A Multi-camera video data set for research on High-Definition surveillance

Athira Nambiar, Matteo Taiana, Dario Figueira, Jacinto Nascimento and Alexandre Bernardino

Computer and Robot Vision Lab, Institute for Systems and Robotics
Instituto Superior Técnico
Lisbon, Portugal

Abstract: We present a fully labelled image sequence data set for benchmarking video surveillance algorithms. The data set was acquired from 13 indoor cameras distributed over three floors of one building, recording simultaneously for 30 minutes. The data set was specially designed and labelled to tackle the person detection and re-identification problems. Around 80 persons participated in the data collection, most of them appearing in more than one camera. The data set is heterogeneous: there are three distinct types of cameras (standard, high and very high resolution), different view types (corridors, doors, open spaces) and different frame rates. This diversity is essential for a proper assessment of the robustness of video analytics algorithms in different imaging conditions. We illustrate the application of pedestrian detection and re-identification algorithms to the given data set, pointing out important criteria for benchmarking and the impact of high-resolution imagery on the performance of the algorithms.

1 Introduction

Video surveillance algorithms have to deal with a great diversity of operating conditions due both to the acquisition device characteristics (optics, electronics, compression) and to the observed scenario (illumination, clutter, perspective). Since it is difficult to model these conditions analytically, most algorithms adopt a data driven learning approach, i.e., they tune and test performance on data sets reflecting the intended operating conditions. From the developer point of view it is of great importance to have access to good benchmarking tools in order to compare different algorithms and perform the necessary modifications to improve their performance. From the end user point of view it is also fundamental to be able to choose, among the existing alternatives, the algorithms with the best performance for a particular scenario. In both perspectives, the availability of rich and diverse data sets is a key for the advance of the state of the art on video surveillance technology. This has been acknowledged along the years in a series of works and several benchmark data sets have been published. Very often the data sets are designed with some end application in mind, for instance Pedestrian Detection (PD), re-identification (RE-ID), activity recognition, etc. Depending on the application, different types of labelling and annotations are necessary.

Works on PD require bounding boxes localising each person in the images. In RE-ID data sets such information is combined with a unique label that identifies each person. In activity recognition, time windows in the video sequence are annotated with the type of activity being performed.

A distinguishing feature of our data set with respect to the state of the art is the availability of several high resolution videos with labels indicating both person location and identity in the images. This makes this data set well suited for tackling the whole re-identification problem, including person detection and matching/tracking across cameras.

About 80 persons were recorded by 13 cameras during 30 minutes inside the three floors of the three floors of the building of the Institute for Systems and Robotics in Lisbon (ISR-Lisboa). More than 64,000 annotations were performed on a total of more than 75000 frames. Annotations include person localisation bounding boxes, person identification labels, occlusion and crowd information tags. This acquisition and labelling effort was performed within the HDA (High-Definition Analytics) project[a], dedicated to the study of the impact of high resolution imagery in the performance of video analytics algorithms. As such, several of the acquired image sequences are in the high resolution range (1280×800 and 2560×1600), which makes the presented data set the first one to include labeled video sequences of such resolution. The data set and scripts utilised during labelling can be downloaded at http://vislab.isr.ist.utl.pt/hda-dataset.

The rest of the paper is organised as follows: the state of the art is reviewed in the following section. Our data set is described in some detail in section 3. Section 4 and 5 are devoted to the benchmarking of PD and RE-ID algorithms respectively, either in the proposed or on alternative datasets. Section 6 draws conclusions and proposes ideas for future work.

2 Related Work

Many data sets targeting visual surveillance scenarios were published over the years. Typical applications include person detection and tracking, event detection, activity recognition, inter-camera tracking and re-identification. In this work we target specially person detection (PD) and re-identification (RE-ID) applications and we review some important related data sets here. A first notable example is the MIT pedestrians data set [1], introduced in 1997, to train pedestrian detectors. It contains 924 frontal and rear views of pedestrian in single frames acquired from multiple video and photographic sources. The views are cropped to fixed-size rectangular images designed to contain a person with height of around 100 pixels. Only positive examples are provided. In 2005, the INRIA person data set [2] was introduced by Dalal and Triggs for similar purposes. It is divided in training and test sets, providing both positive and negative examples.

The CAVIAR [3] data set, which was introduced in 2004, was one of the first to provide video sequences instead of single frames. It was also the first to provide annotations of people location, identity and activity, making this data set useful for many problems. The CAVIAR 1^{st} set was acquired with a single camera at INRIA Labs Grenoble for activity recognition. CAVIAR's 2^{nd} set was acquired in a shopping mall in Lisbon using two cameras with overlapping fields of view, making it more interesting for RE-ID and multi-camera tracking. Some sequences were customised for the RE-ID problem and collected in the data

[a] http://www.hd-analytics.net/

set CAVIAR4REID [4] later in the year of 2011, which contains a total of 72 individuals: 50 appearing in two camera views and 22 appearing in just one.

The ETH pedestrians data set [5] was introduced in 2007. It was recorded with a mobile platform moving along a sidewalk, equipped with a stereo camera. It presents a scenario typical for a mobile robot. The TUD-MotionPairs/TUD-Brussels data set [6] and the Caltech pedestrian data set [7] were introduced in 2009 and contain sequences of images taken in automotive scenarios. Another data set also introduced in 2009 was PETS [8]. It was recorded in public space outdoor scene at University of Reading, UK. The data set addressing person tracking consists of 3 subclasses based on their subjective difficulty level/ density of the crowd. 3 sequences with 8256 frames make up this dataset.

Recently in 2011, [9] compiled a collection of 90 videos from publicly available data sets into the PDds data set. Such videos were labeled uniformly for PD and object classification tasks. Overall the collection contains 28358 frames, divided in 16 different subclasses depending on the complexity of background texture, as well as on people appearance variability and people/object interactions. This is probably one of the most complete data sets for PD but, it lacks scenarios in which the same persons are imaged from different cameras. This hinders its usefulness for the RE-ID and multi-camera tracking community. Besides CAVIAR4REID, other data sets have been designed specially for the RE-ID problem. i - LIDS (2009) for re-identification [10] contains appearances of 119 people and was built from iLIDS Multiple-Camera Tracking Scenario. The presence of occlusions and quite large illumination changes make the RE-ID task challenging. The VIPeR data set [11] introduced in 2007, contains two views of 632 pedestrians captured from different viewpoints. The RE-ID oriented data sets usually contain cropped images of persons instead of full frames because prior detection of their localisation is assumed. A comparison of relevant data sets according to their main characteristics is provided in Table 1.

Name	#CA	#SE	#FR	#BB	#PE	Max. Res.	Main Applications
MIT-cbcl [1]	Many	0	0	924	(NA)	64×128 (C)	PD
INRIA [2]	Many	0	902 (P)	1774 (P)	(NA)	1280×960	PD
CAVIAR 2^{nd} Set	2	26	71392	180694	72	384×288	PD, RE-ID, Tracking, Activity Recognition
TUD-MotionPairs	1	1	1092 (P)	1776	(NA)	720×576	PD
TUD-Brussels	1	1	508 (P)	1326	(NA)	640×480	PD
Caltech [7]	1	1	250000	350000	2300	640×480	PD
PDds [9]	Many	90	28358	(NA)	(NA)	720×576	PD
PETS [8]	8	3	8256	(NA)	(NA)	768×576	PD
iLIDS4REID [10]	2	0	0	476	119	304×153 (C)	RE-ID
VIPeR [11]	2	0	0	1264	632	128×48 (C)	RE-ID
CAVIAR4REID [10]	2	0	0	1220	72	384×288	RE-ID
Ours	13	13	75207	64028	85	2560×1600	PD, RE-ID, Tracking

Table 1: Main characteristics of the surveyed data sets. We compare the number of cameras in the data set (#CA), the number of video sequences (#SE), the number of video frames (#FR), the number of person bounding box labels (#BB), the number of person identity labels (#PE), the maximum video resolution available, and the main application envisaged for the data set. Data sets whose number of video sequences is 0 are composed of independent photographies. Data sets providing cropped images instead of full frames are indicated with 0 in the number of frames. In these cases the maximum resolution refers to the size of the cropped images and is followed by symbol (C). (P) indicates the number of positive examples of the training set. (NA) means data not available.

3 The HD Data Set

CAM	02	17	18	19	40	50	53	54	55	56	57	58	59	60
640x480	✓	✓	✓	✓	✓									
1280x800						✓	✓	✓	✓	✓	✓	✓	✓	
2560x1600														✓
fps	5	5	5	5	5	2	2	2	2	2	2	2	2	1
floor	6	8	8	8	8	7	7	7	7	7	7	7	8	7

Table 2: Details of the labelled camera network.

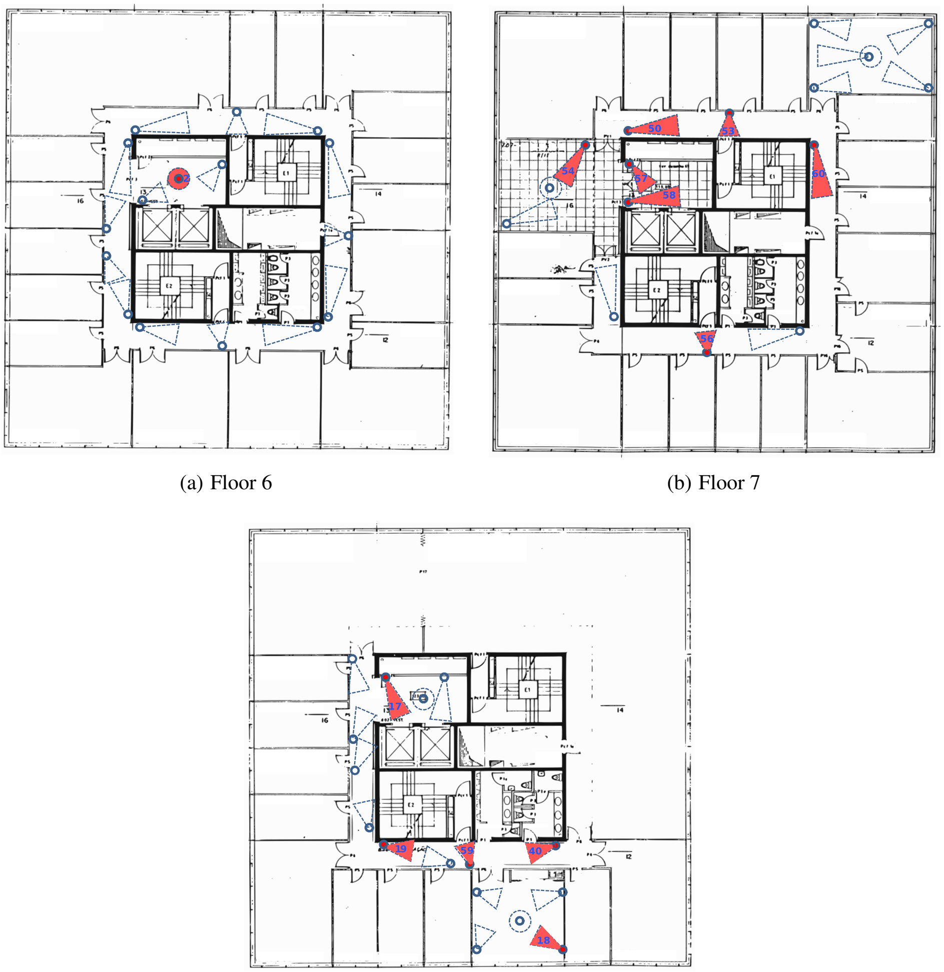
In the context of the HDA project, several video sequences including high and very high resolution video were acquired at the HD camera network (HD-CAMNET). The HD-CAMNET is an IP network with power-over-ethernet cameras installed in the 3 floors of the ISR-Lisboa facilities (floor 6, 7 and 8), whose purpose is to support research in distributed sensor networks and video surveillance. It is composed of 48 access points where IP cameras can be connected and their data acquired simultaneously. The data recordings made for the HDA project involved 13 AXIS cameras, some with standard VGA resolution (AXIS 211, AXIS 212PTZ, and AXIS 215PTZ) some with 1MPixel resolution (AXIS P1344) and one of 4MPixel resolution (AXIS P1347). To save bandwidth, storage and labelling time, the sequences were not acquired at high speed, but at rates of 5Hz, 2Hz and 1Hz for the VGA, the 1MPixel and the 4MPixel resolution respectively. Each camera is identified by its unique ID viz., cam02, cam17, etc. Table 2 describes the camera network details in brief. The camera poses in the three floors are depicted in Figure 1.

This HD data set includes labelled annotations of people in 13 cameras. The image sequences were acquired simultaneously over a period of half an hour in the public spaces of a typical indoor office building (corridors, leisure areas, halls, entries/exits). One frame for each sequence is shown in Figure 2. Besides different resolution, one can also notice very different illuminations, color balance, depth range and camera perspective.

We labelled all frames in the sequences of 13 cameras. Labelling consists in specifying an enclosing bounding box (BB) for each object/person in the image and assigning an identifier (ID) to it. The bounding boxes provide ground truth for Pedestrian Detection algorithms. The unique ID provides the ground truth for re-identification algorithms. And this ID plus the starting frame and ending frame of each person appearance in a video sequence provides ground truth to benchmark tracking algorithms. In the process of labelling, we made use of the following software tools: MATLAB® with the Image Processing Toolbox, Piotr Dollár's Toolbox [12] and Detection Code [13].

The labelling rules are given below:

1. The bounding box is drawn so that it completely and tightly encloses the person.
2. If the person is occluded by something (except image boundaries), the bounding box is drawn by estimating the whole body extent.
3. People partially outside the image boundaries have their BB's cropped to image limits.
4. Partially occluded people and people partially outside the image boundaries are marked as 'occluded'.
5. A unique ID is associated to each person, e.g., 'person01'. In case the identity is ambiguous even for the human eye, the special ID 'personUnk' is used.



(a) Floor 6

(b) Floor 7

(c) Floor 8

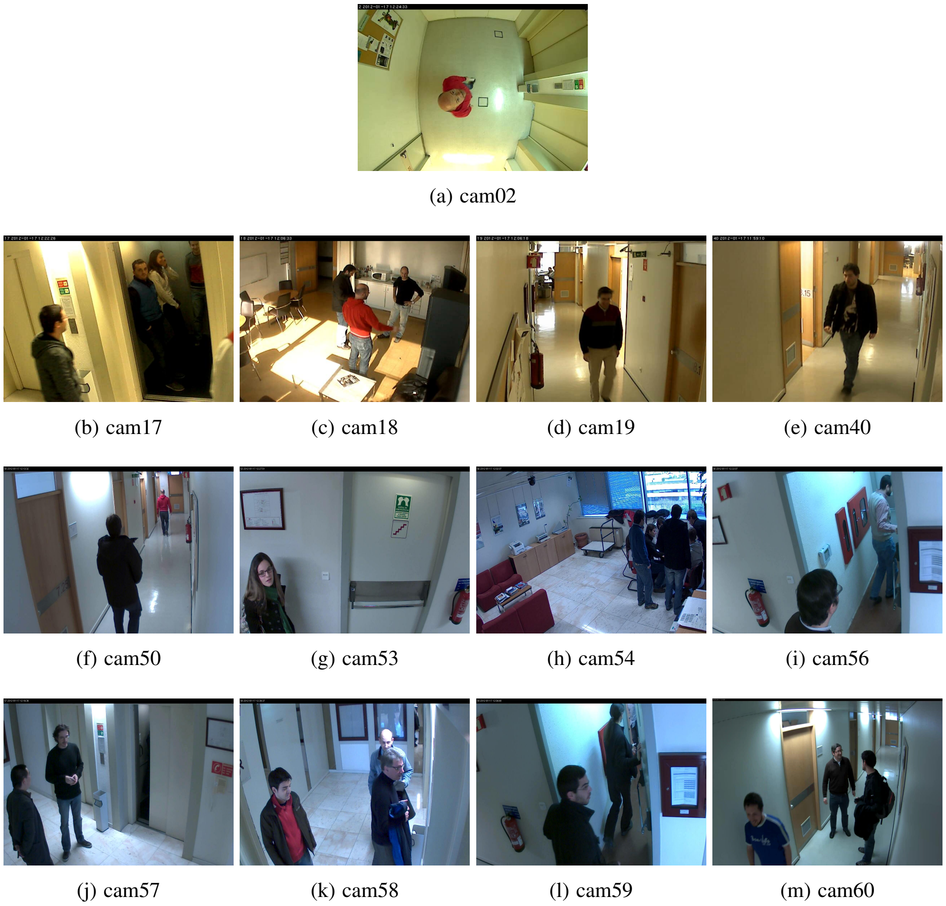
Figure 1 Camera poses: the cameras marked with a red circle and an orange field of view are the cameras used to acquire data.

6. Groups of people that are impossible to label individually are labelled collectively as ‘crowd’. People in front of a ’crowd’ area are labeled normally.

These labelling rules give us the possibility at evaluation time to ignore detections on crowded regions and failures on occluded persons, that could otherwise be ambiguously accounted as false positives and negatives. We thus provide information that can be used to prevent over-penalizing some algorithms in scenarios they were not designed to tackle.

We show labelling examples in Figure 3. The person ID is displayed on top of each BB. Moreover, the annotations contain an occlusion bit and track information – beginning and starting frame of a track and positions of each BB for each frame in between. This makes the dataset applicable as a benchmark for video-tracking algorithms as well.

The data set comprises annotations of 85 persons, of which 70 are men and 15 are women. A statistical characterization of the data is presented in table 3 and Figure 4. We



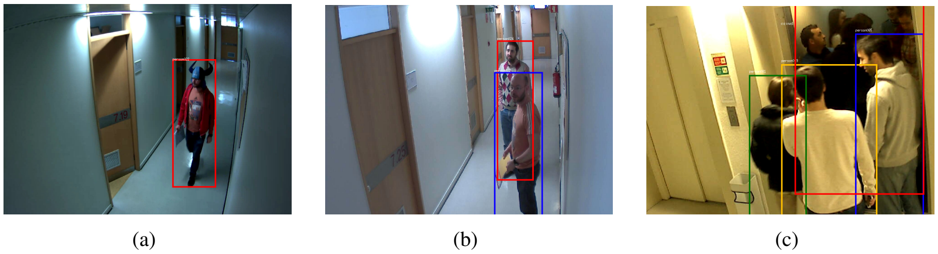
(a) cam02

(b) cam17 (c) cam18 (d) cam19 (e) cam40

(f) cam50 (g) cam53 (h) cam54 (i) cam56

(j) cam57 (k) cam58 (l) cam59 (m) cam60

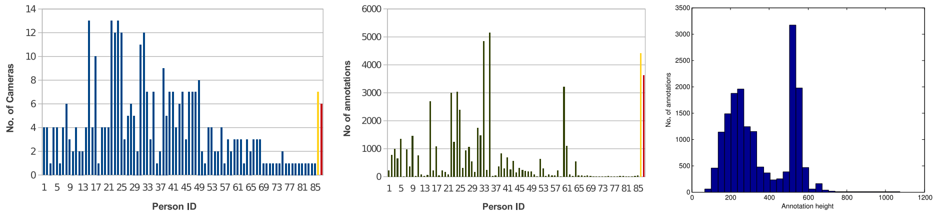
Figure 2 Snapshots of the sequences acquired in the HD-CAMNET data set.



(a) (b) (c)

Figure 3 Labelling examples. (a) An unoccluded person. (b) Two occluded people. (c) A crowd with three occluded people in front of it.

would like to highlight the large range of peoples' BB heights from 69 to 1075 pixels (see Figure 4(c)), which contributes to make this data set one of the most challenging for PD and RE-ID applications.



(a) (b) (c)

Figure 4 (a) Number of sequences each person appears in. Person 86 (yellow) and 87 (red) correspond to the labels ‘personUnk’ and ‘crowd’. (b) Number of BB’s for each person. (c) Histogram of BB height for the unoccluded people. The peaks of the VGA and the high resolution distributions are visible. The BB’s span heights between 69 and 1075 pixels.

CAM	02 LR	17 LR	18 LR	19 LR	40 LR	50 HR	53 HR	54 HR	56 HR	57 HR	58 HR	59 HR	60 VHR
Total number of video frames	9819	9897	9883	9878	9861	2227	3521	3424	3798	3780	3721	3670	1728
Number of Bounding Boxes	1832	3865	13113	18775	7855	1288	465	8703	576	3190	2291	894	1182
Min. unoccl. BB height [pix]	-	310	90	71	71	158	69	153	619	384	395	598	212
Max. unoccl. BB height [pix]	-	463	338	403	408	606	681	608	717	688	681	775	1075
Number of Persons	9	26	32	34	39	20	19	12	34	43	34	34	20

Table 3: Data on the number of frames, the number of annotations and the number of people for each sequence. The minimum and maximum height of unoccluded BB’s are also reported. Camera 02 does not have person height information due to its unconventional overhead perspective.

4 Person Detection Benchmarking

Detecting people in a surveillance context can be accomplished using two main classes of algorithms: background subtraction-based methods and the pattern recognition methods common to PD algorithms. In this work we focus on the latter, because of their applicability to a wider range of scenarios, including moving/shaking cameras and single frame images. We evaluate the performance of two PD systems on the proposed data set: the Fastest Pedestrian Detector in the West [14] (FPDW) and the Grammar Models detector [15] (GM), also known as Discriminatively Trained Deformable Part Models, release 5. These two systems are state-of-the-art representatives of two distinct paradigms for PD: monolithic (human as a whole) and part-based (human as a composition of parts), respectively. We use the code provided by the authors [16] for GM, with a model trained on the Pascal VOC 2010 data set [17], while for FPDW we use our own implementation, based on the original authors’ code trained on the INRIA person data set.

In order to assess the performance of PD systems, we need to perform a matching between the detection BB’s computed by the detection algorithm against the ground truth (GT) BB’s. The possible outcomes of a match are: True Positive (TP), False Positive (FP) and Missed Detection (MD). The matches are typically computed using the Pascal VOC criterion (see [18]), based on a measure of overlap between pairs of BB’s.

Our data set annotation is consistent with the behaviour of the evaluation algorithm for most of the labels. However, because our GT BB’s are designed to enclose the full extent of the projection of a person on one image, this can lead to BB’s that are not horizontally centered on the targets, mainly when the pose of a person's arms or legs is very asymmetrical. In such cases the matching algorithm can report a FP or a MD instead of a TP.

For evaluation we use a customized version of the code provided by [7]. Detections and missed detections on a 'crowd' area of the image are not penalized.

4.1 Performance on different types of scenarios

In order to evaluate the characteristics of our data set with respect to PD algorithms, we partition the video sequences based on characteristic views. We form the groups 'long range' (sequences of corridors, 19, 40, 50 and 60), 'mid range' (sequences of big rooms 18 and 54) and 'short range' (sequences of cameras pointed towards doors or inside small rooms 17, 53, 56, 57, 58 and 59). We leave sequence 02 out (see top image in Fig. 2), as in that case the camera is pointed down from the ceiling: the projections of people onto the image plane are so different from the typical pedestrian projections that both detectors completely fail at the recognition task.

Considering that only GM handles occlusions explicitly, we evaluate the detections using two modes: in the 'base' mode we consider all the BB's composing the GT, while in the 'full visibility' mode we only consider the BB's that are completely visible. Moreover, GM estimates a BB enclosing the full person even when it observes only a part of it. This can lead to detection BB's with parts well outside the image. Such bounding boxes would not match the GT of people only partially inside the image, as in the GT the BB's are bounded to the image limits. We thus crop each detection so that it lies inside the image.

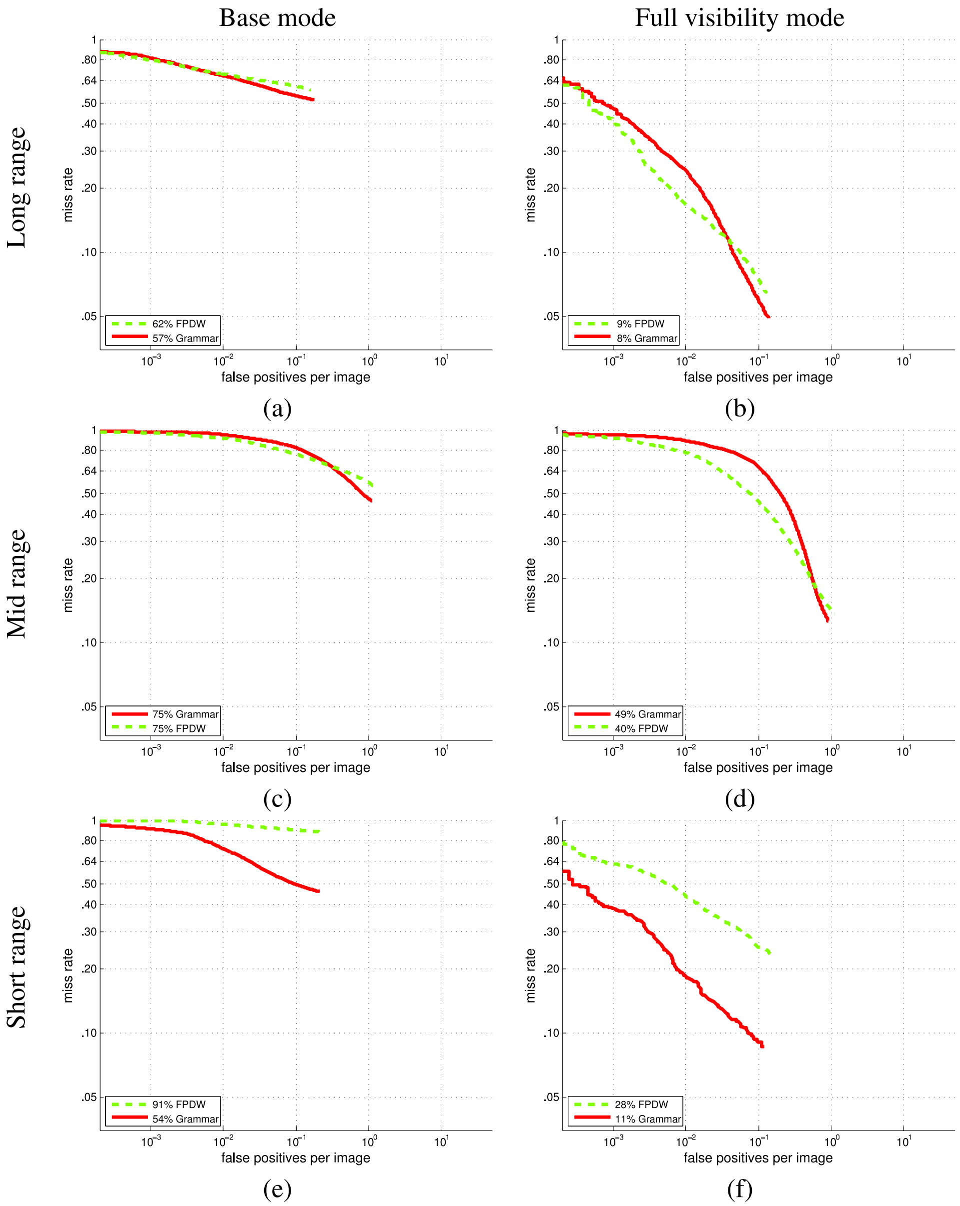
The performances of FPDW and GM are quite similar for the long and the mid range sequences, with a little advantage for FPDW in the fully visible mode. For the short range setup instead, GM is the clear winner. We speculate that the advantage of GM in the short range sequences depends on its part-based structure, which can accommodate the perspective deformations between training and test data better than the monolithic structure of FPDW. We present the results in the form of miss rate/False Positives Per Image (FPPI) plots in figure 5. The overall performance indicator shown is the the log-average miss rate, the average miss rate computed between 10^{-2} and 10^{0} FPPI (see [7]).

4.2 Performance on Different Image Resolution

In the second experiment we assess the impact of image resolution on the PD task. The imaged height of a pedestrian has a strong correlation with the difficulty of detection: smaller pedestrians are more difficult to detect. Most state-of-the-art methods struggle with heights under 80 pixels and even humans start having difficulties for heights under 30 pixels (see [7]). This grants a clear advantage to detection performed on higher resolution images: some people with imaged height under the 30 pixels threshold with a VGA camera would have height above that threshold if imaged with a higher resolution camera.

We downsample the images of the high and medium resolution sequences to a resolution of 640×400, using bilinear interpolation, to compare the performance of the detectors on the two versions. We set the detectors up so that they are able to detect pedestrians at all of the imaged heights that appear in the data set, both at the original and at the downsampled resolution.

Two effects contribute to change the performance of a detector when it is run on the lower resolution version of an image. First, the detector generates more Missed Detections due to the aforementioned phenomenon. Second, the "sliding window" paradigm of the

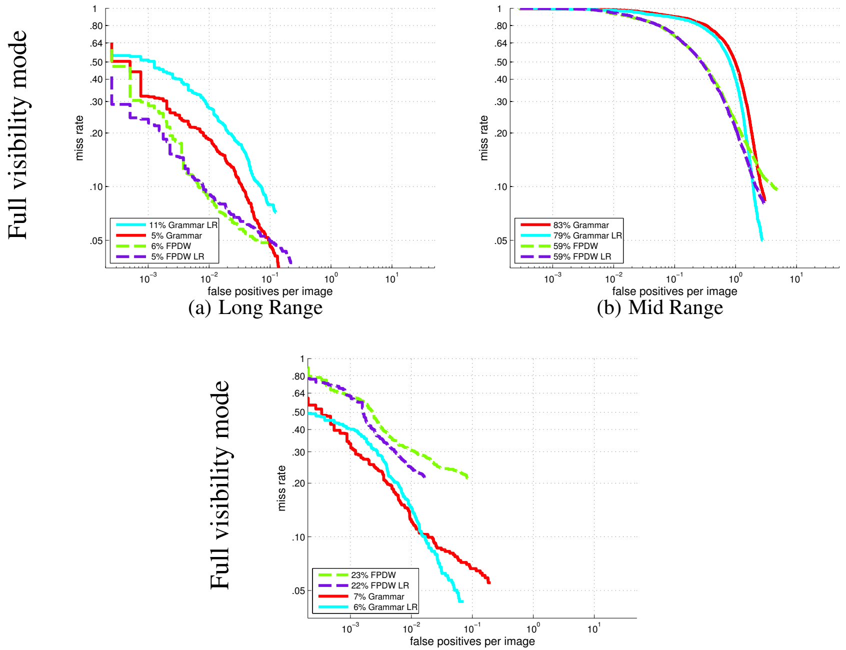


(a) (b) (c) (d) (e) (f)

Figure 5 Miss rate/FPPI plots for different groups of image sequences (rows) and for two different evaluation modes (columns). Lower curves and curves more on the left side of the plots indicate better performance.

PD algorithms implies that a detector has to evaluate a much smaller number of windows when run on the lower resolution version of an image. Statistically, this leads to the detector generating less False Positives, while the number of True Positives remains constant. The two contributions create a balancing effect.

Figure 6 graphically presents the results. The performance of GM on the far range sequences degrades when run on lower resolution images, while the performance of FPDW doesn’t change significantly (see Fig. 6(a)). We observe that the farthest and thus, smallest pedestrians are the ones where the difference is felt. We speculate that GM suffers more than FPDW from the reduction in resolution because it relies on part detectors which require finer image details to work at their best. For the mid range sequence, the performance of both algorithms changes very little when changing the resolution (see Fig. 6(b)). We speculate this happens because the persons in this sequence are not small enough to trigger the phenomenon we observe in the long range sequences. For the short-range sequences we observe a clear reduction in False Positives (especially for FPDW) at high FPPI rates/low detection confidence. We ascribe this reduction to the smaller number of windows classified, as well as to the smoothing of image compression artifacts performed by the bilinear interpolation: image areas containing artifacts which are classified as persons with a low confidence in the original images, are classified as background in the LR images. Camera 59 was excluded from the short range set because of a specific occurrence: in the original size images, the fire extinguisher sign was very often detected as a person by FPDW, while this did not happen in the subsampled images. This peculiar event had a big influence on the results, but was deemed not to be of general interest.



(a) Long Range

(b) Mid Range

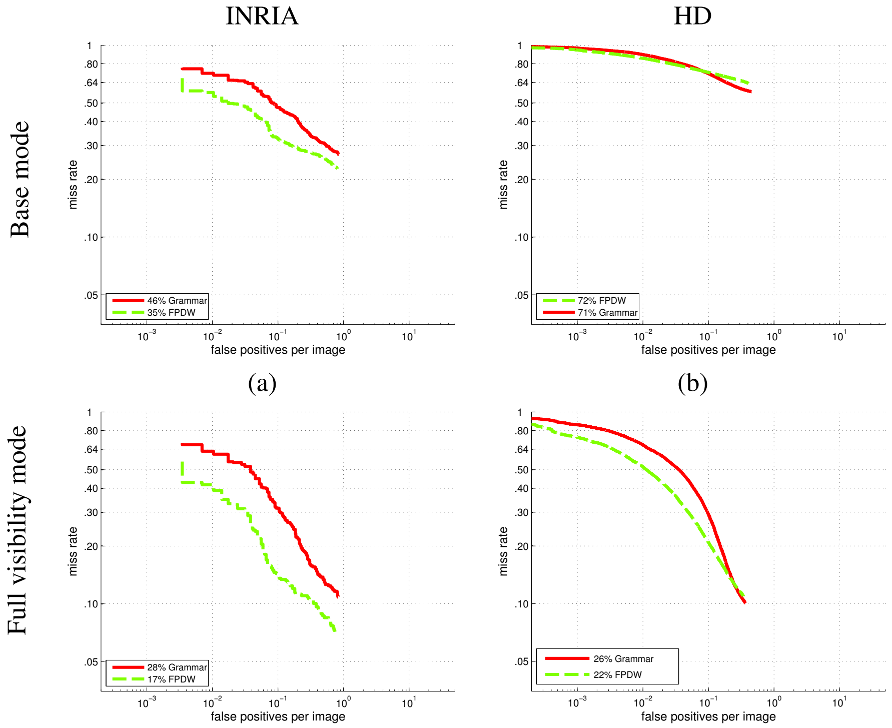
(c) Short Range

Figure 6 Miss Detection rate/FPPI plots comparing the performance of the detectors on the original and the Low Resolution (LR) version of the images, for the three ranges we defined. Long range comprises camera 50 and 60, mid range is just camera 54, while short range includes cameras 53, 56, 57 and 58.

The results show that HR cameras can be advantageous for detecting people, especially at greater distances, but at the same time, this approach has a tendency for generating more False Positives. Given a specific scenario, the practitioner should set each camera to the lowest possible resolution that enables the detection of people at the farthest visible point. This ensures the least amount of Missed Detections while also minimizing False Positives. Another insight given by these results is that scene geometry and camera calibration can be exploited – a detector can be set to evaluate bigger detection windows close to the camera and smaller detection windows further away, thus limiting even more the amount of False Positives.

4.3 Performance on Different Datasets

In another experiment we compare the performances of FPDW and GM on the proposed data set and on the INRIA person data set [2]. We use a recently proposed annotation for the INRIA data set [19], which includes information on the degree of occlusion that affects each person. It can be noted that the HDA data set as a whole is harder than the INRIA data set for the PD task: the performance of the algorithms is considerably worse on it than on the INRIA data set both on the Base and the Fully visible evaluation mode (see Figure 7). Considering the large difference in the results between the Base and the Fully visible evaluation mode, on both datasets, we confirm that occlusion poses a severe challenge to PD algorithms.



(a) (b)

(c) (d)

Figure 7 Miss detection rate/FPPI plots comparing the performance of FPDW and GM on the INRIA and the HDA data sets. HD is clearly more challenging than INRIA even on the case of not-occluded pedestrians. Results on other datasets for FPDW and GM are publicly available (see [7] and [15], respectively).

The proposed dataset presents a significant number of these cases thus proving an excellent benchmarking base for evaluating next generation algorithms.

5 Person Re-identification Benchmarking

The problem of person re-identification (RE-ID) consists in finding the same person in different cameras or in multiple appearances in the same camera. Many recent state-of-the-art algorithms tackle RE-ID as a matching problem, comparing the information contained in the BB's enclosing images of pedestrians. Colour histograms are typically employed to characterize different persons [4, 20, 21], since they compare favorably with respect to other types of features. The spatial arrangement of colour features inside the detection BB's is often taken into account by characterizing the colour distribution in different cells within the BB. Some divide each BB in six horizontal stripes of fixed relative size [22, 20, 21, 23]. The best results however come if a body-part detector is used to detect head, torso and legs, and colour features are computed only on such parts [4]. Following this approach, we use two different part detectors, Andriluka's [24] (PS) and Fenzenswalb's [16] (DTDPM v5) to detect body parts. We then extract a colour histogram feature from each part, concatenate such histograms, and perform nearest-neighbour classification between each test image and the training image set.

To evaluate the performance of RE-ID algorithms, the Cumulative Matching Characteristic curve (CMC) is the popular method of choice. It shows how often, on average, the correct person ID is included in the best K matches against the training set for each test image. The overall algorithm performance is measured by the nAUC – the normalized Area Under the CMC curve.

5.1 Performance on Different Image Resolutions

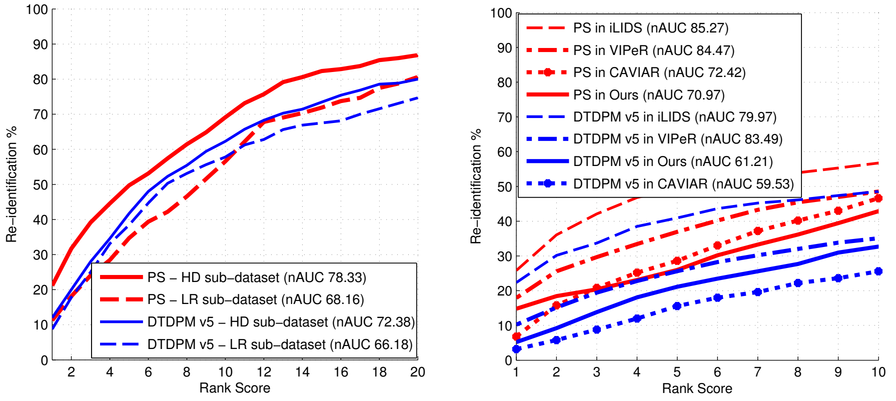
The first experiment tries to assess the influence of high-resolution cameras on the RE-ID results. In principle, high resolution would allow for better feature description of the information contained in the body parts which would benefit re-identification algorithms. In this section we designed two experiments to test this hypothesis.

First, we selected two different sub-sets of non-occluded pedestrian images: one with only high resolution images (HR) and another with only standard resolution images (LR). The HR sub-set is composed by one manually selected detection per person per HR camera (cameras 50 and above – see table 3) totaling 150 detections and 35 pedestrians. The LR sub-set also had one hand-picked detection per person per non-HR camera (cameras 40 and below) totaling 100 detections from 32 pedestrians. Consistent with previous RE-ID work, these data sets were randomly partitioned 100 times taking 1 image per pedestrian for training and 1 image per pedestrian for testing – and displaying the average re-identification performance in the 100 different runs. In Figure 8(a) we can observe that both PS and DTDPM v5 perform consistently better in high resolution cameras.

Second, we downsampled the images of the HR set to VGA resolution, and performed the same analysis. However, in this case, no conclusive difference in RE-ID performance was found. Therefore, no clear evidence exists for supporting the use of high-resolution images for re-identification with the used algorithm. The improvements obtained in the experiments with the different sub-sets may be due to better image quality in general, or simpler environment conditions. Anyway, we believe that more discriminative features than the ones used in this work (simple colour histograms) can exploit in a more effective way the high-resolution contained in the images and improve re-identification algorithms, and this will be the focus of our future work.

5.2 Performance on Different Datasets

The second experiment illustrates how challenging our data set is for RE-ID in comparison to three other data sets (CAVIAR4REID, iLIDS4REID and VIPeR – described in table 1). For this more general case, the data set is comprised of both HR and LR subsets, totaling 250 detections from 52 pedestrians. The selection of the training and testing images was done in the same way as before (by random partition, 100×). Results are shown in Figure 8(b). We observe that, together with CAVIAR4REID, the presented data set is one of the most challenging. The mixture of different resolution cameras, different perspectives and ranges (including one overhead camera), the presence of harsh illumination changes, severe occlusions, and the fact that several subjects add or remove items of clothing from one view to the next (i.e., put jackets on – a notable example in Figure 3(a) and 3(b)) make it one of the most challenging RE-ID data sets up to date. CAVIAR seems to be as challenging as our dataset. Because of the low resolution of its images, it is simply very hard to see and differentiate people. The HD dataset, instead, introduces differences in scale, lighting and pose more drastic than in related datasets. Also it introduces an aspect that state-of-the-art re-identification algorithms are not yet able to tackle in a robust way: differences in people's clothing from one camera to the other.



(a) (b)

Figure 8 a) CMC curves comparing performances on High Resolution (HR) and Low Resolution (LR) subsets of the HDA data set (higher curves correspond to better performance). HR images prove to be better for re-identification. b) Comparing performances between data sets: HD proves to be one of the most challenging data sets.

6 Conclusions

In this paper we presented a new data set for research on video surveillance algorithms using high resolution cameras. The data set contains several video sequences of a typical office building at rush hour, with many people entering and leaving, crossing corridors and appearing in many of the cameras in the network. Additionally, there are quite different views under a variety of acquisition conditions (i.e. illumination changes, occlusions, clothing variations, perspective). We showed that indeed this diversity is of crucial importance to characterise the performance of recent state-of-the-art video surveillance algorithms. More specifically, we focused on pedestrian detection and re-identification methods. In terms of pedestrian detection, we illustrated how the GM and FPDW methods have complementary performances in different settings (e.g. view, resolution). Our results suggest that the selection of the most appropriate methodology depends on the viewing conditions and these should be taken into account during benchmarking. In terms of re-identification, we showed that our data set is currently the most complete and challenging. It includes an unprecedented combination of views, persons and frames, with hard conditions such as clothing changes, partial occlusion and huge resolution variation. The inclusion of high and low resolution imagery is also unique in our data set. This allows for work comparing the performance of algorithms in both resolutions and, most importantly, to assess how high resolution impacts on current surveillance algorithms. Although in a limited sense, we showed that both person detection and re-identification technology benefit, in terms of accuracy, from the use high resolution imagery. Further work will focus on the exploitation of the richer information contained in our data set to improve the quality of the algorithms, mainly by establishing which image features are the most suited for pedestrian detection and re-identification methods in the high resolution world.

7 Acknowledgments

This work was partially supported by the FCT project [PEst-OE/EEI/LA0009/2013], the European Commission project POETICON++ (FP7-ICT-288382), the FCT project VISTA [PTDC/EIA-EIA/105062/2008] and the project High Definition Analytics (HDA), QREN - I&D em Co-Promoção 13750.

References

[1] M Oren, C Papageorgiou, P Sinha, E Osuna, and Tomaso Poggio. Pedestrian detection using wavelet templates. CVPR, 1997.

[2] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. CVPR, pages 886–893, 2005.

[3] Caviar dataset. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[4] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In BMVC, pages 68.1–68.11. BMVA Press, 2011.

[5] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and Appearance for Mobile Scene Analysis. ICCV, pages 1–8, 2007.

[6] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. CVPR, pages 794–801, June 2009.

[7] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian Detection: An Evaluation of the State of the Art. PAMI, pages 1–20, July 2012.

[8] J.Ferryman. and A. Shahrokni. An overview of the pets2009 challenge. In 11th PETS Workshop. IEEE, 2009.

[9] Álvaro García-Martín, José M. Martínez, and Jesús Bescós. A corpus for benchmarking of people detection algorithms. PRL, 33(2):152–156, January 2012.

[10] W. S. Zheng, Shaogang Gong, and T. Xiang. Associating groups of people. In BMVC, 2009.

[11] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In ECCV, Jul 2008.

[12] Piotr's image and video matlab toolbox (pmt). http://vision.ucsd.edu/˜pdollar/toolbox/doc/index.html.

[13] Caltech pedestrian detection evaluation code. http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/DollarEvaluationCode.

[14] Piotr Dollár, Serge Belongie, and Pietro Perona. The Fastest Pedestrian Detector in the West. BMVC, pages 68.1–68.11, 2010.

[15] R Girshick, Pedro Felzenszwalb, and D McAllester. Object detection with grammar models. PAMI, pages 1–9, 2011.

[16] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/.

[17] Visual object classes challenge 2010. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/.

[18] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. IJCV, 88(2):303–338, 2010.

[19] Matteo Taiana, JacintoC. Nascimento, and Alexandre Bernardino. An improved labelling for the inria person data set for pedestrian detection. In JoãoM. Sanches, Luisa Micó, and JaimeS. Cardoso, editors, Pattern Recognition and Image Analysis, volume 7887 of Lecture Notes in Computer Science, pages 286–295. Springer Berlin Heidelberg, 2013.

[20] Tamar Avraham, Ilya Gurvich, Michael Lindenbaum, and Shaul Markovitch. Learning implicit transfer for person re-identification. In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, Computer Vision – ECCV 2012. Workshops and Demonstrations, volume 7583 of Lecture Notes in Computer Science, pages 381–390. Springer Berlin Heidelberg, 2012.

[21] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(3):653–668, 2013.

[22] Chunxiao Liu, Shaogang Gong, ChenChange Loy, and Xinggang Lin. Person re-identification: What features are important? In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, ECCV Workshop, volume 7583 of Lecture Notes in Computer Science, pages 391–401. Springer Berlin Heidelberg, 2012.

[23] Ryan Layne, TimothyM. Hospedales, and Shaogang Gong. Towards person identification and re-identification with attributes. In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, ECCV Workshop, volume 7583 of Lecture Notes in Computer Science, pages 402–412. Springer Berlin Heidelberg, 2012.

[24] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. CVPR, 2009.