International Journal of u- and e- Service, Science and Technology
Vol.8, No. 6 (2015), pp.249-256
http://dx.doi.org/10.14257/ijunesst.2015.8.6.24

Locate and Detect Persons in Crowded Scenes Aided by Objectiveness Measure

Shilin Zhang and Xunyuan Zhang

North China University of Technology
zhangshilin@126.com

Abstract

Locating persons in crowded scenes is very difficult due to multi-resolution and complex environment. The other difficulty in pedestrian detection domain is the real time requirement, because the camera installed on the crossing road is in high definition. In this paper, we presented a multi-task pedestrian detection framework boosted by Bing feature. We firstly trained upright full-body, multi-person, half-body and head models, then we compute the object-ness score and generate 1000 proposals by Bing feature, and at last we apply different model to different aspect ratio of the detection proposals. The experiment results on the PASCAL VOC 2007 show that our method outperforms all the other methods and achieved lower miss rate than the state-of-the-art. The computation time cost is just the half of state-of-the-art method.

Keywords: *Bing feature, Pedestrian Detection, Deformable Part-based Models, Multi-pedestrian Detector*

1. Introduction

Pedestrian detection has been one of the most important topics [1-10] in pattern recognition for decades, for its importance in real time applications, especially in transportation environment, such as driving assistance and video surveillance [11]. In recent years, especially due to the popularity of gradient features, pedestrian detection field has achieved impressive progresses in both effectiveness and efficiency [3, 7]. The leading detectors can achieve satisfactory performance on high resolution benchmarks [20]; however, they encounter difficulties for the low resolution pedestrians [4]. Unfortunately, the low resolution pedestrians are often very important in traffic scenes, as can be seen from Figure 1. For example, the driver assistance systems need detect the low resolution pedestrians to provide enough time for reaction.

In real application, the video frames captured by the high definition camera are sometimes 5 million-pixels (2590*1920). So the processing time per frame must be less than 1/15 per second. Otherwise, they will not satisfy the real application requirement. The Bing feature designed by Cheng [36] can be used for efficient object-ness estimation, which requires only a few atomic operations (*e.g.,* ADD, BITWISE SHIFT, *etc.*,) Bing feature efficiently (300fps on a single laptop CPU) generates a small set of high quality object windows, yielding 96.2% object detection rate (DR) with 1,000 proposals. Increasing the numbers of proposals and color spaces for computing BING features, the performance can be further improved to 99.5% DR. Another salient aspect of the proposed approach is that it is general enough to characterize a variety of detection objects.

ISSN: 2005-4246 IJUNESST
Copyright ⓒ 2015 SERSC

Figure 1. Pedestrian Crowd in Traffic Scenes

So we borrow the idea of the Bing feature to boost the pedestrian detection speed to satisfy the real application requirement. Because Bing feature can generate just 1000 proposals that perhaps containing pedestrians and it is more efficient than the sliding window method [28]. But the detection window is diversified with different aspect radio, and they can be single upright full body person, upper body person and head, or shoulder-to-shoulder multi-pedestrian. So we must devise different models to capture the visual patterns. A multi-pedestrian window found by a multi-pedestrian detector can guide the detection of each pedestrian in this window. When pedestrians walk side by side, they form the shoulder-to-shoulder visual pattern. Then the multi-pedestrian detection results are used to reinforce the evidence of detecting each of the every pedestrian [25, 26].

The contribution of this paper can be summarized in two-fold. 1) We implement a fast and efficient pedestrian detection method that can fulfill the transportation detection task 2) We integrated HOG+SVM, HAAR and DPM algorithms together to process different scenarios. With a fast computation approach, our method outperforms other algorithms with respect to speed and accuracy. Experiments carried on the challenging PASCALVOC 2007 datasets show that the detection accuracy is 96.5% and the computation speed is two times faster than the-state-of-the-art method.

2. Related Works

We review the related works from two aspects: Objectiveness measure and pedestrian detection, and we will discuss them separately in detail in the coming paragraphs.

2.1. Objectiveness Measure

Objectiveness proposal [30-35] generation methods avoid making decisions early on, by proposing a small number of pedestrian position proposals that are expected to cover all persons in an image. Producing rough segmentations [32] as pedestrian proposals has been shown to be an effective way of reducing search spaces for category specific classifiers, whilst allowing the usage of strong classifiers to improve accuracy. However, these two methods are computationally expensive, requiring 2-7 minutes per image. Alexe, *et al.*, [30] proposed a cue integration approach to get better prediction performance more efficiently. Zhang, *et al.*, [32] proposed a cascaded ranking SVM approach with orientated gradient feature for efficient proposal generation. Uijlings, *et al.*, [34] proposed a selective search approach to get higher prediction performance. Cheng, *et al.*, propose a simple and intuitive method which generally achieves better detection performance than others, and is 1,000+ times faster than the other most popular alternatives [31-33].

In addition, for efficient sliding window pedestrian detection, keeping the computational cost feasible is very important [17, 34]. However, it can only be used to

 Copyright ⓒ 2015 SERSC

speed up classifiers that users can provide a good bound on highest score. Also, some other efficient classifiers [24] and approximate kernels [26] have been proposed. These methods aim to reduce computational cost of evaluating one window, and naturally can be combined with objectiveness proposal methods to further reduce the cost.

In this paper, we use the Bing feature to measure the objectiveness of every video frame and get a list of position proposals. Then the detection procedure is resorted to the next strong multi-task pedestrian detection scheme.

2.2. Pedestrian Detection

There is a long history of research on pedestrian detection. Most of the modern detectors are based on statistical learning and sliding-window scan. Large improvements came from the robust features, such as HOG, HAAR, DPM and *etc.*, There are some papers fused HOG with other features to improve the performance. Some papers focused on special problems in pedestrian detection, including occlusion handling, speed, and detector transfer in new scenes. We refer the detailed surveys on pedestrian detection to [19]. The progress on object detection has been achieved by the investigation on classification approaches, features and articulation handling approaches. 1) Classification approaches used include various boosting classifiers [35], SVM classifiers, and grammar models [15] and deep model 3) Features under investigation include HAAR-like features, edge lets [22], shape lets [27], histogram of gradients (HOG) [4], bag-of-words [18], integral histograms [26], color histograms, covariance descriptors [13], co-occurrence features, local binary patterns [18], color-self-similarity [14], depth [15], segmentation [11], features learned from training data and their combinations 3) Articulation handling approaches under investigation include Deformable part based models (DPM) [12, 22, 28], pictorial structures [15], pose let [21, 27] and mixture of parts [16].

Resolution related problems have attracted attention in recent evaluations. The pedestrian detection performance depends on the resolution of training samples. The paper [14] pointed that the pedestrian detection performance drops with decreasing resolution. However, there are very limited works proposed to tackle this problem. The most related work is [33], which utilized root and part filters for high resolution pedestrians, while only used the rigid root filter for low resolution pedestrians. [20] Proposed to use a single model per detection scale, but the paper is focused on speedup. The pedestrian detector is built on the popular DPM (deformable part model), which combined rigid root filter and deformable part filters for detection.

In this paper, we firstly generate 1000 proposals from every image according to objectiveness score computed by Bing feature, and then we proposed a multi-task model to handle dataset bias. The multi-task idea in this paper is motivated by works on face recognition across different domains, such as [28, 5].

3. Methodology

There are two steps to detect pedestrian in video frames. In the first step, we trained a 2 stage SVM objectiveness pedestrian position proposals. In the next stage, we apply the strong pedestrian detection algorithms to filter out the false proposals.

3.1. Objectiveness Measure by Bing Feature

Similar to human vision system which efficiently perceives objects before identifying them [36], we introduce a simple 64D norm of the gradients (NG) feature (Section 3.1), as well as its binary approximation, *i.e.*, binarized normed gradients feature (Section 3.3),for efficiently capturing the objective-ness of an image window. To find pedestrians within an image, we scan over a predefined quantized window sizes (scales and aspect ratios). Each window is scored with a linear model:

Copyright ⓒ 2015 SERSC

$$s_l = \langle w, g_l \rangle \tag{1}$$

$$o_l = v_i * s_l + t_i \tag{2}$$

Where s_l, g_l, l, are filter score, NG feature and location n of a window respectively. Using non-maximal suppression (NMS), we select a small set of proposals from each size i. Some sizes are less likely than others to contain an object instance. Thus we define the objective-ness score (*i.e.*, calibrated filter score) as (2). The parameters are learnt coefficient and a bias terms for each quantized size i.

To make use of recent advantages in model binary approximation [23-25], Bing feature namely binarized normed gradients are an accelerated version of NG feature, defined as:

$$\langle w,b \rangle \approx \sum_{j=1}^{N_w} \beta_j (2\langle a_j^+, b \rangle - |b|) \tag{3}$$

N_w denotes the number of basis vectors.

$a_j \in \{-1-1\}^{64}$ denotes a basis vector.

$\beta_j \in R$ denotes the corresponding coefficient.

$a_j^+ \in \{0-1\}^{64}$ is the binary version of the original presentation.

The detailed implementation, please refer to the paper [36].

3.2. Multi-task Pedestrian Detection

In this step, we will filter out the false positives, and find the true pedestrian locations. According to the aspect ratio of the detection window, we classify the proposals into 3 classes: the upright full body, the multi-pedestrian and the half body or the head. So we will train 3 different models. In all the models, a multi-resolution detection method is adopted, which considers the relationship of samples from different resolutions, including the commonness and the differences. By using the resolution aware transformations, we map features from different resolutions to a common subspace, in which they have similar distribution. A shared detector is trained in the resolution-invariant subspace by samples from all resolutions, to capture the structural commonness. In DPM, pedestrian consists of parts, and every part consists of HOG cells. The only difference among different resolution lies in the feature vector of every cell, so that the resolution aware transformations P_L and P_H are defined on it.

The P_L and P_H are of the dimension $n_d * n_f$, and they map the low and high resolution samples from the original n_f dimensional feature space to the n_d dimensional subspace. The features from different resolutions are mapped into the common subspace, so that can share the same detector. We still denote the learned appearance parameters in the mapped resolution invariant subspace as W_a, which is a $n_d * n_c$ matrix, and of the same size with $P_H \Phi_a(I,L)$, which is the matrix based representation for the DPM. The spatial prior parameter is denoted as w_s. The objective function is formulated as:

$$\arg \min_{W_a, w_s, P_{H,P_L}} \frac{1}{2} w_s^T w_s + f_{I_H}(W_a, w_s, P_H) + f_{I_L}(W_a, w_s, P_L) \tag{4}$$

I_L and I_H denote the high and low resolution training sets, including both pedestrian and background. f_{IH} and f_{IL} are used to consider the detection loss and regularize the parameters P_H, P_L and W_a. We need to find an optimal combination of w_s, P_H, P_L and W_a, however the above equation is not convex. It can be transformed into a standard SVM problem, and can be optimized the two sub-problem iteratively.

 Copyright ⓒ 2015 SERSC

4. Experiments and Comparison

We extensively evaluate our method on PASCAL VOC 2007 dataset [32] by split the samples containing persons into 2 sets, one as the training set (total 2016 images) and the other as the test sets (total 1024 images). We compare our results with 12 state-of-the-art methods in terms of miss rate and efficiency.

4.1. Weighting Scheme and Spatial Predicate

We evaluate DR-#WIN on VOC2007 test set, which consists of 1,024 images with bounding box annotation for the pedestrian instances. The pedestrians' viewpoint, scale, position, occlusion, and illumination, make this dataset very suitable to our evaluation as we want to find all pedestrians in the images. As observed from Figure 2, increasing the divergence of proposals by collecting the results from different parameter settings would improve the DR at the cost of increasing the number of proposals (#WIN). By simply collecting the results from 3 color spaces: RGB, HSV, and GRAY, the recall achieves more DR using only 5,000 proposals.

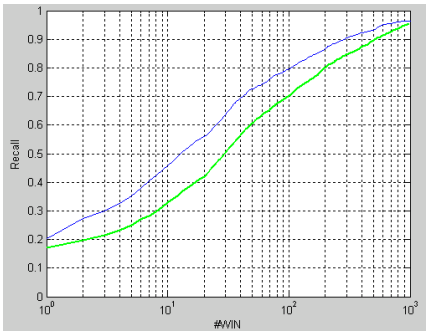

Figure 1. Pedestrian Detection Recall

As can be seen from Figure 2, the DR increases as the #WIN. The blue line is the all objects DR, and the other line denotes the pedestrian recall line. The pattern learned from the dataset is shown as Figure 3.

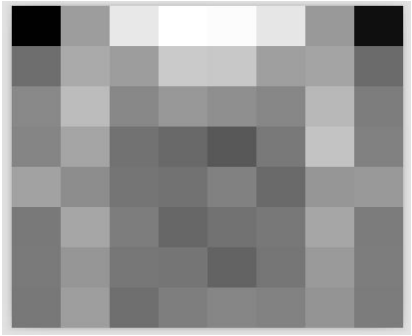

Figure 1. Pedestrian NG Feature Pattern

4.2. Experiment Setup

We investigate 13 state-of-the-art detectors in this experiment on PASCAL VOC 2007.The methods includes VJ [15], Shapelet [24], Poselet [27], LatSVM-V1(and V2) [29], HIKSVM [19], HOG [4], MultiFtr [19], HogLbp [18], Pls [19], FPDW [17], ChnFtrs [27] and ours. The miss rate is adopted to evaluate each method's performance. The comparison is illustrated in Figure 4.

Copyright ⓒ 2015 SERSC

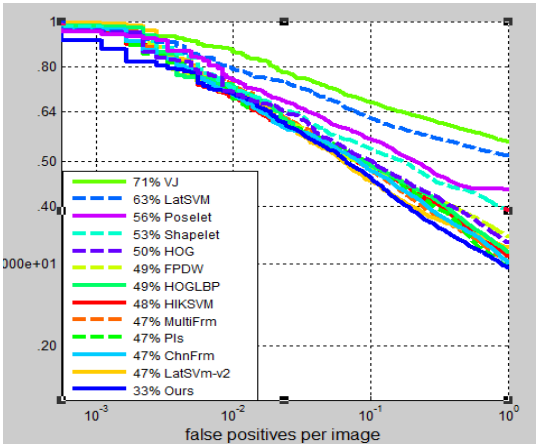

Figure 1. Performance Comparison

As can be seen from the above graph, our method achieves the best performance at the cost processing each image 3 times. Thanks to the fast computation time in the first step, we can do the pedestrian detection efficiently at the detection stage. Some of the result image is shown in Figure 5.

Figure 1. Pedestrian Detection Results

In Figure 5, the images' background is complex. The distant and low resolution pedestrians are detected precisely. The occlusion, pose variation, and distance small object detection problems are also achieved good results.

Copyright ⓒ 2015 SERSC

5. Conclusion

In this paper, we propose a new framework for efficient pedestrian detection aided by Bing feature based objective-ness measure and proposal generation. DPM is used to learn the multi-pedestrian detector which effectively captures the unique visual patterns appearing in multiple nearby pedestrians. Detection performance is improved by fast objective-ness proposal generation. We incorporate state-of-the-art pedestrian detection method to filter the proposals and in the end get the real pedestrian locations.

Extensive experimental evaluation shows that the proposed framework can significantly improve all the state-of-the-art pedestrian detection approaches, and that the detection speed is much faster than all of the other methods. 13 methods are evaluated on the PASCAL VOC 2007 dataset. The lowest miss rate is reduced from 43% to 33% on the dataset.

Our future work is to explore the spatial-temporal information and extend the proposed models to general object detection task.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No.61403004 and the Beijing Higher Education Young Elite Teacher Project under Grant No.YETP1421.

References

[1] A. Bar-Hillel, D. Levi, E. Krupka and C. Goldberg, "Part-based feature synthesis for human detection", European Conference on Computer Vision, **(2010)**, Crete, Greece.
[2] C. Beleznai and H. Bischof, "Fast human detection in crowded scenes by contour integration and local shape estimation", Proceedings of computer vision and pattern recognition, **(2009)**, Florida, USA.
[3] R. Benenson, M. Mathias, R. Timofte and L. Van Gool, "Pedestrian detection at 100 frames per second", Proceedings of computer vision and pattern recognition, **(2012)**, Portland, USA.
[4] S. Biswas, K. W. Bowyer and P. J. Flynn, "Multidimensional scaling for matching low-resolution face images", IEEE Trans. Pattern Anal. Mach. Intell., vol. 5, no. 21, **(2012)**.
[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", Proceedings of computer vision and pattern recognition, **(2005)**, New York, USA.
[6] Y. Ding and J. Xiao, "Contextual boost for pedestrian detection", Proceedings of computer vision and pattern recognition, **(2012)**, Portland, USA.
[7] P. Doll´ar, C. Wojek, B. Schiele and P. Perona, "Pedestrian detection: An evaluation of the state of the art", IEEE Trans. Pattern Anal. Mach. Intell., vol. 2, no. 21, **(2012)**.
[8] M. Enzweiler and D. Gavrila, "Monocular pedestrian detection: Survey and experiments", IEEE Trans. Pattern Anal. Mach. Intell., vol. 1, no. 13, **(2009)**.
[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman, "The pascal voc results, vol. 5, **(2012)**.
[10] P. Felzenszwalb, R. Girshick and D. McAllester, "Cascade object detection with deformable part models", Proceedings of computer vision and pattern recognition, **(2010)**, San Francisco, CA.
[11] D. Geronimo, A. Lopez, A. Sappa and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems", IEEE Trans. Pattern Anal. Mach. Intell., vol. 2, no. 2, **(2010)**.
[12] R. B. Girshick, P. F. Felzenszwalb and D. McAllester, "Discriminatively trained deformable part models", release 5, http://people.cs.uchicago.edu/ rbg/latentrelease5/.
[13] Z. Lin and L. Davis, "A pose-invariant descriptor for human detection and segmentation", European Conference on Computer Vision, **(2008)**, Marseille, France.
[14] D. Park, D. Ramanan and C. Fowlkes, "Multiresolution models for object detection", European Conference on Computer Vision, **(2010)**, Crete, Greece.
[15] M. Sadeghi and A. Farhadi, "Recognition using visual phrases", Proceedings of computer vision and pattern recognition, **(2011)**, Colorado, USA.
[16] S. Tang, M. Andriluka and B. Schiele, "Detection and tracking of occluded people", British Machine Vision Conference, **(2012)**, Surrey, UK.
[17] P. Viola, M. Jones and D. Snow, "Detecting pedestrians using patterns of motion and appearance", International Journal of Computer Vision, vol. 3, no. 22, **(2005)**.
[18] X. Wang, T. Han and S. Yan, "An hog-lbp human detector with partial occlusion handling", International Conference on Computer Vision, **(2010)**, Kyoto, Japan.

Copyright ⓒ 2015 SERSC

[19] C. Wojek and B. Schiele, "A performance evaluation of single and multi-feature people detection", Symposium of the German Association for Pattern Recognition, **(2008)**, Munich, Germany.
[20] J. Yan, Z. Lei, D. Yi and S. Z. Li, "Multi-pedestrian detection in crowded scenes: A global view", Proceedings of computer vision and pattern recognition, **(2012)**, Portland, USA.
[21] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities", Proceedings of computer vision and pattern recognition, **(2010)**, San Francisco, CA.
[22] B. Wu and R. Nevatia, "Detection and tracking of multiple partially occluded humans by bayesian combination of edgelet based part detectors", International Journal of Computer Vision, vol. 75, no. 2, **(2007)**.
[23] S. Tang, M. Andriluka and B. Schiele, "Detection and tracking of occluded people", British Machine Vision Conference, **(2012)**, Surrey, UK.
[24] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features", Proceedings of computer vision and pattern recognition, **(2007)**, Florida, USA.
[25] W. Ouyang, X. Zeng and X. Wang, "Modeling mutual visibility relationship in pedestrian detection", Proceedings of computer vision and pattern recognition, **(2013)**, Portland, Oregon USA.
[26] M. Moussaid, N. Perozo, S. Garnier, D. Helbing and G. Theraulaz, "The walking behaviour of pedestrian social groups and its impact on crowd dynamics", PLoS ONE, vol. 5, no. 4, **(2010)**.
[27] Z. Lin and L. Davis, "A pose-invariant descriptor for human detection and segmentation", European Conference on Computer Vision, **(2010)**, Crete, Greece.
[28] C. Lampert, M. Blaschko and T. Hofmann, "Beyond sliding windows: object localization by efficient subwindow search", Proceedings of computer vision and pattern recognition, **(2008)**, Anchorage, Alaska, USA.
[29] M. Enzweiler and D. M. Gavrila, "A multilevel mixture-of experts framework for pedestrian classification", IEEE Trans. Image Process, vol. 20, no. 10, **(2011)**.
[30] B. Alexe, T. Deselaers and V. Ferrari, "Measuring the objectness of image windows", IEEE TPAMI, vol. 34, no. 11, **(2012)**.
[31] I. Endres and D. Hoiem, "Category independent object proposals", European Conference on Computer Vision, **(2010)**, Crete, Greece.
[32] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman, "The PASCAL Visual Object Classes Challenge results, **(2007)**, http://www.pascalnetwork.org.
[33] R. P. and K. J. R. E, "Generating object segmentation proposals using global and local search", Proceedings of computer vision and pattern recognition, (**2014**), Columbus, Ohio.
[34] J. Uijlings, K. van de Sande, T. Gevers and A. Smeulders, "Selective search for object recognition", International Journal of Computer Vision, vol. 5, no. 22, **(2013)**.
[35] J. Z. Zhang, J. Warrell and P. H. Torr, "Proposal generation for object detection using cascaded ranking svms", Proceedings of computer vision and pattern recognition, **(2011)**, Colorado, USA.
[36] M. M. Cheng, "BING: Binarized Normed Gradients for Objectness Estimation at 300fps", Proceedings of computer vision and pattern recognition, (**2014**), Columbus, Ohio.

Authors

Shilin Zhang, He was born in Shandong province of China on 1980 and graduated from Chinese Academy of Sciences and received his PhD degree in computer science on 2012 in China.

He is now associated with North China University of Technology and his current research interests include image processing, pattern recognition and so on. He is a member of Chinese Association of Automation.

Xunyuan Zhang, He was born in Shandong province of China on 1989 and graduated from Qufu normal University of China and received his bachelor degree in Automation Science on 2011.

He is now pursuing his master degree in North China University of Technology, and his research interests include image processing, pattern recognition and so on.

 Copyright ⓒ 2015 SERSC