Deformable Part Model based Multiple Pedestrian Detection for Video Surveillance in Crowded Scenes

Lu Wang, Xiaoli Ji, Qingxu Deng and Mingxing Jia
College of Information Science and Engineering, Northeastern University, Shenyang, China

Keywords: Deformable Part-based Model, Multiple Pedestrian Detection, Crowd Detection, Video Surveillance.

Abstract: Pedestrian detection is a challenging task for video surveillance. The problem becomes more difficult when occlusion is prevalent. In this paper, we extend a deformable part-based pedestrian detector to pedestrian detection in crowded scenes by considering both body part detection responses and detections' mutual spatial relationship. Specifically, we first decompose the full body detector into several body part detectors, whose detection responses can be computed efficiently from the response of the full body detector. Then, given the detection responses of the body part detectors, hypotheses are nominated by considering both detection scores and responses' mutual spatial relationship. Finally, a local optimization process is applied to make the final decision, where an objective function encouraging detections with high confidence, high discriminability and low conflict with other detections is proposed to select the best candidate detections. Experimental results show the effectiveness of the proposed approach.

1 INTRODUCTION

Pedestrian detection is a very important task for video surveillance. It is difficult due to pose articulations, appearance variations, low figure-ground contrast and etc. Recently, significant advance has been made on detecting well separated individual pedestrians through training detectors using statistical machine learning methods and running the detectors on the detection window that slides over image positions and across scale levels (Dollar, 2012). However, when applied to the detection of crowds, their performance degrades significantly due to ambiguous appearance caused by heavy occlusions.

The deformable part-based model (DPM) trained using latent support vector machine (Felzenszwalb, 2010) has been proved to be one of the most powerful object detectors. It runs detection on individual parts and then sum up the responses to form the final detection score. DPM has a good potential to apply to crowd detection because parts can be flexibly removed from and added to the model to deal with occlusion. There are some works that apply the DPM models to deal with occlusion (Ouyang, 2012); (Shu, 2012); (Yan, 2012). However, (Ouyang, 2012) and (Shu, 2012) focus on improving the responses in a detection window without considering detection responses of neighboring windows; only Yan, 2012 determines the visibility of part by simultaneously considering the appearance and mutual spatial relationship. Therefore, the aim of this work is to adapt a DPM based full body pedestrian detector to crowd detection in surveillance scenarios by considering both body part detection responses and detections' mutual spatial relationship.

In this paper, we assume the camera looks down onto a ground plane and no camera parameter is known. Specifically, we first propose to decompose the original whole body detector trained on the INRIA pedestrian dataset into several body part detectors, whose responses are computed efficiently, and the bias term for each part detector is estimated from the training data so that the same threshold can be used to select responses from different body part detectors. Then, given the detection responses of the body part detectors, hypotheses that may correspond to genuine pedestrians are nominated by considering both detection scores and responses' mutual spatial relationship. Finally, a local optimization process is applied to make the final decision, where an objective function encouraging detections with high confidence, high discriminability and low conflict with other detections is proposed to select the best detections from the mutually overlapped hypotheses.

Wang L., Ji X., Deng Q. and Jia M..
Deformable Part Model based Multiple Pedestrian Detection for Video Surveillance in Crowded Scenes.
DOI: 10.5220/0004739105990604
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 599-604
ISBN: 978-989-758-004-8

2 RELATED WORK

For pedestrian detection in crowded scenes, there are two categories of works. The first category deals with occlusion from the detector's respect. For example, in Wang, 2009, full body detector based on HOG and LBP is first applied and the classification score of each block is used to infer whether occlusion occurs and where it occurs. Ouyang et al., (2012) designed overlapping body parts and verify the visibility of a part by the scores of overlapping parts and the correlation among parts is modeled by a discriminative deep model. Duan et al., (2010) proposed a structural filter consists of a set of detectors which is able to infer what parts are visible in a test window. Shu et al., (2012) designed a DPM based detector that deals with partial occlusion by selecting the subset of parts that maximizes the average score of parts. The disadvantage of this category of methods is that the responses of other detection windows is not considered.

The second category approaches use body part detectors to nominate a set of candidates and then perform optimization over an objective function to select the best candidate subset as the final detection result. As the number of possible combinations of candidates is quite large, efficient optimization method must be developed. For example, Wu, 2005, (Lin, 2007); (Beleznai, 2009) assumed the occlusion order is known and used greedy methods for optimization. Global optimization methods such as Expectation-Maximization (EM) (Rittscher, 2005); (Tu, 2008) and Markov Chain Monte Carlo (MCMC) (Zhao, 2003); (Ge, 2009) have also been developed. In Rujikietgumjorn, 2013, the best set of candidates are determined by applying quadratic programming to maximize the objective function composed of unary detection scores and pairwise mutual overlap constraints. Wang et al. proposed to compromise greedy optimization and global optimization by considering a small portion of mutual overlapped candidates each time (Wang, 2012). In this work, we apply the optimization strategy similar to Wang, (2012).

3 THE BODY PART MODELS

The original full body person model we consider in this work consists of one root filter (F_0) and eight body part filters (F_1, ..., F_8), as shown in Figure 1(a), where each green box corresponds to one body part filter F_i, and the combination of the gray and green areas constitute the root filter F_0. A deformation cost coefficients d_i (i=1,...,8) is also defined for each body part. The features for body part filters are extracted at twice the resolution of the root filter for both training and detection.

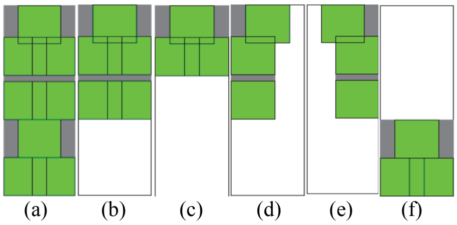

Figure 1: Illustration of the body part models: (a) full body model; (b) upper body model; (c) head shoulder model; (d) left upper body model; (e) right upper body model; (f) lower body model.

The score of each detection window is defined by the location p_0 of the root filter as

$$S(p_0) = \max_{p_1,\dots,p_n} S(p_0,\dots,p_n)\text{, with}$$
$$S_k(p_{k,0},\dots,p_{k,n_k}) = F'_{k,0}\cdot\phi(p_{k,0}) + \textstyle\sum_{j=1}^{n_k}[F'_{k,j}(p_{k,j}) - d_{k,j}\cdot\phi_d(dx_{k,j},dy_{k,j})] + b_k \quad (1)$$

where p_i=(x, y, l) specifies a position(x, y) in the l^{th} level of the feature pyramid; $l_i = l_0 - \lambda$ for i>0 (λ is the number of levels in an octave of the feature pyramid); $\phi(p_i)$ is the feature vector extracted from the feature pyramid with top-left corner at p_i; $\phi_d(dx_i,dy_i)$ are the deformation features.

To deal with the mutual occlusion exists prevalently in crowded scenes, we derive five body part detectors {D_1, ..., D_5} from the full body model, as shown in Figure 1 (b)-(f), namely the upper body, the head shoulder, the left/right upper body and the lower body detectors. Among the five body part detectors, upper body, head shoulder and lower body detectors are widely used in crowd detection works (e.g. Wu, 2005). The left/right upper body parts are applied here to deal with more severe occlusion where only one shoulder is visible.

In each derived body part detector D_k, the constitutional body part filters (green boxes) {$F_{k,1}$,…, F_{k,n_k}} (n_k<8) are a subset of {F_1, ..., F_8}. Similarly, the deformation coefficients {$d_{k,1}$,…, d_{k,n_k}} are a subset of {d_1, ..., d_8} and the root filter $F_{k,0}$ (the combination of the gray and green areas) is a subarray of F_0. Thus, similar to Eq. (1), the response $S_k(p_{k,0})$ of a part detector D_k at the root filter location $p_{k,0}$ can be computed by

$$S_k(p_{k,0}) = \max_{p_{k,1},\dots,p_{k,n_k}} S_k(p_{k,0},\dots,p_{k,n_k})\text{, with}$$
$$S_k(p_{k,0},\dots,p_{k,n_k}) = F'_{k,0}\cdot\phi(p_{k,0}) + \textstyle\sum_{j=1}^{n_k}[F'_{k,j}(p_{k,j}) - d_{k,j}\cdot\phi_d(dx_{k,j},dy_{k,j})] + b_k \quad (2)$$

which is completely part of the calculation in Eq. (1), except that the bias b_k needs to be estimated. To avoid redundant calculation, an efficient way for calculating $F'_{k,0} \cdot \phi(p_{k,0})$ for all D_k's is to: 1) partition F_0 into subfilters, 2) compute filtering responses for each resulting subfilter, and 3) sum up the responses accordingly for each D_k. To achieve this, we divide F_0 into a minimum number of 7 subfilters $\{F_0^1$, ..., $F_0^7\}$ according to the configuration of part detectors, as shown in Figure 2, which requires the minimum computational time and the memory space for saving the responses.

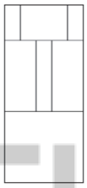

Figure 2: Partition of the root filter F_0 into 7 subfilters for calculating the part detectors' root filters.

To estimate the bias term b_k for each body part detector D_k so that the same threshold can be used to measure responses from different body part detectors, we learn it from the training data similar to the way proposed in Wang, 2009 as follows.

We revisit the INRIA person data and apply the full body detector to find the best deformation configuration of parts (i.e. p_1, ...p_8 given p_0) on both positive and negative examples. Then for each example, we record the score $F'_i \cdot \phi(p_i) - d_i \cdot \phi_d(dx_i, dy_i)$ of each body part i (i = 1, …, 8), as well as the score $F_0^q \cdot \phi(p_0^q)$ of each subfilter F_0^q (q=1, …, 7) of F_0. Considering the linearity of the dot product operation, the score $f(\mathbf{x})$ of an example $\mathbf{x}$ can be written as

$$f(\mathbf{x}) = \alpha\mathbf{x} + b = \sum_{i=1}^{8+7}(\alpha_i B_i + \beta_i)\text{, with}$$
$$\alpha_i = \begin{cases} F_0^i & i = 1, \ldots, 7 \\ (F_{i-7}, d_{i-7}) & i = 8, \ldots, 15' \end{cases}$$
$$B_i = \begin{cases} \phi(p_0^i) & i = 1, \ldots, 7 \\ (\phi(p_{i-7}), -\phi_d(dx_{i-7}, dy_{i-7})) & i = 8, \ldots, 15 \end{cases} \text{ and} \tag{3}$$
$$b = \sum_{i=1}^{7+8} \beta_{i\,.}$$

Then, according to Wang, 2009, β_i can be estimated by

$$\beta_i = D\alpha_i\left(C\sum_{u=1}^{N^+} B_{u,i}^+ + \sum_{u=1}^{N^+} B_{v,i}^-\right), \tag{4}$$

where $B_{u,i}^+$ ($B_{v,i}^-$) denotes the i^{th} block of the u^{th} (v^{th}) positive (negative) example; N^+ ($N^{-)}$ is the number of positive (negative) examples; C is the negative of the ratio between sum of the positive example scores and sum of the negative example scores; D equals to -1/(CN^++N^-). Then the bias b_k for body part detector D_k can be calculated by

$$b_k = \textstyle\sum_{\{q|F_0^q \in F_{k,0}\}} \beta_q + \sum_{\{i|F_i \in D_k\}} \beta_{i+7}, \tag{5}$$

To get a concept of the discriminability of the derived part detectors, the ROC curves of these detectors on the INRIA training data set are shown in Figure 3. We can see that the discriminability of body part detectors is significantly lower than that of the full body detector, and the less number of parts contained, the lower the detector's discriminability is. This result can be expected because less information is made use of by body part detectors.

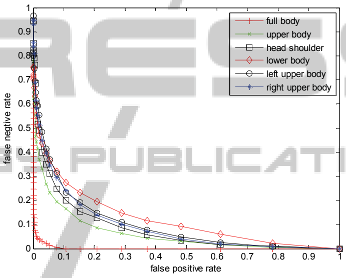

Figure 3: The ROC curves of the part detectors' performance on the INRIA training data.

4 DETECTION

To find the optimal combination of detection responses, we first merge responses of the same type which are likely to correspond to the same pedestrian. Then detection scores and responses' spatial relationship are considered to further exclude unlikely responses. Finally, a local optimization process is applied to make the final decision that selects detections with high confidence, high discriminability and low conflict with other detections.

To select the likely candidate detections, two quantities, namely the attraction force $F_{att}(H, H')$ and the exclusion force $F_{exc}(H,H')$, that can describe the consistency and confliction between any two hypotheses H and H' are calculated. F_{att} calculates the overlapping degree by

$$F_{att}(H, H') = \sum_{i=0}^{8} \frac{area(A(H,i) \cap A(H',i))}{area(A(H,i) \cup A(H',i))} w_i, \tag{6}$$

where $A(H, i)$ represents the image region occupied by the i^{th} part of H (if H does not contain part i, then $A(H, i)$ = 0); ∩ and ∪ denote the intersection and union of two regions; w_i is the weight of the i^{th} part and is set to be $-\beta_i$. Two hypotheses with large F_{att} are likely to correspond to the same pedestrian. F_{exc} is defined as

$$F_{exc}(H,H') = \sum_{i=0}^{8} \sum_{i'=0,i'\neq i}^{8} \frac{area(A(H,i) \cap A(H',i'))}{area(A(H,i) \cup A(H',i'))}. \quad (7)$$

Two hypotheses with large F_{exc} are unlikely to be true simultaneously.

4.1 Hypotheses Formation

Given a global threshold θ, responses of all the body part models with detection score greater than θ are taken as possible hypotheses. To reduce the number of responses to deal with, non-maximum suppression is performed for responses from the same detector by setting the bounding box overlap threshold to be 0.7 (the threshold is set relatively higher to avoid missing genuine detections). After that, we further use F_{att}>1.5 as a criterion to merge responses from the same part detector with bounding box overlapping ratio less than 0.7 but having significant parts overlap. The advantage of F_{att} over bounding box overlap is that weighted body part overlap is also taken into account.

To exclude those hypotheses unlikely to be true, we conduct hypotheses formation as described below.

(1) All the hypotheses produced by the full body detector are added to the list of hypotheses (LoH) as they are most reliable.

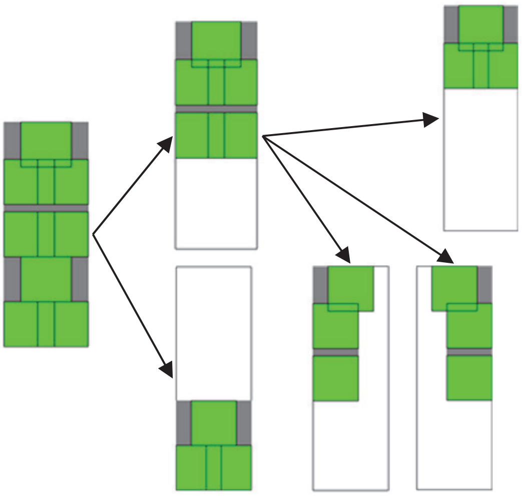

Figure 4: Definition of the parent detector. The whole body detector is the parent detector of upper body detector and lower body detector; the head shoulder detector is the parent detector of head shoulder, left upper body and right upper body detector.

(2) A hypothesis nominated by a body part model is added to LOH if the lacking portion compared with its parent detector (e.g. for the head shoulder detector, the lacking portion is the waist) is significantly occluded by other hypotheses in LoH or the image border. Definition of parent detection is shown is Figure 4.

In our experiment, we found the bottom border of the image can produce a strong shoulder effect, resulting in false positives when the corresponding head position has weak edge response. Therefore, for the head shoulder detector, its detection threshold is set to be 0.2 greater than θ.

4.2 Optimization

Given the list of hypotheses LoH, we use a local optimization method to make the final decision. In each iteration of the optimization process, we consider the hypotheses that might be the lowest in terms of the vertical position together with hypotheses which have significant overlap with them. Then from these hypotheses, the one that best matches the criteria is accepted and unqualified hypotheses are rejected. The optimization process terminates when all the hypotheses are either accepted or rejected. Details of this process are described as follows.

Due to occlusion and the lack of scale constraint, it is not easy to decide which hypothesis is the lowest. Therefore, we take the hypotheses whose bounding boxes' bottom, or center, or top are the lowest as possible lowest hypotheses. Then the hypotheses that overlap significantly with the lowest hypotheses, i.e. F_{att}>a (0.2 is used in our experiment), are also selected for consideration, ensuring that no true detections are neglected during the optimization. After that, we choose from them the one that is most likely to be true according to the following equation

$$H^* = \arg\max_H [score(H) + \textstyle\sum_{\{i|F_i \in D(H)\}} w_i - \sum_{H' \in C_{acc}} F_{exc}(H,H')] \quad (8)$$

where $score(H)$ is the detection score of hypothesis H, $D(H)$ is the type of the body part detector of H, and C_{acc} represents the set of accepted hypotheses. Eq. (8) selects the hypothesis with higher score, more parts (meaning higher discriminability), and less conflict with other accepted hypotheses. From our experiment, Eq. (8) makes correct decisions in most cases. However, sometimes a hypothesis with good confidence may be missed by (8) due to the smaller number of parts it contained. Therefore, we further consider the hypotheses whose tops are lower than H^* and meanwhile with detection scores greater than H^*, and choose the best one from them accord ing to Eq. (8) again.

The hypothesis H^* chosen above is accepted and added to C_{acc} except for the following two cases:

(1) It has less than two basic parts visible or its head is invisible. This is to ensure that the accepted hypothesis has enough visible ratio to support its existence.

(2) It is a body part detector D_k's response, and meanwhile body parts of its parent detector D_j in the same detection window p are all visible but with detection score $S_j(p) < \theta$. This is the case that the occlusion information is not sufficient to explain the existence of the part detector response.

5 EXPERIMENTAL RESULTS

We evaluate the performance of our proposed approach on two data sets, i.e. CAVIAR (http://homepages.inf.ed.ac.uk/rbf/CAVI-RDATA1) and PETS 2009 (http://www.cvg.rdg.ac.uk/PETS20-09/a.html). We select the crowded sequence *OneStopMoveEnter1cor* (1590 frames with resolution being 384×288) from the CAVIAR data set, and the S2L1_1 sequence (221 frames with resolution being 768×576) from the PETS 2009 data set, as the testing data. We compare the performance of our proposed approach with two deformable part based person detectors trained using Latent SVM (Felzenszwalb, 2010). The first one is the full body detector from which we derive our approach, and the second one is a mixture of separately trained full body and body part detectors (the part detectors are the upper body and head shoulder detectors), and is trained on the VOC2007 person dataset. Both detectors are provided online (http://www.cs.uchicago.edu/~pff/latent/).

All the three detectors need to perform nonmaximum suppression, in which the bounding box overlap threshold needs to be determined. We set this parameter to be 0.7 for our approach as stated above. For the other two detectors, we experimentally select the optimal overlap threshold for them and the resulting parameter is 0.6 for the INRIA full body model and 0.5 for the VOC2007 body part model.

The detection performance evaluation criterion used is the commonly applied intersection over union greater than 0.5, under the constraint that the detection and the ground truth are in one to one correspondence.

Figure 5 shows the recall-precision curves of the three detectors on the CAVIAR data set, from which we can see that our proposed approach consistently outperforms both detectors because of the application of body part detectors while performing occlusion reasoning at the same time. Figure 6 illustrates some examples of the detection results of our approach when the threshold θ is set to be 0.

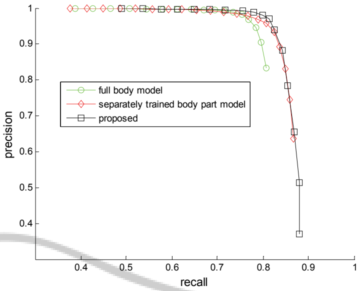

Figure 5: The precision-recall curves of the three detectors on the tested Caviar sequence.

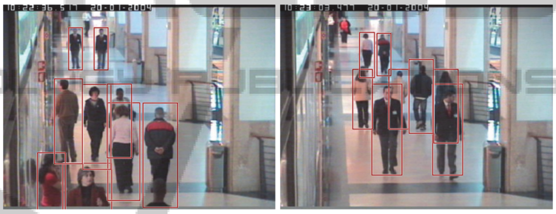

Figure 6: Illustration of the detection results of the proposed approach on the Caviar data set.

Figure 7 demonstrates the performance of the three detectors on the PETS 2009 data set. It can be seen that the body part model performs the worst, whereas our proposed approach gives better result when the recall rate is less than 0.7. The reason that the performance of our approach at high recall rate is not satisfactory is that if two detections share many parts, although their bounding box do not overlap significantly, we take the two detections as conflicting, discarding more true positives when body parts are not accurately localized whereas the non-maximum suppression applied by the full body model does not do this action.

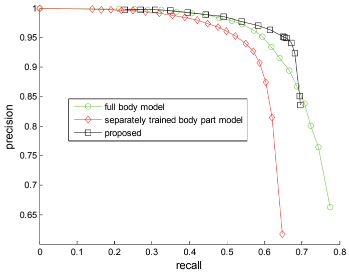

Figure 7: The precision-recall curves of the three detectors on the tested PETS 2009 sequence.

Figure 8 illustrates some examples of the detection results of our approach on the PETS 2009 data set when the threshold θ is set to be 0.

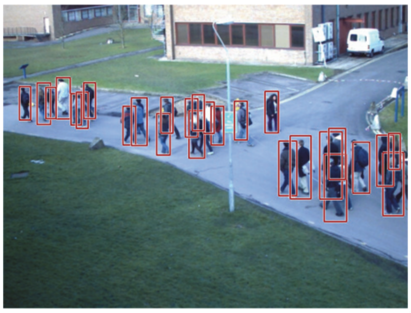

Figure 8: Illustration of the detection results of the proposed approach on the PETS 2009 data set.

For each frame, our proposed approach takes about 15% more time to calculate the part detection scores than the full body detector, whereas the computational time for the hypotheses formation and optimization is quite short and can be neglected. The VOC 2007 body part model cost more than twice the time cost by our approach, due to that body part filters are not well shared among different body part detectors.

6 CONCLUSIONS

We have developed an approach to adapting the deformable part based pedestrian detector to crowded scenes by considering both body part detection responses and detections' mutual spatial relationship, without enforcing much additional computational overhead through part response sharing, while improving the detection results significantly.

Our future wok includes enriching the part detectors with higher discriminability; estimating the pedestrians' size distribution over the image online so that size constraint can be enforced on the detection of future coming frames; designing an efficient global optimization method that considers both conflicts between overlapping visible parts and detection scores so that decisions can be made more properly.

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China 61202258, Program for New Century Excellent Talents in University of China NCET-09-0071, and National Science and Technology Support Program of China 2013BAK02B01-02.

REFERENCES

Beleznai, C., Bischof, H., 2009. Fast human detection in crowded scenes by contour integration and local shape estimation. In *CVPR*, pp. 2246-2253.

Duan, G., Ai, H., Lao, S., 2010. A Structural Filter Approach to Human Detection. In *ECCV*, pp. 238-251.

Dollar, P., Wojek, C., Schiele, B., Perona, P., 2012. Pedestrian detection: an evaluation of the state of the art. *TPAMI* vol. 34, pp. 743-761.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., Ramanan, D., 2010. Object Detection with Discriminatively Trained Part-Based Models. *TPAMI*, vol. 32, pp. 1627-1645.

Ge, W., Collins, R., 2009. Marked point processes for crowd counting. In *CVPR*, pp. 2913-2920.

Lin, Z., Davis, L. S., Doermann, D., DeMenthon, D., 2007. Hierarchical part-template matching for human detection and segmentation. In *CVPR*, 2007, pp. 1-8.

Ouyang, W., Wang, X., 2012. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, pp. 3258-3265.

Rittscher, J., Tu, P. H., Krahnstoever, N., 2005. Simultaneous estimation of segmentation and shape. In *CVPR*, pp. 486-493.

Rujikietgumjorn, S., Collins,R. T., 2013. Optimized pedestrian detection for multiple and occluded people. In *CVPR*.

Shu, G., Dehghan, A., Oreifej, O., Hand, E., Shah, M., 2012. Part-based multiple-person tracking with partial occlusion handling. In *CVPR*, pp. 1815-1821.

Tu, P., Sebastian, T., Doretto, G., Krahnstoever, N., Rittscher, J., Yu, T., 2008. Unified crowd segmentation. In *ECCV*.

Wang, L., Yung N. H. C., 2012. Three-dimensional model-based human detection in crowded scenes. *TITS*, vol. 13, pp. 691-703.

Wang, X., Han, T., Yan, S., 2009. An HOG-LBP human detector with partial occlusion handling. In *CVPR*, pp. 32-39.

Wu, B., Nevatia, R., 2005. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In *CVPR*, pp. 90-97.

Yan, J., Lei, Z., Yi, D., Li, S. Z., 2012. Multi-pedestrian detection in crowded scenes: a global view. In *CVPR*, pp. 3124-3129.

Zhao, T., Nevatia, R., 2003. Bayesian human segmentation in crowded situations. In *CVPR*, pp. 459-466.