PEDESTRIAN DETECTION IN HIGHLY CROWDED SCENES USING "ONLINE" DICTIONARY LEARNING FOR OCCLUSION HANDLING

Bing Wang, Kap Luk Chan, Gang Wang, Haijian Zhang

School of Electrical and Electronic Engineering
Nanyang Technological University

ABSTRACT

Pedestrian detection is one of the most important task for video analytics of an intelligent surveillance system. In this paper, we propose a framework to improve the detection performance of a generic pedestrian detector for highly crowded scenes. The generic offline-trained pedestrian detectors usually cannot handle the problem of detecting pedestrians in highly crowded scenes due to the severe mutual occlusions of the pedestrians. In our approach, we firstly enhance the head detection and suppress the detections of other body parts in the deformable part-based model because the heads of pedestrians less likely to be occluded in highly crowded scenes. Then we propose to utilize multiple-instance dictionary learning to refine the previous detection responses. Compared to other related work, our approach builds a data-adaptive dictionary (codebook) for the heads of pedestrians, hence it can better handle the problem of detecting pedestrians in highly crowded scenes. The experiments on three datasets containing video clips of crowded scenes demonstrated the effectiveness of our proposed approach, significantly improving the state-of-the-art detector.

Index Terms— pedestrian detection, dictionary learning, occlusion handling, crowded scenes

1. INTRODUCTION

With the explosive growth of the deployment of surveillance cameras nowadays, the demand for automatically detecting pedestrians in crowded scenes is high in order to enable rich video analytics for a wide variety of applications. The state-of-the-art pedestrian detectors [1, 2, 3, 4] have achieved very good performance when detecting pedestrians in relatively less crowded scenes, where the occlusions are not severe. However, their performance on crowded scenes is poor, due to the severe mutual occlusions of the pedestrians, and sometime the problem of small target size.

There has been plenty of literature that addresses occlusion handling in object detection [5, 6, 7, 8, 9], which can further promote the tracking-by-detection based multi-target tracking methods [10, 11] within surveillance applications. Modeling occlusions as regions which are inconsistent with target statistics is one common approach. Girshick *et al.* [5] propose to utilize an occluder part in the grammar model when some of the parts are occluded. Wang *et al.* [6] develop a pedestrian detection approach capable of handling partial occlusions by combining Histograms of Oriented Gradients (HOG) and Local Binary Pattern (LBP) as the feature set. Meger *et al.* [7] propose to use depth inconsistency from 3D data to handle the occlusions. Hsiao *et al.* [8] propose to explicitly model occlusions by reasoning 3D interactions of objects. Shu *et al.* [9] propose to select the subset of parts, which maximizes the probability of detection, in pedestrian detection for occlusion handling. Nevertheless, these methods cannot handle severe occlusions in crowded scenes without any 3D information.

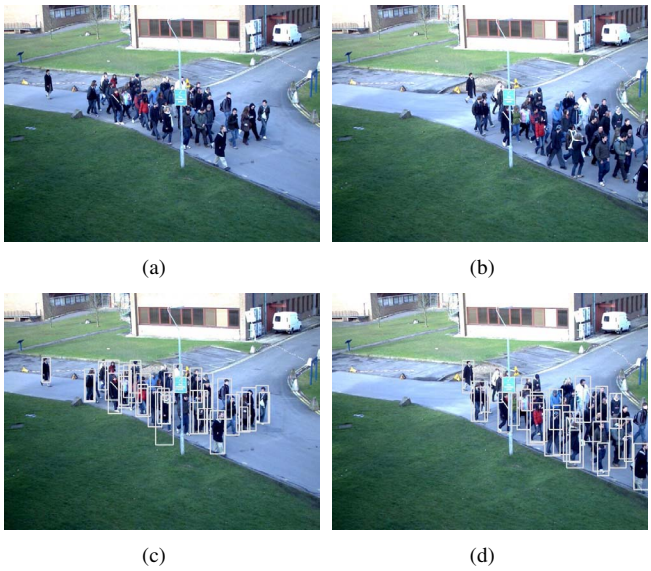


Fig. 1. Examples of crowded scenes and their corresponding detection results with the proposed approach are presented in row 1 and row 2, respectively.

In this paper, we address such difficulties by proposing the use of "online" multiple-instance dictionary learning to refine a generic detector for enhancing head detection. The intuition is that, in a real crowded scene captured by a surveil-

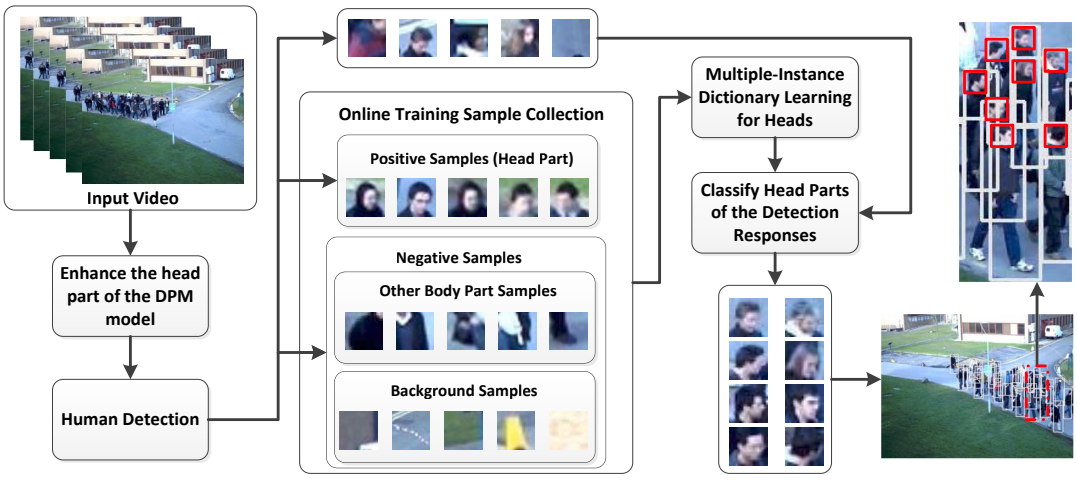


Fig. 2. The framework of the proposed approach.

lance camera, individual target is usually mutually occluded by others, while the head of each individual is less likely to be occluded or at most slightly occluded, as shown in Fig. 1. Based on this assumption, we increase the weight of the head detection in the deformable part-based model (DPM) [3] and learn a discriminative dictionary for heads in an "online" manner to improve detection performance.

Learning is an important part in computer vision tasks, such as stereo image representation [12], image classification [13, 14], image retrieval [15], object categorization [16], object tracking [17, 18], face recognition [19] and human gender recognition [20]. We choose dictionary learning in this work. The dictionary (codebook) enables us to enforce the explicit representations rather than individual features or simple combinations of the features. The dictionary can also facilitate hierarchical representations. Moreover, dimensionality reduction can be achieved through the quantization process.

Our detection framework consists of the steps illustrated in Fig. 2. First, we increase the weight of the head detection in the deformable part-based model [3] and apply this modified detector with a relatively low detection threshold on every frame of a video sequence. The reliable detection responses are selected by their high detection confidences. We use the head parts of the reliable detection responses as the positive samples and the other body parts of the reliable detection responses as one of the two types of the negative samples. The other type of negative samples are collected automatically from background, where there is no detection responses. Second, we use the above three different sample types to learn a discriminative dictionary. Third, we use the online learned dictionary to refine the head parts of the initial detection responses. Finally, we obtain the refined pedestrian detection results.

The rest of our paper is organized as follows. In section 2, the initial pedestrian detection with enhanced head detection is introduced. Section 3 presents the "online" dictionary learning. Section 4 shows the experimental evaluations on several challenging crowded video sequences. Finally, section 5 concludes the paper.

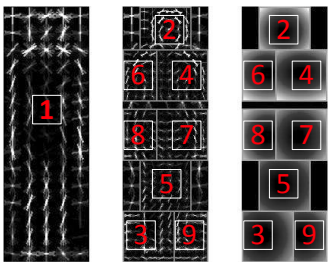


Fig. 3. The deformable part-based model from [3].

2. INITIAL DETECTION WITH ENHANCED HEAD DETECTION

The deformable part-based model for pedestrian detection similar to [3] often fail when the pedestrians are severely occluded. This is because the detection score in [3] is computed from all the body parts, without considering that most of the parts may be occluded in crowded scenes. The detection score at location (x, y) is computed in [3] as:

$$S_1(x, y) = b + \sum_{i=1}^{i=n} s(p_i) \quad (1)$$

where b is a bias term, n is the number of the body parts, and $s(p_i)$ is the score of body part i.

In this formulation, it is obvious that even if a body part was occluded, its corresponding score still makes equal con-

tribution to the final detection score. This can drastically influence the detection performance especially when dealing with highly crowded video sequences. To address this problem, we propose to increase the weight of the head part score in equation (1). As shown in Fig. 3, the second part represents the head in the DPM model. Hence, equation (1) can be reformulated as:

$$S_2(x,y) = b + s(p_1) + \alpha \cdot s(p_2) + \sum_{i=3}^{i=n} s(p_i) \qquad (2)$$

where α is a weighting coefficient of the head part ($\alpha = 2.5$ in our implementation).

Nevertheless, relying on DPM detector with head part enhanced is not sufficient to obtain satisfactory detection results. Hence, in the next section, we propose to use the "online" dictionary learning to further refine the initial detections.

3. "ONLINE" DICTIONARY LEARNING

The objective here is to "online" learn a discriminative dictionary while keeping the computational complexity low. The learning involves "online" training sample collection, feature representation and "online" learning.

The initial detections with scores higher than a certain threshold ω ($\omega = 3$ in our implementation) are selected as the reliable detections. Then perform learning as outlined in Fig. 2 and the positive samples are collected from the reliable detections while the negatives samples are collected from other body parts and background.

For feature representation, image patches are sampled at an interval of 4 pixels in horizontal and vertical directions. At each sampling location, patches of sizes 8×8 and 16×16 pixels are taken. We resize each image patch to 8×8 and compute five types of features to represent it. We choose the patch size for feature extraction according to the actual sizes of the heads of the pedestrians in PETS dataset [21]. The computed features are HOG [1], LBP [6], GIST [22], encoded SIFT [23] and LAB color histograms.

The discriminative dictionary is "online" learned by the max-margin multiple-instance dictionary learning (MMDL) algorithm in [24]. This discriminative dictionary comprises a set of linear classifiers (G-code classifiers) for different patch clusters from the three sample classes. For each sample class, we use the training images in this class as positive samples, and the rest training images from other classes as negative samples. Suppose $K + 1$ G-code classifiers are learned through MMDL. Given a test image, patch-level features are densely extracted. We define x as a patch feature vector, whose response is given by the ith G-code: $w_i^T x, i \in \{0, 1, ..., K\}$. Hence, a response map for each G-code classifier can be obtained. A three-level spatial pyramid representation [25] is utilized for each response map, resulting in $(1^2+2^2+4^2)$ grids. In each grid, the maximal response for each G-code classifier is calculated. Thus, $3 \times (K + 1)$ length feature vector is obtained for each grid. Therefore, the test image is compactly represented by the concatenation of features in all grids.

Note that the feature encoding by using G-code involves no more than a dot product operation. Hence, its computation complexity is very low. This can significantly accelerate the speed of the classification process of our pedestrian detection framework.

Subsequently, we use the "online" learned G-code classifiers for feature encoding of each head image. Then, the refined head parts of the initial detection responses are obtained. Finally, we project the refined head part images back into the video sequence and generate the final pedestrian detection output.

4. EXPERIMENTS

The proposed pedestrian detection framework has been evaluated on three publicly available sequences with crowd scenes: PETS S1.L2, PETS S2.L2 and PETS S2.L3. We choose the medium density crowd and dense crowd sequences for our experiments. The PETS dataset S1 is used for person count and density estimation, which means that the density level is higher than the other PETS datasets. In the S1 dataset, we choose the high density crowd sequence S1.L2 for the evaluation. In all the experiments, we just use the first camera view sequences in PETS dataset. Moreover, we only use the visual information of the sequences and no other prior knowledge such as the camera calibration or the statistic obstacles are used.

We compare our method with the original DPM detector and our modified DPM detector with head part enhanced. To study the effect of sampling on dictionary learning, we compare the learning performance in two settings: learning with head parts and back ground, and learning with head parts, body parts and background.

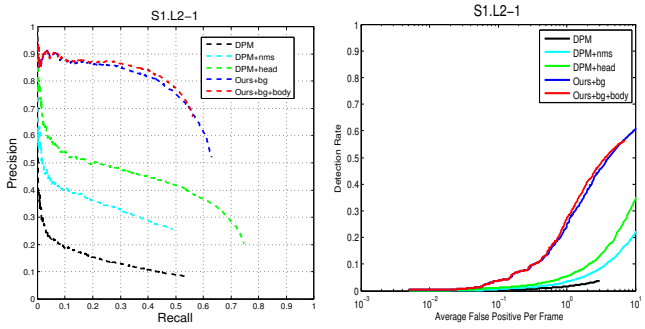
The criterion of the PASCAL VOC challenge [26] for evaluations is used in our experiments. A detection, which has more than 0.5 overlap with the groundtruth, is determined as true positive in the evaluations. The performance of our detection framework is analyzed by computing Precision-Recall curves for all three sequences. Furthermore, we also use detection accuracy (as measured by detection rate and average false positives per frame) as our evaluation criterion. The evaluation code of detection accuracy is from [27].

In our implementation, we set a detection threshold $t_d = 0.3$ for the modified DMP detector to achieve a high recall. For the online dictionary learning, we set the number of randomly selected positive training samples $N_{pos} = 200$ and the number of randomly selected negative training samples $N_{neg} = 400$.

As we can see in Fig. 4, 5, 6, our approach has achieved the best performance in all three crowded sequences. "DPM+

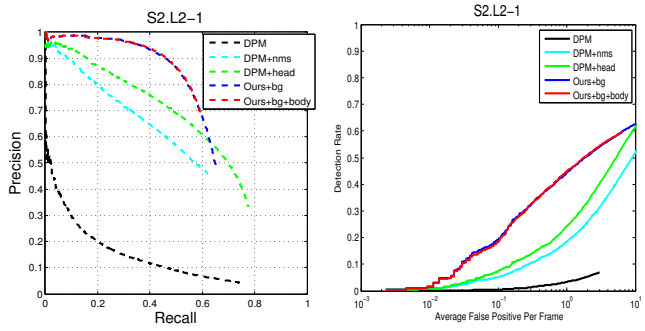
nms" and "DPM" denote the results of the original DPM detector with and without non-maximum suppression, respectively. "DPM+head" denotes the reuslts of the DPM detector with head part enhanced. "Ours+bg+body" and "Ours+bg" denote the results of the training sample classes with and without body part class within the proposed online-learning framework, respectively. For the Precision-Recall curve, the closer to the top right corner, the better performance it is. For the detection accuracy curve, the closer to the top left corner, the better performance it is. In the three sequences, PETS S2.L2 sequence with medium density crowd is less crowded than the other two sequences. The less significant improvement of PETS S2.L2 sequence, which is shown in Fig. 5, demonstrates that our proposed approach is more capable of handling pedestrian detection in highly crowded scenes.

In addition, the computation speed of our approach is relatively fast. Although the initial modified DPM detector takes about 15 second for each frame, our extra steps take only 10 second on average for each frame on a 3GHz PC with 8 GB memory. Moreover, the proposed approach is implemented in Matlab, which can be further optimized.



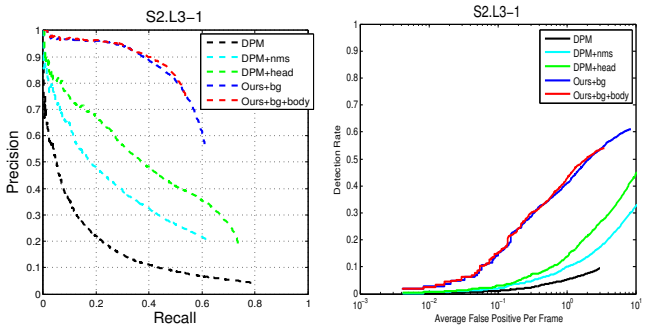
(a) (b)

Fig. 4. Performance comparison on PETS S1.L2 sequence.



(a) (b)

Fig. 5. Performance comparison on PETS S2.L2 sequence.



(a) (b)

Fig. 6. Performance comparison on PETS S2.L3 sequence.

5. CONCLUSION

In this paper, we propose an effective detection framework to improve the performance of generic detectors in crowded scenes. We reformulate the score computation of body parts in the original DPM detector to enhance the head part of the deformable part-based model to make it more suitable to the crowded sequences and use the "online" learned dictionary to refine the detection responses. The experimental results on three benchmark sequences demonstrate the superiority and effectiveness of our approach in detecting pedestrians with occlusions handling in crowded scenes.

6. REFERENCES

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

[2] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[3] P. Felzenszwalb, R. 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.

[4] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in European Conference on Computer Vision (ECCV), 2010.

[5] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Object detection with grammar models," in Conference on Neural Information Processing Systems (NIPS), 2011.

[6] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in IEEE Con-

ference on Computer Vision and Pattern Recognition (CVPR), 2009.

[7] D. Meger, C. Wojek, B. Schiele, and J. J. Little, "Explicit occlusion reasoning for 3d object detection," in British Machine Vision Conference (BMVC), 2011.

[8] E. Hsiao and M. Hebert, "Occlusion reasoning for object detection under arbitrary viewpoint," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[9] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[10] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Online multi-person tracking-by-detection from a single, uncalibrated camera," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 9, pp. 1820–1833, 2011.

[11] B. Wang, G. Wang, K. L. Chan, and L. Wang, "Tracklet association with online target-specific metric learning," in CVPR, 2014.

[12] Ivana Tosic and Pascal Frossard, "Dictionary learning for stereo image representation," IEEE Transactions on Image Processing, vol. 20, no. 4, pp. 921–934, 2011.

[13] Pablo Sprechmann, Ignacio Ramirez, and Guillermo Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[14] Shu Kong and Donghui Wang, "A dictionary learning approach for classification: separating the particularity and the commonality," in European Conference on Computer Vision (ECCV), 2012.

[15] G. Wang, D. Hoiem, and D. Forsyth, "Learning image similarity from flickr groups using fast kernel machines," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, pp. 2177–2188, 2012.

[16] G. Wang, D. Forsyth, and D. Hoiem, "Improved object categorization and detection using comparative object similarity," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 10, pp. 2442–2453, 2013.

[17] Naiyan Wang, Jingdong Wang, and Dit-Yan Yeung, "Online robust non-negative dictionary learning for visual tracking," in IEEE International Conference on Computer Vision (ICCV), 2013.

[18] Junliang Xing, Jin Gao, Bing Li, Weiming Hu, and Shuicheng Yan, "Robust object tracking with online multi-lifespan dictionary learning," in IEEE International Conference on Computer Vision (ICCV), 2013.

[19] J. Lu, Y. P. Tan, and G. Wang, "Discriminative multi-manifold analysis for face recognition from a single training sample per person," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 39–51, 2013.

[20] J. Lu, Y. P. Tan, and G. Wang, "Human identity and gender recognition from gait sequences with arbitrary walking," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 9, no. 1, pp. 51–61, 2014.

[21] PETS 2009, "Pets 2009 benchmark data," http://www.cvg.rdg.ac.uk/PETS2009/a.html.

[22] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International Journal of Computer Vision, vol. 42, no. 3, pp. 145–175, 2001.

[23] David G. Lowe, "Object recognition from local scale-invariant features," in IEEE International Conference on Computer Vision (ICCV), 1999.

[24] Xinggang Wang, Baoyuan Wang, Xiang Bai, Wenyu Liu, and Zhuowen Tu, "Max-margin multiple-instance dictionary learning," in International Conference on Machine Learning (ICML), 2013.

[25] S. Lazebnik, C. Schmid, Ponce, and J. Beyond, "bags of features: Spatial pyramid matching for recognizing natural scene categories," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006.

[26] M. Everingham, V. Gool, L. Williams, C. K. I., J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," International Journal of Computer Vision, vol. 88, no. 2, pp. 303–338, 2010.

[27] H. Pirsiavash, D. Ramanan, and C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.