not clear how much improvement comes from occlusion handling or other factors. In order to specifically compare pedestrian detection algorithms under occlusions, we construct the CUHK occlusion dataset that mainly include occluded images. This dataset contains 1063 images from the datasets of Caltech, ETHZ, TUD-Brussels, INRIA, Caviar and our recorded images from surveillance cameras. The composition of the dataset is shown in Table 1. Images are strictly selected according to the following criteria.

1. Each image contains at least one occluded pedestrian.
2. Datasets Caviar and ETHZ are video sequences with high frame rate, e.g. 25 frames per second for Caviar. In these datasets, the current frame may be very similar to the next frame. In our dataset, the frame rate is reduced to ensure variation among selected images.
3. The image shall not contain sitting humans, since it is potentially controversial whether they should be detected as pedestrian or not.

Each pedestrian is labeled with a bounding box and a tag indicating whether the pedestrian is occluded or not. Since a lot of occluded pedestrians in datasets like INRIA, ETHZ and TUB-Brussels are not considered positive testing samples, the occluded pedestrians are relabeled in our dataset. Occluded pedestrians have been labeled in Caltech dataset, their labels are unchanged in our dataset.

We evaluate the performance of our approach on occluded pedestrians and unoccluded pedestrians separately and compare with two part-based models proposed by Zhu et al. [34] and LatSVM-V2 in [12] in Fig. 7. Our approach has similar performance with [34] and [12] on unoccluded pedestrians and achieved 5% improvement on occluded pedestrians. To investigate the effectiveness of using the deep model to estimate the visibility of parts, we also test our part model that directly sums up detection score using (4) and exclude the deep model. It has comparable performance as [34] and [12] on occluded pedestrians.

[2] Available on www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

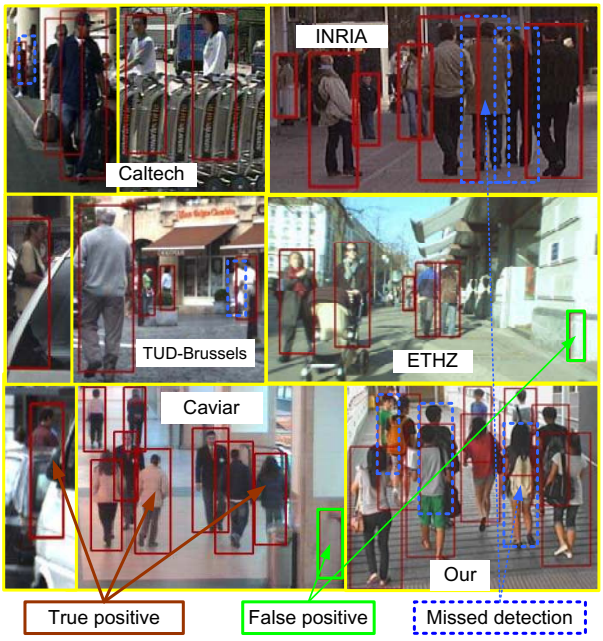


Figure 6. Selected detection results using our framework on the constructed dataset. The sources of images are given.

In order to investigate various schemes for integrating the part detection scores, we conduct another set of experiments in Fig. 7(c)-(f). They all use our parts model and therefore have the same detection scores as input. *Our-P* in Fig. 7 is the weighted mean of part scores and the weights are trained by linear SVM. Fig. 7(c) and (d) show the results of estimating the visibility by thresholding the detection scores, i.e. part score s_i is ignored if $s_i < T_i$. Using the same T_i for all the parts is not optimal. Therefore, we assume that different parts have different threshold T_i and obtain T_i from training data. For each part, T_i is chosen such that certain percentage $\epsilon(= 0.1\%, 1\%, 5\%, 10\%, 20\%, 40\%, 50\%)$ of parts on the positive training samples are considered as occlusions. The approach in [9] defines rule for estimating visibility of parts and integrating detection scores. We use the same rules proposed in [9] to integrate our part scores. As shown in Fig. 7 (c) and (d), the rule based integration does not work well on our parts model although it has reported satisfactory results on the parts model in [9]. This may be due to the fact that we use different features and different parts model from [9]. We cannot exactly obtain the results in [9] on our dataset because its implemenation is not available. The *DBN* in Fig. 7 arranges all part detection scores as the bottom visible layer and 3 layers of hidden units on top of the visible layer as shown in Fig. 3(c). The approach in [15] is then used for training parameters and classifying whether an input window is a pedestrian or not. Fig. 7(e) and (f) show the results of taking $k = 1, 2, 4, 8, 10, 15, 18$ maximum part scores for computing the weighted mean. The experimental results show that all the schemes discussed above perform worse than our deep model (represented by $Ours - D$).

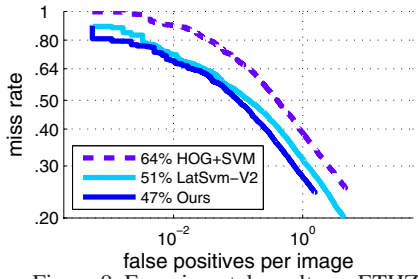


Figure 9. Experimental results on ETHZ.

Since our deep model has only 20 hidden variables in all for 3 layers, training and inference time for the deep model is much less than that for the parts model.

5.2. Experimental Results on Caltech

Since most other relevant publications [28, 1] test on Caltech training dataset and use other datasets as training datasets, we choose the Caltech training dataset as our testing set and the INRIA training dataset as our training set to be consistent with them. In Fig. 8, we compare with HOG+SVM and LatSVM-V2, whose results were published in [8], under varying levels of occlusion. Compared with LatSVM-V2, our approach has 8%, 9% and 3% log-average miss rate improvement for pedestrians with no occlusions, partial occlusions and heavy occlusions respectively. Compared with the 14 state-of-the-art approaches evaluated in [8] (excluding those using motions), our approach ranks as the third, the second and the first for pedestrians with no occlusions, partial occlusions and heavy occlusions respectively. The two approaches [28, 7], which performed better than ours in the cases of no occlusions and partial occlusions, both used a large number of extra features such as color self-similarity, local sums, histograms, Haar features and their various generalizations beside HOG. Only HOG+SVM, LatSVM-V2 and our approach used the same features. With more features being included, the performance of our approach can be further improved.

5.3. Experimental Results on ETHZ

The experimental results on the ETHZ testing sequences are shown in Fig. 9. It is reported in [8] that LatSvm-V2 has the best performance among the 14 state-of-the-art approaches evaluated on the ETHZ dataset. It can be seen that our approach has 4% improvement over LatSVM-V2. The ETHZ dataset consists of 3 testing video sequences. Table 2 shows the miss rates at 1 FPPI for the 3 sequences. The results of ISF are obtained from [9]. The results of HOG+SVM and LatSvm-V2 are obtained from [8] using the results and code provided online by Dollár et al.

5.4. Experimental Results on Daimler Occluded Pedestrian Dataset

The experimental results on the Daimler benchmark testing data in [10] are shown in Fig. 10. Since the dataset is

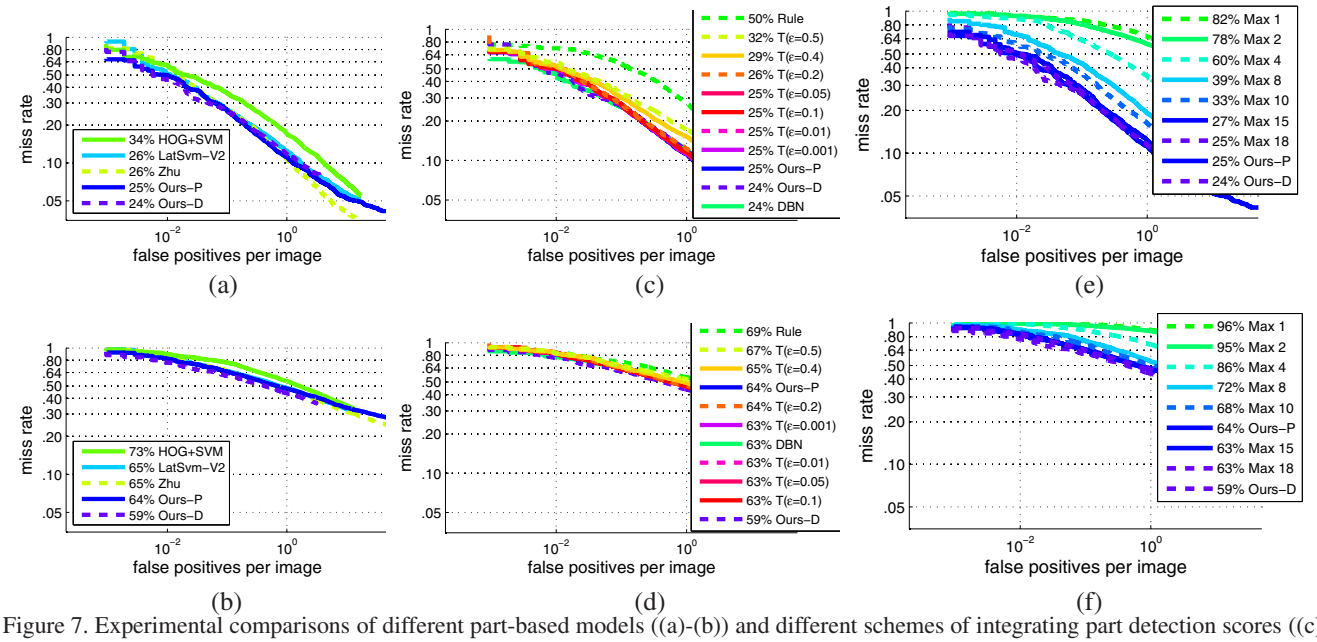


Figure 7. Experimental comparisons of different part-based models ((a)-(b)) and different schemes of integrating part detection scores ((c) - (f)) on our dataset for pedestrians *without occlusions* (upper row) and *with occlusions* (bottom row). *Zhu* denotes results using the parts model proposed by Zhu et al. in [34]. *Ours-P* denotes results of using our parts model in Fig. 4 and directly summing up detection score however without the deep model. In this case, it is equivalent to computing the weighted mean of part scores. *Ours-D* denotes the results of using our parts model and the discriminative deep model introduced in Section 4.2. *DBN* denotes the results of replacing our deep model by DBN. *Rule* denotes results of using the rule in [9] for integrating our part scores. *T(ϵ=ϵ_0)* denotes the results the estimating visibility by hard-thresholding. T_i is learned from the training data such that ϵ_0 percentage of parts in the positive training samples are considered as occlusions. *Max k* denotes taking the k maximum part scores for computing the weighted mean.

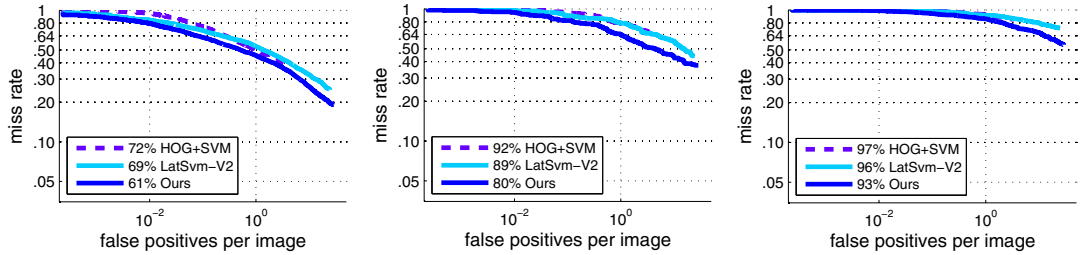


Figure 8. Experimental results on Caltech for pedestrians under *no occlusions* (left), *partial occlusions* (center) and *heavy occlusions* (right). The ratio of occluded area is larger than 0.65 for *partial occlusions* and [0.2 0.65] for *heavy occlusions*. The log-average miss rate of our model is 61% for no occlusions and 80% for partial occlusions.

Table 2. Miss rate at 1 FPPI for different approaches. Seq 1 has 999 frames, Seq 2 has 450 frames and Seq 3 has 354 frames.

	Seq 1	Seq 2	Seq 3
ISF [9]	47%	38%	52%
HOG+SVM [5]	34%	44%	44%
LatSvm-V2 [12]	30%	34%	32%
Ours	24%	33%	29%

used for occluded pedestrian classification instead of detection, false positive versus detection rate is used for evaluation. Since our focus is on detection for single images, we only use the image intensity for all evaluated algorithms. Compared with LatSVM-V2, our approach has similar performance on unoccluded pedestrian, and our approach achieves about 20% detection rate improvement for occluded pedestrian. LatSVM-V2, HOG+SVM and our approach in Fig. 10 are trained on INRIA for consistency with previous experimental results. Since all results in [10] are trained on the Daimler training data and have different implementation of HOG feature from ours, we did not show the results in [10]. For example, the HOG+SVM trained on INRIA using the code in [12] have quite different result from the HOG+SVM trained on Daimler training data reported in [10].

6. Conclusion

This paper describes a probabilistic framework for pedestrian detection with occlusion handling. It effectively estimates the visibility of parts at multiple layers and learns their relationship with the proposed discriminative deep model. Since it takes the detection scores of parts as input,

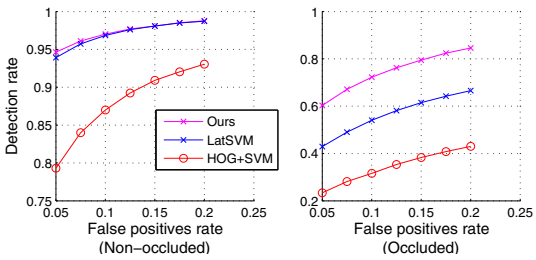


Figure 10. Experimental results on Daimler occlusion dataset.

it is very flexible to incorporate with new features and other deformable part-based models. Through extensive experimental comparison on multiple datasets, various schemes of integrating part detectors are investigated. Our approach outperforms the state-of-the-arts especially on pedestrian data with occlusions.

Acknowledgment: This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (project No. CUHK417110 and CUHK417011) and National Natural Science Foundation of China (project no. 61005057).

References

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

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

[3] M. Bergtholdt, J. H. Kappes, S. Schmidt, and C. Schnorr. A study of parts-based object class detection using complete graphs. *International Journal of Computer Vision*, 87(1-2):93–117, 2010.

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

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2005.

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

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

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

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

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

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

[12] P. Felzenszwalb, R. B. Grishick, D.McAllister, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 32:1627–1645, 2010.

[13] R. Fisher. homepages.inf.ed.ac.uk/rbf/caviar/.

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

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

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

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

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

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

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

[21] C. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV*, 2004.

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

[23] M. Pedersoli, J. Gonzlez, A. D. Bagdanov, and J. J. Villanueva. Recursive coarse-to-fine localization for fast object detection. In *ECCV*, 2010.

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

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

[26] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009.

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

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

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

[30] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR*, 2011.

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

[32] 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.

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

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