Pattern Recognition Letters 33.2 (2012): 152–156

DOI: http://dx.doi.org/10.1016/j.patrec.2011.09.038

A corpus for benchmarking of people detection algorithms

Álvaro García-Martín, José M. Martínez, Jesús Bescós

Video Processing and Understanding Lab - Escuela Politécnica Superior

Universidad Autónoma de Madrid, E-28049 Madrid, Spain

e-mail: {alvaro.garcia, josem.martinez, j.bescos}@uam.es

Abstract

This paper describes a corpus, dataset and associated ground-truth, for the evaluation of people detection algorithms in surveillance video scenarios, along with the design procedure followed to generate it. Sequences from scenes with different levels of complexity have been manually annotated. Each person present at a scene has been labeled frame by frame, in order to automatically obtain a people detection ground-truth for each sequence. Sequences have been classified into different complexity categories depending on critical factors that typically affect the behavior of detection algorithms. The resulting corpus, which exceeds other public pedestrian datasets in the amount of video sequences and its complexity variability, is freely available for benchmarking and research purposes under a license agreement.

Keywords: People detection, ground-truth, corpus, dataset, surveillance video.

1. Introduction

Due to the rise in popularity of video surveillance systems over the last years, people detection has gradually experienced a great development. The ability to detect people in video is the key to a number of multiple applications, not only in video surveillance, but also in different areas like robotics, video games, intelligent vehicles, etc. In parallel, interest on reliable strategies to assess the quality of people detection has also grown. Nowadays there are several public datasets that try to evaluate the performance of people detection algorithms. These datasets are necessary to fairly evaluate algorithms under different conditions and to compare new algorithms with existing ones.

Most reported people detection datasets are just based on sets of images [1, 2, 3, 4]. There are also many video datasets in the video security domain, but most of them do not include ground-truth annotations for people detection [5]; they just include annotations for action recognition [6, 7, 8]. A majority of the datasets including ground-truth annotations for people detection are designed only for specific surveillance applications: driver assistance systems [4, 9, 10], people detection walking through a busy pedestrian zone [11], very specific scenarios [12] or even very general video security systems [13].

Based on our previous experience in the field of people detection in video sequences [14, 15], we here describe a set of videos and annotations designed specifically for the people detection task. We have analyzed the critical factors that influence the detection and generated a corpus in which they are specifically considered. Table 1 provides a detailed comparison of existing public people detection datasets.

As opposed to people detection datasets based on images, the availability of a sequence of images inherent to a video dataset allows to consider motion information and to evaluate tracking algorithms. Additionally, according to the study and

Name	Content	Numbers		Ground Truth	Complexity[2]
		Images[1]	Videos		
MIT[1]	Color images	924-pos	-	Cut-outs images	Low: F/B views
INRIA[2]	Color images	902-pos 1671-neg	-	Bounding box (PASCAL format[16])	Low: F/S/B views
DCI[3]	Gray-scale images	24000-pos 25000-neg	-	Cut-outs images	Low: F/S/B views
TUD-Brussels[4]	Color pair-images	508-pos	-	Bounding box (non-standard format)	Medium: F/S/B views, occlusions and multiple scales
TUD-MotionPairs[4]	Color pair-images	1310-pos	-	Bounding box (non-standard format)	Medium: F/S/B views, occlusions and multiple scales
TUD-Pedestrians[12]	Color images/videos	860-pos	2 videos (272 frames)	Bounding box (non-standard format)	Medium: F/S/B views and occlusions
DCII[9]	Color images/videos	15560-pos 6744-neg	1 video (21791 frames)	Bounding box and 3D localization (non-standard format)	High: F/S/B views, occlusions, multiple scales and non-static camera
Caltech[10]	Color videos	-	1 video (250000 frames)	Bounding box (vvb file format)	High: F/S/B views, occlusions, multiple scales and non-static camera
ETH[11]	Stereo-Color videos	-	4 videos (2293 frames)	Bounding box (non-standard format)	High: F/S/B views, occlusions, multiple scales and non-static camera
VPU-Lab(Ours)	Color videos	-	90 videos (28358 frames)	Bounding box (Viper xml format[17])	Low,Medium and High: F/S/B views, occlusions, multiple scales, interactions, backgrounds and static/non-static camera

Table 1: Public people detection datasets. [1]Number of positive (pos) and negative (neg) examples. [2]Views: front (F), side (S) and back (B).

identification of critical factors affecting people detection techniques, we have designed a dataset that includes different background and people classification complexity levels (low, medium and high). The described dataset mainly excels in the amount of sequences (90 videos) and variability of sequences. It includes a great variability of scenarios: outdoor/indoor surveillance scenes with different background complexities (textural, lighting changes, multimodal, etc) and it also includes a great variability of people appearance and interactions: scenes with one or multiple persons, pose changes, scale variations, people wearing different clothes, people carrying different objects and with multiple interactions with objects and/or persons.

The structure of this paper is as follows. Section 2 presents a number of design considerations necessary to achieve a representative set of video-sequences from a people detection point of view. The sequences definition and annotation procedure are discussed in Sections 3 and 4, whilst some examples are provided in Section 5. Finally, in Section 6, the main conclusions are summarized.

2. Ground-truth design: critical factors in people detection

In order to obtain meaningful evaluation results, a corpus should include a set of representative video sequences, ranging from low to high complexity situations. The term "complexity" will be used hereinafter to express the degree of difficulty for a particular people detection algorithm to yield accurate results.

The people detection task [18, 19] consists mostly of, firstly, the design and training of a person model based on characteristic parameters (motion [20], dimensions [21], silhouette [22], etc.); and, secondly, the adjustment of this model to the candidate objects in the scene. All candidates that adjust to the model will be detected/classified as person, whilst all the others will not. Therefore, people detection can be split up into the localization of initial object candidates in the scene (detection) and their subsequent classification (verification). Starting from these ideas, global sequence complexity has been found to be strongly dependent on a series of specific properties of objects [10], on background complexity [23]

and on some relationships among these elements [24]. These dependencies have been designated as "critical factors", emphasizing their influence on the algorithms' results. Since specific settings for these factors can significantly increase (low complexity settings) or decrease (high complexity settings) detection accuracy, they seem a convenient mechanism to regulate sequence complexity.

Table 2 summarizes the critical factors concerning foreground and background that we have considered. We next describe them including a brief discussion of their influence on the overall sequence complexity.

Background						**Classification**					
Textural complexity			Variability			Appearance variability			People/Object interactions		
Not textured	Slightly textured	Textured	Lighting changes	View changes	Multimodal	Pose variations	Different clothes	Carry objects	Objects	People	Objects & People

Table 2: Critical factors in people detection.

2.1. Background complexity

We here define background complexity as the difficulty to detect in the scene the initial objects candidate to be person, due to the presence of edges, multiple textures, lighting changes, reflections, shadows and any kind of background variation. The following critical factors have been identified:

Textural complexity. Scenarios including an important amount of textured areas can make highly difficult the localization of initial object candidates. In fact, depending on the algorithm used, highly textured background areas can be easily wrongly detected as objects. Consequently, low textured background areas correspond to lower complexity situations and vice versa.

Variability. This refers to the property of some backgrounds to undergo variations usually produced by external factors (light and point of view changes) or multimodal backgrounds (such as twinkling water, swaying trees or glowing flames). Static scenarios with less variations correspond with low complexity levels, while scenarios with multiple variations correspond with more challenging situations.

2.2. People classification complexity

We here define it as the difficulty to verify the object candidates to be person in the scene. It is related to the number of objects, their velocity, partial occlusions, pose variations and interactions between different people and/or objects. We have grouped these elements into two fundamental critical factors:

Appearance variability. People appearance exhibits very high variability since they are non-rigid objects, they can change pose, they can also wear different clothes and carry different objects, they have a considerable range of sizes and shapes mainly due to the point of view and the relative situation with the camera. People with limited appearance variability (no pose changes, no sizes variations, etc) entail low complexity levels, while the cases with high appearance variability entail a more complex classification.

People/Object interactions. People must be identified in real-life scenarios, that is they must be detected in the context of the environment surrounding them. People present interactions with objects and/or with other people. These interactions make more difficult their identification and classification. In order to identify all persons involved in these situations, it is necessary to deal with occlusions. Occlusions resulting from objects, other persons or visibility of the camera limits the visible appearance of the person occluded.

3. Description of the ground-truth

In the previous Section, high, medium and low complexity settings for every critical factor have been identified. They have all been considered in the ground-truth design, thus making the resulting set of sequences specially useful to identify weak-points of a specific algorithm. We have grouped all the test sequences into different complexity categories depending on these critical factors. A description of complexity levels for the associated content is shown in Table 3, whilst

Figure 1 shows two sequence examples of each category. The videos have been collected from several public datasets related with the people detection/object classification task [5, 23], AVSS 2007 dataset (available at [6]), PETS 2006 dataset (available at [7]) and TRECVID 2008 dataset (available at [8]).

Overall, sequences include both non-rigid (people, clothes, ropes...) and rigid objects (boxes, rucksacks, toys...) differing in size, motion (slow and fast displacements, rotations, chaotic motion) and textural appearance. These objects are involved in a number of interactions (intersecting and not intersecting trajectories, merging and splitting, partial and complete occlusions...), and in different contexts, like typical every-day situations (runners taking over each other, object being thrown, people dancing,...). or surveillance video scenarios (office scenarios, subway platform,...). Regarding the backgrounds, sequences include indoor and out-door scenarios. Additionally, different background complexities were considered by controlling the influence of homogeneous areas, external factors variations and multimodal motion.

4. Sequences annotation

In addition to video frames, a description of the detected people (frame number and bounding box) are also required in order to have the corpus ground-truth. Therefore we have manually annotated 90 sequences (see Table 3). To carry out the annotation task, we have used the Viper tool [17] that outputs XML files with the description (frame by frame people location, width and height). We have decided to use the Viper tool because it is one of the most popular in the research community, it is easy to manage, it has associated performance evaluation tools and offers a variety of metrics for performing comparison between video metadata files.

Performance can be evaluated at two levels: sequence sub-unit (frame, window, etc) or global sequence. Sub-unit performance is usually measured in terms of sensitivity and specificity, and it is usually visualized in terms of ROC curves [2, 9, 10]. Global performance is also measured in terms of sensitivity and specificity (ROC) [1, 3], but it can be alternatively measured in terms of recall and precision [4, 12]. The first level gives us information of the classification stage, while the second one provides overall system performance information. In order to evaluate a video surveillance system, it is more interesting to compare the overall performance.

In complex environments with multiple people and partial occlusions, it is often not obvious where to draw the line and decide whether a person should be annotated or not. In our set of sequences, people "occur" in every state of occlusion, from fully visible to just one single body part visible. We therefore decided to annotate all those cases where a human could clearly detect the person, without human reasoning. As a consequence, all people were annotated as a single entity (blob) covering the visible part of them whenever at least the head or most of the torso is visible.

5. Examples

Figure 1 shows some example frames from several sequences of the corpus including annotated blobs, just to offer an idea of sequences appearance and their corresponding annotations. The complete set of sequences along with their description, associated category, and the annotation ground-truth can be downloaded from http://www-vpu.eps.uam.es/PDds/ (Figure 2 shows a screenshot of the web site). It is freely available for research purposes (after completing a license agreement form).

6. Conclusions

This paper compiles the motivations and considerations applied to the generation of a corpus (dataset and associated ground-truth) for the evaluation of people detection algorithms in video sequences. Both the wide range of considered critical factors and the development of an accurate ground-truth for the presented corpus, makes it especially suitable for algorithm's tuning, results

Figure 1: Sequence examples. Every example shows three random frames from a sequence.

Sequence	Category	Subcategory	Background		Classification	
			Textural complexity	Variability	Appearance variability	People/Object interactions
1-4	C1	C1-a	Low	Low	Low	Low
5-6	C1	C1-b	Low	Medium	Low	Low
7-8	C2	C2-a	Low	Low	Medium	Low
9-10	C2	C2-b	Low	Low	Medium	Medium
11-12	C2	C2-c	Low	Medium	Low	Medium
13	C3	C3-a	Medium	Medium	Medium	Low
14-16	C3	C3-b	Medium	Medium	Medium	Medium
17-18	C4	C4-a	Low	Low	Medium	High
19-20	C4	C4-b	Low	Low	High	Medium
21	C4	C4-c	Low	Low	High	High
22-24	C5	C5-a	Medium	High	Medium	High
25	C5	C5-b	Medium	High	High	Medium
26	C5	C5-c	High	High	Medium	High
27-33	C5	C5-d	High	High	High	Low
34-65	C5	C5-e	High	High	High	Medium
66-90	C5	C5-f	High	High	High	High

Table 3: Critical factors on video corpus.

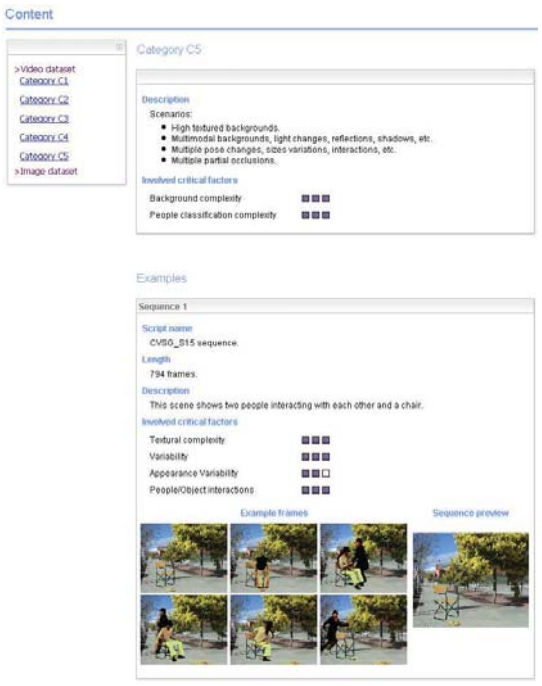

Figure 2: Screenshot of the public web.

evaluation and comparison. A more complete people detection corpus in surveillance scenarios than the ones available in the state of the art has been developed, providing a common framework for the evaluation of people detection algorithms under different complexity conditions.

In the future, we will try to extend the contents of the dataset and make use of sequences recorded in a chroma studio and composed with different backgrounds [23], in order to analyze independently the background and foreground factors.

7. Acknowledgments

This work has been partially supported by the Cátedra UAM-Infoglobal ("Nuevas tecnologías de vídeo aplicadas a sistemas de video-seguridad") and by the Universidad Autónoma de Madrid ("FPI-UAM: Programa propio de ayudas para la Formación de Personal Investigador").

8. References

References

[1] C. Papageorgiou, T. Poggio, A trainable system for object detection, International Journal of Computer Vision 38(1) (2000) 15–33.

[2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. of CVPR, Vol. 1, 2005, pp. 886–893.

[3] S. Munder, D. M. Gavrila, An experimental study on pedestrian classification, IEEE Trans. on Pattern Analysis and Machine Intelligence 28(11) (2006) 1863–1868.

[4] C. Wojek, S. Walk, B. Schiele, Multi-cue onboard pedestrian detection, in: Proc. of CVPR, 2009.

[5] R. Vezzani, R. Cucchiara, Annotation collection and online performance evaluation for video surveillance: The visor project, in: Proc. of AVSS, 2008, pp. 227–234.

[6] International conference on advanced video and signal based surveillance, in: http://www.avss2007.org, 2007.

[7] International workshop on performance evaluation of tracking and surveillance, in: http://www.pets2006.net/, 2006.

[8] Trecvid 2008 evaluation for surveillance event detection, in: http://www-nlpir.nist.gov/projects/tv2008/tv2008.html, 2008.

[9] M. Enzweiler, D. M. Gavrila, Monocular pedestrian detection: Survey and experiments, IEEE Trans. on Pattern Analysis and Machine Intelligence 31(12) (2009) 2179–2195.

[10] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: A benchmark, in: Proc. of CVPR, 2009.

[11] A. Ess, B. Leibe, L. V. Gool, Depth and appearance for mobile scene analysis, in: Proc. of ICCV, 2007.

[12] M. Andriluka, S. Roth, B. Schiele, People-tracking-by-detection and people-detection-by-tracking, in: Proc. of CVPR, 2008, pp. 1–8.

[13] A. T. Nghiem, F. Bremond, M. Thonnat, V. Valentin, Etiseo, performance evaluation for video surveillance systems, in: Proc. of AVSS, 2007.

[14] A. Garcia-Martin, J. Martinez, Robust real time moving people detection in surveillance scenarios, in: Proc. of AVSS, 2010, pp. 241–247.

[15] A. Garcia-Martin, A. Hauptmann, J. Martinez, People detection based on appearance and motion models, in: Proc. of AVSS, 2011.

[16] The pascal visual object classes challenge 2005 development kit, in: http://pascallin.ecs.soton.ac.uk, 2005.

[17] Viper-gt, the ground truth authoring tool, http://vipertoolkit.sourceforge.net/docs/gt/.

[18] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews 34(3) (2004) 334–352.

[19] M. Valera, S. A. Velastin, Intelligent distributed surveillance systems: a review, IEE Proc. on Visual Image Signal Processing 152(2) (2005) 192–204. doi:10.1049/ip-vis:20041147.

[20] R. Cutler, L. S. Davis, Robust real-time periodic motion detection, analysis, and applications, IEEE Trans. on Pattern Analysis and Machine Intelligence 22(8) (2000) 781–796.

[21] P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, N. Papanikolopoulos, Estimating pedestrian counts in groups, Computer Vision and Image Understarding 110(1) (2008) 43–59.

[22] F. Xu, K. Fujimura, Human detection using depth and gray images, in: Proc. of AVSS, 2003, pp. 115–121.

[23] F. Tiburzi, M. Escudero, J. Bescos, J. M. Martinez, A ground truth for motion-based video-object segmentation, in: Proc. of ICIP, 2008, pp. 17–20.

[24] B. Wu, R. Nevatia, Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors, in: Proc. of ICCV, 2005, pp. 90–97.