SPID: Surveillance Pedestrian Image Dataset and Performance Evaluation for Pedestrian Detection

Dan Wang[1], Chongyang Zhang[1](✉), Hao Cheng[1], Yanfeng Shang[2], and Lin Mei[2]

[1] Institute of Image Communication and Network Engineering,
Shanghai Jiao Tong University, Shanghai, China
sunny_zhang@sjtu.edu.cn

[2] The Third Research Institute of The Ministry of Public Security,
Shanghai, China

Abstract. Pedestrian detection is highly valued in intelligent surveillance systems. Most existing pedestrian datasets are autonomously collected from non-surveillance videos, which result in significant data differences between the self-collected data and practical surveillance data. The data differences include: resolution, illumination, view point, and occlusion. Due to the data differences, most existing pedestrian detection algorithms based on traditional datasets can hardly be adopted to surveillance applications directly. To fill the gap, one surveillance pedestrian image dataset (SPID), in which all the images were collected from the on-using surveillance systems, was constructed and used to evaluate the existing pedestrian detection (PD) methods. The dataset covers various surveillance scenes and pedestrian scales, view points, and illuminations. Four traditional PD algorithms using hand-crafted features and one deep-learning-model based deep PD methods are adopted to evaluate their performance on the SPID and some well-known existing pedestrian datasets, such as INRIA and Caltech. The experimental ROC curves show that: The performance of all these algorithms tested on SPID is worse than that on INRIA dataset and Caltech dataset, which also proves that the data differences between non-surveillance data and real surveillance data will induce the decreasing of PD performance. The main factors include scale, view point, illumination and occlusion. Thus the specific surveillance pedestrian dataset is very necessary. We believe that the release of SPID can stimulate innovative research on the challenging and important surveillance pedestrian detection problem. SPID is available online at: http://ivlab.sjtu.edu.cn/best/Data/List/Datasets.

1 Introduction

Pedestrians are the primary surveillance objects in security systems, and thus pedestrian detection is becoming the fundamental research area in intelligent surveillance systems. In the practical surveillance systems, pedestrian detection (PD) is still a challenging problem due to the visual appearance differences

© Springer International Publishing AG 2017

C.-S. Chen et al. (Eds.): ACCV 2016 Workshops, Part III, LNCS 10118, pp. 463–477, 2017.
DOI: 10.1007/978-3-319-54526-4_34

sunny_zhang@sjtu.edu.cn

caused by the large-scale variations of surveillance scenes. Most PD algorithms require pedestrian datasets to train classifiers or learn discriminative features using machine learning. The training dataset is a strong dependency of PD algorithms. In the past few years, an increasing number of benchmarks have been proposed to push forward the performance of pedestrian detection, e.g., INRIA [1], ETH [2], Caltech [3] and KITTI datasets [4]. However, most existing pedestrian datasets are collected from non-surveillance videos. The comparisons of the self-collected data and practical surveillance data show that there exist significant data differences. The properties, like resolution, illumination, view point, and occlusion of pedestrians differ greatly. Despite an extensive set of ideas has been explored for pedestrian detection, most existing algorithms are trained on traditional datasets. Consequently the accuracy and robustness of the most existing PD algorithms may perform not so well in the real surveillance systems.

This paper introduces the surveillance pedestrian image dataset (SPID) that aims to fill the gap between the existing datasets and real surveillance data. We collected approximately 297 surveillance video clips (20 min per clip) from on-using surveillance cameras. The videos are collected from different areas and contain 8 typical surveillance scenes. The multiple scenes in SPID are shown in Fig. 2, including the highways, campuses, city roads and rural areas. Videos are recorded continuously 24 h per day, cover 6 various illumination conditions from morning to night. The differences of pedestrian distribution and appearance under various scenes are huge. We extracted the frames with at least one pedestrian, set up the annotation standard and labeled several properties for each pedestrian. After selection, about $110k$ ($k = 10^3$) pedestrian objects were collected and labeled. The pedestrian properties include: illumination, view, size, pose, attachment, occlusion and appearance. Our goal is to complement existing benchmarks by providing real-world surveillance data.

The main contributions of our work are three-fold: (a) The construction of a typical and diverse surveillance dataset, which requires significant efforts in collection and annotation. As far as we know, this dataset is the first released surveillance pedestrian dataset. SPID is available online at: http://ivlab.sjtu.edu.cn/best/Data/List/Datasets. We believe SPID can stimulate innovative research on the challenging and important surveillance PD problem; (b) Four traditional PD algorithms using hand-crafted features and one deep-learning-model based deep PD methods are evaluated using existing pedestrian dataset (such as INRIA [1] and Caltech [3]) and SPID in this work. Our experiments show that the PD performance on SPID is worse than that on traditional datasets. (c) From experiments, we also validate that the data differences between SPID and other pedestrian datasets influence the performances greatly. Main factors include scale, view point, illumination and occlusion.

The rest of the paper is organized as follows. Related works are reviewed in Sect. 2. Section 3 introduces the collection, annotation and properties of SPID. In Sect. 4 we choose seven PD methods to evaluate on INRIA dataset, Caltech dataset and SPID. We show the result ROC curves for each algorithm

sunny_zhang@sjtu.edu.cn

and analyze the performance differences, as well as some improvement methods. Finally, Sect. 5 summarizes our work and suggests future directions with SPID.

2 Related Works

Multiple public pedestrian datasets have been collected over the past decades, including commonly used INRIA, ETH [2], TUD-Brussels [5], Daimler [6], Caltech and KITTI datasets [4]. TUD-Brussels and ETH are medium-sized video datasets. Daimler is not frequently used by PD methods because it only contains grayscale images. Below are the datasets we consider in the paper. Table 1 shows the comparisons of several pedestrian detection datasets. The first column define the imaging setup method, and the next four columns indicate number of pedestrian and images in the training and test sets. Properties column summarizes additional characteristics of the datasets. The strengths and weaknesses of INRIA, Caltech and KITTI dataset are discussed in detail.

Table 1. Comparison of public pedestrian datasets

		Testing			Training			Properties								
Dataset	imaging setup	# pedestrians	# pos. images	# neg. images	# pedestrians	# pos. images	# neg. images	color images	occlusion labels	view labels	illumi. labels	pose labels	video seqs.	no select. bias	view point	publication
INRIA[1]	photo	1208	614	1218	566	288	453	√							hori.	2005
ETH[2]	mobile cam.	2388	499	-	12k	1804	-	√					√	√	-	2007
Daimler[6]	mobile cam.	15.6k	-	6.7k	56.6k	21.8k	-						√		-	2009
Caltech[3]	mobile cam.	192k	67k	61k	155k	65k	56k	√	√				√	√	hori.	2009
KITTI[4]	mobile cam.	-	7518	-	4445	7481	-	√					√	√	hori.	2012
SPID	fixed cam.	53.7k	15.4k	-	56.2k	14.5k	-	√	√	√	√	√	√	√	bird	2016

INRIA. INRIA dataset is the oldest pedestrian dataset with high quality annotations and high images resolution. Most pedestrians are captured horizontally in day time. The illumination is fine and contours of pedestrians are clear. The diverse scenes contain outdoor landscapes like city road, mountain, beach and indoor environment. However the number of images is not large. Figure 1(a) shows the height distribution of INRIA test set.

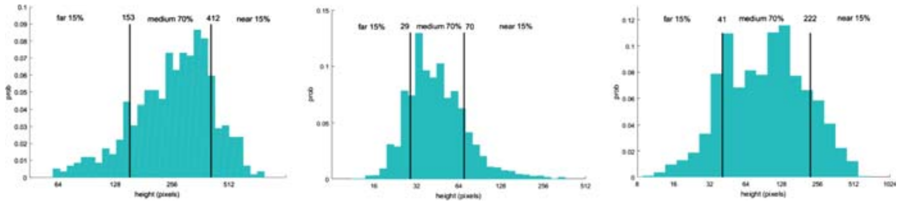
Caltech. Caltech dataset is one of the most popular PD datasets. The videos were captured by a vehicle driving through U.S. urban streets in sunny days. Camera is set on a moving car, therefore the view of pedestrians are horizontal and the sizes of pedestrians are small compared to INRIA. The test set contains 67k positive images and 192k pedestrians. Although Caltech is the largest pedestrian dataset, its images are extracted each frame per video, the sampling frequency is high to 30 fps. The difference between two sequential frames is small,

sunny_zhang@sjtu.edu.cn

therefore usually every 30th image in the Caltech dataset is used for training and testing [7]. The medium pedestrian height is [30, 80] pixels. Figure 1(b) shows the height distribution of Caltech test set.

KITTI. KITTI dataset contains videos captured by a moving car around city streets with good weather conditions, therefore the pedestrians are also captured at eyelevel. This object detection benchmark consists of 7481 training images and 7518 testing images, comprising a total of 80256 labeled objects. KITTI provides both flow and stereo data.

KITTI and Caltech are the predominant datasets for PD. The scales of these two benchmarks are both large and challenging. A large number of PD algorithms have been evaluated on INRIA and Caltech, meanwhile KITTI is being gradually adopted. Figure 1 shows the height distribution on INRIA, Caltech, and SPID test sets.



(a) INRIA (b) Caltech (c) SPID

Fig. 1. Pedestrian height distribution on INRIA, Caltech, and SPID test sets

The state-of-the-art pedestrian detectors developed in recent years mainly take two features: hand-crafted features and machine-learned deep features.

Hand-Crafted Features. In 2005 Dalal and Triggs introduced the HOG detector [1], which was a landmark for pedestrian detection. Later in 2009 Felzenswalb et al. put forward the classic deformation part based model (DPM) [8], a pedestrian is expressed as several parts with a deformable configuration. In 2009 Caltech dataset was introduced and a new evaluation method was proposed. Dollár et al. used FPPI (False Positive per Image) to compare the performance of the detectors. In the same year Dollár et al. proposed integral channel feature (ICF) method [9], in which Harr-like features are computed over multiple channels for each pedestrian. In 2014 the k-poselets [10] method was introduced as an improvement of DPM and poselets [11]. Aggregated channel features (ACF) [12] is an alternate approach to exploit approximate multi-scale features using ICF. Locally decorrelated channel features (LDCF) [7] is accomplished utilizing the ACF detector. LDCF uses the deeper trees and a denser sampling of the pedestrian data. In 2014, Informed Haar-like [13] features improve pedestrian detection. Pedestrian shapes are geared by three rectangles as models for different body parts. In 2015 [14] points out the link between ACF, ChnFtrs,

sunny_zhang@sjtu.edu.cn

Informed Haar, and LDCF, and generates a series of filtered channel features. Checkerboards [9] is a naive set of filters that covers the same sizes, which get the best result on Caltech test set among all the algorithms up to 2015.

Deep Learning Method. The success of deep learning techniques in image classification has promoted researchers to try it on PD. Object detectors that out-perform others are generally based on variants of R-CNN model. A reduced set of detection proposals is created for an image, the proposals are evaluated by a convnet. Convolutional neural networks (CNN) perform best among the deep learning models. CNN optimizes the feature representation automatically in the detection task and regularizes the neural network. Chen et al. [15] used pre-trained deep CNN with ImageNet dataset to generate candidate windows, together with ACF detector to get final features. CifarNet [16] is a small network designed to solve the CIFAR-10 classification problem. AlexNet [17] is a network designed to solve the ILSVRC2012 classification problem. These two networks are both re-implemented in Caffe project [18]. GoogLeNet [19] was responsible for ILSVRC2014. Ren et al. designed Faster R-CNN [20], which used the Simonyan and Zisserman very deep model (VGG) [21]. Faster R-CNN achieved state-of-the-art object detection accuracy on both PASCAL VOC 2007 and 2012.

3 SPID: Surveillance Pedestrian Image Dataset

The images in Surveillance Pedestrian Image Dataset are extracted from the videos recorded by daily used surveillance cameras. All the pedestrians are well labeled in several different properties. Each image contains at least one pedestrian. The various collecting conditions cover various intensities, different time periods, multiple scenes and so on. SPID contains 29989 well-labeled images and about 110k pedestrian objects. We split the dataset into training and test sets roughly in half. The test set contains 14550 images and the training set contains 15439 images as well. For detailed statistics about the data, see the bottom row of Table 1. Although SPID is only second to Caltech dataset in terms of scale. Caltech contains all adjacent frames in the video while SPID dataset does not. Considering the effective size (containing not-so-similar frames), SPID may be even bigger than Caltech.

3.1 Dataset Collection

The images in the dataset are all collected from on-using surveillance systems. Some of them are collected from the monitoring cameras used in companies, university campus; some of them are got from public security cameras. The surveillance cameras are set along the streets in the cities and campus, or beside the squares and highways. Most cameras in cities are placed 10 m above the ground, the other cameras like those set at campus entrance are placed 1 m high. We gather 297 pieces of videos, totally about 61 h long. Basically each piece of the videos is about 20 min long, we extract the videos to images per second and

sunny_zhang@sjtu.edu.cn

pick up the images with at least one pedestrian to label. The resolution of the images are various from 352*288, 720*576, 1280*960 to 1920*1080. Our dataset contains both low resolution and HD images. Figure 1 shows multiple diverse examples of pedestrians in SPID.

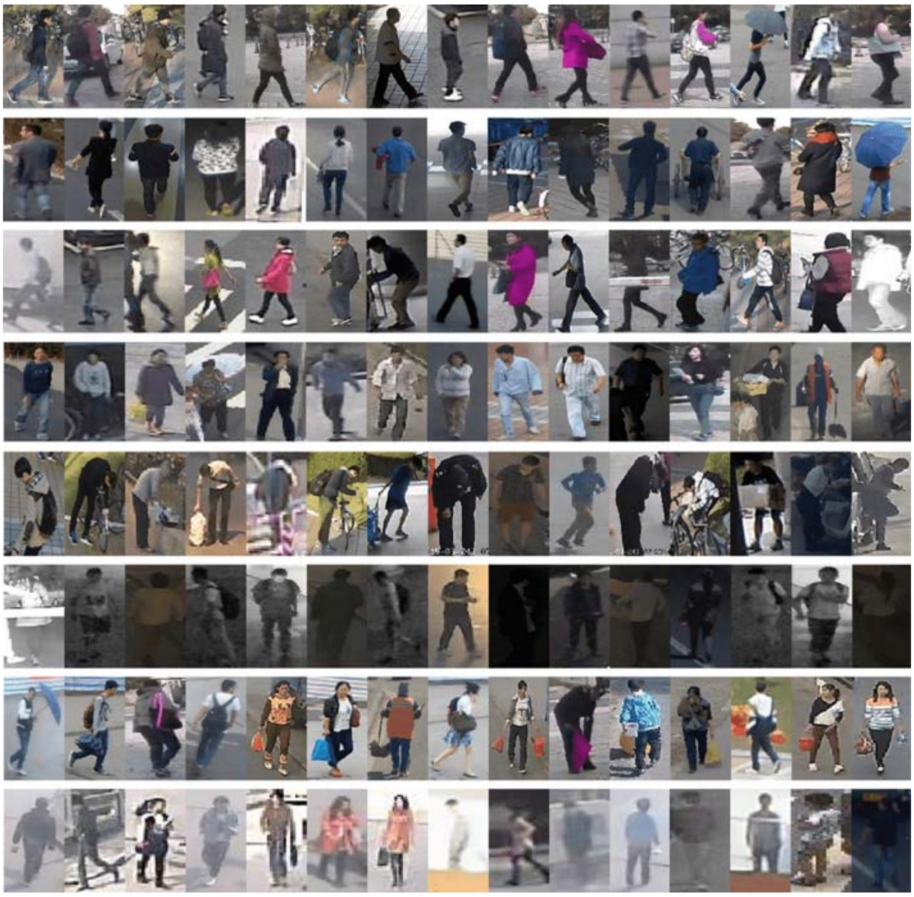


Fig. 2. Examples of diverse pedestrians in SPID. The first four rows show the various view of pedestrians (right side, back side, left side and front side). The 5^{th} row shows special pose pedestrians (stoop). Next row shows the pedestrians captured during night and illumination is bad. The 7^{th} row shows pedestrians with various attachments. The last row shows the low resolution pedestrians.

3.2 Ground Truth Annotation

After referencing the annotations of the widely used datasets, like INRIA, Caltech, ETH and USC, the ground truth annotation file format for surveillance dataset is properly settled. Each annotation file, which is written in xml format,

sunny_zhang@sjtu.edu.cn

corresponds to an image. The xml file contains the total visible pedestrian number of a single image and the corresponding bounding box (BB) coordinate of each pedestrian. If a pedestrian is occluded, the BB only contains the visible part. If an image contains multiple pedestrian objects, we label them individually, i.e. this dataset only has person label and no group label.

Each BB describes a specific pedestrian object. For each pedestrian, we also provide the annotation file with several useful properties: the name of its source image, the specific size and level, scene, view, pose, intensity condition, occlusion level and attachment information. In addition, the top color, bottom color, top style, bottom style of pedestrian clothes are also labeled.

3.3 Dataset Statistics

Scene. The cameras are settled at different areas in cities and campuses. The scenes of the video are diverse, which contain roads, city highways, campus roads and school entrances (see Fig. 3). Therefore the scenes of our dataset represent the typical surveillance scenes in daily life.

Illumination. The videos we collected are recorded during continuous time periods, the basic illumination is high during the day and gradually becomes low during the night, as shown in Fig. 3. We divide the illumination into several different situations: normal, foggy, rainy, cloudy, dusk, night and other. Figure 4(a) shows specific statistics of illumination.

Scale. Due to the various height of the cameras, the distance between pedestrians and cameras covers a wide range. Figure 1 shows the pedestrians height distribution on INRIA, Caltech and SPID test sets. Pedestrian scales influence the detection performance greatly, as shown in [3].

View. Due to the variation of camera positions, pedestrians show various angles of view, such as front side, left side, back side and right side. The property of view is more useful for the granular pedestrian detection, due to that appearance features are different when a pedestrian stand with different views. Pedestrians

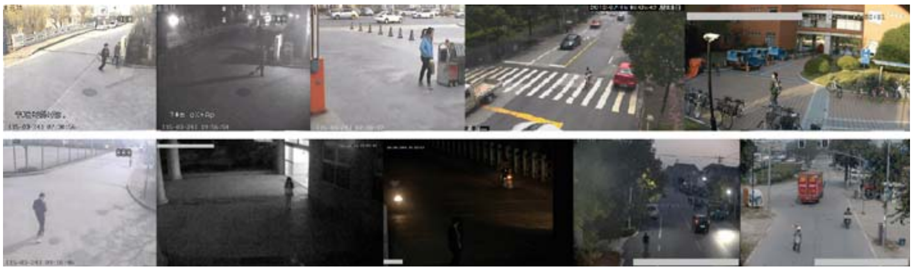
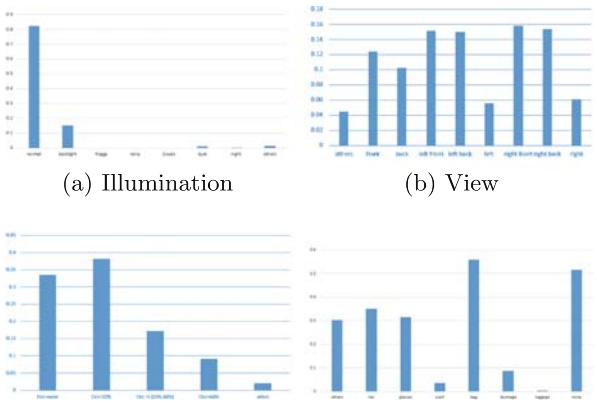


Fig. 3. Multiple scenes in SPID

sunny_zhang@sjtu.edu.cn



(a) Illumination (b) View

(c) Occlusion (d) Attachment

Fig. 4. Detail statistics of SPID data set.

with different view looks discrepant visually. For instance, when a pedestrian is standing at the front side, his face is a key info to detect the person (as detection in PASCAL). However, at the other three side, face is occluded. The division and label of this property may help to design some other features for this condition. Figure 4(b) shows the detail distribution of pedestrian views.

Pose. Pose is an important property for pedestrian detection, usually upright pedestrian is easy to be detected. In other pedestrian dataset like INRIA and Caltech, most pedestrian objects are also upright, barely contains other pedestrian poses. We consider the pose variation of surveillance pedestrians and fill the blank. Once we label the different poses, we could train specific features for the non-upright pedestrians. In our dataset we label different poses including upright, squat, bent, jump, lie down, swing, sit and others.

Occlusion. Occlusion influences the accuracy of pedestrian detection rapidly in [3]. Caltech dataset is the only one that labels this property. Unlike Caltech, we divide the occlusion levels by the ratio of occluded parts to the whole pedestrian body. Occluded ratio is grouped into 4 levels: Occ = none, Occ < 33%, Occ in [33%–66%], Occ > 66%, others. The levels are measured by human subjective vision, see Fig. 4(c) for detail.

Attachment. On one hand, basically other objects in surveillance videos do not contain attachment. Attachments may change the contour of pedestrian, which reduces the robustness of detection methods. On the other hand, we could use object (like bags, bicycles) detection to help with pedestrian detection. Common attachments of pedestrians contain hat, glasses, scarf, bag, dunnage, luggage and others, see Fig. 4(d).

sunny_zhang@sjtu.edu.cn

Appearance. The color and style of clothes are important attributes for person identification. We choose some fundamental color and style to make a rough division. Colors include 11 types such as black, white, red, yellow and so on. The styles contain purity, horizontal stripes, vertical stripes, checks and pattern.

4 Evaluation of Pedestrian Detection Algorithms

We made an extensive evaluation of five pedestrian detection algorithms under various scenarios and different datasets to increase the scope of experiments. In Sect. 4.1 we introduce the selection and validation of algorithms. Next we give a brief description of the evaluation standard in Sect. 4.2 and report experiment performance curves in Sect. 4.3. In Sect. 4.4 we analyze and emphasize the influence of dataset discrepancy by showing the evaluation results under different conditions.

4.1 Algorithms Selection and Validation

The selection of the PD algorithms satisfies the rules that algorithms must be published and open source on the Internet. To compare the state-of-the-art PD methods, both hand-crafted detectors and deep learning methods are chosen. For hand-crafted features, we chose to evaluate the pretrained detectors with default parameters, which were obtained from online source codes. For Faster R-CNN, we used original network designed for multiple classification problems. The evaluated hand-crafted detectors are LDCF, ACF, DPM, k-poselets. ACFCaltech+ and LDCFCaltech indicates these two detectors pre-trained with Caltech training set. Similarly ACFInria and LDCFInria represents for detectors trained on INRIA training set. As for the deep learning methods, original Faster R-CNN, Faster R-CNN finetuned with Caltech training set and network finetuned with SPID training set are tested. Evaluation datasets include the test sets of INRIA, Caltech and SPID. The evaluation results are computed with the latest online released codes.

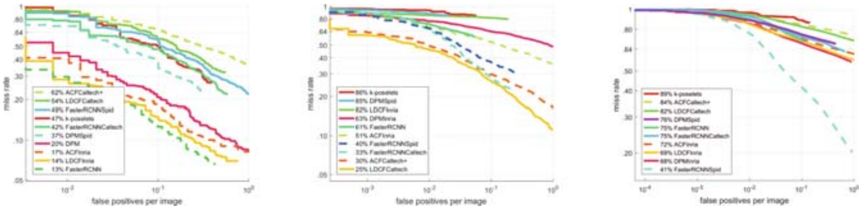
4.2 Evaluation Standard

Since Dollár et al. [22] proposed the evaluation methodology to use ROC curves compare the performance of different PD algorithms, we adopt this standard as well. The ROC curves show the relationship between miss rate and FPPI (False Positive Per Image). In INRIA dataset only the testing positive 288 images are considered. The test set (set06–set10) of Caltech is used, we extracted images with 30 frames interval and obtained totally 4024 images. The SPID test set contains 15439 images. For each image, the ground truth bounding boxes (BB_{gt}) and detected bounding boxes (BB_{dt}) are loaded. The aspect ratio of all the boxes is preprocessed to 0.41 and the overlap threshold of bouding boxes is set to 0.5. Specific accuracy calculation method is the same as that in [22].

sunny_zhang@sjtu.edu.cn

4.3 Experiment Results

The experiments are grouped into two aspects: one is the comparison result of the pedestrian detection methods trained using existing datasets, the other is the comparison result of these methods retrained using SPID training data. The evaluation result ROC curves of five representative pedestrian algorithms on three proposed datasets under different settings are shown in Fig. 5.



(a) INRIA (b) Caltech (c) SPID

Fig. 5. Evaluation results under reasonable condition on three test sets. (a) and (b) shows evaluation results on INRIA and Caltech. (c) performances upgrade on SPID.

Figure. 5a–c show performance for three test sets under reasonable setting. This setting serves as a filter, which selects pedestrians over 50 pixels tall to evaluate the performance. In Fig. 5 a Faster R-CNN and LDCFInria perform best with log-average miss rate of 13–14% On INRIA. ACFInria and DPM achieve the log-average miss rate about 20%. In Fig. 5b LDCFCaltech achieves the best result on Caltech-Test dataset and k-poselets is the worst one. Figure 5c plots performance on the entire SPID test set. Faster R-CNNSpid outperforms the other detectors remarkably. DPMInria and LDCFInria have close log-average miss rates about 68% on 0.1FPPI.

For all algorithms, performance is best on INRIA among all test sets. INRIA contains high-resolution pedestrians, with Faster R-CNN, LDCFInria and ACFInria achieving log-average miss rates of 13–17% (See Fig. 5a). Performance is also fairly high on Caltech (see Fig. 5b) with 25% log-average miss rate obtained by LDCFCaltech. This possibly due to that LDCFCaltech and ACFCaltech+ are both trained on Caltech training set. Faster R-CNNSpid and DPMInria perform better on SPID than Caltech-Test, which means the model of DPM and Faster R-CNN may be fit for SPID pedestrians when trained well.

The comparison result of the pedestrian detection methods using released codes (k-poselets, ACFCaltech+, LDCFCaltech, Faster R-CNN, ACFInria, LDCFInria and DPMInria), show the large performance decreasing when applying the existing methods on SPID directly; The other comparison result of these methods retrained (Faster R-CNNCaltech, DPMSpid, Faster R-CNNSpid), verifies that, the existing methods retrained using SPID or Caltech can't work very well on surveillance images. In other words, due to the remarkable discrepancy between the surveillance images and non-surveillance datasets, developing of specific pedestrian detection algorithms are needed for the surveillance applications.

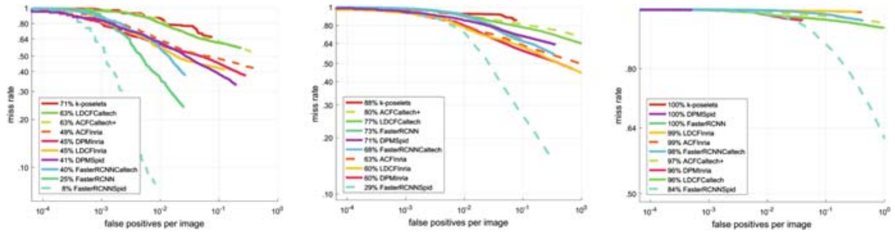
sunny_zhang@sjtu.edu.cn

4.4 What are the Factors Influencing the Performances?

We rank detector performance with multiple test sets to assess whether the discrepancy between data are influential to the results. Several data discrepancies are considered, including the pedestrian scale variation: near scale, medium scale and far scale; view point difference: horizontal and bird's eye view. We evaluate these two factors in detail respectively.

Scale. Figure 1 plots the height distribution for pedestrians on three test sets. However, the height distributions are somewhat dissimilar. Heights of SPID pedestrians have wider variations, basically cover the range of both INRIA and Caltech. About 70% pedestrians are 41–222 pixels tall. The wider pedestrian scale range makes SPID more challenging than INRIA and Caltech.

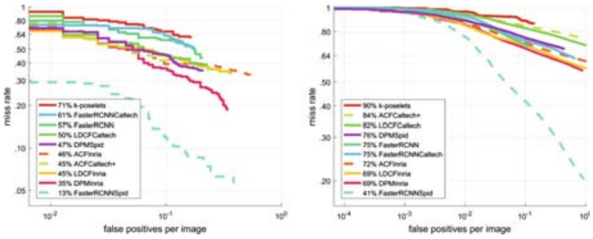
The results of Fig. 5 show that all the algorithms perform best on INRIA, which may due to the high resolution pedestrians in INRIA dataset. From Fig. 1, the discrepancy of pedestrian height among three datasets is large. We group pedestrians of SPID test set by their heights in pixels into three scales: near (283 or more pixels), medium (between 81–283 pixels), and far (81 pixels or less). This division to three scales is motivated by the distribution of heights in SPID, human performance and surveillance system requirements. To verify the influence of scale variation, we tested SPID pedestrians with three height constraints. Results for large, near and medium scale pedestrians, are shown respectively in Fig. 6. For near scale, Faster R-CNNSpid performs best; while other detectors, DPMSpid, LDCFInria, DPMInria and ACFInria achieve log-average miss rates about 40–50%. Under medium scale, performance degrages at least 15% for most algorithms. While Faster R-CNNSpid still achieves the best result. All algorithms degrade most on the far scale. ACFInria and LDCFInria are detectors trained on INRIA, the height of training pedestrians are tall, therefore for small pedestrians on SPID, the performances of these two detectors degrade rapidly. Faster R-CNNSpid achieves the best relative performance under this condition, but absolute performance is quite poor with log-average miss rate of 84%.



(a) Near scale (b) Medium scale (c) Far scale

Fig. 6. Evaluation results of various pedestrian scales on SPID. (a) shows the performance on unoccluded pedestrians over 283 pixels tall. (b) degrades for pedestrians about 81–283 pixels tall. (c) degrades the most for 32–80 pixel high pedestrians.

sunny_zhang@sjtu.edu.cn



(a) Horizontal view (b) Bird's eye view

Fig. 7. Evaluation results on SPID test set under camera angle variation.

View Point. During our experiment we notice that INRIA and Caltech pedestrians are all captured at eyelevel (i.e., the angle of a pedestrian head and camera is nearly horizontal). On the other hand, most cameras in SPID are set 10 m above the ground (i.e., the camera is like the bird's eye). Pedestrian body parts and ratios change greatly at different views. Up to 98.9% pedestrians in SPID are captured at bird's eye view, while none in INRIA and Caltech. Motivated by the great angle difference, we separate SPID to SPID-bird and SPID-horizontal. The evaluation results under view variation on SPID test set is shown in Fig. 7. Except for Faster R-CNN and k-poselets, other algorithms degrade severely, the log-average miss rate of DPM degrades 18%. LDCFCaltech, ACFInria, LDCFInria and ACFCaltech+ degrade most, these detectors are trained by pedestrians at eyelevel, thus perform not well on another view.

Summary. We evaluated 4 hand-crafted detectors and 1 deep learning methods on INRIA, Caltech-Test and SPID. The results show that all algorithms perform best on INRIA. ACFCaltech and LDCFCaltech, which are trained on Caltech training set, perform better on Caltech test set than SPID. While Faster R-CNNSpid performs best on SPID. This section shows the different influence of different properties. Figure 8 shows some typical detection results. FasterRCNNSpid and DPMInria works better under bad conditions, while LDCFCaltech misses most pedestrians on low illumination images. However, on HD images with complex background, all the algorithms degrade remarkably.

Performance is far from perfect on SPID, even under the favorable conditions. Table 2 shows miss rates of all detectors under various conditions on SPID test set. The performance decrease is caused by the data differences of multiple datasets. Main factors include scale and view point, as well as illumination and occlusion. All have obvious influences of algorithm performances.

(a) 8% of all pedestrians are missed even at the near scale. For smaller pedestrians, performance degrades catastrophically. Table 2 shows that nearly half the detectors reach 100% log-average miss rate at far scales.
(b) The special bird's eye view, caused by surveillance cameras, also increases the detection difficulty, see Table 2.

sunny_zhang@sjtu.edu.cn

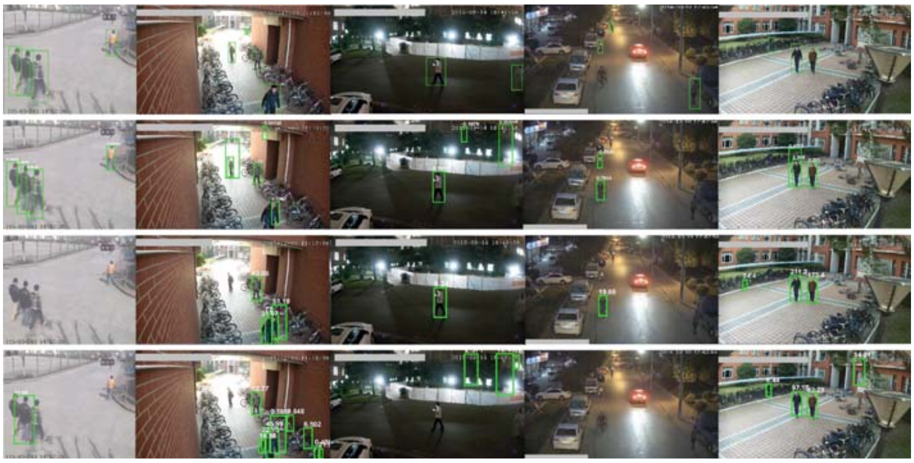


Fig. 8. Typical results of top four best detectors on SPID. The first row shows typical results of FasterRCNNSpid. And the second row shows results of DPMInria, the third row represents for LDCFInria, and the last row indicates the ACFInria performance.

Table 2. Miss rates of all detectors under various conditions on SPID test set.

Detector	Training dataset	Scales			View points	
		Near	Medium	Far	Horizontal	Bird's eye
FasterRCNNSpid	SPID	8%	29%	84%	13%	41%
FasterRCNN	ImageNet [23]	25%	73%	100%	57%	75%
FasterRCNNCaltech	Caltech [3]	40%	68%	50%	61%	75%
DPMSpid	SPID	41%	71%	100%	47%	76%
LDCFInria	INRIA [1]	45%	60%	99%	45%	69%
DPMInria	INRIA	45%	60%	96%	35%	69%
ACFInria	INRIA	49%	63%	99%	47%	72%
ACFCaltech+	Caltech	63%	80%	97%	45%	84%
LDCFCaltech	Caltech	63%	77%	96%	50%	82%
k-poselets	SPID	71%	88%	100%	71%	90%

(c) Low illumination interferes the detection result greatly, which increases the hardness for algorithms on surveillance data captured during night.

(d) Performance is abysmal under heavy occlusion, nearly all the pedestrians are missed even at high false positive rates.

The experiment results verify that, due to the dataset discrepancy, the algorithms trained on traditional datasets do not perform well on surveillance pedestrian detection problems, especially under far scale, bird's eye view, low illumination and heavy occlusion. However, the retrained algorithms on SPID training set still perform not good enough. There is considerable room for improvement in pedestrian detection. In addition, we validate the effectiveness and necessity of SPID.

sunny_zhang@sjtu.edu.cn

5 Conclusion

This paper introduces the SPID: Surveillance Pedestrian Image Dataset, which contains multiple scenes from videos captured by on-using surveillance systems. The pedestrians in SPID are well labeled. Five latest pedestrian detection algorithms including hand-crafted detectors and deep learning methods are tested on INRIA, Caltech and SPID. Evaluation results show that the data differences, such as scale, view, illumination and occlusion between existing public datasets and SPID have large impact on the detection performances. Throwing new light on existing datasets, we hope that the proposed benchmarks will complement the gap. Some algorithms perform well under favorable conditions, however when it comes to complicated situations in surveillance images, performance degrades significantly. Our intention is to improve the algorithm performance on SPID.

Acknowledgement. This work was partly funded by NSFC (No. 61571297, No. 61527804), 111 Project (B07022), and China National Key Technology R&D Program (No. 2012BAH07B01). The authors also thank the following organizations for their surveillance data supports: SEIEE of Shanghai Jiao Tong University, The Third Research Institute of Ministry of Public Security, Tianjin Tiandy Digital Technology Co., Shanghai Jian Qiao University, and Qingpu Branch of Shanghai Public Security Bureau.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
2. Ess, A., Leibe, B., Gool, L.V.: Depth and appearance for mobile scene analysis. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, pp. 1–8. IEEE (2007)
3. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: l benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 304–311. IEEE (2009)
4. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
5. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 794–801. IEEE (2009)
6. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: survey and experiments. IEEE Trans. Pattern Anal. Mach. Intell. **31**(12), 2179–2195 (2009)
7. Nam, W., Dollár, P., Han, J.H.: Local decorrelation for improved detection. arXiv preprint arXiv:1406.1134 (2014)
8. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
9. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features (2009)

sunny_zhang@sjtu.edu.cn

10. Gkioxari, G., Hariharan, B., Girshick, R., Malik, J.: Using k-poselets for detecting people and localizing their keypoints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3582–3589 (2014)
11. Bourdev, L., Malik, J.: Poselets: body part detectors trained using 3D human pose annotations. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1365–1372. IEEE (2009)
12. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE Trans. Pattern Anal. Mach. Intell. **36**(8), 1532–1545 (2014)
13. Zhang, S., Bauckhage, C., Cremers, A.B.: Informed Haar-like features improve pedestrian detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 947–954 (2014)
14. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1751–1760 (2015)
15. Chen, X., Wei, P., Ke, W., Ye, Q., Jiao, J.: Pedestrian detection with deep convolutional neural network. In: Jawahar, C.V., Shan, S. (eds.) ACCV 2014. LNCS, vol. 9008, pp. 354–365. Springer, Heidelberg (2015). doi:10.1007/978-3-319-16628-5_26
16. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images (2009)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
18. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
19. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
22. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **34**(4), 743–761 (2012)
23. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: ImageNet: a large-scale hierarchical image database, pp. 248–255 (2009)

sunny_zhang@sjtu.edu.cn