



Nighttime FIR Pedestrian Detection Benchmark Dataset for ADAS

Zhewei Xu¹, Jiajun Zhuang², Qiong Liu¹(✉), Jingkai Zhou¹, and Shaowu Peng¹

¹ School of Software Engineering, South China University of Technology,
Guangzhou 510006, China

{se_xuzhewei,201510105876}@mail.scut.edu.cn,

{liuqiong,swpeng}@scut.edu.cn

² College of Computational Science, Zhongkai University of Agriculture
and Engineering, Guangzhou 510225, China

zhuangjiajun@zhku.edu.cn

Abstract. Far infrared (FIR) pedestrian detection is an essential module of the advanced driver assistance system (ADAS) at nighttime. Recently, a wave of deep convolutional neural networks (CNN) has taken the visible spectrum pedestrian detection benchmarks top ranks. However, due to the lack of dataset, we could not evaluate the performance of CNN methods on FIR images. In this paper, we introduce a nighttime FIR pedestrian dataset, which is the largest nighttime FIR pedestrian dataset. The dataset contains fine-grained annotated video, recorded from diverse road scenes and we provide detailed statistical analysis. We selected three kinds of advanced pedestrian detection methods as the baseline and evaluated their performance. Benefit from training data volume and diversity, the experimental results show that CNN-based detectors obtained good performance on FIR image. We also propose three suggestions for improving performance, which reduces the average miss rate of the vanilla Faster R-CNN by 12.97% and 9.77% on KAIST and our dataset respectively. The dataset will be public online.

Keywords: FIR pedestrian detection
Convolutional neural networks · Dataset

1 Introduction

Far infrared (FIR) pedestrian detection is an essential module of the advanced driver assistance system (ADAS) [8, 12, 14, 28]. It aims to alert drivers about a possible collision with pedestrians, especially at nighttime scenes where it is hard to ensure enough illumination. Physically, pedestrians are more visible in FIR cameras than in visible spectrum cameras at night.

Recently, a wave of deep convolutional neural networks (CNN) [3, 16, 22, 23, 26] has taken the visible spectrum pedestrian detection benchmarks top ranks. Comparing with visible spectrum, pedestrian detection from FIR imagery present as objects with lower resolution and less texture information is still a

challenging problem. However, due to the lack of large nighttime FIR pedestrian detection dataset, we could not evaluate the performance of CNN methods on FIR images. Besides most of the presented FIR pedestrian detectors [18–20] using handcraft features give performance comparing on some small dataset, which often leads to some bias because the handcraft features depend on some prior knowledge from a specific dataset.

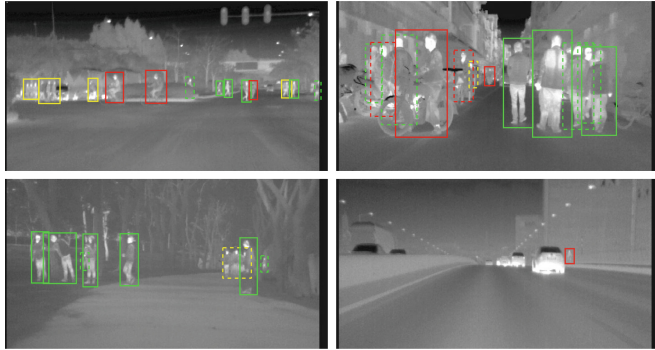


Fig. 1. Example images (cropped) and annotations. The solid green, red, yellow boxes denote ‘walk_person’, ‘ride_person’ and ‘people’ respectively. The dashed boxes denote occluded objects.

With this in mind, we proposed a large FIR pedestrian detection dataset to promote relativity study. The main contributions of this paper are: (1) we introduce a nighttime FIR pedestrian dataset, which is the largest nighttime FIR pedestrian dataset with fine-grained annotated video. Figure 1 shows some examples of this dataset. The pedestrians vary widely in appearance, pose and scale. (2) We provide a detailed statistical analysis of the dataset. (3) Benefit from training data volume and diversity, the experiment results show that CNN-based detectors obtained good performance on FIR image. (4) We also propose three modifications for improving CNN-based detector performance, which reduces the average miss rate of the vanilla Faster R-CNN by 12.97% and 9.77% on KAIST and our dataset respectively.

This paper is organized as follows. Section 2 briefs the related work. Section 3 introduces SCUT dataset and carries out corresponding statistical analysis. Section 4 introduces three kinds of advanced pedestrian detection methods and a modified Faster R-CNN fitted for FIR pedestrian detection. In Sect. 5, we report the performance evaluation results by experiments under several different conditions on SCUT dataset. The final is the summary of this work.

2 Related Work

2.1 Pedestrian Detection Datasets

Several early pedestrian datasets in visible spectrum include INRIA [4], ETH [11], TudBrussels [24], and Daimler [10]. But they are superseded by larger

and richer datasets such as Caltech [8] and KITTI [12]. Recently, Zhang et al. [28] proposed a new diverse dataset namely CityPersons which make CNN show strong generalization ability as training set.

Table 1. Comparing pedestrian datasets. The horizontal lines divide the datasets based on the image types (e.g. color, thermal, and color-thermal). The first four columns indicate pedestrian number and image number in training and testing dataset ($k = 10^3$). Properties column summarizes some dataset characteristics.

	Training		Testing		Properties						
	# pedestrians	# images	# pedestrians	# images	# total frames	color	thermal	occ. labels	videos	moving cam	publication
Caltech [9]	192k	128k	155k	121k	250k	✓		✓	✓	✓	'09
KITTI [12]	12k	1.6k	-	-	80k	✓		✓	✓	✓	'12
CityPersons [28]	20k	3k	11k	1.6k	5k	✓		✓		✓	'17
KAIST [14]	42k	50k	45k	45k	95k	✓	✓	✓	✓	✓	'15
CVC [13]	4.8k	3.5k	4.3k	1.4k	5k	✓	✓	✓	✓	✓	'16
OSU-T [5]	984	1.9k	-	-	2k		✓		✓		'05
LSI [1]	10.2k	6.2k	5.9k	9.1k	15.2k		✓		✓	✓	'13
TIV [25]	-	-	-	-	63k		✓		✓		'14
SCUT(Our)	175k	108k	177k	103k	211k		✓	✓	✓	✓	'18

There are few of available FIR pedestrian detection datasets. OSU-T [5] is probably the first benchmark dataset. But the images were captured by a static camera mounted on a building in campus. Hence, it lacks diversity and reasonable background clutter and not suitable for on-road pedestrian detection. LSI dataset [1] is captured from a vehicle under different illumination and temperature scenes, but contains only 15K image frames with a low resolution. TIV dataset [25] provides a high-resolution FIR images with rich annotation of person (e.g. walk person, running person and bicyclist) and other scenario participants (e.g. vehicles and motorbikes). It is more appropriate for the task in video monitoring scene. Recently, multispectral (color-thermal) datasets [13, 14] are proposed for all day pedestrian detection. But only 30K frames and 37K bounding boxes (BB) recorded from nighttime. Table 1 provides an overview of above datasets.

2.2 Pedestrian Detection Evaluating

For improving on-road FIR pedestrian detection, the early work mainly focuses on various reformative hand-crafted features [14, 18–20]. Olemda et al. [19] propose phase congruency feature to resist illumination change, Liu et al. [18] design pyramid entropy weighted HOG to highlight object profile and Qi et al. [20] adopt sparse representation to get rich semantic context. Following the success of integral channel features, Hwang et al. [14] proposed multispectral ACF in which intensity and HOG are used as extend thermal feature channels.

Recently, multifarious convolutional neural network variants achieve top ranks on Caltech benchmark [3, 16, 22, 23]. Most of them are custom architectures

derived from Faster R-CNN. Cai et al. [3] propose a multi-scale CNN (MSCNN) with a multiple output layers proposal sub-network, so that receptive fields can match the objects of different scales. Li et al. [16] propose scale-aware CNNs (SA-FastRCNN), capturing features for pedestrians of different image sizes by a scale gate function. Zhang et al. [26] proposed RPN+BF, combining RPN (the first part of Faster RCNN) with a following boosting forest. RPN generates region proposals, confidence scores and features, all of which are used to train a cascaded Boosted Forest classifier (BF). The bootstrapping strategy used in BF largely promotes pedestrian detection accuracy. The result of RPN+BF on Caltech reaches 9.6% log-average miss rate.

Further, Liu et al. [17] proposed a multispectral detector built upon Faster R-CNN and performed 37% miss rate on KAIST. König et al. [15] proposal RPN+BDT classifier for reducing potential false positive detection. As baseline model, it is hopeful to adapt Faster R-CNN for on-road FIR pedestrian detection as well.

3 SCUT FIR Pedestrian Dataset

We introduce SCUT dataset as a benchmark of on-road FIR pedestrian detection for researcher and engineer of this field. The image sequences are collected from several driving scenarios over one month in Guangzhou, China. A fine-grained set of high-quality annotations and corresponding statistics are presented. SCUT dataset is highlighted in data volume, data diversity and a wide range of imaging distance. Besides, the work of this paper has excellent practical significance to boost ADAS or intelligent vehicle in China because Chinese road traffic occupies a large market share and the road environment is more complex than some other countries.

3.1 Data Collection and Annotation

Data Capture. Image sequences in SCUT dataset are captured by a monocular FIR camera mounted on a car (Fig. 2). The spatial resolution of the camera is 384×288 with 13 mm focal length, and the field of view is $28^\circ \times 21^\circ$. The output resolution is resized to 720×576 pixels by an image acquisition card for better observation and annotation. We collect about 11 hours-long image sequences ($\sim 10^6$ frames) at 25 fps by a vehicle driving through diverse traffic scenarios at



Fig. 2. Camera setup

a speed less than 80 km/h. The driver is independent of the authors. The image sequences all include 11 road sections under 4 kinds of scenes, i.e. downtown, suburbs, campus and expressway (Fig. 1).

Ground Truth Annotation. Piotr’s Computer Vision Toolbox [6] is adopted to annotate ground truth for pedestrian in a image frame. If a pedestrian or person group is visible, a tight bounding box (BB) is drawn around the object. For occluded pedestrian, a BB involves estimating the location of hidden parts. Among all, we annotated 211,011 frames for a total of 477,907 BBs around 7,659 unique pedestrians. Newly, an annotation protocol [27] is presented by drawing a center line from head to the central point between both feet and then generate a BB with a fixed aspect ratio. Although this procedure ensure the BB is well centered on the subject, it may also lose some parts of the limbs.

SCUT dataset provides a set of fine-grained labels to divide all BBs into six categories by following rules. An individual person when walking, running or standing posture is labeled as ‘walk_person’. An individual person when sitting or squatting is labeled as ‘squat_person’. An individual person when riding bicycle or motorbike is labeled as ‘ride_person’. A person group that are hard to distinguish each other is labeled as ‘people’. In addition, an individual person and a person group who is ambiguous or occluded area $>2/3$ are labeled as ‘person?’ and ‘people?’ respectively.

Training and Testing Data. The annotated image sequences randomly divided into training and testing dataset. There are 21 subsets, each video recorded in one of 11 road section. We divide the data randomly in half, S0~S10 as training set, S11~S20 as testing set. The total number of both image frames and BBs in each dataset is similar. Details about the amount of SCUT training/testing dataset can be seen in Table 1.

Table 2. Dataset summary

Label	Frames with anno.	BB	Occluded	Unique	Avg. frames per obj
walk person	92,278	193,765	57815	3,136	61.79
ride person	83,672	157,994	17386	1,824	86.62
squat person	50,483	71,930	18708	1,259	57.13
people	30,267	39,303	15702	1,138	34.54
person?	10,254	12,470	4061	250	49.88
people?	2,330	2,445	508	36	67.92
Summary	148,132	447,907	114180	7,659	62.40

3.2 Dataset Statistics

The statistics on SCUT dataset is discussed here based on Table 2. As a whole, about 70% frames have at least one BB annotation and about 78% BBs contain walk person or ride person, who must be detected in ADAS-oriented applications. A walk person appears 2.47 seconds and a ride person appears 3.46 seconds averagely. We analyze further some sub-theme, i.e. scale and distance, data diversity, pedestrian occlusion and pedestrian position etc. The statistics may be basic supporting when establishing a road FIR pedestrian detection system.

Scale and Distance. Similar to Dollár et al. [9], we group pedestrians by dividing the pixel height of BBs into three scales: near (more than 80 pixels), medium (30~80 pixels) and far (less than 30 pixels). The statistics of histogram distribution in the pixel height of BBs is investigated respectively on walk person

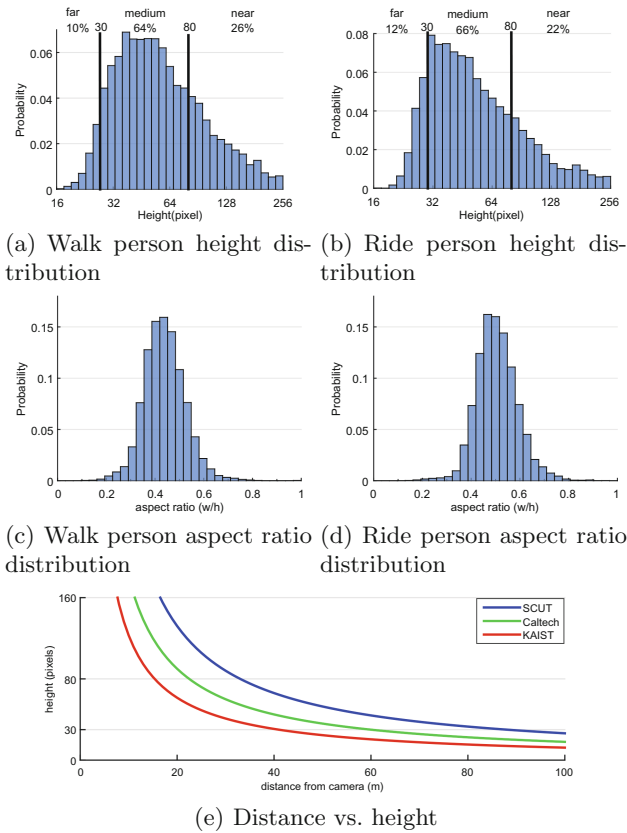


Fig. 3. (a) and (b) Distribution of walk person and ride person pixel heights respectively. (c) and (d) Distribution of walk person and ride person aspect ratio respectively. (e) Pixel height h as a function of distance d .

and ride person, as shown in Fig. 3(a) and (b), which is similar each other. Cut-off for near/far scale is marked respectively. Most observed walk person ($\sim 64\%$) and ride person ($\sim 65\%$) lie in medium scale. At far distance region, the number of pedestrian decreases sharply because it is difficult to identify reliably a small pedestrian. Furthermore, the statistics of BBs aspect ratio histogram is shown in Fig. 3(c) and (d) respectively on walk person and ride person. The log-average aspect ratios of walk person, ride person and the both are 0.43, 0.50 and 0.46 respectively.

Medium and far pedestrian is usually more important than near because it is necessary to have enough reaction time for a driver when alerting him to a possible collision. The focal length in pixels of our FIR camera is 1554 (due to $576/2/f = \tan(21^\circ/2)$). Using a pinhole camera model, an object observed pixel height h a pedestrian observed is inversely proportional to the distance d from the camera: $h \approx Hf/d$, where H is a true height of a pedestrian. Assuming $H \approx 1.7$ m, we gain $d \approx 2641.8/h$ m. Figure 3(e), compare the relationship between the pixel height of a pedestrian and the corresponding distance in meter on SCUT dataset, KAIST, and Caltech.

Diversity. The data diversity on SCUT dataset and KAIST can be seen in Table 3. Comparing with KAIST (4 labels and 103k BBs), we provide fine-grained data category labels and a larger number of BBs (6 labels and 448k BBs), which is the first difference of data diversity.

Following common practice for a pedestrian, the minimum high is 20 pixels and the maximum high is the image resolution [9]. Due to the camera with a longer focal length, the distance range from the camera is 4.6~132 m on SCUT dataset, but only 2.4~61 m on KAIST. So, the sampling space on SCUT dataset is around two times larger than that of KAIST, which is the third difference of data diversity.

In addition, the image sequences are collected from 11 different road sections under four kinds of scenes on SCUT dataset, but only from 3 different road scenes on KAIST, which is the fourth difference of data diversity. The last difference of data diversity works in that total number of frame is 211k on SCUT but only 95k on KAIST.

Table 3. Comparison of data diversity on SCUT and KAIST datasets

	KAIST	SCUT
# frame	95k	211k
# label	4	6
# bounding box	103k	448k
# unique person	1182	7659
# pedestrian distance	2.4 m~61 m	4.6 m~132 m
# road scene	3	4

Pedestrian Occlusion. Since a camera is usually at a horizontal perspective in a traffic scene, a pedestrian may be occluded by another pedestrian or object. We individualize every BB by adding an attribute tag for occlusion. An unoccluded object (person or people) is tagged as ‘no occlusion’. An occluded object is tagged as ‘occlusion’. Among all, $\sim 25\%$ BBs are marked as ‘occlusion’. Also, among all walk persons, the occluded BB accounts for $\sim 30\%$ while occluded BB only accounts for $\sim 11\%$ in ride persons. A walk person could be occluded by trees, parking cars, and another pedestrian when appearing on a sidewalk. The ratio of walk person occluded is larger than the ride person because there are a lot of walk persons on a sidewalk, as shown in Fig. 1.

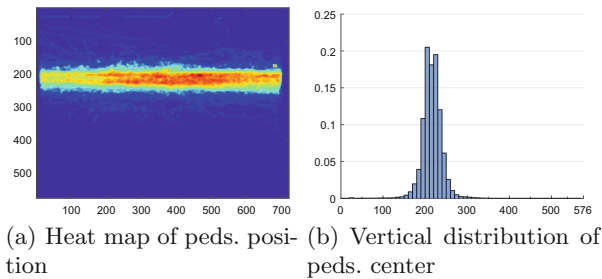


Fig. 4. (a) Center position heat map of pedestrian (walk person and ride person) BBs, which are log-normalized. (b) Vertical center position histogram of pedestrian BBs.

Pedestrian Center Position. Figure 4(a) shows log-normalized heat map to annotate pedestrian center position. Viewpoint and ground plane geometry constrain a pedestrian appearing only in a narrow band running horizontally across the center of the image. Because the vehicle drives under the right-handed traffic condition, more pedestrians appear on the right side of the image. Corresponding pedestrian vertical (y-coordinate) distribution is shown in Fig. 4(b). The average y-coordinate is 216, and about 97% of the pedestrians are in the range of 166 – 266 y-coordinate.

4 Experiments Results

According to the baseline verified in Sect. 3, we benchmark SCUT dataset. In Sect. 4.1, we evaluate performance under different conditions using SCUT dataset. Next, we report the impact of training data volume in Sect. 4.2.

We employ the evaluation strategy proposed by Dollar et al. [8]. It stated a detected BB (BB_{dt}) and a ground truth BB (BB_{gt}) have an IoU ratio ≥ 0.5 . And each BB_{dt} and BB_{gt} may be matched at most once and the ignored BB_{ig} need not to be matched. In our dataset, six types of BBs are always set to ignore: any BB under 20 pixels high or truncated by image boundaries, containing a ‘person?’, ‘people?’, ‘people’ or ‘squat_person’. Detections within these regions will not affect performance.

Table 4. Step by step improvements on KAIST FIR image from vanilla Faster R-CNN to TFRCN, we gain 12.97 MR points total.

Detector aspect	MR	Δ MR
vanilla Faster R-CNN	56.01	-
+ M1 Handling class imbalance	49.20	6.81
+ M2 Adjusting anchors distribution	48.55	0.65
+ M3 Reducing the feature stride	43.04	5.51
TFRCN-ours	43.04	12.97

4.1 Baseline Detectors

To evaluate the training and testing effect for the benchmark, we select three promising pedestrian detectors as baselines, i.e., ACF-based [2], Faster R-CNN [21] and RPN+BF [26]. ACF represents the Integral Channel Feature detector family [7]. Faster R-CNN stands for the CNN-based detector. RPN+BF acts like a typical hybrid detector, which achieves the state-of-the-art performance on pedestrian detection. As vanilla Faster R-CNN failing to handle class imbalance and small pedestrians, we propose a modified Faster R-CNN model to promote performance. We show the step-by-step improvements in Table 4 on the KAIST FIR datasets under the ‘reasonable’ setting [14]. In total, the MR of TFRCN reduces from 56.01% to 43.04% on KAIST dataset.

4.2 Benchmarking on SCUT Dataset

We sampled images from train subset videos with 2-frame skips (53976 images) for training CNN-based detector and sampled 75-frame skip (1413 images) for training ACF-T+ detector family. The BB_{gt} for training exclude occluded, truncated and small pedestrian (<50 pixels). We plot the miss rate using a per-image evaluation scheme (FPPI) and summarize the performance with a single value by using log-average miss rate (a short for MR) over the range of $[10^{-2}, 10^0]$. In all experiments, the image interval of the testing dataset is set to 25 frames. Figure 5 shows the evaluation results for the various subsets of the test set described below.

Reasonable. For this experiment, we use a representative subset of the proposed dataset, named reasonable all. The reasonable all subset (Fig. 5(a)) consists of walk person and ride person which are taller than 50 pixels. This subset is also divided into reasonable walk person and reasonable ride person subsets based on the label. In Fig. 5(a-c), RPN+BF perform best and TFRCN achieves the second rank. We noticed that the CNN-based detectors on ride person subset (Fig. 5(c)) perform better than walk person (Fig. 5(b)). According to Sect. 2.2, we believe this is due to there are more occluded walk persons than ride persons.

Scale. Reasonable subset doesn’t cover small pedestrian. As discussed in Sect. 2.2, we have enough high-quality data to group pedestrians by pixel height

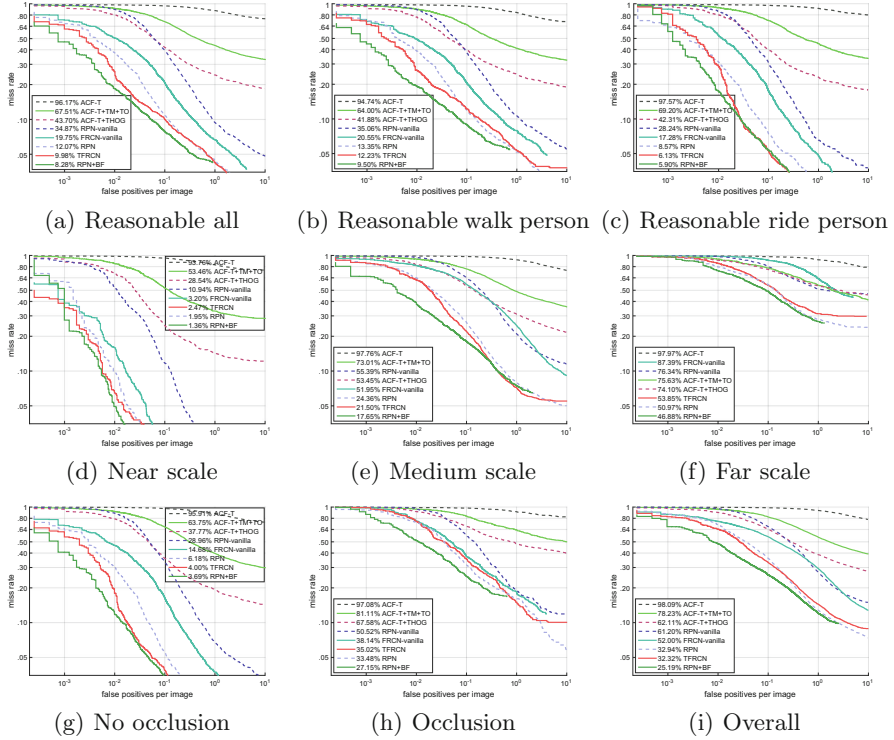


Fig. 5. False positive per image (FPPI) versus miss rate in various conditions.

into the near (80 or more pixels), medium (30–80 pixels) and far (30 pixels or less) scales. Results for each scale, on unoccluded pedestrians only, are shown in Fig. 5(d)–(f). In general, as the height of a pedestrian gets smaller, the performance degrades dramatically. FRCN-vanilla and RPN-vanilla are more sensitive to scale. At the far scale, there are more negative samples and smaller convolutional feature size. With our modification, TFRCN achieves 33.54% MR improvement.

Occlusion. The impact of occlusion on detecting pedestrians with a minimum height of 50 pixels is shown in Fig. 5(g)–(h). For this experiment, we build two subsets based on the occlusion tags: no-occlusion, occlusion (0~2/3 of area occluded). Performance drops significantly under occlusion situation in all detectors.

Overall. In this experiment, we examine detectors on entire dataset in Fig. 5(i). It is most similar to the real driving scene. RPN+BF outperforms the remaining methods. With our modification, TFRCN achieves the second rank, with a MR of 32.32%. FRCN-vanilla and ACF-T+ family detector performance are poor, with a MR of over 50%.

Summary. CNN-based detectors achieve better performance. Benefiting from bootstrapping strategy, the RPN+BF excels the other methods. ACF-T+ family detectors perform poorly on all subset. FRCN-vanilla and RPN-vanilla likely suffer from feature size and unfit anchor setting. With our modification, TFRCN gets better performance than RPN, except for two special cases (far scale and occlusion).

5 Summary

In this paper, we introduced a large FIR pedestrian detection dataset, which was collected from several driving scenarios over one month in Guangzhou, China. The dataset provides a large-scale of fine-grained annotations with high-level data diversity. As we know, SCUT dataset is the largest FIR dataset providing occlusion labels and temporal correspondences captured from non-static real traffic scenes.

After the detailed analysis of the basic statistics of SCUT dataset, we evaluated the performance using ACF-based and CNN-based models and benchmarked several promising CNN-based detectors. The experimental results demonstrated that the CNN-based detectors achieved better performance on SCUT dataset, especially provided with a larger training set of high-level diverse examples. We expect that the proposed dataset can promote the development of FIR detection method.

Acknowledgement. This paper are supported by the Science and Technology Planning Project of Guangdong Province (2017A020219008), the Project on the Integration of Industry, Education and Research of Guangdong Province (2017B090901047) and the Science, and the Technology Program of Guangzhou Province (201607010069).

References

1. Pedestrian classification and detection in far infrared images. *Integr. Comput. Aided Eng.* **20**(4), 347–360 (2013)
2. Fast feature pyramids for object detection. *IEEE Trans. Patt. Anal. Mach. Intell. (TPAMI)* **36**(8), 1532–1545 (2014)
3. Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *ECCV 2016*. LNCS, vol. 9908, pp. 354–370. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_22
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *CVPR*, pp. 886–893. IEEE Press (2005)
5. Davis, J.W., Keck, M.A.: A two-stage template approach to person detection in thermal imagery. In: *IEEE Workshops on Application of Computer Vision*, pp. 364–369. IEEE Press (2005)
6. Dollár, P.: <http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html>
7. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: *The British Machine Vision Conference (BMVC)*, pp. 91.1–91.11. BMVA Press (2009)

8. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. *IEEE Trans. Patt. Anal. Mach. Intell. (TPAMI)* **34**(4), 743–761 (2012)
9. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: a benchmark. In: *CVPR*, pp. 304–311. IEEE Press (2009)
10. Enzweiler, M., Gavrilu, D.M.: Monocular pedestrian detection: survey and experiments. *IEEE Trans. Patt. Anal. Mach. Intell. (TPAMI)* **31**(12), 2179–2195 (2009)
11. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: *CVPR*, pp. 1–8. IEEE Press (2008)
12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the Kitti vision benchmark suite. In: *CVPR*, pp. 3354–3361. IEEE Press (2012)
13. González, A., et al.: Pedestrian detection at day/night time with visible and FIR cameras: a comparison. *Sensors* **16**(6), 820 (2016)
14. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: benchmark dataset and baseline. In: *CVPR*, pp. 1037–1045. IEEE Press (2015)
15. König, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., Teutsch, M.: Fully convolutional region proposal networks for multispectral person detection. In: *CVPR Workshops*, pp. 243–250. IEEE Press (2017)
16. Li, J., Liang, X., Shen, S., Xu, T., Feng, J., Yan, S.: Scale-aware fast R-CNN for pedestrian detection. *IEEE Trans. Multimedia* **20**(4), 985–996 (2018)
17. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. In: *The British Machine Vision Conference (BMVC)*, pp. 1–13. BMVA Press (2016)
18. Liu, Q., Zhuang, J., Ma, J.: Robust and fast pedestrian detection method for far-infrared automotive driving assistance systems. *Infrared Phys. Technol.* **60**, 288–299 (2013)
19. Olmeda, D., de la Escalera, A., Armingol, J.M.: Contrast invariant features for human detection in far infrared images. In: *IEEE Intelligent Vehicles Symposium*, pp. 117–122. IEEE Press (2012)
20. Qi, B., John, V., Liu, Z., Mita, S.: Pedestrian detection from thermal images: a sparse representation based approach. *Infrared Phys. Technol.* **76**, 157–167 (2016)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: *NIPS*, pp. 91–99. Curran Associates, Inc. (2015)
22. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: *CVPR*, pp. 5079–5087. IEEE Press (2015)
23. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: *ICCV*, pp. 1904–1912. IEEE Press (2016)
24. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: *CVPR*, pp. 794–801. IEEE Press (2009)
25. Wu, Z., Fuller, N., Theriault, D., Betke, M.: A thermal infrared video benchmark for visual analysis. In: *CVPR*, pp. 201–208. IEEE Press (2014)
26. Zhang, L., Lin, L., Liang, X., He, K.: Is faster R-CNN doing well for pedestrian detection? In: *Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS*, vol. 9906, pp. 443–457. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_28
27. Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: *CVPR*, pp. 1259–1267. IEEE Press (2016)
28. Zhang, S., Benenson, R., Schiele, B.: CityPersons: a diverse dataset for pedestrian detection. In: *CVPR*, pp. 3213–3221. IEEE Press (2017)