PEDESTRIAN DETECTION IN CROWDED SCENES VIA SCALE AND OCCLUSION ANALYSIS

Lu Wang *Lisheng Xu* *Ming-Hsuan Yang*

Northeastern University, China University of California at Merced, USA

ABSTRACT

Despite significant progress in pedestrian detection has been made in recent years, detecting pedestrians in crowded scenes remains a challenging problem. In this paper, we propose to use visual contexts based on scale and occlusion cues from detections at proximity to better detect pedestrians for surveillance applications. Specifically, we first apply detectors based on full body and parts to generate initial detections. Scale prior at each image location is estimated using the cues provided by neighboring detections, and the confidence score of each detection is refined according to its consistency with the estimated scale prior. Local occlusion analysis is exploited in refining detection confidence scores which facilitates the final detection cluster based Non-Maximum Suppression. Experimental results on benchmark data sets show that the proposed algorithm performs favorably against the state-of-the-art methods.

Index Terms— Pedestrian detection, crowded scenes, scale prior, occlusion analysis

1. INTRODUCTION

Pedestrian detection is an important task for numerous applications including visual surveillance and intelligent vehicles as it is a prerequisite for object tracking and event recognition. While significant progress has been made to detect pedestrians by considering each one independently at a time [1] [2] [3] [4] [5] [6], these methods are less effective in crowded scenes with heavy occlusion [7]. In this paper, we propose an algorithm to detect pedestrians in crowded scenes by considering scale prior and mutual occlusion of all part detections of the scenes.

Estimating the scale of each person correctly is challenging due to heavy occlusion especially in crowded scenes. However, pedestrians can be better detected when approximate scale of pedestrians can be estimated [8].

In this work, we exploit scale and occlusion cues to detect pedestrians in crowded scenes by weight averaging estimates from detection response maps, thereby alleviating the problems with results dominated by maximal confidence scores. For scale estimates, we use multiple detectors based on full body and parts to deal with heavy occlusion in crowded scenes. To better integrate the prior and detection response, we propose a scheme where the detection confidence scores can either be penalized or rewarded based on the degree of scale consistency in local neighborhood.

Different from existing methods which rely on either global [9] [8] or greedy optimization [10] [11] for occlusion analysis, we propose a novel local occlusion analysis method in which the potential occluder is first searched for each part detection and used as prior for the detection confidence score.

The Non-Maximum Suppression (NMS) scheme is used to generate final detection results in which similar detections are analyzed and clustered. As occlusion relationship is encoded into the confidence scores of individual detections, the fast greedy NMS can be used to generate the final solution. This plays an essential role in detecting pedestrians that are heavily occluded. If the occluder is incorrectly removed from the final solution, the occluded pedestrians can still be detected by the proposed scheme, whereas in such cases methods based on greedy optimization fail.

Experimental results on benchmark data sets demonstrate that the proposed algorithm performs favorably against the state-of-the-art pedestrian detection methods in crowded scenes.

2. RELATED WORK

Numerous methods have been developed for pedestrian detection under occlusion in crowded scenes. Existing approaches can be generally categorized into three types. The first category of methods is based on single object occlusion analysis. Wang et al. determine occlusion maps from the responses of block-wise HOG features [12] for pedestrian detection. In [13], occlusion map is estimated from depth and motion discontinuities such that pedestrians can be effectively detected in crowded scenes. Most recently, Ouyang et al. develop a deep learning model to implicitly infer the visibility of each body part to deal with occlusion for pedestrian detection [14].

The second type of methods use context information for occlusion handling. Part-based object detection methods have been used to collect local evidence, which is followed by optimization processes for occlusion reasoning. Global optimization methods based on Markov Chain Monte Carlo

[9], Expectation Maximization [15], binary integer programming [8], and greedy approaches [10] [11] have been used for pedestrian detection based on local cues. As responses within local neighborhood are dependent, methods based on greedy optimization are less effective because one false detection may result in more false detections. Furthermore, global optimization methods are mostly computationally expensive.

Detectors that consider two or more overlapped objects as one complex object have been developed [16] [17] [18] in which typical appearance representations of object mutual occlusions are modeled and used for disambiguation. Other cues such as ground planes or head pose estimation from detection responses have also been applied to reduce false alarms [19] [20] in crowd scenes. In [8], scale prior is estimated from detection responses locally to improve detection accuracy, without assuming that ground planes or head poses are known.

3. PROPOSED ALGORITHM

3.1. Full-body and part detectors

We use the the Deformable Part Model (DPM) [2] trained on the INRIA data set [1] as our full-body detector. To detect pedestrians that are heavily occluded, we decompose the root filter F_0 of the full-body detector into blocks and use these blocks and deformable parts to construct part detectors. Similar to [21], the bias term of the linear classifier for each part detector is learned from the training data by distributing the bias term of the full body detector to each component (blocks and deformable parts) proportionally.

3.2. Scale prior estimation

Suppose we have a set of detection responses $D = \{D_1, D_2, \ldots, D_n\}$ for an image I. Each D_i is a five tuple $(x_i, y_i, s_i, l_i, c_i)$, with (x_i, y_i) the detection location, s_i the scale, $l_i \in \{FB, HS, UB\}$ the type of the detector (FB, HS and UB represent full body, head-shoulder and upper body respectively), and c_i the detection confidence score given by detectors [2] [21]. The scale effect of each detection D_i at an image location (x, y) is defined as

$$F_{x,y}(D_i) = \begin{cases} c_i w(l_i) \exp(-\frac{1}{1+s_i/s_0}[\frac{(x-x_i)^2}{\sigma_x^2} + \frac{(y-y_i)^2}{\sigma_y^2}]), & c_i > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $w(l_i)$ is a weighting factor proportional to the number of deformable parts in one detection (as a detection with more detected parts is more discriminative and reliable), s_0 is the scale of a standard sized bounding box, i.e. 120×40 pixels, and σ_x and σ_y control the neighborhood extent to which the scale of a detection can be propagated across the image. Thus, a detection with a higher confidence score and more deformable parts plays a more important role in affecting the detections within its neighborhood. The detection response is not propagated if its detection confidence score is negative.

The scale prior at an image location is estimated by

$$s_p(x, y) = \frac{\sum_{i=1}^{n} F_{x,y}(D_i) s_i}{\sum_{i=1}^{n} F_{x,y}(D_i)} \quad (2)$$

which is the average scale of all the detections weighted by their respective scale influence at that location. Given the estimated scale prior s_p and scale effect function $F_{x,y}(D_i)$, the detection confidence score is refined by

$$c_p(D_i) = c_i + \alpha \cdot \sum_{j=1}^{n} F_{x,y}(D_j) \cdot \left(\left(\min \left(\frac{s_i}{s_p(x_i, y_i)}, \frac{s_p(x_i, y_i)}{s_i} \right) \right)^2 - \beta \right) \quad (3)$$

where α is a weighting factor and β controls the scale consistency. If the scale of a detection is significantly different from the estimated scale prior, the detection confidence score is reduced, and vice versa.

3.3. Mutual occlusion analysis

Mutual occlusion analysis is used to detect pedestrians under heavy occlusion based on part detectors. In this paper, two detections D_i and D_j are called an occludee-occluder pair, as shown in Fig. 1, if D_i is a part detection and the image area of the body part excluded by D_i is mostly occupied by D_j, and D_i and D_j could not correspond to the same person (i.e. the overlap ratios between corresponding deformable parts are small).

(a) occlusion pair (b) occluder (c) occludee

Fig. 1: Illustration of an occluder-occludee pair.

The optimal occluder $D_{o*}(D_i)$ of D_i is one of its occluders with the highest detection confidence score

$$D_{o*}(D_i) = \underset{D_j \in D}{\operatorname{argmax}} \left(c_p(D_j) | O(D_i, D_j) = 1 \right) \quad (4)$$

The final confidence score of a part detection D_i is defined as

$$c_f(D_i) = \min(c_p(D_{o*}(D_i)), c_p(D_i)) \quad (5)$$

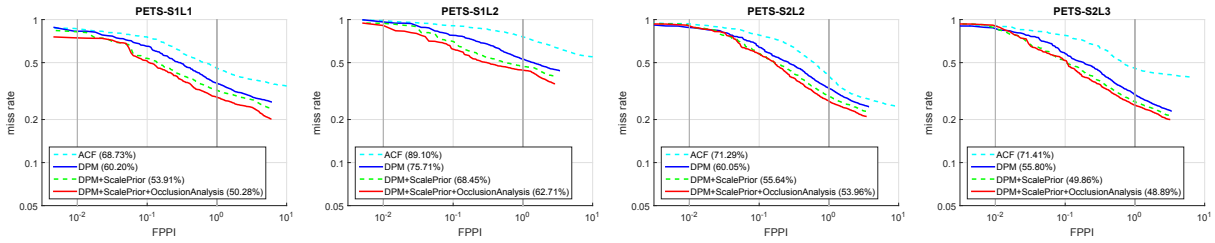

Fig. 2: Results on the PETS-S1L1, PETS-S1L2,PETS-S2L2, PETS-S2L3 sequences.

The operation in (5) is based on the assumption that D_i can be positive only when $D_{o*}(D_i)$ is positive. If a part detection does not have an occluder, its possibility of being a false positive becomes high. As a penalty, we set its c_f to be c_p minus the ratio of the area of the body part it excludes to the area of the full body. In this way, the final detection confidence score of each part detection encodes its occluder's information, which facilitates the following simple NMS procedure for final detection determination. For a full body detection $c_f(D_i) = c_p(D_i)$.

3.4. Detection cluster based Non-Maximum Suppression

Multiple responses of the same person are likely to occur due to the nature of sequential processes at different locations and scales. In addition, although part detectors can deal with occlusion to certain extent, the localization and scale accuracy is low. As such, we cluster responses from D into independent pedestrian hypotheses to determine the final detection results.

Two detections are clustered into one group if (a) the detected head parts have sufficient overlap and at least two other deformable parts overlap with more than a certain threshold (e.g., 50%), or (b) one of the two detections has more than half number of its deformable parts overlap with the corresponding parts of the other detection. With these heuristic rules, each cluster is treated as one pedestrian hypothesis that consists of multiple instances although they differ from each other in terms of location, scale and parts.

For NMS, detections are sorted according to their final confidence scores c_f in the descending order and the overlap ratio is calculated for the effective region of a detection instead of the full bounding box (e.g. for a upper body detection, the effective region is the upper 60% of the full bounding box) so that heavily occluded detections still have chances to be accepted as positives. Furthermore, in contrast to existing methods that use NMS, once an instance of a detection cluster is determined to be positive, the remaining instances within that cluster are not further processed. When an instance is accepted as positive, its bounding box is taken as the representative bounding box of the cluster, i.e. confidence score weighted average of all the bounding boxes in that cluster.

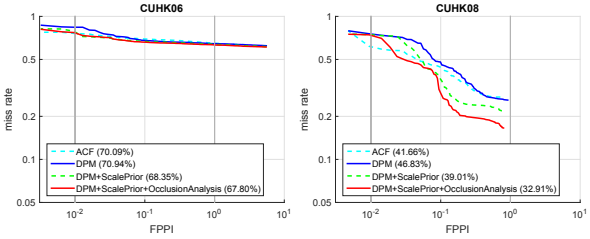

Fig. 3: Results on the CUHK06 and CUHK08 sequences.

4. EXPERIMENTAL RESULTS

The proposed multiple pedestrian detection algorithm is evaluated on the PETS 2009 and the CUHK benchmark data sets. Both data sets consist of image sequences with multiple pedestrians. For the PETS 2009 data set, we use the S1L1, S1L2, S2L2 and S2L3 sequences where multiple pedestrians are heavily occluded. Similarly, the CUHK06 and CUHK08 sequences from the CUHK data set are used for evaluation. For all the experiments, the parameters are fixed to be $\sigma_x = 50$ and $\sigma_y = 25$ in (1) to account for scale variations in different directions. In addition, α and β in (3) are set to be 0.35 and 0.95 respectively, which indicate that if the scale of a detection response deviates from the estimated scale for less than 5%, its detection confidence score will be increased; otherwise, its detection confidence score will be decreased. Our sensitivity analysis on the parameters shows that the proposed algorithm performs robustly with these settings.

The detection results are evaluated in terms of Log-Average Miss Rate (LAMR) and we compare the proposed approach with the state-of-the-art ACF detector [3] and the baseline DPM detector [2]. For the PETS data set, as shown in Fig. 2, the LAMR is reduced by 4.41% to 7.26% with only scale prior, and the LAMR is further reduced by 0.97% to 5.74% with occlusion analysis. The most significant improvement of 13.00% is achieved on the most crowded S1L2 sequence while the least improvement of 6.09% is obtained on the least crowded S2L2 sequence. For the CUHK sequences, as depicted in Fig. 3, the improvement is 3.14% for the less occluded CUHK06 sequence, whereas the improvement is 13.92% for the heavily occluded CUHK08 sequence.

Fig. 4 shows some qualitative results of the evaluated

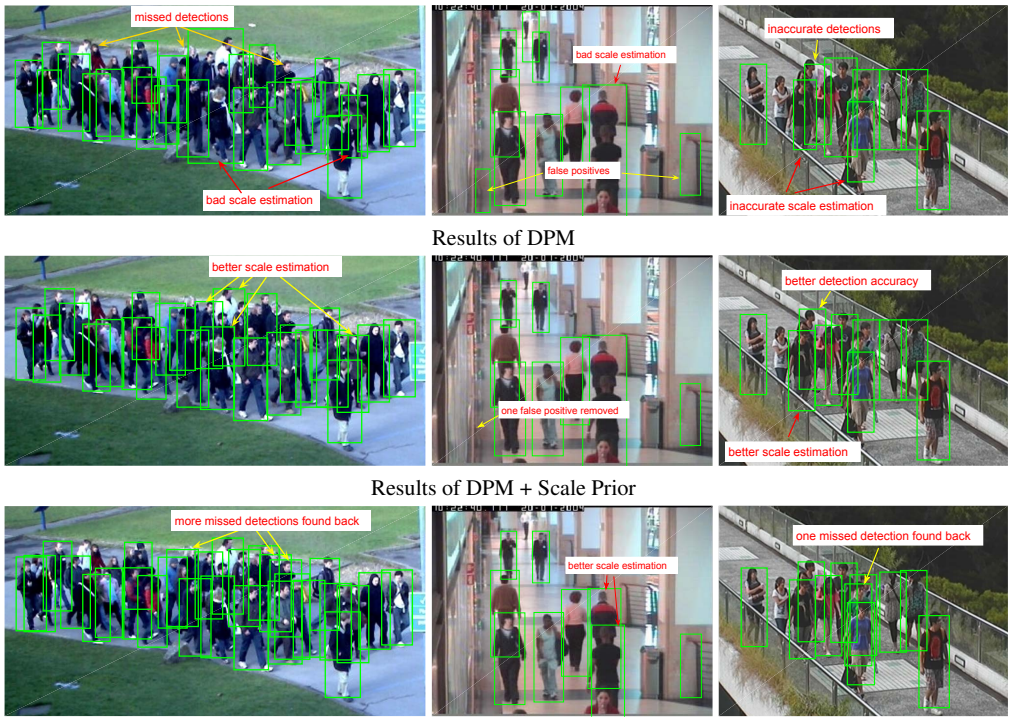

Fig. 4: Detection results on images from the PETS S1L2, CUHK06 and CUHK08 data sets.

algorithms on images from the benchmark data sets. While there are many missed detections and some wrong scale false alarm detections for the DPM method due to inaccurate scale estimation and heavy occlusion in crowded scenes, the proposed algorithm is able to detect more pedestrians and remove some detections of incorrect scales.

We compare the proposed algorithm with two learning based multiple pedestrian detection methods [11] [17] on the PETS and Parking Lot data sets. As the LAMR scores are not reported and the source codes are not available, we compute the Average Precision (AP) as discussed in [11] and draw the 1-precision vs recall curve as presented in [17]. Table 1 shows that the proposed algorithm achieves higher AP than [11] on the two PETS sequences without learning from examples with ground truth labels. Fig. 5 shows that at the recall rate of 0.9, the proposed algorithm achieves precision higher than 0.7, whereas the recall rate of [17] is lower than 0.9 even when the precision is lower than 0.2. We note that the proposed algorithm is not evaluated against [8] as the source code is not available.

Table 1: Evaluation with [11] on the PETS sequences

	PETS-S2L2	PETS-S2L3
Yan et al. [11]	0.7555	0.6557
Proposed algorithm	**0.7683**	**0.6857**

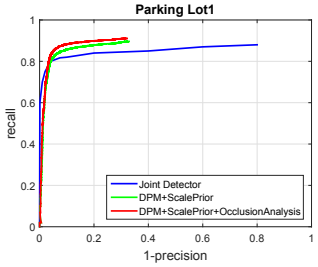

Fig. 5: Comparison with [17] on the Parking Lot1 sequence

5. CONCLUSION

We propose an algorithm to exploit scale prior and occlusion analysis to detect pedestrians in crowded scenes. Scale prior at each image location is estimated based on information provided by neighboring detections, and the confidence score of each detection is refined according to its consistency with the estimated scale prior. Local occlusion analysis is proposed to encode occlusion information into detection confidence scores, which facilitates the final fast detection cluster based NMS. Experimental results on benchmark data sets show that the proposed algorithm performs favorably against the state-of-the-art methods.

Acknowledgement Lu Wang is supported in part by Chinese Scholarship Council, NSFC #61202258 and the Fundamental Research Funds for Central Universities of China #N130404016.

6. REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[3] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[4] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1751–1760.

[5] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5079–5087.

[6] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 82–90.

[7] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[8] H. Idrees, K. Soomro, and M. Shah, "Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning," *IEEE Transactions on Pattern Analysis and Machine Intelligences*, vol. 37, no. 10, pp. 1986–1998, 2015.

[9] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 459–466.

[10] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Proceedings of IEEE International Conference on Computer Vision*, 2005, pp. 90–97.

[11] J. Yan, Z. Lei, D. Yi, and S. Z Li, "Multi-pedestrian detection in crowded scenes: A global view," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3124–3129.

[12] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proceedings of IEEE International Conference on Computer Vision*, 2009, pp. 32–39.

[13] M. Enzweiler, A. Eigenstetter, B. Schiele, and D.M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 990–997.

[14] W. Ouyang, X. Zeng, and X. Wang, "Partial occlusion handling in pedestrian detection with a deep model," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[15] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu, "Unified crowd segmentation," in *Proceedings of European Conference on Computer Vision*, 2008, pp. 691–704.

[16] C. Arteta, V. Lempitsky, J A. Noble, and A. Zisserman, "Learning to detect partially overlapping instances," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3230–3237.

[17] S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele, "Learning people detectors for tracking in crowded scenes," in *Proceedings of IEEE International Conference on Computer Vision*, 2013, pp. 1049–1056.

[18] W. Ouyang, X. Zeng, and X. Wang, "Single-pedestrian detection aided by two-pedestrian detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1875–1889, 2015.

[19] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *Proceedings of European Conference on Computer Vision*, 2010, pp. 241–254.

[20] I. Ali and M. N. Dailey, "Multiple human tracking in high-density crowds," *Image and Vision Computing*, vol. 30, no. 12, pp. 966–977, 2012.

[21] L. Wang, X. Ji, Q. Deng, and M. Jia, "Deformable part model based multiple pedestrian detection for video surveillance in crowded scenes," in *Proceedings of the International Conference on Computer Vision Theory and Applications*, 2014, pp. 599–604.