Electronics Letters 51.7 (2015): 559-560

DOI: http://dx.doi.org/10.1049/el.2014.3795

PDbm: People detection benchmark repository

A. Garcia-Martin, B. Alcedo and J. M. Martinez

Following the approach of the Change Detection Challenge, in order to facilitate the evaluation of new algorithms for people detection, we present a people detection benchmarking repository. It includes realistic sequences, people detection ground truth and an evaluation framework. It will be updated based on received feedback, and will maintain a comprehensive ranking of submitted methods for years to come.

Introduction: People detection is one of the most challenging problems in computer vision and video processing. Typical applications include video surveillance (e.g., people counting, people density estimation, anomaly detection, action recognition), smart environments (e.g., room monitoring, fall detection), and image/video retrieval (e.g., activity localization and tracking). Although subsequent processing may be different in each case, typically one has to start with the localization of objects of interest which, in all previous scenarios, are people.

To date, many people detection algorithms have been developed that perform well in some types of videos or specific and constrained scenarios. There is no single algorithm today able to deal with every typical real-world challenge (e.g., appearance variability, illumination or background variations, occlusions, etc).

Following the idea in [1] for change detection, in order to easily compare any people detection approach from the state of the art over not specific or constrained scenarios, and to share the results with the research community, we present a realistic, large-scale dataset that covers a range of challenges present in the real world and includes accurate ground truth for people detection, named People Detection benchmark (PDbm) [2].

Dataset: The chosen dataset has been extracted from the Change detection dataset 2012 [1]. It provides a realistic, camera-captured, diverse set of videos. The video sequences have been chosen in order to cover typical people detection challenges. The dataset includes traditional indoor and outdoor scenarios in computer vision applications: video surveillance, smart cities, etc. The Change detection dataset 2012 includes the following challenges: dynamic background, camera jitter, intermittent object motion, shadows and thermal signatures.

The proposed People detection challenge includes 16 selected sequences from the whole original dataset (31 sequences). We have selected all the sequences including people (currently excluding thermal cameras because detection algorithms rarely consider thermal images). Each sequence is accompanied by a newly developed accurate people detection ground-truth.

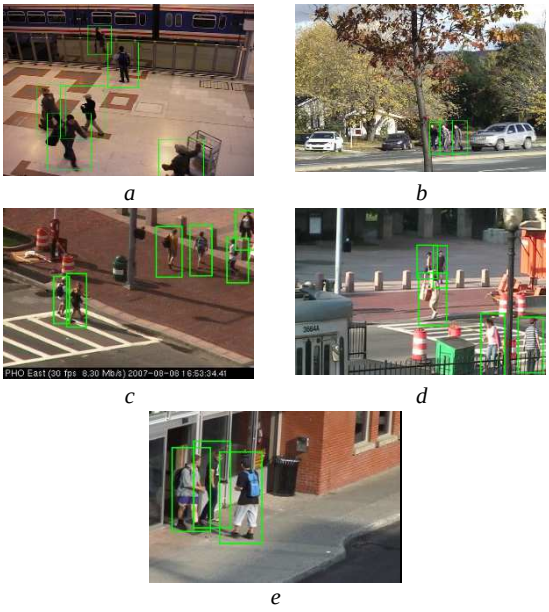
The test sequences have been classified into different complexity categories depending on two aspects [3]: the people classification and background complexity. The people classification complexity is defined as the difficulty to classify moving and temporally stationary people in a scenario. It includes three complexity levels: low, medium or high. They are related with the camera point of view, the presence of partial occlusions and pose variations. The background complexity is classified according to the already mentioned Change detection dataset 2012 challenge [1].

Table 1 includes a description of each video sequence in terms of complexity and length. Figure 1 shows sample frames of each category with the manually annotated ground truth.

Ground Truth: In our dataset, there is a great variability of people physical appearances, scales, poses, partial occlusions and camera point of views. For these reasons, it is not always clear how to determine whether a person should be annotated or not. We have decided to manually annotate every single person as a single entity (blob) that meets certain requirements: the person is fully visible, at least half of the person is visible including the head or the person is fully visible with the exception of the head. The chosen annotation tool have been the Video Image Annotation Tool (via – http://sourceforge.net/projects/via-tool/).

Table 1: *Video sequences summary according to the background and classification complexity. Length in terms of number of frames.*

	Video	Background complexity	Classification complexity	#Frames
1	office	Baseline	Low	2050
2	pedestrians	Baseline	Low	1099
3	PETS2006	Baseline	Medium	1200
4	fall	Dynamic Background	High	4000
5	overpass	Dynamic Background	Medium	3000
6	badminton	Camera Jitter	Medium	1150
7	sidewalk	Camera Jitter	High	1200
8	abandonedBox	Intermittent Object Motion	High	4500
9	sofa	Intermittent Object Motion	Low	2750
10	tramstop	Intermittent Object Motion	High	3200
11	winterdriveway	Intermittent Object Motion	High	2500
12	backdoor	Shadow	Low	2000
13	busStation	Shadow	Low	1250
14	copyMachine	Shadow	High	3400
15	cubicle	Shadow	Medium	7400
16	peopleInShade	Shadow	Low	1199
			Total	41898



a *b* *c* *d* *e*

Fig. 1 *Sample frames of each category extracted from [1] with annotated ground truth: (a) baseline "PETS2006", (b) dynamic background "fall", (c) camera jitter "sidewalk", (d) intermittent object motion "tramstop" and (e) shadow "busStation".*

Evaluation metrics: In order to evaluate different people detection approaches, we need to quantify the different performance results. Global sequence performance has usually been described in terms of Precision-Recall (PR) curves [4]. For each value of the detection confidence, Precision-Recall curves compute Precision and Recall as equations (1) and (2).

$$Precision = \frac{\text{TruePositivePeopleDetections}}{\text{TruePositivePeopleDetections}+\text{TruePositivePeopleDetections}} \quad (1)$$

$$Recall = \frac{\text{TruePositivePeopleDetections}}{\text{TruePositivePeopleDetections}+\text{FalseNegativePeopleDetections}} \quad (2)$$

Precision and Recall evaluate the detection decision or classification task. However, the people detection evaluation should also take into account the detection performance in terms of location and size of the

detected person. For this reason, we also use the three evaluation criteria defined by [5]: relative distance, cover and overlap. According to [5], a detection is considered true if $d_r \leq 0.5$ (i.e., maximum deviation up to 25% of the annotated object size) and cover and overlap are both above 50%. More than one hypothesis per object is considered as a false positive.

The integrated Average Precision (AP) is generally used to summarize the overall performance, represented geometrically as the area under the PR curve (AUC-PR).

Results: In this section, we evaluate five different people detection approaches from the state of the art. We have selected five diverse people detection approaches: HOG (Histogram of Oriented Gradients) [6], ISM (Implicit Shape Model) [5], Edge [7], DTDP (Discriminatively Trained Deformable Parts) [9] and ACF (Aggregate Channel Features) [8].

All the approaches use the default settings proposed by their respective authors: the HOG results have been obtained using the available software (http://pascal.inrialpes.fr/soft/olt/), the ISM results have been obtained using the available software (http://www.vision.ee.ethz.ch/~bleibe/index.html), the Edge results have been obtained with the original code, the DTDP results have been obtained using the available software (voc-release4, http://www.cs.berkeley.edu/~rbg/latent/) and the ACF (Inria and Caltech versions) results have been obtained using the available software and respective person models (http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html).

Firstly, we present the experimental results for each video sequence (see Table 2), and, afterwards, the results for each proposed different complexity categories: background complexity (see Table 3) and classification complexity (see Table 4). In addition, we make use of the average AUC and algorithm ranking along each evaluation.

Table 2 shows the results over each video sequence. The results show clearly how the performance is quite different between approaches and between sequences. In general, almost all the approaches present good results in some sequences and poor results in some others. For example, the best detector in average is the ACF Caltech detector. It has a 92.9% performance over sequence number 2 and a 24.7% performance over sequence number 5.

Table 3 show the average results for each background complexity category. The results show the background complexity effect over people detection performance. It is logical that the baseline sequences have the lowest complexity and the highest performance (77.7%). However, it is clear the higher complexity on those scenarios with background variations (around 30% performance).

Table 4 show the results for each classification complexity category. The results show the classification complexity effect over people detection performance. It is logical that the lowest complexity ones have the highest performance (76.8%). However, the complex ones have the lowest performance (19.9%).

Conclusion: In this work a people detection benchmarking repository has been presented. We provide an online platform to allow comparison with state of the art methods. We make use of different types of realistic videos, with accurate ground truth annotations. As future work, the ground truth will be extended with new sequences and we will maintain a comprehensive ranking of submitted methods for years to come.

Table 4: *People detection average performance in terms of AUC for each classification complexity and average ranking.*

Classification complexity	HOG	ISM	Edge	DTDP	ACF Inria	ACF Caltech	Average
Low	62.3	73.6	73.3	84.9	**86.1**	80.7	76.8
Medium	53.3	50.5	37.9	**74.4**	64.4	55.3	56.0
High	6.9	9.5	21.5	13.4	13.3	**54.5**	19.9
Average	40.8	44.5	44.2	57.6	54.6	**63.5**	50.9
Ranking	5.33	4.67	4.33	**2.00**	2.33	2.33	

Table 2: *Experimental results. People detection performance in terms of area under the Precision-Recall curve (AUC-PR) and average ranking.*

Video	HOG	ISM	Edge	DTDP	ACF Inria	ACF Caltech	Average
1	89.3	71.4	84.5	96.7	**99.3**	86.4	87.9
2	63.2	82.9	90.2	66.3	77.1	**92.9**	78.8
3	55.6	**75.7**	71.7	69.7	68.9	56.0	66.3
4	10.1	1.0	5.4	13.2	33.9	**59.1**	20.5
5	61.3	71.2	7.1	**85.1**	51.6	24.7	50.2
6	49.9	34.6	31.7	**74.9**	67.2	54.4	52.1
7	0.0	3.0	7.2	11.0	2.7	**89.4**	18.9
8	0.0	12.2	21.4	0.0	4.6	**33.4**	11.9
9	47.4	65.9	74.4	**90.4**	72.2	76.1	71.1
10	7.2	5.5	14.3	0.7	8.7	**59.5**	16.0
11	10.7	5.9	33.7	11.4	8.6	**34.8**	17.5
12	82.0	76.2	70.5	92.4	91.1	**92.6**	17.5
13	70.9	73.6	59.3	80.2	**87.2**	81.6	84.1
14	13.6	29.2	46.9	44.1	21.3	**51.1**	75.4
15	46.5	20.5	41.2	67.9	70.0	**86.2**	34.3
16	21.1	71.5	60.9	83.3	**89.6**	54.5	55.4
Average	39.3	43.8	45.0	55.4	53.4	**64.5**	50.2
Ranking	4.81	4.25	3.81	2.81	3.13	**2.19**	

Table 3: *People detection average performance in terms of AUC for each background complexity and average ranking.*

Background complexity	HOG	ISM	Edge	DTDP	ACF Inria	ACF Caltech	Average
Baseline	69.4	76.7	**82.1**	77.6	81.8	78.4	77.7
Dynamic Background	35.7	36.1	6.3	**49.2**	42.8	41.9	33.3
Camera Jitter	25.0	18.8	19.4	42.9	34.9	**71.9**	35.5
Intermittent Object Motion	16.3	22.4	35.9	25.6	23.5	**51.0**	29.1
Shadow	46.8	54.2	55.8	**73.6**	71.8	73.2	62.6
Average	38.6	41.6	39.9	53.8	51.0	**63.3**	48.0
Ranking	5.40	5.00	3.60	2.20	2.80	**2.00**	

Acknowledgments: The Spanish Government has partially supported this work ("TEC2011-25995 EventVideo").

A. García-Martín, B. Alcedo and J.M. Martínez (Video Processing and Understanding Lab (VPU), Escuela Politécnica Superior (EPS), Universidad Autónoma de Madrid (UAM), Ciudad Universitaria de Cantoblanco, 28049 Madrid, Spain). E-mail: alvaro.garcia@uam.es

References

1.N. Goyette, P. Jodoin, F Porikli, J. Konrad, P. Ishwar, "Changedetection.net: A new change detection benchmark dataset," in Proc. Of CVPRW, 2012, pp. 1-8.

2.A. Garcia-Martin, People Detection Benchmark, http://www-vpu.eps.uam.es/PDbm/, last accessed October 2014.

3.A. Garcia-Martin, J. M. Martinez, J. Bescós, A corpus for benchmarking of people detection algorithms, Pattern Recognition Letters, January 2012,Vol. 33(2), pp. 152-156.

4.B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," IJCV, vol. 77(1-3), pp. 259–289, 2008.

5.B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In Proc. of CVPR, 2005, pp. 878-885.

6.N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Proc. of CVPR, 2005, pp. 886-893.

7.A. Garcia-Martin and J. M. Martinez, "Robust real time moving people detection in surveillance scenarios," in Proc. of AVSS, 2010, pp. 241–247.

8.P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. PAMI, September 2010, Vol. 32(9), pp. 1627-1645.

9.P. Dollar, R. Appel, and W. Kienzle, "Crosstalk cascades for framerate pedestrian detection," in Proc. of ECCV, no. 645-659.