Neurocomputing ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Pedestrian detection based on hierarchical co-occurrence model for occlusion handling

Xiaowei Zhang [a], Hai-Miao Hu [a,b,*], Fan Jiang [a], Bo Li [a,b]

[a] Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[b] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 10 March 2015
Received in revised form
6 May 2015
Accepted 9 May 2015

Keywords:
Pedestrian detection
Partial occlusions
Co-occurrence relations
Visibility status

ABSTRACT

In pedestrian detection, occlusions are typically treated as an unstructured source of noise and explicit models have lagged behind those for object appearance, which will result in degradation of detection performance. In this paper, a hierarchical co-occurrence model is proposed to enhance the semantic representation of a pedestrian. In our proposed hierarchical model, a latent SVM structure is employed to model the spatial co-occurrence relations among the parent–child pairs of nodes as hidden variables for handling the partial occlusions. Moreover, the visibility statuses of the pedestrian can be generated by learning co-occurrence relations from the positive training data with large numbers of synthetically occluded instances. Finally, based on the proposed hierarchical co-occurrence model, a pedestrian detection algorithm is implemented to incorporate visibility statuses by means of a Random Forest ensemble. The experimental results on three public datasets demonstrate the log-average miss rate of the proposed algorithm has 5% improvement for pedestrians with partial occlusions compared with the state-of-the-arts.

1. Introduction

Pedestrian detection is an important topic for practical applications, such as video surveillances [9,29], intelligent vehicles [10], and robot sensing. State-of-the-art algorithms have been used for achieving progress on pedestrian detection. However, the presence of partial occlusions causes significant degradation of performance, even for part-based algorithms that are supposed to be robust in that respect [11]. Therefore, pedestrian detection is still a challenge [1–8].

A general part-based hierarchy approach is introduced for detection of partially occluded objects in [15]. The algorithm requires a design for hierarchical object-parts to place parts for sharing weak features. Enzweiler et al. [16] presented a mixture of experts to focus on the unoccluded region applied to depth and motion images for handling partial occlusion. Javier et al. [17] describe a general algorithm for building a robust classifier ensemble by random subspace algorithm against partial occlusions. Zhang et al. [28] proposed a latent hierarchical model with varying structures to represent the behavior with multiple groups, and employ a multi-layer-based inference method to infer the group affiliation. Girshick et al. [6] proposed an extension of the deformable part-based detector [18] with occlusion handling. Specifically, the algorithm tries to place different body parts over the window. However, most previous approaches rely only on the detection score of a part for estimating its visibility and do not consider spatial co-occurrence relations among body parts.

Recently, Duan et al. [19] proposed a structural filter approach to human detection to deal with occlusions and articulated poses. The method manually defines the rules to describe the relationship between the visibility of a part and its overlapping larger and smaller parts. However, the visibility status of a part is obtained by hard thresholding of its detection score. Quyang et al. [20] presented a probabilistic pedestrian detection framework to learn the visibility relationship among overlapping parts at multiple layers. However, this method subjectively designs seven visibility parts to integrate in the last layer and thus is unable to model complex occlusion statuses for pedestrian detection.

The above algorithms have improved the performance for pedestrian detection to some extent. However, these algorithms rely only on their respective detection scores of parts for estimating visibility or depend on spatial consistency among the adjacent visibility parts. These algorithms fail to capture strong correlations among random visible parts, especially complex dependencies among nonadjacent visible parts. Moreover, complex occlusion

* Corresponding author at: Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China.
E-mail address: frank0139@163.com (H.-M. Hu).

http://dx.doi.org/10.1016/j.neucom.2015.05.038
0925-2312/

patterns are often inevitable due to various viewpoint changes. These algorithms just manually design several visibility statuses to integrate adjacent parts and not learn from training data. Thus, the algorithms fail to represent complex occlusion patterns for pedestrian detection.

As shown in Fig. 1, single-part detectors are imperfect, and such visibility estimation is inaccurate. Part detection score is relatively low when its visual cue does not fit the part detector well no matter whether the partial occlusion occurs or not. Furthermore, it is a key issue that how to integrate the inaccurate scores of part detectors and to estimate its location when there is partial occlusion in the sliding window. For example, many part-based deformable models in [6,21] summed the scores of part detectors. A pedestrian existing input window is considered as having a high sum for its score. However, when one part is occluded, the score of its part detector will be relatively low. Consequently, the summed score will be low. If the part-based deformable model [6] is used to detect the image, these occluded pedestrians may be mistaken for negative examples with low summed scores.

Considering the problems faced by the approaches discussed above, this proposes a hierarchical co-occurrence model to automatically learn complex dependencies among different parts for occlusion handling. Spatial consistency is built among the parent–child pairs of nodes from multiple layers, which fully explore complex correlation among visible parts. Moreover, the co-occurrence relations among random visibility parts within the same layer are modeled as latent variables of the structural SVM to generate visibility statuses. Finally, the random forest is used to combine visibility statuses to build a classifier ensemble robust against partial occlusions. Based on the proposed hierarchical co-occurrence model, a pedestrian detection algorithm is implemented for partial occlusion handling. Experimental results on three public datasets demonstrate that the proposed algorithm improves the log-average miss rate by around 5% for pedestrians with partial occlusions compared with the state-of-the-art algorithms.

2. Hierarchical co-occurrence model

2.1. Representation of the hierarchical co-occurrence model

To deal with the variations that cannot be tackled by a monolithic model, approaches to learning multiple parts model have been introduced in [19,20]. Integrating the advantage of part-based detectors in occlusion handling requires solving two key issues for successful detection of partially occluded pedestrians. The first issue is the decision if partial occlusion occurs in a scanning window and which body parts are occluded. The second issue is integrating inaccurate scores of part detectors and estimating their locations if partial occlusion is found in the sliding window. Therefore, the major challenges are the modeling of the correlation of the visibilities of different parts and the proper combinations of the results of part detectors according to the estimation of component visibility.

Fig. 2 shows the proposed hierarchical co-occurrence model with latent variables. The top layer has a wide variety of visibility status, which represents the possibility of the appearance of learned behavior from positive training data with large numbers of synthetically occluded instances. A visibility status is obtained by randomly combining one or more parts in the middle layer. The second layer has 12 part nodes in a 4×3 grid layout. Each of which represents one part of an object. Each node at the second layer has 4 child nodes at the bottom layer that contains 32 block nodes in an 8×4 grid layout. The nodes at lower layers capture more detailed appearance.

Note that in our model, we followed the implementations of [7] to calculate HOG descriptors. Fig. 2 shows the weights of HOG

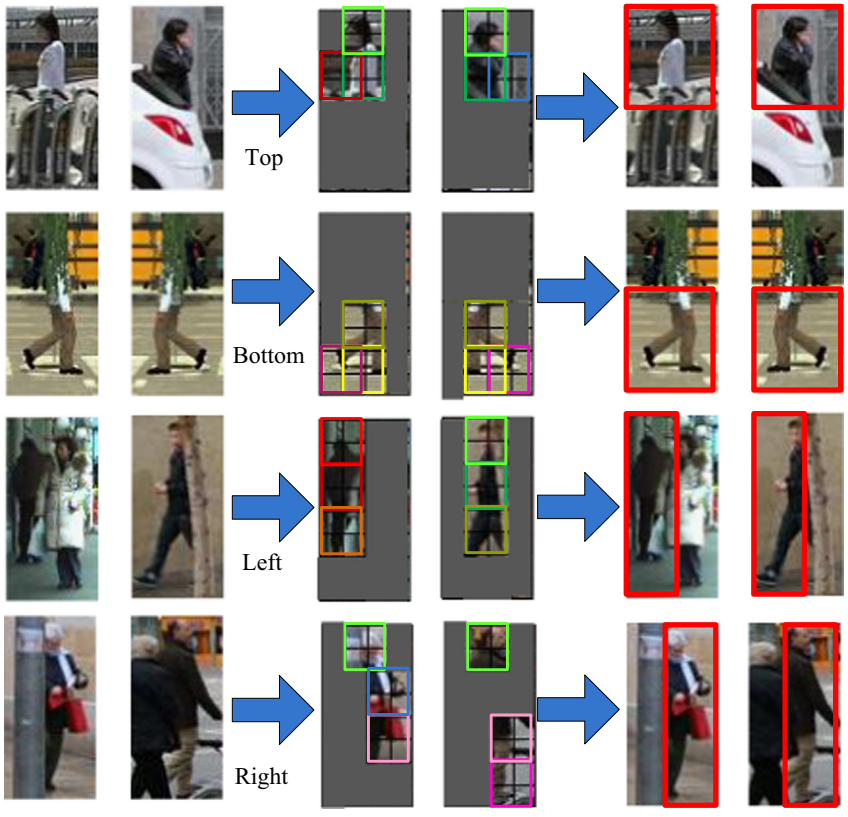

Fig. 1. Complex occlusion patterns and estimation of visibility of a part from its detection score or from its correlated components. Black regions represent the estimated parts that are not visible.

Please cite this article as: X. Zhang, et al., Pedestrian detection based on hierarchical co-occurrence model for occlusion handling, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.038

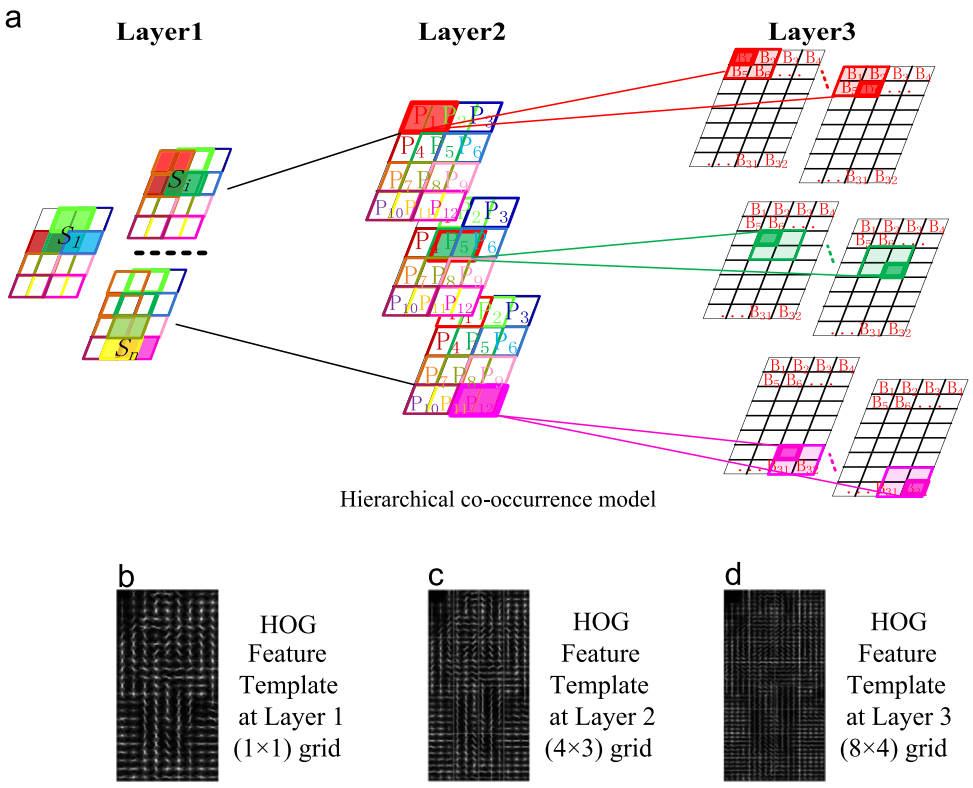

Fig. 2. Hierarchical co-occurrence model (The model is composed of a collection of HOG filters at different layers. HOG filters form a parent–child hierarchy where connections control the relative displacement of the parts).

descriptors at different layers. The features at different layers capture object appearance in a coarse-to-fine way.

In our model, we assume the model contains M visibility statuses in layer 1. The labels of all the blocks are denoted as $\boldsymbol{b}=(b_1, b_2,\ldots, b_{32})$ in layer 3, where b_i is the label of the i-th block. Similarly, the labels in layer 2 can be defined as $\boldsymbol{p}=(p_1, p_2,\ldots, p_{12})$, where p_i is the label of the i-th part. The labels in layer 1 can be signified as $\boldsymbol{s}=(s_1, s_2,\ldots,s_M)$, where s_i is the label of the i-th subspace. Let B, P, and S be the spaces of all the possible label set for the blocks, parts, and subspaces, respectively (i.e., $b \in B$, $p \in P$, and $s \in S$), as shown in Fig. 2.

Inspired by [20,28,31], we propose four types of co-occurrence relations to represent the interaction among the nodes in our hierarchical model. These four types of co-occurrence relations are block–block dependencies (in layer 3), block–part dependencies (between layers 3 and 2), part–part dependencies (in layer 2), and part–subspace dependencies (between layers 2 and 1), as shown in Fig. 3. The visibility status of a subspace is obtained by randomly combining one or several parts in the middle layer. The co-occurrence relations in our model allow parts to be switched off and thus are robust to partial occlusion parts or hard to detect parts. The visibility of one part is also correlated with the visibility of other parts at the same layer by modeling the co-occurrence relations among the parent–child pairs of nodes with sharing parents.

Let $h=(\boldsymbol{b}, \boldsymbol{p}, \boldsymbol{s})$ signify co-occurrence relations. We then construct a score function to evaluate the compatibility of candidate labels. To generalize the structural SVM formulation, we extend our joint feature vector $\Phi(x, y)$ with an extra argument h to $\Psi(x, y, h)$ to describe the relation among input x, output y, and latent variable h. Given a training set of input–output structure pairs $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, the score function $f_\omega(x, y)$ with latent variables h can be rewritten as

$$f_\omega(x,y) = \arg\max_{y,h} \omega^T \cdot \Psi(x,y,h) \tag{1}$$

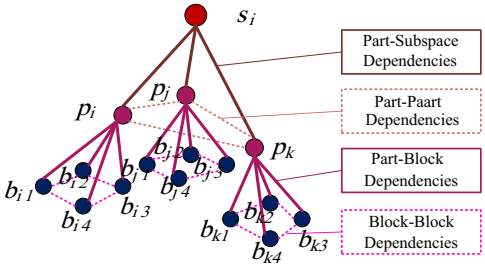

Fig. 3. Structure illustration of co-occurrence relations. Dashed lines and nodes indicate that the interaction and the node labels are latent.

Note that in our hierarchical model, given the extracted image features X, $\omega^T \cdot \Psi(x,y,h)$ can be calculated by Eq. (2)

$$\begin{aligned}\omega^T\Psi(x,y,h) = &\sum_{b_i\in B}\omega_1^T\phi_1(b_i,x_i)+\sum_{k=1}^{K}\sum_{\substack{(b_i^k,b_j^k)\in B,\\ p_k\in P}}\omega_2^T\phi_2(b_i^k,b_j^k,p_k)\\ &+\sum_{k=1}^{K}\sum_{\substack{b_i^k\in B,\\ p_k\in P}}\omega_3^T\phi_3(b_i^k,p_k)+\sum_{p_i\in P}\omega_4^T\phi_4(p_i,x_i)\\ &+\sum_{k=1}^{K}\sum_{\substack{(p_i^k,p_j^k)\in B,\\ s_k\in S}}\omega_5^T\phi_5(p_i^k,p_j^k,s_k)+\sum_{k=1}^{K}\sum_{\substack{p_i^k\in P,\\ s_k\in S}}\omega_6^T\phi_6(p_i^k,s_k)\Pi\Pi\end{aligned} \tag{2}$$

where $\omega=\{\omega_1,\omega_2,\omega_3,\omega_4,\omega_5,\omega_6\}$ is the concatenation of linear parameters for the interaction among the nodes in the hierarchical model. ω_1 and ω_4 indicate that the image is not only linked to the block level, it also connects the part level. ω_2, ω_3, ω_5 and ω_6 represent the four types of co-occurrence relations among the

Please cite this article as: X. Zhang, et al., Pedestrian detection based on hierarchical co-occurrence model for occlusion handling, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.038

nodes in our hierarchical model. $\Psi=\{\phi_1,\phi_2,\phi_3,\phi_4,\phi_5,\phi_6\}$ is the concatenated vector of all features given feature X and candidate labels ($\boldsymbol{b}$, $\boldsymbol{p}$, $\boldsymbol{s}$).

Therefore, the detection process with our hierarchical model is to learn a suitable model parameter vector ω and to infer optimal labels ($\boldsymbol{b}$, $\boldsymbol{p}$, $\boldsymbol{s}$).

2.2. Model learning with latent variables

To learn the structure SVM model and parameter vector ω, the model is optimized by minimizing the objective function, following the structure SVM formulation of [22]:

$$L(\omega)=\min_{\omega}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\left[\max_{\hat{y}_i}\left[\omega^T\psi(x_i,\hat{y}_{ii})+\Delta(y_i,\hat{y}_i)\right]-\omega^T\psi(x_i,y_i)\right] \tag{3}$$

where C is the regularization parameter and $\Delta(y_i,\hat{y}_i)$ is the 0–1 loss function, $\Delta(y_i,\hat{y}_i)=1$ if $y_i\neq\hat{y}_i$, and $\Delta(y_i,\hat{y}_i)=0$.

In our hierarchical co-occurrence model, the input–output relationship is characterized by (x, y) pairs in the training set and depends on a set of unobserved latent variables $h=(\boldsymbol{b},\boldsymbol{p},\boldsymbol{s})$. As [30] used the structure SVM to learn the latent hierarchical model, the structural SVM formulation with latent variables adopted for learning is as follows:

$$L(\omega)=\min_{\omega}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\left(\max_{\hat{y}_i,\hat{h}}\left[\omega^T\cdot\psi(x_i,\hat{y}_i,\hat{h})+\Delta(y_i,\hat{y}_i,\hat{h})\right]\right)-C\sum_{i=1}^{n}\left(\max_{h}\omega^T\cdot\psi(x_i,\hat{y}_i,\hat{h})\right) \tag{4}$$

Note that parameters $h=(\boldsymbol{b},\boldsymbol{p},\boldsymbol{s})$ define the co-occurrence relations of the image and are considered as latent variables. It is easy to observe that the above formulation reduces to the usual structural SVM formulation in the absence of latent variables. The variable models spatial consistency among the parent–child pairs of nodes from multiple layers.

A latent SVM leads to a non-convex optimization problem. However, a latent SVM is semi-convex, and the training problem becomes convex once latent information is specified for positive training examples. To solve the optimization problem of Eq. (4), the incremental concave–convex procedure (iCCCP) [12] can be rewritten as the difference of two convex functions

$$L(\omega)=\min_{\omega}\{f(\omega)-g(\omega)\} \tag{5}$$

where f and g are both convex, but $f(\omega)-g(\omega)$ is not. The cutting plane algorithm [30] is employed to solve the standard structural SVM optimization problem inside the iCCCP loop. The implementation procedures of the iCCCP are shown in Algorithm 1.

Algorithm 1. Incremental concave–convex procedure algorithm

Initialization: $t=0$; $S=\{x_i,y_i\}^+\cup\left\{x_j,y_j\right\}^-$, $i=1,2...,N^+$, $j=1,2...,N^-$

Repeat $t=t+1$

1. Fill in latent variables $(x_i,y_i)\in S$:

$h_i^*=\arg\max_h\omega_t\cdot\Psi(x_i,y_i,h)$

2. Solve the structure SVM problem over S (given h, estimate ω):

$\omega_{t+1}=\mathrm{argmin}_{\omega}f(\omega)-C\sum_i\omega\cdot\Psi(x_i,y_i,h_i^*)$

3. $S=S\cup\left\{x_j,y_j\right\}^-$, $j=N^-K^{t-1}+1,N^-K^{t-1}+2,...,N^-K^t$

Until $[f(\omega_t)-g(\omega_t)]-[f(\omega_{t-1})-g(\omega_{t-1})]<\delta$.

In our algorithm, the iCCCP starts from a small number $N^-=30$ (set by hand) of negative images, learns ω given the hard negative examples selected from 30 images, and proceeds to update ω by incrementally adding more negative images into the training set. The scaling factor K of new negative images is 1.15. The tolerance parameter is $\delta=10^{-6}$. After several rounds, the number of hard negative examples decreases exponentially.

We learn a latent SVM model ω to predict the visibility statuses of an unseen image window through giving a set of training samples $X=\{x_1,x_2,...,x_n\}$ and their labels $Y=\{y_1,y_2,...,y_n\}$, where each $y_i=\{-1,1\}$, $i=\{1,2,...,n\}$. We also use latent parameters $h=(\boldsymbol{b},\boldsymbol{p},\boldsymbol{s})$ to model the spatial configuration and co-occurrence relations among the components of image windows for the estimation of visibility statuses.

2.3. Model inference for visibility estimation

Generally, partial occlusions can vary considerably in terms of shape and size. Hence, a flexible model is needed to handle various occlusion statuses. Fortunately, the valuable visibility status subspace s_k can be generated by modeling the co-occurrence relations among component p_i as latent variable, which can have a more discriminative performance compared with the scores of part detectors.

Fig. 4 shows that discriminative visibility statuses are built by combining three pairs of subspace against partial occlusions. This step will result in weak classifiers being more concentrated in non-occluded areas than when training a non-constrained classifier. The key insight of this work is the possibility of changing the spatial distribution of the regions selected by the random subspace for estimating the label of the sliding image window accurately.

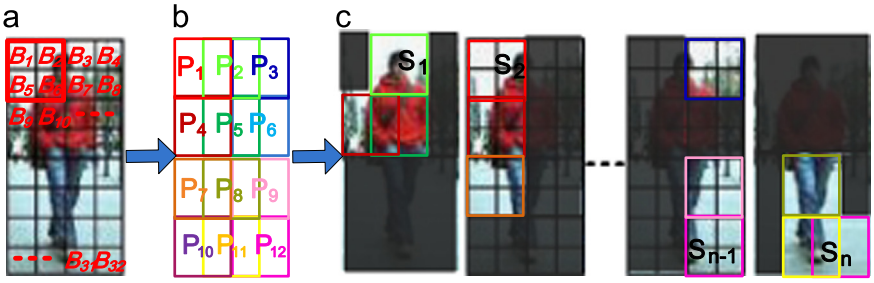

Fig. 4. Visibility status by modeling the co-occurrence relations among components. The features at different layers capture pedestrian appearance in a down-to-up way. The features at lower layers capture more detailed appearances. (a) Block Representation, (b) Part Representation and (c)Visibility Statuses Representation.

Please cite this article as: X. Zhang, et al., Pedestrian detection based on hierarchical co-occurrence model for occlusion handling, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.038

To infer visibility statuses, spatial consistency relationship among components is built to decide the visibility of the object region. Each block b_i is the smallest unit at the bottom layer and is the basis to build the upper layer structure detector. Every four adjacent block b_i is modeled as a consistency part unit p_i through building the co-occurrence relations among their discriminative local classifier, as shown in Fig. 4. Visibility part p_i is well provided with anti-interference ability.

In the case of partial occlusion, visibility part unit p_i is relatively less discriminative. For example, as shown in the first row in Fig. 1, only a few parts are considered as visible regions, whereas others are viewed as occlusion or background. If naively using the few visibility parts to decide whether partial occlusion occurs or not, the pedestrian can be incorrectly classified. Considering inherent correlation between the whole and the parts and among the parts, the co-occurrence relations in part unit p_i is further established to integrate all part detectors for improving performance. Every three pairs of random visibility parts p_i are modeled as spatial subspace s_k through building the co-occurrence relations among their parts p_i, as shown in Fig. 4. Spatial subspace s_k models inherent co-occurrence correlations among visibility parts to integrate inaccurate scores of part detectors. The visibility of spatial subspace s_k is defined as follows:

$$p(y*|s_k)=\sum_{b_i\in B}\omega_1^T\phi_1(b_i,x_i)+\sum_{\substack{(b_i^k,b_j^k)\in B,\\ p_k\in P}}\omega_2^T\phi_2(b_i^k,b_j^k,p_k)+\sum_{\substack{b_i^k\in B,\\ p_k\in P}}\omega_3^T\phi_3(b_i^k,p_k)+\sum_{p_i\in P}\omega_4^T\phi_4(p_i,x_i)+\sum_{\substack{(p_i^k,p_j^k)\in B,\\ s_k\in S}}\omega_5^T\phi_5(p_i^k,p_j^k,s_k)+\sum_{\substack{p_i^k\in P,\\ s_k\in S}}\omega_6^T\phi_6(p_i^k,s_k) \quad (6)$$

where $p(y^*|s_k)$ represents the visibility of $k-$th random spatial subspace, and ω_i^T represents the co-occurrence relations among the nodes in our hierarchical model.

In our algorithm, visibility status s_k is represented as a decision tree of random subspace. The random forest [23] can be used to build a robust classifier ensemble against partial occlusions, which consist of multiple trees constructed by pseudo-randomly selecting subsets of components of the feature vector. Each decision tree returns a probability distribution $p(y^*|s_k)$ of visibility status s_k for a given test sample x and the final class label y is calculated via

$$y=\arg\max_{y*}\frac{1}{T}\sum_{k=1}^{T}p(y*|s_k) \quad (7)$$

where T represents the number of trees in random forest F, and s_k denotes the visibility status of a random subspace.

In pedestrian detection, the vast majority of training examples are negative. This result makes it infeasible to consider all negative examples at a time. Instead, a common practice is to construct training data consisting of positive instances and "hard negative" instances. The positive examples are constructed from the unoccluded training examples (as labeled in the INRIA data) and we use random subwindows from negative images to generate negative examples. Hard negatives are data mined from the very large set of possible negative examples. For this purpose, we propose to use an efficient procedure that consists of the following steps, as shown in Algorithm 2.

Algorithm 2. Random forest algorithm for visibility estimation

Initialization: *Training Set* $S=P\cup N$, $P=\{x_i,y_i\}^+$, $N=\{x_j,y_j\}^-$;
Random Forest $F=\phi$;
Output: *Score of visibility estimation score*
1. For $i=1,..,n$ do:
(a) Train M new trees T of visibility statuses s_k using training set S.
$F:=F\cup T_k$, $k=1,2...,M$;
(b) Use the current random forest F for detecting sliding window.
If the sliding window $x_i\in$ false positive, *then*
Consider x_i as negative samples and add them to training set S.
(c) Use the new training set S to update probabilities $p(y^*|s_k)$ for all trees
in random forest F.
2. Calculate the score of sliding window: $score=\frac{1}{M}\sum_{t=1}^{M}p(y*|s_k)$.
3. While $score<\theta$ and $k<T$do:
$score=\frac{1}{k+1}(score\cdot k+p(y*|s_{k+1}))$ and $t=t+1$.
4. If $score<\theta$ reject sliding window; otherwise, output is the $score$.

Our algorithm allows reducing the number of hard negatives obtained at each iteration. This is mainly due to the fact that at each iteration, more detection trees of random subspace are responsible for classification. In addition, all the probabilities $p(y^*|s_k)$ of the detection trees are updated by using the entire training set, in which slight increments indicate their discriminative ability. Moreover, training time is reduced to the smaller number of negative samples introduced at each iteration.

In particular, a testing sample is positive if at least one decision tree in the random forest gives a positive decision at last. Hence, computing the probability for all the trees of the random forest on these windows is not needed because the scores of sliding windows are visible at early stages of the cascade. Note that the number of features in visibility subspace is randomly generated from visibility statuses, unlike the original RSM [3], which has a fixed number of randomly selected features from the original space having the same dimension.

3. Pedestrian detection based on hierarchical co-occurrence model

In this paper, we propose a novel hierarchical co-occurrence model and a corresponding coarse-to-fine inference process, which is extremely efficient. The hierarchical structure can be used for better modeling and for accelerating inferences. All the best performing part-based models incorporate hierarchical structure [11,26,32]. The speed of our model results from the placement of higher layer visibility parts, which is guided by the placement of lower layers. This feature yields high computational savings, but makes inferences more sensitive to partial occlusion or other sources of noise.

Therefore, pedestrian detection based on hierarchical co-occurrence model is proposed to prune the search top-down step. The model starts the inference from the root filter and propagates only solutions that are locally more promising. With the holistic object detected, we simply use the root filter to detect the bounding box

Please cite this article as: X. Zhang, et al., Pedestrian detection based on hierarchical co-occurrence model for occlusion handling, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.038

for the holistic object hypothesis as the detection output. We use the configuration of their body parts to generate complex visibility statuses for pedestrian detection.

We present a general method for handling partial occlusions (as shown in Fig. 5). In such a design, if the confidence given by the holistic classifier falls into an ambiguous range, then an occlusion inference process is applied by using the responses of the hierarchical co-occurrence model. Finally, if the inference process determines a partial occlusion, a random forest ensemble classifies the window. Otherwise, the final output is given by the holistic classifier. To obtain a more accurate decision, we apply the ensemble only when partial occlusion is suspected.

Our established co-occurrence relations are modeled as hidden variables with the multi-layers of the hierarchical model, which integrates the inaccurate scores of part detectors based on the visibility probabilities of parts with multiple sizes when occlusions exist. Through learning co-occurrence relations among visible parts in the hierarchical model, the visibility patterns can be generated to build the

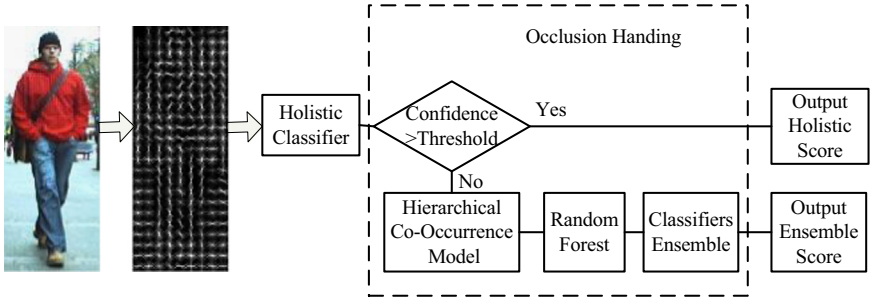

Fig. 5. Pedestrian detection scheme based on hierarchical co-occurrence model. (The resulting feature vector is first evaluated by the holistic classifier for accelerating pedestrian detection. Then, an occlusion inference process can be executed to incorporate visibility statuses when the confidence given by the holistic classifier falls into an ambiguous range.)

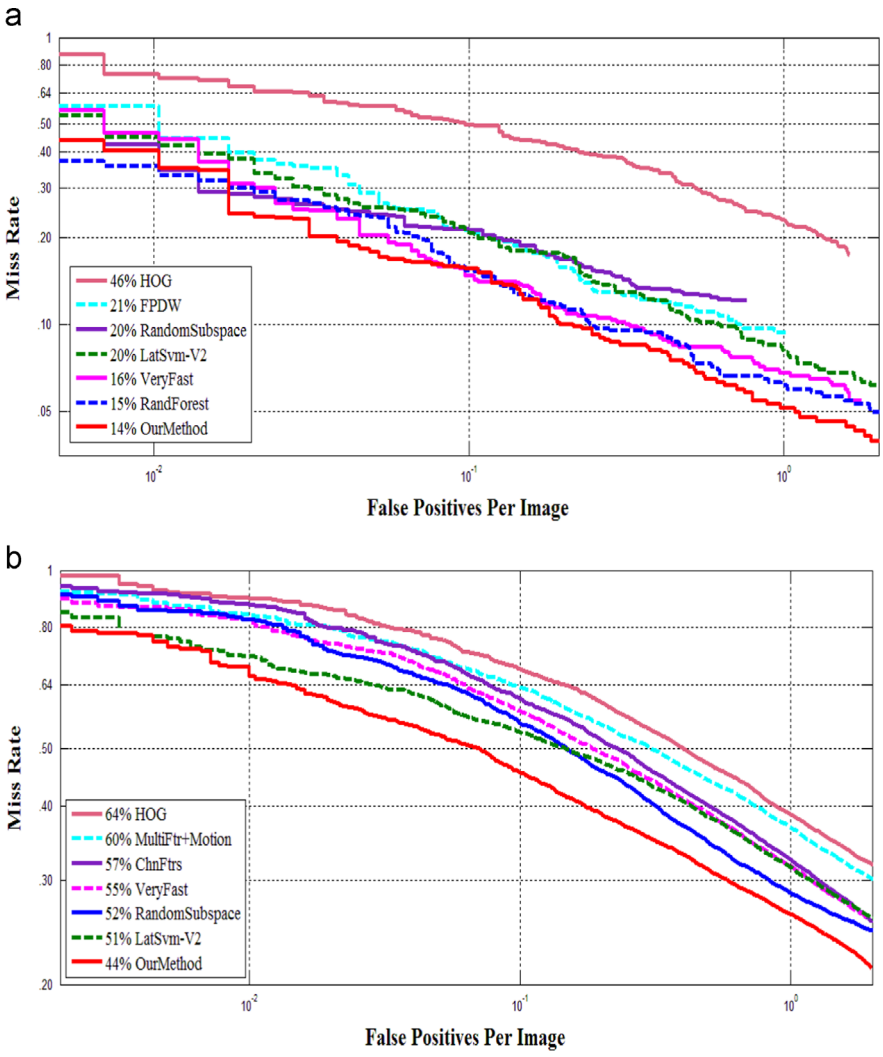

Fig. 6. Experimental comparisons of different part-based models on INRIA and ETHZ datasets. (a) INRIA dataset and (b) ETHZ dataset.

Please cite this article as: X. Zhang, et al., Pedestrian detection based on hierarchical co-occurrence model for occlusion handling, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.038

robust classifier ensemble against partial occlusions. By including multiple layers, the proposed hierarchical model achieves better variational lower bound on the training data and has more reliable visibility estimation.

4. Experiments and discussion

In this section, the performance of proposed algorithm is fully evaluated based on three datasets, namely, INRIA [7], ETH [24], and Caltech [25] datasets, which are publicly available. In the following experiments, the state-of-the-art approaches can be compared with our approach, as they are also trained from the INRIA dataset. The labels and evaluation code provided by Dollar et al. online are used for evaluating the criteria proposed in [10]. We focus on the reasonable subset, i.e., images with 50 pixels or larger, and non-occluded or partially occluded pedestrians. The performance of our proposed hierarchical model is compared with other relevant state-of-the-art approaches, which are HOG [7], HOGLBP [13], MultiFtr [5], ChnFtrs [4], LatSVM-V2 [6], VeryFast [2], MT-DPM [1], Random Forest [3], FPDW [27], and Random Subspace [17]. As [10] proposed the evaluation criteria, log-average miss rate is used to summarize the detector performance. The performance is computed by averaging miss rate at FPPI rates evenly spaced in log-space within the range of 10^{-3}–10^{0}. The experiments demonstrate that the proposed hierarchical co-occurrence model outperforms the state-of-the-art algorithms, especially on pedestrian data with partial occlusions.

4.1. Experimental results on ETHZ dataset

Experimental results indicate the performance of our proposed algorithm on public datasets. Studies in [2,3,5,6] report that state-of-the-art algorithms have the best performance when evaluated on the ETHZ dataset. As most approaches are trained on the INRIA training dataset (as shown in Fig. 6(a)), our proposed model is also trained on the INIRA training dataset. It can be seen that the log-average miss rate of our approach has 6% improvement over LatSVM-V2 by Felzenszwalb et al. [6] on the ETHZ dataset, as shown in Fig. 6(b).

The pedestrian detection results of our hierarchical co-occurrence model are shown in Fig. 7 on the ETH dataset. A detected image sliding window is represented by the green dotted bounding box if the area overlaps the detected window, and the ground truth (green solid bounding box) exceeds 50%, or by the red dotted bounding box if otherwise. Fig. 7 shows the green dotted boxes demonstrating the performance of our proposed hierarchical model.

4.2. Experimental results on the Caltech dataset

Similar to other relevant publications previously [3,11], we use the Caltech Training Dataset as training data and test our proposed model on the Caltech Testing Dataset. Our hierarchical co-occurrence model matches or outperforms the state-of-the-art algorithms on the Caltech datasets. Fig.8 shows the comparison of the log-average miss rate with HOG [7], HOGLBP [14], MultiFtr [5], ChnFtrs [4], LatSVM-V2 [6], VeryFast [2], MT-DPM [1], and Random Subspace [17] under varying levels of occlusion. Fig. 8(a) indicates that our approach has similar performance for the overall Caltech testing dataset under no occlusion. However, our approach has 5% log-average miss rate improvement for pedestrians with partial occlusions compared with MT-DPM [1], as shown in Fig. 8(b). Fig. 9 shows the detection results of our model on the Caltech dataset. This experiment shows that the usage of the hierarchical co-occurrence model outperforms state-of-the-art algorithms especially for partial occlusion. With more

Fig. 7. Detection results using our hierarchical structure model on the ETHZ dataset. (The green dotted bounding box means correct detection; the red solid bounding box is missed detection; and the red dotted bounding box is false positive.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Please cite this article as: X. Zhang, et al., Pedestrian detection based on hierarchical co-occurrence model for occlusion handling, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.038

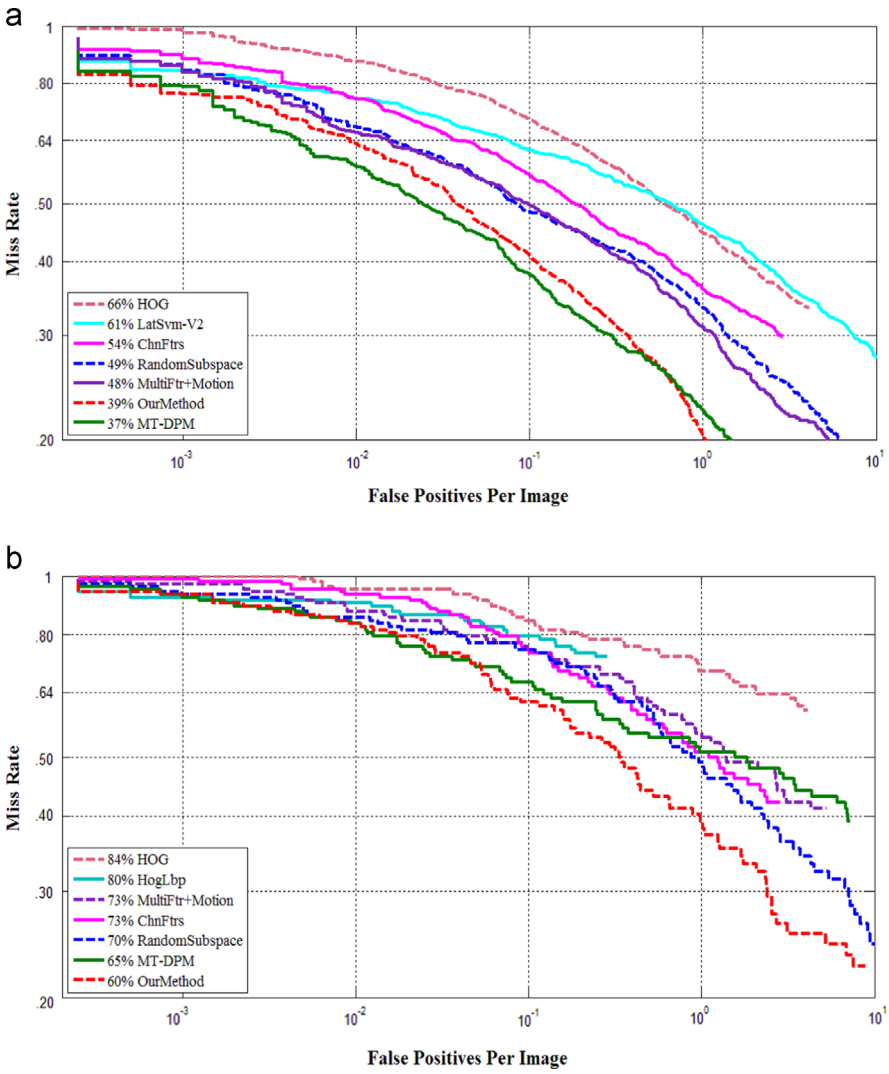

Fig. 8. Per-image evaluation. Experimental comparisons of different part-based models on Caltech testing dataset. (a) Non-Occlusion and (b) Partial Occlusion.

features being included, the performance of our approach can be further improved.

4.3. Computational complexity analysis

Fast detection rates are of the essence in many applications of pedestrian detection, such as video surveillances, intelligent vehicles, robot sensing, and human machine interfaces. Although our study focuses on accuracy, we also conclude by jointly considering accuracy and speed. We measure the runtime of each detector by using images from the Caltech dataset (averaging runtime over multiple frames). To compensate for detectors running on different hardware, all runtimes are normalized to the rate of a single modern machine.

Table 1, which is ordered according to descending log-average miss rate on partial occlusion pedestrians in the Caltech dataset, gives an overview of each detector. In Fig. 8(b), we plot log-average miss rate versus runtime for each detector on 640×480 images. To save runtime in our pedestrian detection algorithm, we employ this method. If the confidence given by the holistic classifier falls into an ambiguous range, then an occlusion inference process is applied by using the responses of the hierarchical co-occurrence model, as shown in Fig. 4.

Although the frame rates may seem low, all tested detectors can be employed as part of a full system. Such systems may employ ground plane constraints and perform region-of-interest selection (e.g., from stereo disparity or motion), which reduces runtime drastically. Moreover, numerous approaches have been proposed for speeding up detection, including increasing the speed of the detector itself, through the use of approximations or special purpose hardware, such as GPUs (for a review of fast detection see [2]). Nevertheless, the above runtime analysis gives a sense of the speed of current detectors.

5. Conclusion

This study describes an effective approach for pedestrian detection with occlusion handling in still images. The approach effectively estimates the visibility of components at multiple layers by learning spatial co-occurrence relations with the proposed hierarchical co-occurrence model. The latent SVM structure modeling with the co-occurrence relations as latent variables is employed to generate visibility statuses. Then, random forest is used to build a robust classifier ensemble for handling various partial occlusions. Through comparing experimental results on multiple publicly datasets, various schemes of integrating component detectors are investigated. The log-average miss rate of our proposed algorithm has 5% improvement for pedestrian detection, especially on pedestrian datasets with partial occlusions compared with the state-of-the-art approaches.

Please cite this article as: X. Zhang, et al., Pedestrian detection based on hierarchical co-occurrence model for occlusion handling, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.038

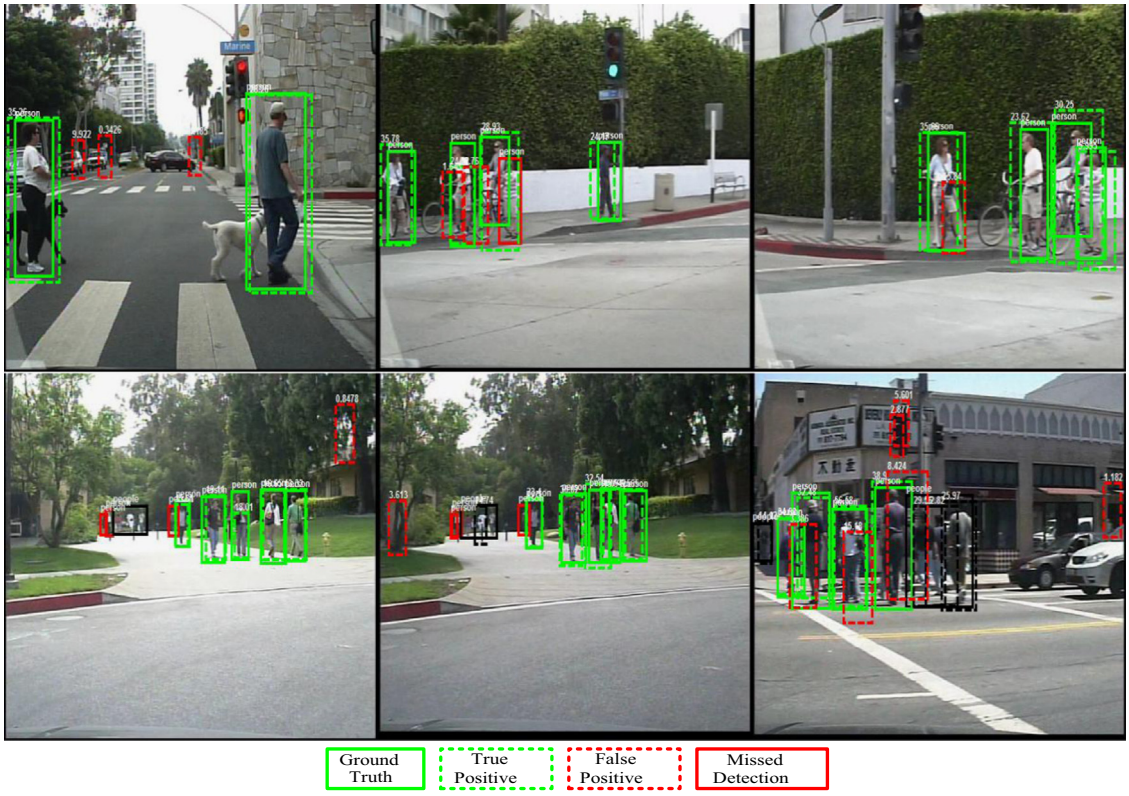

Fig. 9. Detection results using our hierarchical structure model on the Caltech dataset. The green dotted bounding box means correct detection, the red solid bounding box is a missed detection, and the red dotted bounding box is false positive. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Computational complexity comparison on partial occlusion pedestrians in Caltech dataset.

Method	Classifier	Non-maximum suppression	Scales per octave	Frames per second (fps)	Log-average missing rate
HOG [7]	Linear SVM	MS	14	0.239	84%
HOGLBP [14]	Linear SVM	MS	14	0.062	80%
MultiFtr+Motion [5]	Linear SVM	MS	14	0.020	73%
ChnFtrs [4]	AdaBoost	PM	10	1.183	73%
MT-DPM [1]	Latent SVM	MS	10	1.000	65%
Our method	Latent SVM	MS	10	0.589	60%

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 61370121), the National Hi-Tech Research and Development Program (863 Program) of China (No.2014AA015102), and Outstanding Tutors for doctoral dissertations of S&T project in Beijing (No. 20131000602).

References

[1] J. Yan, X. Zhang, Z. Lei, S. Liao, and S.Z. Li, Robust multi-resolution pedestrian detection in traffic scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[2] R. Benenson, M. Mathias, R. Timofte, L. Van Gool, Pedestrian detection at 100 frames per second, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[3] J. Marin, D. Vazquez, A.M. Lopez, J. Amores, B. Leibe, Random Forests of Local Experts for Pedestrian Detection, in: Proceedings of the IEEE Conference on Computer Vision (ICCV), 2013.

[4] V. Dung Hoang, M. Hale, K. Hyun Jo, Hybrid cascade boosting machine using variant scale blocks based HOG features for pedestrian detection, Neurocomputing 135 (5) (2014) 357–366.

[5] S. Walk, N. Majer, K. Schindler, B. Schiele, New features and insights for pedestrian detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[6] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.

[7] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[8] X. Su, W. Lin, X. Zheng, X. Han, H. Chu, X. Zhang, A new local-main-gradient-orientation HOG and contour differences based algorithm for object classification, in: Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), 2013.

[9] L. Weixin, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2014) 18–32.

[10] D. Geronimo, A.M. Lopez, A.D. Sappa, T. Graf, Survey of pedestrian detection for advanced driver assistance systems, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1239–1258.

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

[12] Z. Long, C. Yuanhao, Y. Alan, and F. William, Latent hierarchical structural learning for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[13] M. Norouzi, M. Ranjbar, G. Mori, Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[14] X. Wang TH, and S. Yan, A HOG-LBP human detector with partial occlusion handling, in: Proceedings of the IEEE Conference on Computer Vision (ICCV), 2009.

[15] B. Wu, R. Nevatia, Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors, in: Proceedings of the IEEE Conference on Computer Vision (ICCV), 2005.

Please cite this article as: X. Zhang, et al., Pedestrian detection based on hierarchical co-occurrence model for occlusion handling, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.038

[16] M. Enzweiler, A. Eigenstetter, B. Schiele, D.M. Gavrila, Multi-cue pedestrian classification with partial occlusion handling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.
[17] D.V. Javier Marin, Antonio M. Lopez, et al., Occlusion handling via random subspace classifiers for human detection, IEEE Trans. Syst. Man Cybern. Soc. 44 (3) (2014) 342–354.
[18] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multi-scale, deformable part model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.
[19] G. Duan, H. Ai, S. Lao, A Structural filter approach to human detection, in: Proceedings of the IEEE European Conference on Computer Vision (ECCV), 2010.
[20] W. Ouyang, X. Wang, A discriminative deep model for pedestrian detection with occlusion handling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[21] M. Pedersoli, A. Vedaldi, J. Gonzalez, A coarse-to-fine approach for fast deformable object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
[22] M. Andriluka, S. Roth, B. Schiele, Pictorial structures revisited: people detection and articulated pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.
[23] H. Gu, X. Xie, Q. Lv, Y. Ruan, and L. Shang, Etree: effective and efficient event modeling for real-time online social media networks, in: Proceedings of the Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on, 1, 2011 pp. 300–307.
[24] A. Ess, B. Leibe, L.V. Gool, Depth and appearance for mobile scene analysis, in: Proceedings of the IEEE Conference on Computer Vision (ICCV), 2007.
[25] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: a Benchmark, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.
[26] M. Pedersoli, J. Gonzalez, A.D. Bagdanov, and J.J. Villanueva, Recursive coarse-to-fine localization for fast object detection, in: Proceedings of the IEEE European Conference on Computer Vision (ECCV), 2010.
[27] P. Dollar, S. Belongie, and P. Perona, The fastest pedestrian detector in the west, in: Proceedings of the British Machine Vision Conference, 2010.
[28] C. Zhang, X. Yang, J. Zhu, W. Lin, Parsing collective behaviors by hierarchical model with varying structure, ACM Multimedia (MM), 2012.
[29] H. Hu, X. Zhang, W. Zhang, B. Li, Joint global-local information pedestrian detection algorithm for outdoor video surveillance, J. Vis. Commun. Image R. 26 (1) (2015) 168–181.
[30] C.N.J. Yu and T. Joachims. Learning structural svms with latent variables, in: Proceedings of the International Conference on Machine Learning (ICML), 2009.
[31] Xiaoqin Zhang, Weiming Hu, Nianhua Xie, Hujun Bao, Steve Maybank, A Robust Tracking System for Low Frame Rate Video, International Journal of Computer Vision (2015) 1–26.
[32] Di Wang, Xiaoqin Zhang, Mingyu Fan, Xiuzi Ye, An Efficient Classifier Based on Hierarchical Mixing Linear Support Vector Machines, International Joint Conference on Artificial Intelligence (2015).

Xiaowei Zhang received the B.S. degree from the College of Mathematics and Computer Science, Shanxi Normal University, Linfen, China, in 2009, and the M.S. degree from the institute of information science and engineering, Shandong Normal University, Jinan, China, in 2013, and he is currently pursuing the Ph.D. degree in computer science and engineering from Beihang University, Beijing, China. His current research interests include pattern recognition, computer vision and machine learning.

Hai-Miao Hu received the B.S. degree from Central South University, Changsha, China, in 2005, and the Ph. D. degree from Beihang University, Beijing, China, in 2012, all in computer science. Since 2013, he has been an assistant professor at the School of Computer Science and Engineering, Beihang University. His research interests include video coding and networking, image/video processing, and video analysis and understanding.

Fan Jiang received the B.S. degree in software engineering from the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China, in 2013, and he is currently pursuing the M.S. degree in computer science and engineering from Beihang University, Beijing, China. His current research interests include multimedia retrieval and object recognition.

Bo Li received the B.S. degree in computer science from Chongqing University in 1986, the M.S. degree in computer science from Xi'an Jiaotong University in 1989, and the Ph.D. degree in computer science from Beihang University in 1993.

Now he is a professor of Computer Science and Engineering at Beihang University, the Director of Beijing Key Laboratory of Digital Media, and has published over 100 conference and journal papers in diversified research fields including digital video and image compression, video analysis and understanding, remote sensing image fusion and embedded digital image processor.

Please cite this article as: X. Zhang, et al., Pedestrian detection based on hierarchical co-occurrence model for occlusion handling, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.038