Infrared Physics and Technology 96 (2019) 199–208

Contents lists available at ScienceDirect

Infrared Physics & Technology

journal homepage: www.elsevier.com/locate/infrared

Regular article

Benchmarking a large-scale FIR dataset for on-road pedestrian detection

Zhewei Xu[a], Jiajun Zhuang[b], Qiong Liu[a,*], Jingkai Zhou[a], Shaowu Peng[a]

[a] School of Software Engineering, South China University of Technology, Guangzhou 510006, China
[b] College of Computational Science, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China

ARTICLE INFO

Keywords:
FIR pedestrian detection
Faster R-CNN
Convolutional neural networks
Dataset

ABSTRACT

Far infrared (FIR) pedestrian detection is an essential module of advanced driver assistance systems (ADAS) at nighttime. Although recent deep learning-based detectors have achieved excellent results on visible images in the daytime, their performance on nighttime FIR images is still unidentified, due to the existing nighttime FIR data set is not sufficient to fully train a deep learning detector. To this end, a nighttime FIR pedestrian dataset with the largest scale at present is introduced in this paper, which is called SCUT (South China University of Technology) dataset. The dataset contains fine-grained annotated videos recorded from variable road scenes. In addition, a detailed statistical analysis of the dataset is provided and four representative pedestrian detection methods are evaluated. Benefit from the volume and diversity of training data, the experiment results show that convolutional neural networks (CNN) based detectors obtained good performance on FIR image. To further explore the performance of pedestrian detection on FIR images, four important modifications on Faster R-CNN were studied and a strong baseline for SCUT dataset was proposed, which achieves the best detection result by reducing log-average miss rate from 36.78% to 17.73%. The dataset will be public online (SCUT dataset will be published on: https://github.com/SCUT-CV/SCUT_FIR_Pedestrian_Dataset).

1. Introduction

Far infrared (FIR) pedestrian detection is an essential module of the advanced driver assistance system (ADAS) [1–4]. It aims to alert drivers about a possible collision with pedestrians, especially for traffic scenes at nighttime where it is hard to ensure enough illumination. For instance, about 78 percent of fatal traffic accidents occur at nighttime [5], dawn or dusk due to poor lighting conditions. Physically, pedestrians under weakly illuminated scenarios can be captured with higher visibility using FIR imaging devices.

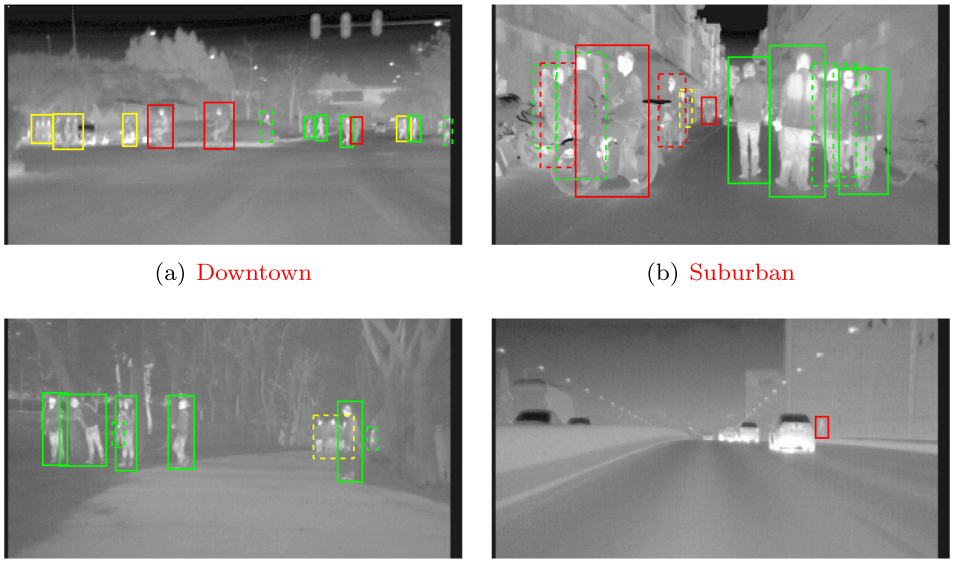
Recently, a wave of deep convolutional neural networks (CNN) based pedestrian detectors) [6–10] have achieved good performance on several high-quality pedestrian benchmarks in the visible spectrum. However, pedestrian detection in the infrared spectrum is still a challenging problem, probably due to two main reasons: (1) the low resolution of existing FIR pedestrian dataset providing less texture information, and (2) the lack of large-scale pedestrian dataset in infrared spectrum to ensure the training of deep learning-based detectors with good generalization performance. Therefore, it is hard to fully evaluate the performance of CNN-based methods on pedestrian detection in the infrared spectrum. In addition, the performance of most reported FIR pedestrian detectors [11–13] using handcraft features which were evaluated on some home-made datasets with small-scale, resulting in some bias because the handcraft features usually depend on some prior knowledge from a specific dataset.

To this end, a large-scale FIR pedestrian detection dataset called SCUT (South China University of Technology) dataset is proposed to explore the performance of representative CNN-based detectors on FIR images. The main contributions of this paper are as follows: (1) we introduce a FIR pedestrian dataset recorded at nighttime, which is the largest FIR pedestrian dataset with fine-grained annotated videos. As illustrated in Fig. 1, the pedestrians vary widely in appearance, pose and scale. (2) We provide a detailed statistical analysis of SCUT dataset and evaluate four representative pedestrian detection methods as the baseline. Benefit from the volume and diversity of training data, CNN-based detectors obtained good performance even with the initialized pre-training via ImageNet classification. (3) We analyze four important modifications on Faster R-CNN and propose a stronger baseline for SCUT dataset which reduces the log-average miss rate from 36.78% to 17.73%. Experimental results show that our modified Faster R-CNN detector achieves the best performance.

This paper is organized as follows. Section 2 briefs the related work. Section 3 introduces the SCUT dataset and carries out corresponding statistical analysis. Section 4 introduces three kinds of advanced

* Corresponding author.
E-mail addresses: zhuangjiajun@zhku.edu.cn (J. Zhuang), liuqiong@scut.edu.cn (Q. Liu), 201510105876@mail.scut.edu.cn (J. Zhou), swpeng@scut.edu.cn (S. Peng).

https://doi.org/10.1016/j.infrared.2018.11.007
Received 26 July 2018; Received in revised form 17 October 2018; Accepted 19 November 2018
Available online 20 November 2018
1350-4495/



(a) Downtown (b) Suburban

(c) Campus (d) Expressway

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.

pedestrian detection methods and a modified Faster R-CNN fitted for FIR pedestrian detection. In Section 5, we report the performance evaluation results by experiments under several different conditions on the SCUT dataset. The final section refers to conclusion and discussion of this work.

2. Related work

We focus on establishing a nighttime FIR pedestrian dataset and evaluating on-road pedestrian detection methods in this paper. The current situation on pedestrian detection datasets and CNN-based evaluating models are briefed in two parts.

2.1. Pedestrian detection datasets

Several visible spectrum pedestrian datasets have been proposed including INRIA [14], ETH [15], TudBrussels [16], and Daimler [17]. But they are superseded by larger and richer datasets such as Caltech and KITTI. Recently, Zhang et al. [18] proposed a new diverse dataset namely CityPersons which shows strong generalization ability of CNN.

There are few of available FIR pedestrian detection datasets. OSU-T [21] 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 [5] is captured from a vehicle under different illumination and temperature scenes, but contains only 15K image frames with a low resolution. TIV dataset [22] 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 more appropriate for the task in video monitoring scene. The KMU dataset [23] contains three types of FIR video sequences, 10 for training and 13 for testing, while varying the speed and activities of the pedestrians in the summer night. The environment temperature is about 30 Celsius which is more challenging than in the winter, spring, and autumn seasons. Recently, multispectral (color-thermal) datasets [3,20] 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

For improving on-road FIR pedestrian detection, the early work mainly focuses on various reformative hand-crafted features [3,12,11,13]. Olemda et al. [11] propose phase congruency feature to resist illumination change, Liu et al. [12] design pyramid entropy weighted HOG to highlight object profile and Qi et al. [13] adopt sparse representation to get rich semantic context. Jeong et al. [23] a cascade random forest with low dimensional Haar-like and OCS-LBPs features to detect sudden pedestrian crossing (SPC) at night. The system infers the SPC based on the likelihood and spatiotemporal features of each pedestrian, such as the overlapping ratio and the direction and magnitude of the pedestrian's movement. Following the success of integral channel features [24], Hwang et al. [3] 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 [7–10]. Most of them are custom architectures derived from Faster R-CNN. Cai et al. [7] propose a multi-scale CNN (MSCNN) with a multiple output layers proposal subnetwork, so that receptive fields can match the objects of different scales. Li et al. [8] propose scale-aware CNNs (SA-Fast R-CNN), capturing features for pedestrians of different image sizes by a scale gate function. Zhang et al. [6] proposed RPN + BF, combining RPN (the first part of Faster R-CNN) 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. [25] proposed a multispectral detector built upon Faster R-CNN and performed 37% miss rate on KAIST. König et al. [26] 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 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 (December) in Guangzhou, China. A fine-grained set of high quality

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 [19]	192k	128k	155k	121k	250k	✓		✓	✓	✓	'09
KITTI [2]	12k	1.6k	–	–	80k	✓		✓	✓	✓	'12
CityPersons [18]	20k	3k	11k	1.6k	5k	✓		✓		✓	'17
KAIST [3]	42k	50k	45k	45k	95k	✓	✓	✓	✓	✓	'15
CVC [20]	4.8k	3.5k	4.3k	1.4k	5k	✓	✓	✓	✓	✓	'16
OSU-T [21]	984	1.9k	–	–	2k		✓		✓		'05
LSI [5]	10.2k	6.2k	5.9k	9.1k	15.2k		✓		✓	✓	'13
TIV [22]	–	–	–	–	63k		✓		✓		'14
KMU [23]	–	7.9k	–	5.0k	12.9k		✓		✓	✓	'17
SCUT(Our)	175k	108k	177k	103k	211k		✓	✓	✓	✓	'18

annotations and corresponding statistics are presented. SCUT dataset is highlighted in data volume, data diversity and a wide range of imaging distance. In addition, the work of this paper has good 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

3.1.1. Data capture

Image sequences in SCUT dataset are captured by a monocular FIR camera mounted on a car (Fig. 2). We used an NV628 model of Guangzhou SAT Infrared Technology Co. Ltd. The spatial resolution of the camera is 384 × 288 with 13 mm focal length, and the field of view is 28° × 21° and the sensitive wavelengths between 8 and 14 μm. The output resolution is resized to 720 × 576 pixels by an image acquisition card for better observation and annotation. We collect about 11 h-long image sequences (~ 10^6 frames) at 25 fps by a vehicle driving through diverse traffic scenarios at 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).

3.1.2. Ground truth annotation

Piotr's Computer Vision Toolbox [27] 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 [28] 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.

3.1.3. Training and testing data

The annotated image sequences are 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. The number of pedestrian and frame in the training/testing set can be seen in Table 1.

3.2. Dataset statistics

The statistics on SCUT dataset is discussed here based on Table 2. The first column is the frames with annotations, that is, the frame has at least one BB annotation. As a whole, the proportion of frames with annotations is about 148,132/211,011 ≈ 70% frames. The second column is the number of BB. In all BBs, about (193765 + 157994)/447907 ≈ 78% is walk person or ride person, who must be detected in ADAS-oriented applications. The last column shows the average frames per object. The average time per object is the average frames per object divided by the frame rate. For example, the walk person appears 61.79/25 ≈ 2.47 s. Similarly, the average time of ride person is 86.62/25 ≈ 3.46. We analyze further some sub-theme, i.e. scale and distance, data diversity, pedestrian occlusion, and pedestrian position, etc. The

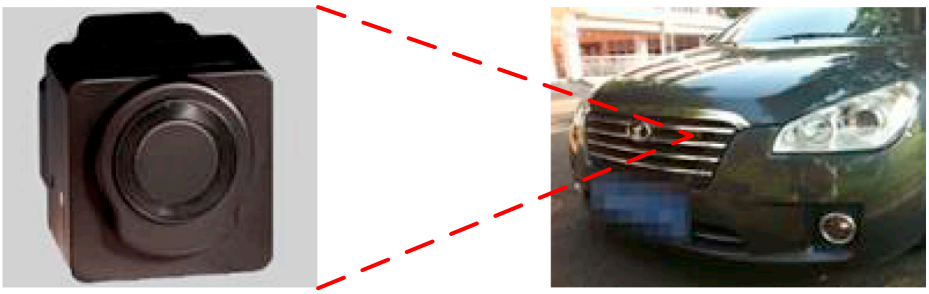


Fig. 2. Camera setup.

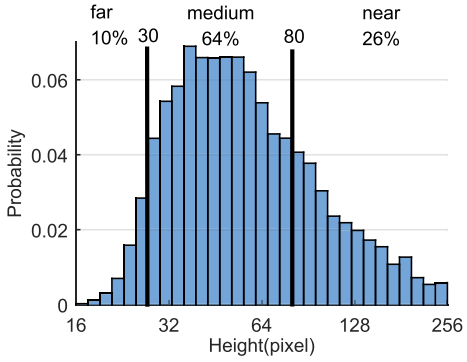
Table 2
Dataset summary.

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

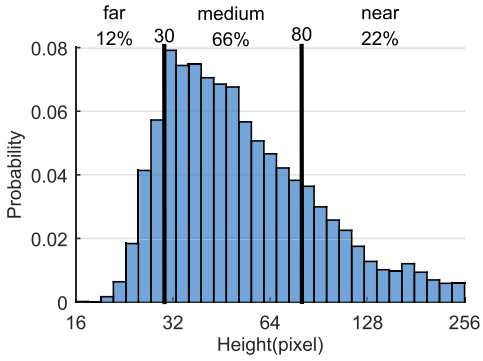
statistics may be basic supporting when establishing a road FIR pedestrian detection system.

3.2.1. Scale and distance

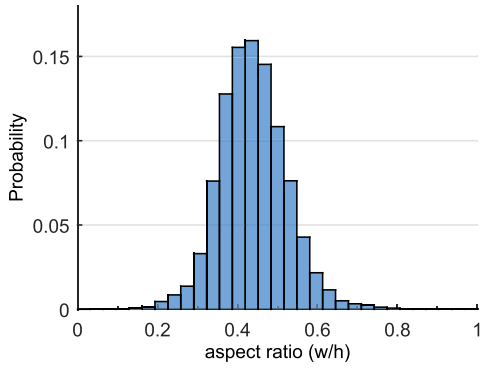
Similar to Dollár et al. [19], we group pedestrians by dividing the pixel height of BBs into three scales: near (more than 80 pixels), medium (30–80 pixels) and far (less than 30 pixels). The statistics of histogram distribution in the pixel height of BBs is investigated respectively on walk person and 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 (~64%) and ride person (~65%) lie in medium scale. At far distance region, the number of pedestrian decreases sharply because it is difficult to identify reliably a



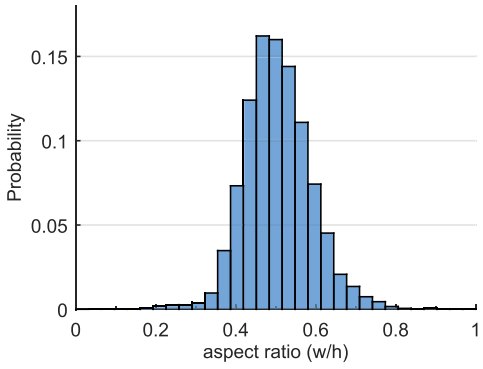
(a) Walk person height distribution



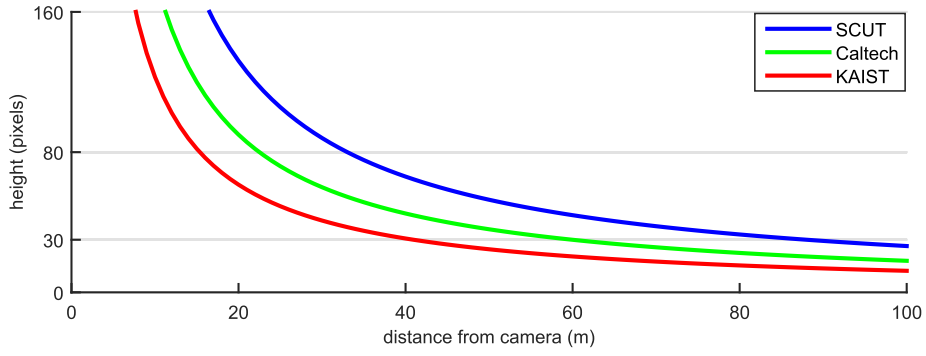
(b) Ride person height distribution



(c) Walk person aspect ratio distribution



(d) Ride person aspect ratio distribution



(e) Distance vs. height

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.

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	2.4–61 m	4.6–132 m
# Road scene	3	4

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 $576/2/f = tan(21°/2)$). 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. Fig. 3(e), compare the relationship between the pixel height of a pedestrian and the corresponding distance in meter on SCUT dataset, KAIST [3] and Caltech [1].

For a pedestrian with same pixel height, the distance from the vehicle (i.e. camera) is farther on SCUT dataset than KAIST. For example, a person in 80 pixels is about ~33 m distance from the vehicle on SCUT dataset, but only ~15 m on KAIST. This indicates that SCUT dataset is more reasonable for on-road pedestrian detection.

3.2.2. 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. The number of unique pedestrian reaches 7659 on SCUT dataset, but only 1182 on KAIST, which is the second difference of data diversity.

Following common practice for a pedestrian, the minimum high is 20 pixels and the maximum high is the image resolution [19]. 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. The pedestrian appearance diversity of distance and posture is shown in Fig. 4.

In addition, the image sequences are collected from 11 different road sections under 4 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.

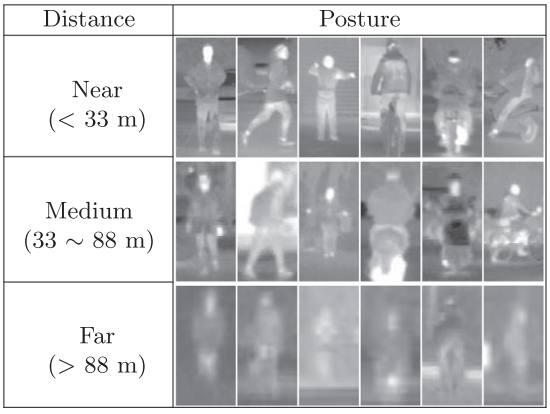


Fig. 4. Examples of pedestrians with no occlusion tag. It shows appearance diversity of distance and posture. Images are cropped and resized for better visualization.

3.2.3. Pedestrian occlusion

Since a camera is usually at 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, ~25% BBs are marked as 'occlusion'. In addition, among all walk persons, the occluded BB accounts for ~30% while occluded BB only accounts for ~11% in ride persons. A walk person could be occluded by trees, parking cars and another pedestrian when appearing on sidewalk. The ratio of walk person occluded is larger than the ride person because there are a lot of walk persons on sidewalk, as shown in Fig. 1.

3.2.4. Pedestrian center position

Fig. 5(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. 5(b). The average y-coordinate is 216, and about 97% of the pedestrians are in the range of 166–266 y-coordinate.

4. Baseline detector

To evaluate the training and test effect for the benchmark, a basic model should be considered to serve as the baseline. First, in Section 4.1, we introduce the vanilla Faster R-CNN [29]. Next, we propose four modifications on Faster R-CNN in Section 4.2. Finally, the performance of the modified Faster R-CNN and other 3 representative pedestrian detectors are evaluated and the comparison results demonstrate the effectiveness of the proposed modifications.

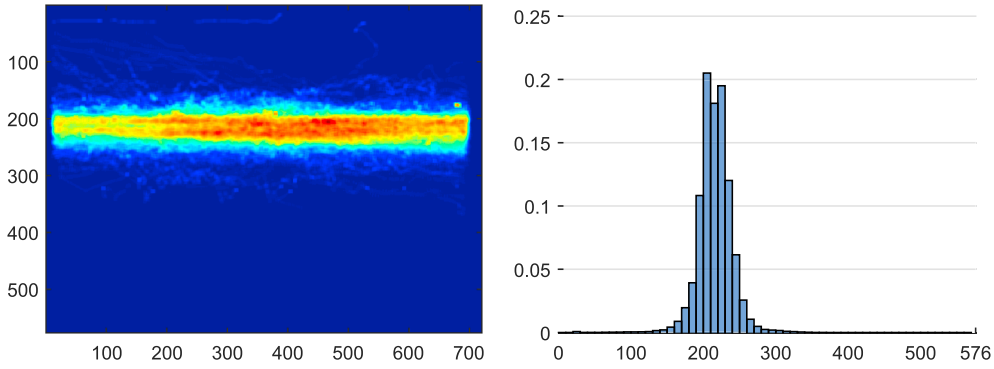
4.1. Vanilla faster R-CNN

The Faster R-CNN, which is consist of a Region Proposal Network (RPN) and a Fast R-CNN head [30], is a competitive detector on general object detection. Comparing with handcraft candidate generator, RPN can benefit from massive training data and speed up detection by GPU acceleration. It is worth noting that Faster R-CNN (FRCN) trained by default settings in [29] will under-perform on the FIR data due to two reasons: (1) different from images in visible spectrum, low resolution of FIR images will lead to a coarse feature map at the last layer, and (2) there is extreme class imbalance between pedestrians and other background targets, which is ignored in vanilla Faster R-CNN. To better explore the performance of pedestrian detection on FIR images, we analyze four important modifications.

4.2. Faster R-CNN for pedestrian

In this section, we explore four key modifications on Faster R-CNN, i.e., sample balance handling, anchor setting, backbone choosing and feature resolution expansion. In Section 5.2, how the above modifications improve the performance is evaluated and thus a strong baseline for SCUT dataset is proposed, which reduces the log-average miss rate from 36.78% to 17.73%.

Sample balance handling. The ratio of foreground to background is 1:3 and the number of samples per mini-batch is 512 in original Vanilla Faster R-CNN. If there are fewer than 128 foreground targets in a mini-batch, background targets will be filled in. However, extreme class imbalance is unavoidable in the task of pedestrian detection, indicating that most of the images contain only background targets or a small



(a) Heat map of position (b) Vertical distribution

Fig. 5. (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.

number of pedestrians. Lowest threshold of intersection over union (IoU), i.e., 0.1 is assigned to avoid scanning empty images in Vanilla Faster R-CNN. Nevertheless, some hard-negative background targets do not have any intersection with foreground ones for the task of pedestrian detection. Therefore, the limitation of minimum threshold is removed, and the ratio of no-empty images (contain at least one pedestrian) to empty images (containing only background targets) is assigned to 1:1 instead, which can maintain a manageable balance between the number of foreground and background targets. In addition, we do not consider the sampling for Regions of Interest (ROIs) at ignore area and use end-to-end training framework.

Anchor Setting. In RPN, the default anchor setting is appropriate for general object detection with only three scales and three ratios (marked **S3R3** in this paper), but it does not adapt to the task of pedestrian detection. Most pedestrians are less than 80 pixels in our dataset (Fig. 3) and our intuition is to fit for medium- and small-scale pedestrians by adding anchor scales. Therefore, we analyze another three anchor settings, i.e., S9R1, S15R1, and S5R3. Detailed settings for the parameters are shown in Table 4.

(1) **S9R1.** In [6], Zhang et al. argue that original anchors of inappropriate aspect ratios are noisy and harmful for improving the detection accuracy. Hence, they adopted anchors of a single average aspect ratio and used 9 different scales, starting from 40 pixels in height with a scaling stride of 1.3×.
(2) **S15R1.** In [18], Zhang et al. split the full-scale range in bins, where each bin contains an equal number of samples. Based on similar idea, we sort all bounding boxes by pixels in height and split the scale range in 16 quantize bins (equal number of samples per bin) and use the 15 split points as RPN scales to generate proposals with fixed a ratio of 1/0.46, which is the same as pedestrian ratio presented in Section 2.2.
(3) **S5R3.** For fitting small pedestrians, we simply add two small scales, i.e., 32 and 64, in the S3R3.

Backbone Choosing. In the task of general object detection, e.g., COCO [31] and Pascal VOC [32], choosing Residual Network (ResNet) [33] as backbone can achieve better performance than VGG16 [34]. However, ResNet is rarely used in pedestrian detection [35–37]. In Section 5.2, we evaluate ResNet as a backbone on SCUT dataset as a comparison. In addition, we use Conv4 as the RoIs pooling layer in ResNet50 and ResNet101 to detect and localize small-scale pedestrians. The input images are upsampled by 1.5× for better matching the ImageNet pre-training appearance distribution [18].

Feature Resolution Expansion. Since original Faster R-CNN is designed for the task of general object detection in Pascal VOC dataset, the feature stride is too large to the task of detecting pedestrians. The default VGG16 has a feature stride of 16, but median size of pedestrians in SCUT dataset is 56 × 26 (Fig. 3). Therefore, in contrast to the object pixels in width, a coarse stride would reduce the chances of searching ROIs with high IoU and force the network to handle the large displacement. To increase the feature resolution, three strategies are considered:

(1) **Remove pool4 (C5-wop4).** In [36,18,38,39], they remove fourth max-pooling layer from VGG16 by reducing the stride to 8 pixels. However, it reduces the receptive field on Conv5, resulting in a domain shift while conducting fine-tuning from a pre-trained network.
(2) **Remove Conv5 (C4).** Removing Conv5 and the fourth pooling layer could avoid domain shift and increase feature map size, but it still reduces the receptive field and losses high-level semantic feature.
(3) **Dilated convolution (C5d2).** To keep the receptive field and high-level semantic feature, we can assign the stride of Pool4 to 1 and dilate all Conv5 filters by a value of 2, which can also increase the feature map by 2×. Nevertheless, dilated convolution [40] will introduce more padding which causes some error.

Table 4
Anchor settings for RPN.

Name	Scale	Ratio	Count
S3R3	[128, 256, 512]	0.5, 1, 2	9
S9R1	[42, 54, 70, 91, 119, 154, 201, 261, 339]	0.46	9
S15R1	[27, 31, 33, 36, 39, 43, 47, 51, 56, 61, 69, 79, 91, 112, 159]	0.46	15
S5R3	[32,64,128,256,512]	0.5,1,2	15

4.3. Other baseline approaches

To understand the difficulties of pedestrian detection on SCUT dataset, we also select other three promising pedestrian detectors for comparison, i.e., ACF-based [41] detector, RPN + BF [6], MSCNN [7] and YOLO [42–44].

ACF-based. The standard ACF [41] has 10 augmented channels (LUV + M + O): LUV deNotes 3 channels of CIELUV color space, M deNotes 1 channel of gradient magnitude, and O deNotes 6 channels of gradient histogram which is a simplified version of histogram of oriented gradients (HOG) [14]. Hwang et al. [3] extend ACF to multispectral data by encoding the thermal intensity channel. They suggest three baselines: ACF-T, ACF-TM + TO, ACF-T + THOG. For the task of FIR pedestrian detection, color channels are abandoned, only the intensity and gradient channels are extracted to characterize objects in

FIR images, where all the parameters are using the default settings.

RPN + BF. The RPN + BF [6] concatenates RPN and boosting framework. RPN generates region proposals, confidence scores, and features, all of which are used to train a cascaded Boosted Forest classifier. Bootstrapping strategy is used for mining hard negative examples. RPN + BF achieves 9.6% MR on Caltech pedestrian dataset, which is the second-best. Hence, it is served as the state-of-the-art baseline detector on SCUT dataset.

MSCNN. The MSCNN [7] is also modified from Faster R-CNN, which consists of a proposal sub-network and a detection sub-network. Detection is performed at multiple output layers in the proposal sub-network, so that receptive fields match objects with different scales. These complementary scale-specific detectors are combined to produce a strong multi-scale object detector.

YOLO. YOLO [42] is a one-stage detector which effectively combined the steps of region proposal and classification. YOLO is extremely fast which make it a real-time detector. Two improved versions were proposed after YOLO. YOLOv2 [43] focus on improving recall and localization while maintaining classification accuracy and real-time. YOLOv3 [44] mainly improved performance with predicting boxes at 3 different scales and using more deep backbone network.

5. Experiments

According to the baseline verified in Section 4, we benchmark SCUT dataset. The evaluation protocol and subsets of ground truth are described in Section 5.1, followed by the analysis of four important modifications of Faster R-CNN. In Section 5.3, we evaluate the performance under different conditions using SCUT dataset.

5.1. Evaluation protocol and subsets

We employ the evaluation protocol proposed by Dollar et al. [1]. It stated a detected bounding box (BB_{dt}) and a ground truth bounding box (BB_{gt}) with an IoU larger than 0.5. Each BB_{dt} and BB_{gt} are matched at most once and it is unnecessary to ensure an ignored bounding box (BB_{ig}) to be matched. In SCUT dataset, six types of *BBs* are always set to be ignored: *BBs* under 20 pixels in height or truncated by image boundaries, and *BBs* containing a 'person?', 'people?', 'people' or 'squat_person'. Detections within these regions will not affect the results of performance evaluation.

The log-average miss rate is adopted to summarize the performance of pedestrian detectors by averaging the corresponding values at 17 false positive per-image evaluation (FPPI) rates evenly spaced in log-space ranging from 10^{-4} to 1. Particularly, the minimum miss rate achieved is used for those curves which terminate without reaching the given FPPI rate.

To further evaluate the performance of the detectors, the test set is divided into multiple evaluation subsets as shown in Table 5. In all the experiments, the image interval of test dataset is set to 25 frames.

Reasonable. A representative subset was chosen and marked as **Reasonable**. The reasonable all subset consists of walk person and ride person with more than 50 pixels in height. This subset is further divided into **Reasonable-walk** and **Reasonable-ride** person subsets based on the label of ground truth.

Table 5
Subset setting of SCUT ground truth.

Subset	Scale	Category	Occlusion
Overall	[20,inf]	walk_person,ride_person	All
Reasonable	[50,inf]	walk_person,ride_person	All
Reasonable-walk	[50,inf]	walk_person	All
Reasonable-ride	[50,inf]	ride_person	All
Near	[80,inf]	walk_person,ride_person	None
Medium	[30,80]	walk_person,ride_person	None
Far	[20,30]	walk_person,ride_person	None
No occlusion	[50,inf]	walk_person,ride_person	None
Occlusion	[50,inf]	walk_person,ride_person	Occluded

Scale. Reasonable subset doesn't cover small-scale pedestrian. However, as discussed in Section 2.2, SCUT dataset contains enough high-quality data to group pedestrians by the pixels in height into three subsets including **Near** (no less than 80 pixels), **Medium** (ranging from 30 to 80 pixels) and **Far** (no more than 30 pixels) scales, which can be used to evaluate the performance of pedestrian detection on different scales. And these subsets exclude occluded pedestrians.

Overall. The performance of pedestrian detectors on the entire test dataset can be evaluated for the real-world scenarios.

Occlusion. Since Reasonable and overall subsets contain both non-occluded and occluded ground truth, two subsets containing pedestrians with no less than 50 pixels in height were formed based on the annotated occlusion tags in order to evaluate the anti-occlusion performance of pedestrian detectors.

Table 6
Improvement by handle imbalance modification.

Method	Reasonable	Overall
Vanilla	36.78%	66.32%
Handle imbalance	**30.15%**	**54.30%**

5.2. Ablation experiment of Faster R-CNN

In this section, how the four modifications described in Section 4.2 affect the performance of FIR pedestrian detection is analyzed. VGG16 is adopted as the backbone in Faster R-CNN framework with 180k training iterations, where the base learning rate is assigned to 0.0025 and decreased by a factor of 10 after the 80k and 160k training iterations. A mini-batch involves 1 image per GPU. The settings of weight decay and momentum are assigned to 0.0001 and 0.9, respectively.

Sample balance handing. Benefit from the balance handling, the MR of detector on reasonable subset decreased from 36.78% to 30.15% as shown in Table 6. Besides, the performance on overall subset also achieves significant improvement, indicating that the handling of sample unevenness and the ignored area plays an important role on detecting FIR pedestrians.

Anchor setting. Table 7 shows the effect of four anchor settings on the missed detection rate of Faster R-CNN. Surprisingly, the S15R1 scheme, which strictly follows the sample distribution, is not the optimal solution. The reason probably lies in that the dense scale will cause a lot of redundant RoIs, leading to generate more false positives. S9R1 achieves the best performance on near-scale subset, indicating that the scale interval is more appropriate for pedestrians with more pixels in height. Overall, S5R3 achieves better performance on in most subsets and reduces the MR of reasonable from 30.15% to 23.28%. Hence, S5R3 is used as the default anchor setting for FIR pedestrian detection on SCUT dataset.

Backbone Choosing. We report the results of performance with different backbones for pedestrian detection. Table 8 shows that VGG16 outperforms ResNet in SCUT dataset. It is well known that the increase in depth for the networks results in the strength of feature

Table 7
Compare four kinds of anchor setting.

Method	Reasonable	Overall	Near	Medium	Far
S3R3	30.15%	54.30%	14.76%	47.24%	74.33%
S9R1	27.03%	52.23%	**10.53%**	47.58%	74.87%
S15R1	29.67%	55.09%	13.72%	50.76%	78.17%
S5R3	**23.28%**	**46.83%**	11.42%	**41.41%**	**68.25%**

Table 8
Compare three kinds of backbone.

Backbone	Parameters	Pooling Layer	Iteration	Reasonable	Overall
ResNet101	39M	Conv4	140k	19.94%	42.71%
ResNet50	20M	Conv4	160k	20.94%	44.57%
VGG16	14M	Conv5	180k	**18.08%**	**39.82%**

representation using a large number of required training data. However, it can be concluded from optimal number of iterations that the deeper network reaches saturation earlier, which means that ResNet is likely to be over-fitting on the task of pedestrian detection. The reason may be that there are too many network parameters for FIR pedestrian detection.

Feature Resolution Expansion. Table 9 shows the results using different settings of RoI feature convolution layers. First of all, increasing image size can significantly improve the performance of detectors. The C5-wop4 achieves the best performance for detecting pedestrian on near-scale subset, whereas the domain shift and less receptive field are harmful to detect pedestrians on medium-scale and far-scale subsets. The C5d2 obtains the best performance on overall and medium-scale subsets. Comparing with C5, C5d2 decreases MR by 1.14 points on overall subsets. Since the medium-scale pedestrians account for the main parts for the task of pedestrian detection in ADAS applications, Faster R-CNN with C5d2 RoI pooling feature layer is adopted as the baseline method.

To further verify the performance of improved Faster R-CNN, the evaluation was also conducted using the KAIST FIR dataset, where only the FIR data were adopted. The comparisons in terms of MR under nighttime reasonable configuration are illustrated in Table 10. In the previous representative studies, Liu et al. [25] proposed a multispectral detector called Halfway Fusion by fusing both the color and FIR data in the KAIST FIR dataset and achieved a MR of 35.49%. However, our improved Faster R-CNN achieves MR of 26.42% using only FIR data. The result demonstrates that the modified Faster R-CNN is a state-of-the-art detector for FIR pedestrian detection.

5.3. Benchmark on SCUT

We analyze the performance under nine conditions on the test subsets in SCUT dataset as mentioned on section Table 5, and the results are illustrated in Fig. 6 where log-average miss rate is used as the common reference value for summarizing the performance.

Reasonable. In Fig. 6(a) and (b), our modified Faster R-CNN (FRCN-our) performs best with MRs of 17.73 and 16.85%, respectively. However, Fig. 6(c) shows that RPN + BF and MSCNN are better than FRCN-our on Reasonable-ride subset. Probably because it benefits from the bootstrapping strategy used in RPN + BF which reduces hard negative samples, and multiple RoI Pooling layers used in MSCNN by handling multiscale pedestrians. YOLOv3 gets better performance than vanilla Faster R-CNN (FRCN-vanilla).

Scale. Results using subsets on three different scales using only unoccluded pedestrians are shown in Fig. 6(d)–(f), respectively. In general, with the decrease in pixels in height, the performance of FIR pedestrian detection degrades dramatically. FRCN-our achieves the best performance on medium- and far-scale subsets, whereas MSCNN achieves the best performance on near-scale subset. Benefit from predictions across scales, YOLOv3 has achieved good performance on medium and small-size pedestrians.

Occlusion. The impact of occlusion on detecting pedestrians with a minimum value of 50 pixels in height is shown in Fig. 6(g) and (h). Performance drops significantly under the occlusion situation for all detectors. MSCNN performs better than FRCN-our on non-occlusion subset, but FRCN-our outperforms the other detectors on occlusion subset. As Reasonable subset, YOLOv3 better than FRCN-vanilla, but YOLOv2 worse than FRCN-vanilla.

Overall. Performance evaluation for different detectors is conducted on the entire dataset containing both the walk person and ride person and the results are shown in Fig. 6(i). Benefit from our modification, FRCN-our achieves the best performance with a MR of 36.68%.

Summary. The experiment results show that benefit from the volume and diversity of training data, CNN-based detectors obtained better performance on FIR data even with initialized via ImageNet classification pre-training. MSCNN and RPN + BF are more appropriate for detecting FIR pedestrians with larger pixels in height and easy conditions, such as the detection on near-scale and non-occlusion subsets. ACF-T + family detectors perform poorly on all subsets because. FRCN-vanilla might suffer from the feature size and improper anchor settings. As a one-stage detector, YOLOv2 is also effective in thermal imaging pedestrian's detection and YOLOv3 surprisingly achieved better performance than FRCN-vanilla on most conditions. Benefit from our modifications, FRCN-our achieves better performance than the other detectors on most test subsets.

Table 10
Comparison of Halfway and our baseline detector reported on nighttime test set of KAIST.

Method	Train	Test	Miss Rate
Halfway fusion [25]	Color + FIR	Color + FIR	35.49%
Faster R-CNN Our	FIR	FIR	**26.42%**

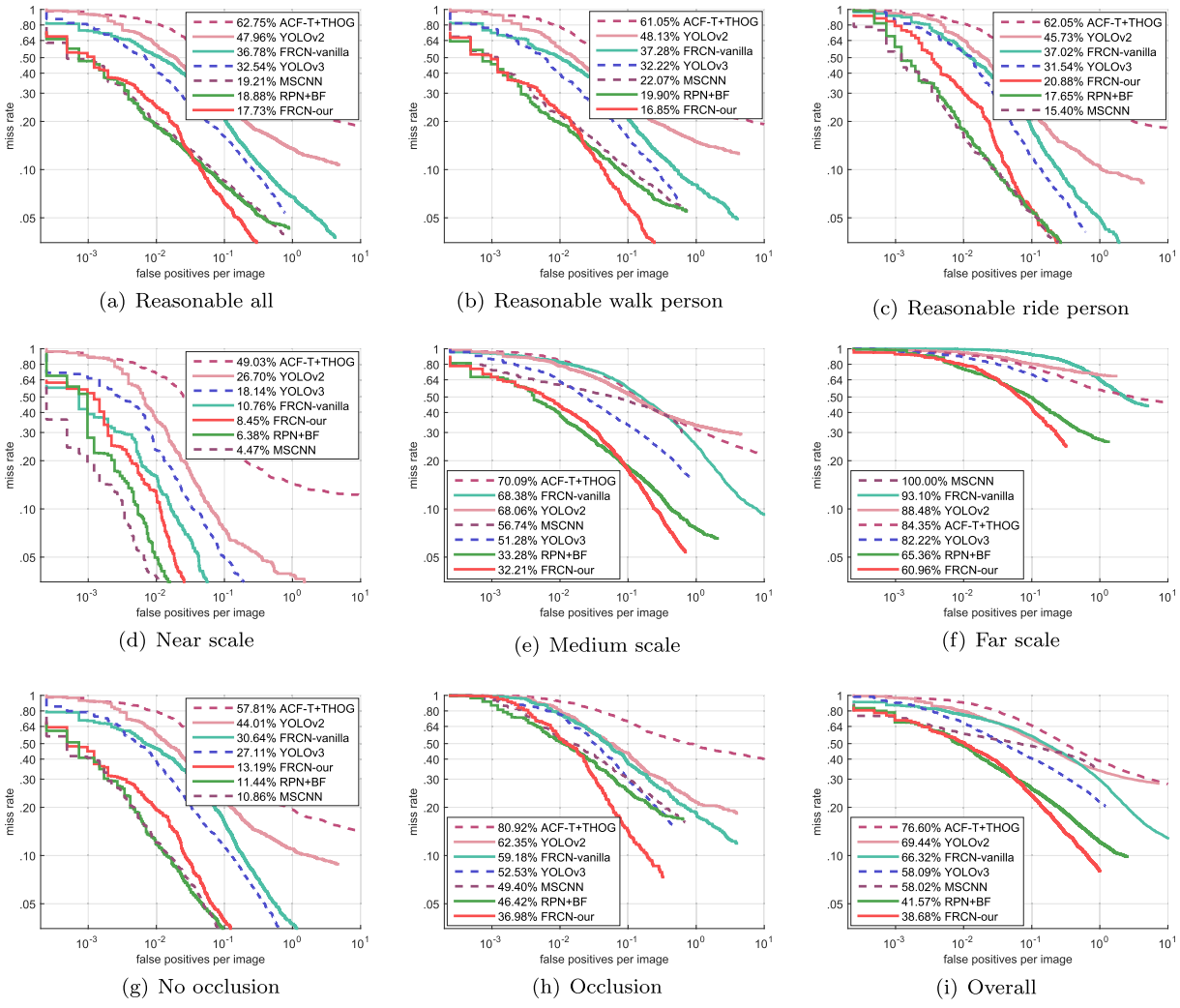
6. Conclusion and discussion

The main progress of pedestrian detection in visible spectrum has been greatly driven by several challenging benchmark datasets over the past few years. In order to motivate more efforts toward the task of on-road FIR pedestrian detection, SCUT dataset is introduced in this paper, which was collected from several variable driving scenarios over one month in Guangzhou, China. The dataset provides a large-scale of fine-grained annotations with high-level data diversity. To the best of our knowledge, SCUT dataset is the largest FIR dataset providing occlusion labels and temporal correspondences captured from non-static real traffic scenes. The dataset will be public online after this paper published. In our future work, we will gradually add daytime and other seasons videos.

After the detailed analysis on the basic statistics of SCUT dataset, we

Table 9
Comparisons of different methods to increase RoI feature convolution layer resolution.

Method	Scale	Feature Size	Reasonable	Overall	Near	Medium	Far
C5	1×	36 × 45	24.67%	49.14%	11.21%	43.94%	70.07%
C5	1.5×	54 × 67	18.08%	39.82%	8.70%	32.83%	60.36%
C5-wop4	1.5×	108 × 135	18.03%	40.54%	**7.10%**	35.27%	62.85%
C4	1.5×	108 × 135	**16.71%**	40.06%	7.17%	33.72%	**60.19%**
C5d2	1.5×	109 × 136	17.73%	**38.68%**	8.45%	**32.21%**	60.96%



(a) Reasonable all (b) Reasonable walk person (c) Reasonable ride person

(d) Near scale (e) Medium scale (f) Far scale

(g) No occlusion (h) Occlusion (i) Overall

Fig. 6. False positive per image (FPPI) versus miss rate in various conditions.

evaluated the performance of FIR pedestrian detection using ACF-based and CNN-based models, and benchmarked with several promising CNN-based detectors. Besides, a properly adapted Faster R-CNN is proposed and used as one of the baseline detectors. The experimental results demonstrated that the CNN-based detectors achieved better performance on SCUT dataset and other FIR datasets, especially provided with a larger training set of high-level diverse examples. State of the art performance is achieved by improving Faster R-CNN detector from serval important modifications including sample balance handling, anchor setting, backbone choosing and feature resolution expansion.

We expect that the proposed SCUT dataset can encourage the development of methodologies for better FIR pedestrian detection. There are two directions still needing further researches.

(1) **Video pedestrian detection.** Fast and accurate object recognition in video data is crucial for ADAS applications. However, most detectors are designed for locating pedestrian in still images. Applying image detectors on individual video frames introduce unaffordable and unnecessary computational cost.

(2) **Context.** For on-road object detection systems, there are many context information available in the scenarios. As mentioned in Section 3.2, the information of viewpoint and ground plane geometry constrains the location of a pedestrian only in a narrow band running horizontally across the center of the image. Utilizing ground plane context will further reduce many constrained false positives.

Acknowledgement

This paper is supported by Science and Technology Planning Project of Guangdong Province, China under Grant No. 2017A020219008. This work is also partially supported by the Natural Science Foundation of Guangdong Province under the Grant No. 2016A030310235.

References

[1] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: an evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 743–761, https://doi.org/10.1109/TPAMI.2011.155.

[2] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The kitti vision benchmark suite, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, IEEE, 2012, pp. 3354–3361, , https://doi.org/10.1109/CVPR.2012.6248074.

[3] S. Hwang, J. Park, N. Kim, Y. Choi, I.S. Kweon, Multispectral pedestrian detection: benchmark dataset and baseline, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 07-12-June, IEEE, 2015, pp. 1037–1045, , https://doi.org/10.1109/CVPR.2015.7298706.

[4] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, Towards reaching human performance in pedestrian detection, IEEE Trans. Pattern Anal. Mach. Intell. doi:https://doi.org/10.1109/tpami.2017.2700460.

[5] D. Olmeda, C. Premebida, U. Nunes, J.M. Armingol, D.L.E. Arturo, Pedestrian classification and detection in far infrared images, Integr. Comput. Aid. Eng. 20 (4) (2013) 347–360 <http://hdl.handle.net/10016/17370> .

[6] L. Zhang, L. Lin, X. Liang, K. He, Is faster R-CNN doing well for pedestrian detection? European Conference on Computer Vision (ECCV), Springer, 2016, pp. 443–457, , https://doi.org/10.1007/978-3-319-46475-6_28.

[7] Z. Cai, Q. Fan, R.S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, European Conference on Computer Vision

(ECCV), Springer, 2016, pp. 354–370, , https://doi.org/10.1007/978-3-319-46493-0_22.
[8] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast r-CNN for pedestrian detection, IEEE Trans. Multimedia (2017) 1, https://doi.org/10.1109/tmm.2017.2759508.
[9] Y. Tian, P. Luo, X. Wang, X. Tang, Deep learning strong parts for pedestrian detection, The IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 1904–1912, , https://doi.org/10.1109/ICCV.2015.221.
[10] Y. Tian, P. Luo, X. Wang, X. Tang, Pedestrian detection aided by deep learning semantic tasks, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 5079–5087, , https://doi.org/10.1109/CVPR.2015.7299143 Available from: 1412.0069.
[11] D. Olmeda, A. de la Escalera, J.M. Armingol, Contrast invariant features for human detection in far infrared images, in: IEEE Intelligent Vehicles Symposium (IV), IEEE, 2012, pp. 117–122. doi:10.1109/ivs.2012.6232242.
[12] Q. Liu, J. Zhuang, J. Ma, Robust and fast pedestrian detection method for far-infrared automotive driving assistance systems, Infrared Phys. Technol. 60 (2013) 288–299, https://doi.org/10.1016/j.infrared.2013.06.003.
[13] B. Qi, V. John, Z. Liu, S. Mita, Pedestrian detection from thermal images: a sparse representation based approach, Infrared Phys. Technol. 76 (2016) 157–167, https://doi.org/10.1016/j.infrared.2016.02.004.
[14] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, IEEE, 2005, pp. 886–893, , https://doi.org/10.1109/CVPR.2005.177.
[15] A. Ess, B. Leibe, K. Schindler, L. Van Gool, A mobile vision system for robust multi-person tracking, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2008, pp. 1–8, , https://doi.org/10.1109/cvpr.2008.4587581.
[16] C. Wojek, S. Walk, B. Schiele, Multi-cue onboard pedestrian detection, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 794–801, , https://doi.org/10.1109/cvpr.2009.5206638.
[17] M. Enzweiler, D.M. Gavrila, Monocular pedestrian detection: Survey and experiments, IEEE Trans. Pattern Anal. Mach. Intell. 31 (12) (2009) 2179–2195, https://doi.org/10.1109/tpami.2008.260.
[18] S. Zhang, R. Benenson, B. Schiele, CityPersons: a diverse dataset for pedestrian detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. doi:https://doi.org/10.1109/cvpr.2017.474.
[19] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: a benchmark, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 304–311, , https://doi.org/10.1109/CVPR.2009.5206631.
[20] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez, J. Xu, A. López, Pedestrian detection at day/night time with visible and FIR cameras: a comparison, Sensors 16 (6) (2016) 820, https://doi.org/10.3390/s16060820.
[21] J.W. Davis, M.A. Keck, A two-stage template approach to person detection in thermal imagery, IEEE Workshops on Application of Computer Vision, vol. 1, IEEE, 2005, pp. 364–369, , https://doi.org/10.1109/ACVMOT.2005.14.
[22] Z. Wu, N. Fuller, D. Theriault, M. Betke, A thermal infrared video benchmark for visual analysis, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 201–208. doi:https://doi.org/10.1109/CVPRW.2014.39.
[23] M. Jeong, B.C. Ko, J.-Y. Nam, Early detection of sudden pedestrian crossing for safe driving during summer nights, IEEE Trans. Circ. Syst. Video Technol. 27 (6) (2017) 1368–1380, https://doi.org/10.1109/tcsvt.2016.2539684.
[24] P. Dollar, Z. Tu, P. Perona, S. Belongie, Integral channel features, Procedings of the British Machine Vision Conference 2009, British Machine Vision Association, 2009, , https://doi.org/10.5244/c.23.91.
[25] J. Liu, S. Zhang, S. Wang, D.N. Metaxas, Multispectral deep neural networks for pedestrian detection, The British Machine Vision Conference (BMVC) (2016) 1–13, https://doi.org/10.5244/c.30.73 Available from: 1611.02644.
[26] D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, M. Teutsch, Fully convolutional region proposal networks for multispectral person detection, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, IEEE, 2017, pp. 243–250, , https://doi.org/10.1109/cvprw.2017.36.
[27] P. Dollár, Piotr's Computer Vision Matlab Toolbox. < http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html > .
[28] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, How far are we from solving pedestrian detection? The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 1259–1267, , https://doi.org/10.1109/CVPR.2016.141 Available from: arXiv: 1602.01237.
[29] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems (NIPS), Curran Associates, Inc., 2015, pp. 91–99, , https://doi.org/10.1109/tpami.2016.2577031.
[30] R. Girshick, Fast R-CNN, The IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 1440–1448, , https://doi.org/10.1109/ICCV.2015.169.
[31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, European Conference on Computer Vision (ECCV), Springer International Publishing, 2014, pp. 740–755, , https://doi.org/10.1007/978-3-319-10602-1_48 Available from: 1405.0312.
[32] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338, https://doi.org/10.1007/s11263-009-0275-4.
[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:https://doi.org/10.3389/fpsyg.2013.00124. Available from: 1512.03385.
[34] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. Available from: 1409.1556.
[35] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, C. Shen, Repulsion Loss: Detecting Pedestrians in a Crowd, 2018. Available from: 1711.07752.
[36] C. Li, D. Song, R. Tong, M. Tang, Illumination-aware Faster R-CNN for robust multispectral pedestrian detection, Pattern Recogn. 85 (2019) 161–171, https://doi.org/10.1016/j.patcog.2018.08.005 Available from: 1803.05347.
[37] D. Huang, S. Huang, H. Wu, N. Liu, Pedestrian detection via structure-sensitive deep representation learning, in: Y. Zhao, X. Kong, D. Taubman (Eds.), International Conference on Image and Graphics, Springer International Publishing, Cham, 2017, pp. 127–138, , https://doi.org/10.1007/978-3-319-71607-7_12.
[38] G. Brazil, X. Yin, X. Liu, Illuminating pedestrians via simultaneous detection & segmentation, in: The IEEE International Conference on Computer Vision (ICCV), 2017. doi:https://doi.org/10.1109/iccv.2017.530.
[39] H. Choi, S. Kim, K. Park, K. Sohn, Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks, International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 621–626, , https://doi.org/10.1109/icpr.2016.7899703.
[40] F. Yu, V. Koltun, T. Funkhouser, Dilated residual networks, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, , https://doi.org/10.1109/CVPR.2017.75 Available from: 1705.09914.
[41] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2014) 1532–1545, https://doi.org/10.1109/TPAMI.2014.2300479.
[42] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, , https://doi.org/10.1109/cvpr.2016.91.
[43] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, , https://doi.org/10.1109/cvpr.2017.690.
[44] J. Redmon, A. Farhadi, Yolov3: An Incremental Improvement. Available from: 1804.02767.