BEST: Benchmark and Evaluation of Surveillance Task

Chongyang Zhang(✉), Bingbing Ni, Li Song, Guangtao Zhai,
Xiaokang Yang, and Wenjun Zhang

Institute of Image Communication and Network Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
sunny_zhang@sjtu.edu.cn

Abstract. Smart/Intelligent video surveillance technology plays the central role in the emerging smart city systems. Most intelligent visual algorithms require large-scale image/video datasets to train classifiers or acquire discriminative features using machine learning. However, most existing datasets are collected from non-surveillance conditions, which have significant differences as compared to the practical surveillance data. As a consequence, many existing intelligent visual algorithms trained on traditional datasets perform not so well in the real world surveillance applications. We believe the lack of high quality surveillance datasets has greatly limited the application of the computer vision algorithms in practical surveillance scenarios. To solve this problem, one large-scale and comprehensive surveillance image and video database and test platform, called Benchmark and Evaluation of Surveillance Task (abbreviated as BEST), is developed in this work. The original images and videos in BEST were all collected from on-using surveillance cameras, and have been carefully selected to cover a wide and balanced range of outdoor surveillance scenarios. Compared with the existing surveillance/non-surveillance datasets, the proposed BEST dataset provides a realistic, extensive and diversified testbed for a more comprehensive performance evaluation. Our experimental results show that, performance of seven pedestrian detection algorithms on BEST is worse than that on the existing datasets. This highlights the difference between non-surveillance data and real surveillance data, which is the major cause of the performance decreases. The dataset is open to the public and can be downloaded at: http://ivlab.sjtu.edu.cn/best/Data/List/Datasets.

1 Introduction

Video surveillance technology plays more and more important roles in the emerging public security management systems. Clearly todays video surveillance systems need to change the security paradigm from "investigation to preemption" [1]. Automatic visual analysis technologies can move today's video surveillance systems from the investigative to preventive paradigm. The aim of developing smart/intelligent visual surveillance is to replace the traditional passive video surveillance, which is proving ineffective as the number of cameras exceeds the

C.-S. Chen et al. (Eds.): ACCV 2016 Workshops, Part III, LNCS 10118, pp. 393–407, 2017.
DOI: 10.1007/978-3-319-54526-4_29

capability of human operators to monitor them. In short, the goal of visual surveillance is not only to put cameras in the place of human eyes, but also to accomplish the entire surveillance task as automatically as possible [2,3], such as awareness of location, identity and activity of objects in the monitored space, preempt incidents or detect abnormal events real-timely, enhance forensic capabilities through content based video/image retrieval. It also has a wide spectrum of promising applications, including access control in special areas, human identification at a distance, crowd flux statistics and congestion analysis, detection of anomalous behaviors, and interactive surveillance using multiple cameras, etc. [3].

Smart/Intelligent Surveillance is the use of computer vision and pattern recognition technologies to analyze information from situated sensors [1–3]. As an active research topic in computer vision, smart visual surveillance attempts to detect, recognize and track certain objects from image sequences, and more generally to understand and describe object behaviors. Recent advances in computer vision, such as deep learning, multi-modal analysis, large-scale spatio-temporal analysis, have shown great potential for some high level understanding tasks in smart video surveillance application. These include high performance human/object detection and tracking, cross camera human identification and re-identification, and action/activity/event detection. These novel techniques require large scale surveillance datasets to model of various visual understanding tasks as well as evaluation of algorithmic performances. Most intelligent visual algorithms require large-scale image/video datasets to train classifiers or learn discriminative features using machine learning. The training dataset is a strong dependency of the machine learning algorithms. However, most existing datasets are collected from non-surveillance videos, which have significant data differences compared to the practical surveillance data. In this way, many existing intelligent visual algorithms are trained on traditional datasets, and thus many of them perform not so well in the real surveillance application systems.

To this end, one large-scale and comprehensive surveillance image and video database platform, called Benchmark and Evaluation of Surveillance Task (with BEST being short), is developed to aim to highlight vision related surveillance tasks. The original images and videos in BEST were all collected using on-using surveillance cameras, and images were captured with significant scenarios, background clutter, occlusions, and viewpoint/illumination variations, which makes the dataset very challenging. Based on this newly collected multi-task multi-camera on-using surveillance databases, this benchmark aims at bringing together cutting-edge researches in the field of surveillance task aware based intelligent surveillance algorithms and applications. The datasets have been released in our website for download and usage: http://ivlab.sjtu.edu.cn/best/Data/List/Datasets.

2 Related Works

In the past decades, an increasing number of benchmarks have been proposed to push forward the performance of computer vision, e.g., ImageNet [4], PASCAL VOC [5] for visual object classification, INRIA [6], ETH [7], Caltech [9] and KITTI [10] datasets for pedestrian detection, KTH [11] and Weizmann [12] datasets that consist of people action videos, PETS [13] and TRECVID [15] used for the object tracking, event detection, or retrievals. Two datasets are available in [16] to provide a realistic, camera-captured, diverse set of videos used for change detection. As for the Person ReID benchmark, the iLIDS-VID dataset [17] contains of 600 image sequences for 300 people in two non-overlapping camera views, and the PRID2011 dataset [18] includes 400 image sequences for 200 people from two cameras. However, most existing datasets are collected from non-surveillance videos. The comparison (Table 1) of the self-collected data and practical surveillance data show that there exist significant data differences. The properties, like object appearance, image resolution, illumination, view point, and occlusion of objects differ greatly. Despite an extensive set of ideas has been explored for intelligent surveillance applications, most existing algorithms are trained on traditional datasets. Consequently the accuracy and robustness of the most existing algorithms may perform not so well in the real surveillance systems.

Table 1. Comparison of BEST and existing datasets.

	Existing non-surveillance datasets [4,6,9,11,12]	Existing surveillance datasets [13,15–17]	BEST
Data resource (Surveillance or not)	Mostly Not (Self-sampling or Internet)	Partly Yes	Yes
Scenario (Surveillance or not)	Mostly Not	Partly Yes (Simulated)	Yes (Real)
Task (Surveillance or not)	Not	Partly Yes (Partly are simulated)	Yes (Partly are simulated)
Diversity	Yes	Not	Yes

We believe the lack of high quality surveillance-oriented datasets greatly limits the application of the computer vision in the practical surveillance domain. To this end, we collect and organize a large-scale and comprehensive surveillance image and video database platform, called Benchmark and Evaluation of Surveillance Task, with BEST being short. The original images and videos in BEST were all collected using on-using monitoring cameras, and they have been selected to cover a wide range of surveillance scenarios and are representative of typical outdoor visual data. Compared with the existing surveillance or non-surveillance datasets, the BEST dataset provide a realistic, camera-captured, diverse set of surveillance images or videos, which is much larger in scale and diversity:

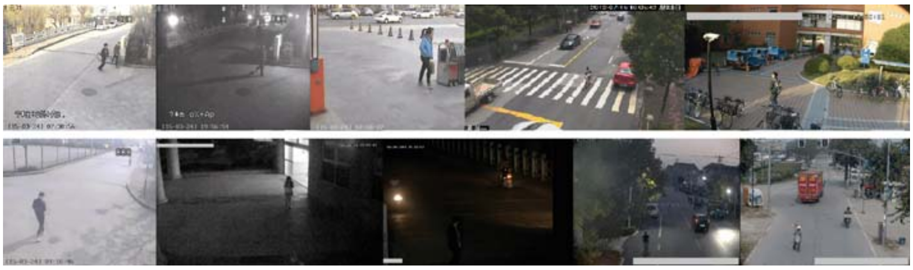


Fig. 1. Different scenarios in BEST

Resource. The original images and videos in BEST were all collected using on-using surveillance cameras.

Scale. The BEST benchmark contains over 10 million original surveillance images and more than 10k surveillance video clips; and the well-labeled images have reached a scale of more than 100 thousands samples.

Diversity.

– Scenarios: The datasets contain more than 20 surveillance scenarios, such as streets, roads, highways, campus, entrances, squares, and so on (see Fig. 1).
– Resolution: Spatial resolutions of the images/videos vary from 320 × 240, 720 × 576, 1280 × 720, to 1920 × 1080.
– Illumination: We divide the illumination into several different situations: normal, foggy, rainy, cloudy, dusk, night and others.
– View: Due to the variation of camera positions, various view angles can be got in the datasets, such as front side, left side, back side and right side. The division and label of this property may help with the object detection and recognition task.
– ...

There are three datasets have been constructed and released in the BEST benchmark:

– SPID: Surveillance Pedestrian Image Dataset
– SHAD: Surveillance Human Action Dataset
– SPRD: Surveillance Person Re-Identification Dataset

These datasets will be revised/expanded from time to time based on feedback from the academia and the industry.

3 SPID: Surveillance Pedestrian Image Dataset

3.1 Description

As a subset of BEST2016 dataset, this dataset was constructed for the pedestrian detection task in surveillance images. The images in Surveillance Pedestrian

Image Dataset (SPID) 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 datas resources satisfy the diversification benchmark including various intensities, different time periods, multiple scenarios and so on. SPID contains about 30k well-labeled images and 110k pedestrian objects. We split the dataset into training and testing sets roughly in half. The testing 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 2.

Table 2. Comparisons between proposed SPID and existing pedestrian detection 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 [6]	photo	1208	614	1218	566	288	453	√							hori.	2005
ETH [7]	mobile cam.	2388	499	-	12k	1804	-	√					√	√	-	2007
Daimler [8]	mobile cam.	15.6k	-	6.7k	56.6k	21.8k	-						√		-	2009
Caltech [9]	mobile cam.	192k	67k	61k	155k	65k	56k	√	√				√	√	hori.	2009
KITTI [10]	mobile cam.	-	7518	-	4445	7481	-	√					√	√	hori.	2012
SPID	fixed cam.	53.7k	15.4k	-	56.2k	14.5k	-	√	√	√	√	√	√	√	bird	2016

The images in SPID 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 6–10 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 pick up the images with at least one pedestrian to label. The resolution of the images are various from 320 × 240, 720 × 576, 1280 × 720, to 1920 × 1080. Our dataset contains both low resolution and HD images. Figure 2 shows multiple diverse examples of pedestrians in SPID.

3.2 Ground Truth Annotation

The SPID images are well-labeled and one .xml format ground truth annotation file can be got for each labeled image and bounding box pedestrian sample. 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

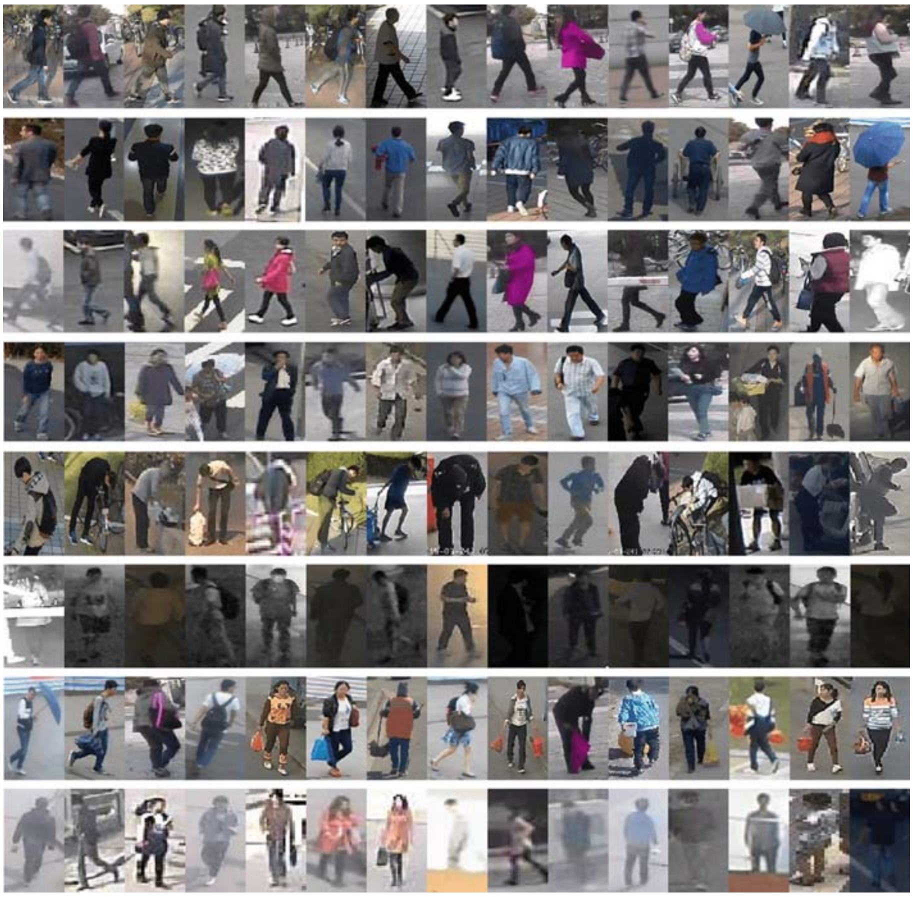


Fig. 2. Examples of diverse pedestrians in SPID. The first four rows show the various view of pedestrians (right side, back side, left side and front side). The fifth row shows special pose pedestrians (stoop). Next row shows the pedestrians captured during night and illumination is bad. The following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

pedestrian objects, we label them respectively, i.e. this dataset only has ‘person’ label and no ‘group’ label.

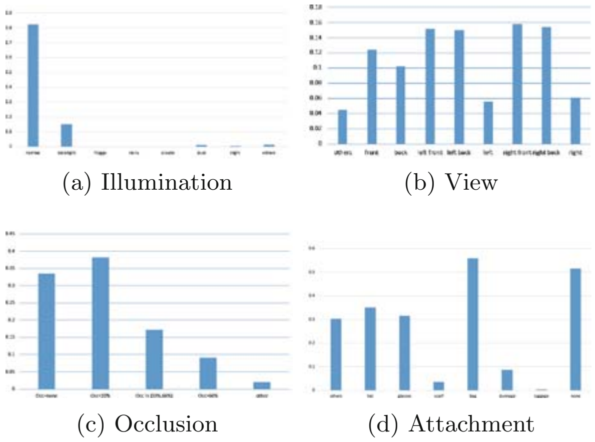
Each BB describes a specific pedestrian object. For each pedestrian sample, we also provide one annotation file with several useful properties: the name of its source image, the specific size and level, scene, view, pose, intensity condition, occlusion level, and attachment information. In addition, the top colors, bottom color, top style, bottom style of this pedestrians cloth are also labeled. The main properties are descripted as follows.

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. 1). 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 of one image or sample in different time and weathers are always different: its is high during the sunny days and maybe low during the night or cloudy weather, as shown in Fig. 1. We divide the illumination into several different situations: normal, foggy, rainy, cloudy, dusk, night and other. The distribution statistics of the illumination in SPID can be seen in Fig. 3(a).

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 the different appearance features with the different views: When a pedestrian is standing as the front side view, his/her face is one key appearance feature to detect the person (as many algorithms do in PASCAL). However, as the other three sides, face is occluded. Totally 9 type views are designed in SPID: front/back side, left/right side, left-front/right-front side, left-back/right-back side, and others. The distribution statistics of the view in SPID can be seen in Fig. 3(b). The division and label of this property may help to design some other features for this condition.

Occlusion. Occlusion influences the accuracy of pedestrian detection significantly [8]. Caltech dataset is the only existing pedestrian dataset that labeled 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 in SPID: Occ = none, Occ < 33%, Occ in [33%–66%], Occ > 66%, others.



(a) Illumination (b) View

(c) Occlusion (d) Attachment

Fig. 3. Statistics of different properties in SPID data set.

The levels are measured by human subjectively when labeling. The distribution statistics of the occlusion in SPID can be seen in Fig. 3(c).

Attachment. In many times, one pedestrian will hold one attachment (like bag, bike, etc.) with him, and the attachment may change the contour of one pedestrian. This will result the decrease of detection performance for the algorithms trained using dataset without any pedestrian-with-attachment. In SPID, the property of with or without attachment is labeled to help for the training of pedestrian detection methods. The common attachments of pedestrians contain hat, glasses, scarf, bag, dunnage, luggage and others. The distribution statistics of the attachment in SPID can be seen in Fig. 3(d).

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 SPRD: Surveillance Person Re-Identification Dataset

4.1 Description

The SPRD (Surveillance Person Re-Identification Dataset) contains 9700 pedestrian images taken from arbitrary viewpoints from 24 real surveillance cameras, also under varying illumination conditions. Each sequence contains about 200 to 300 images from the same person. For each camera view, we record continuous human sequences to facilitate multi-shot person Re-ID. The naming rule of this dataset is as follows. Each folder contains images from a unique person. They might include different camera views. The number of images of the same person varies. The naming rule for each image is as follows. For example, image file "057_02_0001_00562.jpg" denotes camera ID 57, person ID 2, sequence number 1, and frame ID 562. We follow a multi-shot person Re-ID setting. Namely, instead of using only a single image, sequence of images can be used together to represent a person. Basically, a pair of person image sequences of the used for training should be from different viewing cameras (this can be read from the camera ID).

Compared with existing datasets for the purpose of person re-identification, such as iLIDS-VID [17] and PRID2011 [18], the proposed SPRD dataset possesses the following advantages. First, it contains more view variations than other datasets. In particular, the number of viewpoints (cameras) used for capturing one person is usually more than ten. In contrast, previous dataset only contains person images from a limited number of viewpoints, e.g., fewer than three. Second, the proposed SPRD dataset is annotated with part bounding boxes, which makes part-matching based person re-identification algorithms accessible. Last, the proposed dataset has longer sequences than most existing person re-identification datasets, which enables multi-shot based person re-identification algorithms. Detailed comparisons are given in Table 3.

Table 3. Comparisons of various person re-identification datasets.

Dataset	# Camera views	Has part annotation
iLIDS-VID [17]	2	No
PRID [18]	2	No
SPRD	10+	Yes

5 SHAD: Surveillance Human Action Dataset

5.1 Description

The task of surveillance action recognition is to recognize human actions in surveillance videos. Monitoring and understanding surveillance human actions play an important role in smart video surveillance applications. Many surveillance action recognition algorithms have been proposed in recent years; however, most of existing human action datasets used to evaluate proposed algorithms are constructed using self-collected video, not the real on-using surveillance data. Thus, in the paper we introduce the Surveillance Human Action Dataset (SHAD) which is collected from real surveillance videos. The dataset contains 290 high definition video clips recorded by 25 surveillance cameras and selected from about 47 h of video data. Six human action classes that occur frequently in real surveillance are included: walking, sitting, bending, squat, falling and cycling (Table 4). Each action is performed by large number of people and recorded with both cluttered background and varying illumination. Additionally, we provide annotations in which distinct people ID are assigned and bounding boxes are annotated to localize each people in the frame. Meanwhile, for some people the visible human key points are annotated. To the best of our knowledge, SHAD is currently the most challenging dataset for surveillance human action recognition and we hope the release of the dataset will help accelerate the research progress of the field. Comparisons between proposed SHAD and existing human action datasets is given in Table 5.

Table 4. Summarizations of labeled actions for multi-action recognition tasks.

Clip label	Clip number	Number of labeled actions						Action number
		Walking	Sitting	Bending	Squat	Falling	Cycling	
Walking	50	51	1	0	0	1	0	53
Sitting	40	57	40	5	0	0	12	114
Bending	50	49	0	50	0	0	1	100
Squat	50	96	0	1	50	0	3	150
Falling	50	35	0	8	1	50	0	94
Cycling	50	19	0	0	0	0	52	71
Total	290	307	41	64	51	51	68	582

Table 5. Comparisons between proposed SHAD and existing human action datasets.

	KTH [7]	Weizmann [11]	TRECVID [13]	VIRAT [14]	SHAD
Number of action types	6	10	10	23	6
Avg. number of samples per class	100	9	3–1670	10–1500	40–50
Resolution	160 × 120	180 × 144	720 × 576	1920 × 1080	1920 × 1080
Human height	80–100	60–70	20–200	20–180	7–536
Human to video height ration	65–85%	42–50%	4–36%	2–20%	0.65–49.6%
Number of scene	N/A	N/A	5	7	20
Natural background clutter	No	No	Yes	Yes	Yes
Bounding boxes	Cropped	Cropped	No	Yes	Yes
Labeled human ID	No	No	No	No	Yes
Keypoint annotations	No	No	No	No	Yes

5.2 Dataset Properties

Scene. The dataset contains video clips coming from 20 surveillance cameras. For each camera, it captures different scenes. The backgrounds are cluttered with campus buildings, trees, bicycles and so on. As mentioned above, the illumination also varies due to the large record duration.

Viewpoints. Due to the fact that people are not required to perform actions in restricted areas, action performers show various viewpoints, such as front side, back side, left side or right side, towards the cameras as shown in Fig. 5.

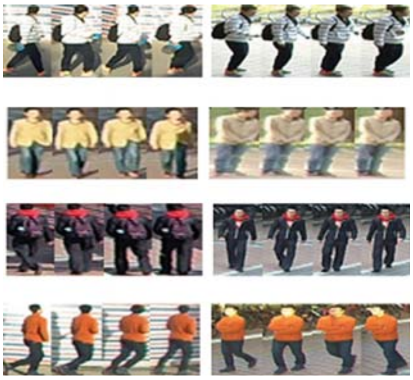
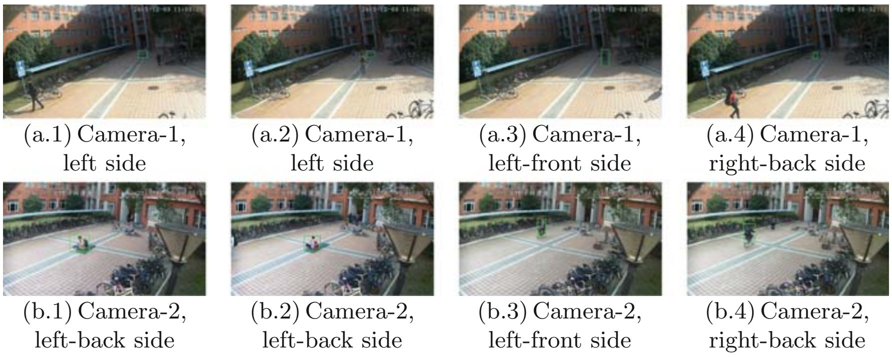


Fig. 4. Examples of pedestrians in SPRD.

Meanwhile, some surveillance cameras are close to each other, which makes the same action recorded by different cameras and captured in different viewpoints.

Scale. Because the surveillance cameras can cover wide range of areas, the heights of people in our dataset vary dramatically, which can be better visualized by Fig. 6 which depicts the distribution of people heights in our dataset (Fig. 6).



(a.1) Camera-1, left side

(a.2) Camera-1, left side

(a.3) Camera-1, left-front side

(a.4) Camera-1, right-back side

(b.1) Camera-2, left-back side

(b.2) Camera-2, left-back side

(b.3) Camera-2, left-front side

(b.4) Camera-2, right-back side

Fig. 5. An example of various viewpoints shown in two different cameras. The target human is bounded by green rectangle box. The rest two columns are examples that the same action is captured by these two cameras.

5.3 Annotations

To give benefits for SHAD dataset users, we also provide some annotations for this dataset. However, annotating a large video dataset presents a challenge on its own. For one video clip, two types of annotations are included. The first annotation is frame-level: for certain frame in one clip, bounding boxes are applied to localize all the visible people in the current frame and each people is assigned with a unique people ID. The people ID for the same people throughout the whole clip will be kept the same which make it possible to keep the trajectory of each people. The second annotation is human-level, which aims at providing a more detailed description on people himself or herself. For example, the locations of visible human key-points (like shoulder, elbow, wrist and so on) are annotated because the human pose can be well represented by the configuration of human key-points. To be specific, there are 13 key-points included: head, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle and right ankle.

6 Performance Evaluation of Pedestrian Detection Using SPID

With the remarkable data difference between the surveillance images and non-surveillance datasets, the performance of intelligent algorithms can vary significantly when applying them into the BEST directly. In this section, performance evaluation of 7 popular pedestrian detection methods using existing non-surveillance dataset and the surveillance pedestrian datasetSPID, is given to show the variance of detection performance.

ROC curves are the most used evaluation methodology to use compare the performance of different PD algorithms [9]. 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 (BBgt) and detected bounding boxes (BBdt) are loaded. The aspect ratio of all the boxes is preprocessed to 0.41 and the overlap threshold of bounding boxes is set to 0.5. Specific accuracy calculation method is the same as that in [19].

The performance comparison of six existing pedestrian detection methods using hand-crafted features: k-poselets [20], ACFCaltech+ (ACF [21] retrain using Caltech dataset), LDCFCaltech (LDCF [22] retrain using Caltech dataset), ACFInria, LDCFInria and DPMInria (DPM [23] retrain using Inria dataset) and one deep-learned-feature based method (Faster R-CNN [24].), is shown in Fig. 7. From Fig. 7, its easy to find that the detection performance decreasing significantly when applying the existing methods on SPID directly. For the performance of these method retrained using SPID dataset (DPMSpid, Faster R-CNNSpid), the retrained methods using SPID or Caltech cant work very well on surveillance images (see the DPMSpid and Faster R-CNNSpid lines in Fig. 7). 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. The detail of this evaluation can be seen in [25] (Fig. 7).

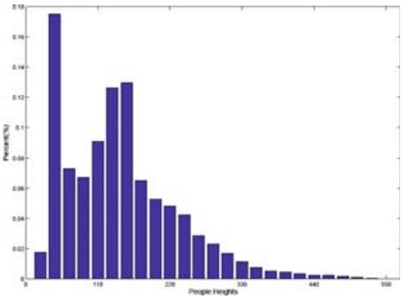
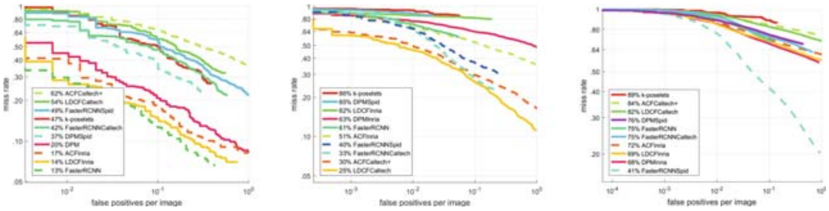


Fig. 6. The distribution of human heights SHAD. A wide range of heights is covered and the majority of heights are less than 220 pixels.



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

Fig. 7. Evaluation results under reasonable condition on three test datasets. (a), (b) and (c) shows the pedestrian detection results on INRIA, Caltech, and SPID, respectively.

7 Conclusion

This paper introduces the BEST platform: Benchmark and Evaluation of Surveillance Task, which contains multiple task-driven datasets using videos and images captured by on-using surveillance systems. The images in BEST are well labeled, and some popular intelligence algorithms, such as pedestrian detection methods, including hand-crafted detectors and deep learning methods, are evaluated using the existing datasets (such as INRIA, Caltech) and the proposed BEST dataset 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. This benchmark encapsulates a rigorous and comprehensive academic benchmarking effort for testing and evaluation existing and new algorithms for surveillance tasks. It will be revised/expanded from time to time based on received feedback, and will maintain a comprehensive ranking of submitted methods for years to come.

Acknowledgement. This work was partly funded by NSFC (No. 61571297, No. 61371146, No. 61527804, 61521062), 111 Program (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. Shu, C.F., Hampapur, A., Lu, M., Brown, L., Connell, J., Senior, A., Tian, Y.: IBM smart surveillance system (s3): a open and extensible framework for event based surveillance. In: IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 318–323 (2005)

2. Hampapur, A., Brown, L., Connell, J., Pankanti, S.: Smart surveillance: applications, technologies and implications. In: Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia, pp. 1133–1138 (2004)
3. Hu, W., Tieniu, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. **34**(3), 334–352 (2004)
4. 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)
5. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
7. Ess, A., Leibe, B., Van Gool, L.: Depth and appearance for mobile scene analysis. In: IEEE International Conference on Computer Vision, pp. 1–8 (2007)
8. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: survey and experiments. IEEE Trans. Pattern Anal. Mach. Intell. **31**(12), 2179–2195 (2009)
9. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: a benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 304–311 (2009)
10. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The kitti vision benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361 (2012)
11. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: 17th International Conference on Proceedings of the Pattern Recognition, (ICPR 2004), vol. 3, pp. 32–36 (2004)
12. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Action as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. **29**(12), 1395–1402 (2005)
13. http://www.cvg.reading.ac.uk/PETS2016/a.html/
14. Oh, S., Hoogs, A., Perera, A., Cuntoor, N.: A large-scale benchmark dataset for event recognition in surveillance video. In: Proceedings of IEEE Computer Vision and Pattern Recognition, pp. 3153–3160 (2011)
15. Over, P., Awad, G.M., Fiscus, J.G., Antonishek, B., Michel, M., Kraaij, W., Smeaton, A.F., Qunot, G.: TRECVID 2015 an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID 2015. NIST, USA (2015)
16. http://www.changedetection.net/
17. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 688–703. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_45
18. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian Conference on Image Analysis, pp. 91–102 (2011)
19. Dollr, 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)
20. Gkioxari, G., Hariharan, B., Girshick, R., Malik, J.: Using k-poselets for detecting people and localizing their keypoints. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3582–3589 (2014)