Omnidirectional Pedestrian Detection by Rotation Invariant Training

Masato Tamura
Hitachi, Ltd.

Shota Horiguchi
Hitachi, Ltd.

Tomokazu Murakami
Hitachi, Ltd.

{masato.tamura.sf, shota.horiguchi.wk, tomokazu.murakami.xr}@hitachi.com

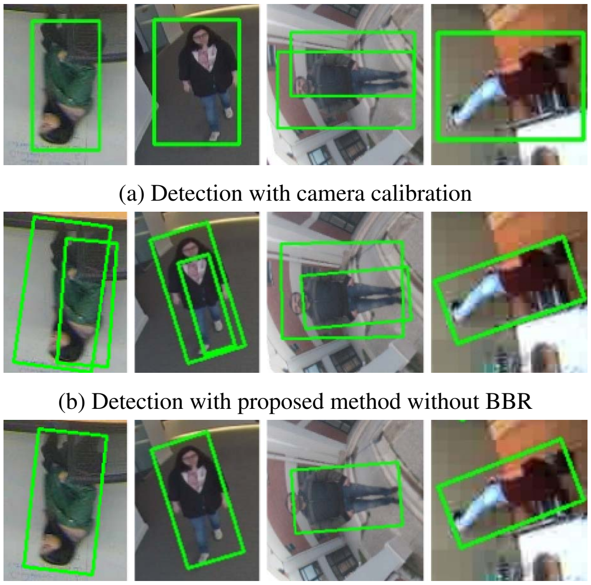
Abstract

Recently much progress has been made in pedestrian detection by utilizing the learning ability of convolutional neural networks (CNNs). However, due to the lack of omnidirectional images to train CNNs, few CNN-based detectors have been proposed for omnidirectional pedestrian detection. One significant difference between omnidirectional images and perspective images is that the appearance of pedestrians is rotated in omnidirectional images. A previous method has dealt with this by transforming omnidirectional images into perspective images in the test phase. However, this method has significant drawbacks, namely, the computational cost and the performance degradation caused by the transformation. To address this issue, we propose a rotation invariant training method, which only uses randomly rotated perspective images without any additional annotation. By this method, existing large-scale datasets can be utilized. In test phase, omnidirectional images can be used without the transformation. To group predicted bounding boxes, we also develop a bounding box refinement, which works better for our detector than non-maximum suppression. The proposed detector achieved a state-of-the-art performance on four public benchmarks.

1. Introduction

Pedestrian detection is a canonical subproblem of object detection that has received much attention thanks to its diverse applications, such as car safety, surveillance, and robotics [4,38]. Although sophisticated detectors have been proposed in the past decades [12, 13, 21, 28, 30, 35–37, 39–41], pedestrian detection has been an active research topic and recent papers still show improvements.

While early pedestrian detection methods relied on handcrafted features [12,13,28,36,37,40], recent CNN-based approaches have enabled progress and demonstrated state-of-the-art performance thanks to their data-driven feature representations, efficient end-to-end learning methods, and the availability of large-scale pedestrian detection datasets [21, 30, 35, 39, 41]. Many of the CNN-based approaches use customized architectures derived from general object detectors [19,25,31–33]. This suggests that proper customization of general object detectors for pedestrian detection is an effective approach.


(a) Detection with camera calibration

(b) Detection with proposed method without BBR

(c) Detection with proposed method

Figure 1: Comparison of three detection methods.

Compared to standard pedestrian detection, there has been much less research on the application of CNNs to omnidirectional pedestrian detection due to the lack of sufficient omnidirectional images to train CNNs. In omnidirectional images, the appearance of pedestrians is deformed in exchange for a wide angle of view. One significant deformation is rotation, which standard pedestrian detectors cannot deal with. To apply CNNs to omnidirectional pedestrian detection, this deformation has to be managed without omnidirectional images. One solution is to generate such data in the training phase by rotating perspective images. In doing so, existing large-scale perspective image datasets such as COCO [23] and VOC [15] can be utilized. However, it is necessary to annotate each rotated person individ-

978-1-7281-1975-5/19/$31.00 ©2019 IEEE
DOI 10.1109/WACV.2019.00216

ually because the original ground truth labels do not fit them anymore. This leads to a significant increase in annotation costs. In addition, the predicted bounding boxes do not always fit tightly to the appearance of people when they are inclined. Another solution is to first transform omnidirectional images into perspective images using calibrated camera parameters and then apply standard pedestrian detectors to the perspective images in the test phase [34]. However, this method has several significant drawbacks, including camera calibration costs, computational costs for transformation, and performance degradation due to interpolation of image transformation. In addition, it needs to suppress a large number of predicted bounding boxes backprojected from highly overlapped multiple perspective images, which induces false positives (Fig. 1a).

To address these issues, we propose a rotation invariant training method for omnidirectional pedestrian detection. In our method, rotated images are generated using perspective images in the training phase as described above, but we use only provided ground truth labels for training so that no additional annotation is necessary. In this way, the predicted bounding boxes become invariant even though pedestrians are rotated. In the test phase, omnidirectional images are input to CNNs without transforming the images to perspective images. With omnidirectional images, the angle of a person is uniquely determined by the person's position. This characteristic enables the determination of the angles of predicted bounding boxes from their positions. Using predicted bounding boxes and angles determined by their positions, regions of pedestrians can be predicted. To further improve the performance, we propose a bounding box refinement (BBR) that replaces non-maximum suppression (NMS). We empirically found that multiple bounding boxes that have a small intersection over union (IoU) are frequently output to one pedestrian, as shown in Fig. 1b. While NMS fails to group these bounding boxes, BBR works well for grouping these bounding boxes of significantly different sizes (Fig. 1c) because it uses detected center points for grouping. In addition, BBR is easier to perform than NMS when bounding boxes are rotated because the IoU calculation for a pair of rotated bounding boxes is quite complicated.

We summarize our contributions as follows:

- We propose a training method that uses only perspective images for omnidirectional pedestrian detection. This method generates a large amount of training data without any annotation. The trained detector can produce tightly fitted bounding boxes to pedestrians.
- We propose BBR, which works better than NMS in the proposed detector.
- We conduct extensive experiments on multiple public benchmark datasets for omnidirectional pedestrian detection and demonstrate state-of-the-art performance.

2. Related work

2.1. Object detection

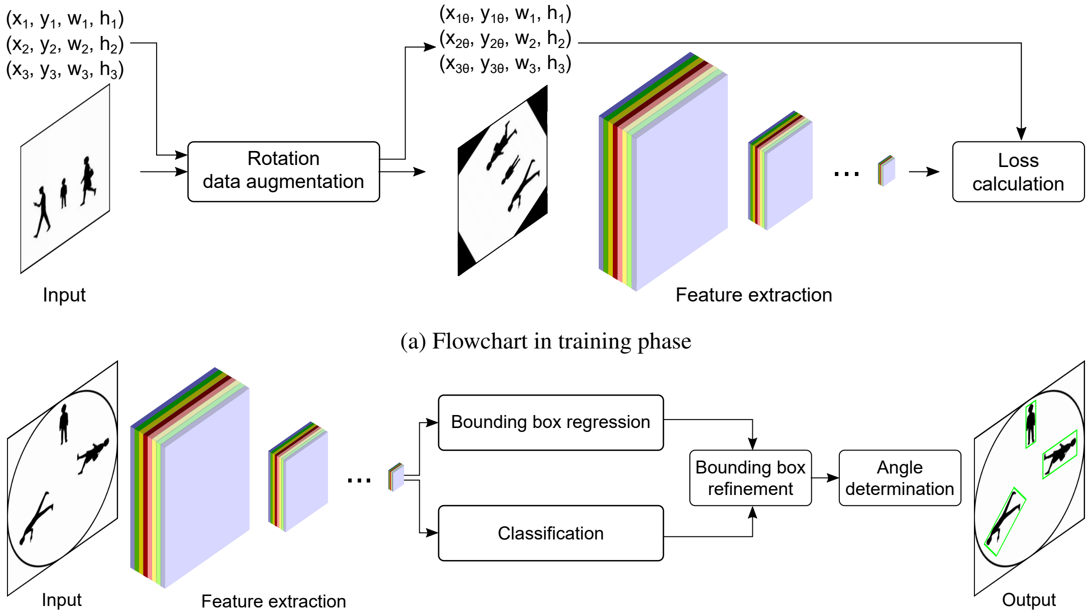
The powerful learning ability of CNNs has resulted in recent object detectors achieving high performance in several object detection benchmarks [19, 25, 31–33]. Almost all of them share a common approach consisting of object category classification and bounding box regression. The bounding box regression is carried out using horizontal anchor boxes that enable the predicted regions of objects to be expressed by horizontal bounding boxes. However, due to the appearance of pedestrians rotated to any angles, horizontal bounding boxes sometimes cannot express the precise regions of pedestrians. For rotated object detection, the prediction of angles is proposed in [24]. The authors add angle regression to SSD [25], which enables tight fitting of bounding boxes to rotated objects. For the prediction of angles, objects must be annotated with angles in training data. Although this method can be applied to omnidirectional pedestrian detection, the prediction of angles is redundant because the angles are uniquely determined on the basis of their positions in omnidirectional images. Utilizing this, our method can output rotated bounding boxes without any effort for annotating pedestrians rotated to any angles.

2.2. Pedestrian detection

Pedestrian detection is one of the most popular object-specific detection tasks that has a wide variety of applications. Before the deep learning era, many integral channel feature (ICF)-based detectors [12, 13, 28, 36, 37, 40] were proposed. ICF is a kind of handcrafted features. Because of the sophistication of the feature, they have dominated the top performance of pedestrian detection benchmarks [8, 14, 18]. Even though the learned features of CNNs outperformed handcrafted features in many computer vision tasks, ICF was used as support for CNNs until recently [21, 37]. Now, however, CNN-based detectors [30, 35, 39, 41] have exceeded the performance of ICF variants. The common approach of these methods is to customize general object detectors for pedestrian detection problems such as low resolution and occlusion [30, 35, 41]. Our method is also based on a general object detector, YOLOv2 [32], but takes omnidirectional images as input.

2.3. Omnidirectional pedestrian detection

Omnidirectional pedestrian detection is a part of pedestrian detection. However, standard pedestrian detectors cannot be directly applied due to the deformation of the appearance of pedestrians in omnidirectional images. Because of this, many omnidirectional pedestrian detectors have been proposed [6, 7, 9, 22, 27, 29, 34]. Early attempts relied on handcrafted features as well as standard pedestrian detection [6, 7, 9, 22, 27]. The common approach of these meth-



(a) Flowchart in training phase

(b) Flowchart in test phase

Figure 2: Flowchart in proposed method. In training, perspective images are input. The images are transformed by rotation data augmentation. In test, omnidirectional images are input. The detector deals with the images without transforming them to perspective images.

ods is approximating the features of the usual appearance of pedestrians from the deformed appearance by transforming images or features. Although this approach simplifies the classification of features, approximation errors, which are common, lead to significant degradation of performance. A few approaches use CNN-based detectors for omnidirectional pedestrian detection [29,34]. In [29], the authors proposed background modeling to enhance the detection performance of YOLO [31]. However, it is obvious that detection performance is highly affected by background modeling. In [34], the authors proposed transforming omnidirectional images into perspective images using calibrated camera parameters and then applying the original YOLOv2 [32] to the perspective images in the test phase. Although this method enables YOLOv2 to detect pedestrians to some extent without any training, it requires a certain amount of computational time to transform images and carry out detection in multiple perspective images transformed from one omnidirectional image. In addition, the detection performance is affected by transformation accuracy. Compared to this method, our method transforms images only in the training phase. This solves the problem described above.

3. Method

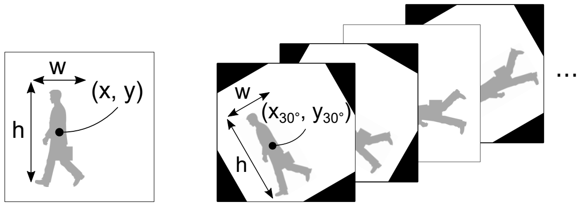
3.1. Overview

The proposed detector is a fully convolutional neural network with an architecture based on YOLOv2 [32]. The flowchart of our proposed method is shown in Fig. 2, where the upper flow is the training flow and the lower flow is the test flow.

In the training phase, perspective images are used to train the detector. The images are augmented by rotating them randomly. In this augmentation, the positions in the ground truth labels are changed but the sizes inside them are not. The rotated images and the converted ground truth labels are used to train the detector (Sec. 3.2)

In the test phase, pedestrians in omnidirectional images are detected. The images are input to the network without transforming them to perspective images. Then bounding box regression and classification are carried out. After that, BBR groups obtained bounding boxes to prevent overdetection, and angles of the grouped bounding boxes are then determined by their positions (Sec. 3.3 and 3.4).

The key novelty of our method is rotation data augmentation and BBR. The proposed rotation data augmentation enables rotation invariant bounding box regression in the test phase. Since the angles of bounding boxes are uniquely determined by their positions in omnidirectional images, rotation invariant bounding box regression is a good fit for omnidirectional pedestrian detection. The proposed BBR uses detected center points to group bounding boxes. This works well for the proposed detector, which outputs various significantly different sizes of bounding boxes for one pedestrian.



(a) Original image (b) Augmented images

Figure 3: Rotation data augmentation.

3.2. Rotation data augmentation

In the training phase, perspective images are randomly rotated before being input to the network. Figure 3 shows an example of an original image and augmented images. The key point of our data augmentation is that the sizes in ground truth labels are not changed even when images are rotated. The positions in ground truth labels are changed according to the rotated angles. The bounding box regression loss is calculated using the unchanged sizes and the changed positions. With the proposed data augmentation, the detector can learn to output a bounding box with the precise size regardless of the angles of pedestrians.

3.3. Bounding box refinement

The common approach to group output bounding boxes is based on NMS. However, we found that NMS does not work well because the proposed detector tends to output multiple bounding boxes to one pedestrian, which has a small IoU. Therefore, as an alternative, we propose BBR, which is based on the mean shift clustering [5, 17] of detected center points. Note that before clustering, we eliminate unreliable detection results whose confidence scores are smaller than a threshold because these noises degrade clustering performance. The confidence threshold is set to 0.05 in this paper.

Suppose there are N predicted bounding boxes after eliminating the unreliable results, denoted by $\mathcal{D} = \{(\boldsymbol{p}_i, \boldsymbol{q}_i, s_i) \mid 1 \leq i \leq N\}$, where $\boldsymbol{p}_i = (x, y)$ is detected center, $\boldsymbol{q}_i = (w, h)$ is detected size, and s_i is confidence score of detection. The proposed BBR consists of the mean shift clustering of box centers and box summarization.

In the mean shift clustering, the mean points are initialized by all the detected centers $\boldsymbol{m}_i^0 = \boldsymbol{p}_i$ $(i = 1, \dots, N)$. Then, each initialized mean point $\boldsymbol{m}_i$ is updated by iterating Eqs. (1) and (2) until convergence:

$$\alpha_{i,j}^t = \begin{cases} 1 & \left(\text{if } \|\boldsymbol{m}_i^t - \boldsymbol{p}_j\| \leq r\right) \\ 0 & (\text{otherwise}), \end{cases} \tag{1}$$

$$\boldsymbol{m}_i^{t+1} = \frac{\sum_{j=1}^{N} \alpha_{i,j}^t \boldsymbol{p}_j}{\sum_{j=1}^{N} \alpha_{i,j}^t}. \tag{2}$$

Here, r in Eq. (1) is the threshold of the nearest neighbor search and is set to 0.04 in normalized image coordinates in this paper. After the convergence, the distances between the mean points are calculated, and the point pairs whose distances are smaller than r in Eq. (1) are integrated, resulting in M clusters $\{\mathcal{C}_i\}_{i=1}^M$. Note that $\{\mathcal{C}_i\}_{i=1}^M$ with $\mathcal{C}_i \neq \emptyset$ gives a partition of $\{1, \dots, N\}$.

Finally, the position, size, and confidence score of the M clusters, namely $\boldsymbol{c}_i = (\boldsymbol{\mu}_i, \boldsymbol{\eta}_i, \gamma_i)$ $(i = 1, \dots, M)$, are obtained as follows: for $i = 1, \dots, M$,

$$\begin{bmatrix} \boldsymbol{\mu}_i & \boldsymbol{\eta}_i \end{bmatrix} = \frac{\sum_{j \in \mathcal{C}_i} s_j \begin{bmatrix} \boldsymbol{p}_j & \boldsymbol{q}_j \end{bmatrix}}{\sum_{j \in \mathcal{C}_i} s_j}, \tag{3}$$

$$\gamma_i = \max\{s_j \mid j \in \mathcal{C}_i\}. \tag{4}$$

3.4. Angle determination

In omnidirectional images, directions of pedestrians are uniquely determined by their positions. Given the center of the detected bounding box (x, y) and that of the input image (x_c, y_c), the angle θ of the bounding box can be calculated as follows:

$$\theta = \begin{cases} 0 & (\text{if } (x, y) = (x_c, y_c)) \\ \angle\left(x - x_c + \sqrt{-1}\,(y - y_c)\right) & (\text{otherwise}), \end{cases} \tag{5}$$

where $\angle z \in (-\pi, \pi]$ denotes the argument of complex number $z = a + \sqrt{-1}b$. The calculated angles are added to the outputs of the refinement process, and the final outputs of the proposed detector are $\mathcal{O} = \{(\boldsymbol{\mu}_i, \boldsymbol{\eta}_i, \gamma_i, \theta_i) \mid 1 \leq i \leq M\}$.

4. Experiments

4.1. Datasets

To evaluate the effectiveness of our proposed method, we performed experiments using four public datasets: MW-18Mar [1], PIROPO [2], Bomni [10], and the CVRG omnidirectional dataset (CVRG) [3]. We manually annotated pedestrian regions with both rotated and horizontal bounding boxes and used the labels of the rotated ones to evaluate the proposed method and those of the horizontal ones to evaluate other methods for comparison. The characteristics of each benchmark dataset are shown in Table 1.

4.2. Implementation details

The basic implementation follows the original YOLOv2 [32]. The difference is that BBR is used instead of NMS, and an angle determination process is added after BBR.

For training, we use the COCO object detection dataset [23]. COCO is one of the largest object detection datasets that is commonly used as a general object detection benchmark. We use this dataset to train the proposed detector because it contains various appearances of people.

Table 1: The characteristics of each dataset.

Dataset	Camera type	Situation	Used sequence	Camera position	#Frames	#GT Labels
MW-18Mar	Fisheye	Indoor	Test	5 Static	481	1,342
PIROPO	Fisheye	Indoor	Test2	4 Static	375	803
Bomni	Fisheye	Indoor	Scenario1	1 Static	337	1,122
CVRG	Catadioptric	Indoor & Outdoor	N/A	Dynamic	30	71

We extract data that are labeled as people from both the training and validation sets. To further improve the performance, we also use depth-based person identification from the top view dataset (DPI-T) [20]. DPI-T contains RGB-D images from top-view, but bounding box labels are not provided. Instead of manually annotating them, we automatically generate them by predicting the positions and sizes of the people from depth maps. Adding these top-view images to training data enables the detector to detect pedestrians right under omnidirectional cameras. In total, 273,469 bounding box labels from COCO and 5,337 bounding box labels from DPI-T are used for training.

In the training phase, we finetune detectors from the weights pretrained on ImageNet [11] using COCO and DPI-T for 30k iterations. We use a stochastic gradient descent optimizer with momentum, which is set to 0.9. The initial learning rate is set to 0.001, and decayed 0.1 every 10k iterations. Weight decay is set to 0.0005. Data augmentation of random translation, cropping, flipping, and color distortion are used in the same way as the original YOLOv2 implementation. In addition, the proposed rotation data augmentation is carried out for rotation invariant bounding box regression, as described in Sec. 3.2. The size of the input image is randomly changed every ten iterations in the same manner as the original implementation. The smallest size is 320×320 and the largest size is 608×608. By this method, the detector learns to predict well across a variety of input sizes, and offers an easy tradeoff between speed and accuracy.

In the test phase, the images are trimmed along their long side to make them square and the visual field is centered. Unless otherwise stated, the trimmed images are resized to 416×416. To improve the performance, augmented images are input to the network. The regression and classification results of augmented images are averaged with those of the original image. Unless otherwise stated, the augmentation is only horizontal flipping.

4.3. Benchmark results

To evaluate the detector performance, miss rates at every false positive per image (FPPI) and log average miss rates (LAMRs) are shown, which is described in [14]. The IoU threshold of associating a ground truth and a detection result is set to 0.5. To show the effectiveness of each method, we first compared four types on the benchmarks. All the types are finetuned on the COCO dataset. The types are as follows:

NMS This is the original YOLOv2. The predicted bounding boxes are grouped by NMS. We use the labels of horizontal bounding boxes to evaluate this method.

RotInv + NMS This is the proposed detector trained with proposed rotation data augmentation for rotation invariant bounding box regression. For calculating the IoU of rotated bounding boxes, which is necessary for NMS, we use the Monte-Carlo method proposed in [26].

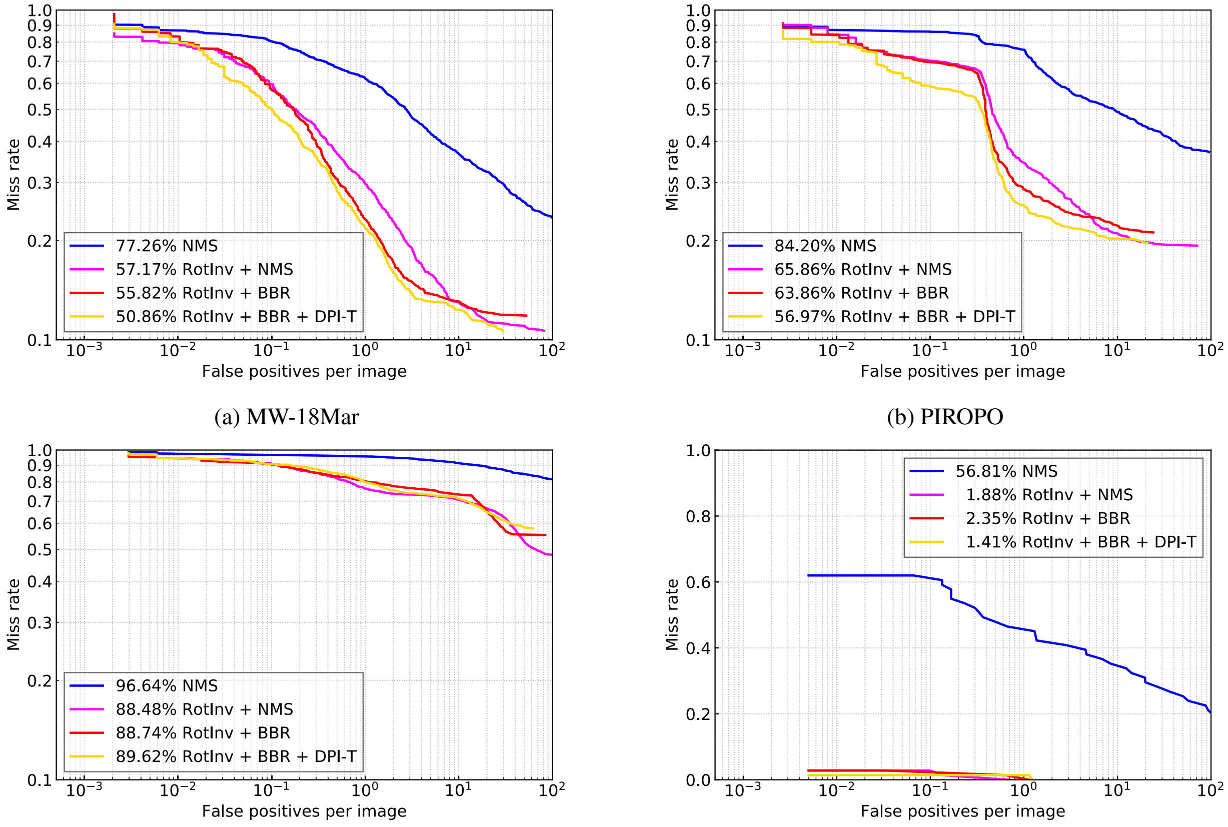
RotInv + BBR This is the proposed detector. Compared to the previous type, this type uses BBR for grouping bounding boxes instead of NMS.

RotInv + BBR + DPI-T This is the proposed detector. Compared to the previous type, this type is finetuned both on COCO and on DPI-T.

Figure 4 shows the evaluation results of the four benchmarks. Note that the miss rate in Fig. 4d is linear scaled, while that in the others is log scaled. The percentages in the legends express the LAMRs. In all the benchmarks, the LAMR of the NMS type is the highest compared to other types because it detects only pedestrians who have an upright appearance. In contrast, with rotation invariant bounding box regression, the LAMR decreases dramatically. This means that the proposed rotation invariant bounding box regression works well for detecting pedestrians at any angle. The effect is remarkable in CVRG: the LAMR decreases from 56.81% for NMS to 1.88% for RotInv + NMS.

In terms of grouping bounding boxes, the results show that BBR is slightly better than NMS on MW-18Mar and PIROPO, as shown in Figs. 4a and 4b. These two datasets have long sequences and diverse appearance of pedestrians, so they are more reliable to evaluate compared to the other two datasets. This means that BBR is more appropriate than NMS for the proposed detector that outputs significantly different sizes of bounding boxes to one pedestrian.

The effect of adding DPI-T to the training data is especially apparent in MW-18Mar and PIROPO. In MW-18Mar and PIROPO, people walk around here and there, which means they have a certain number of ground truth labels



(a) MW-18Mar

(b) PIROPO

(c) Bomni

(d) CVRG

Figure 4: Evaluation results of four benchmarks. Note that miss rate in (d) CVRG is linear scaled, while that in others is log scaled. Percentages in the legends express the LAMRs. RotInv: the detector is trained with the proposed rotation data augmentation. NMS: bounding boxes are suppressed with non-maximum suppression. BBR: bounding boxes are refined with the proposed bounding box refinement. DPI-T: DPI-T is added to training data.

specifying their positions right under the cameras. These results suggest that it is difficult to reproduce the appearance of people right under cameras from perspective images.

In CVRG, the LAMRs are quite low, except for the NMS type. This means that solving the rotation deformation problem is the most important for catadioptric cameras.

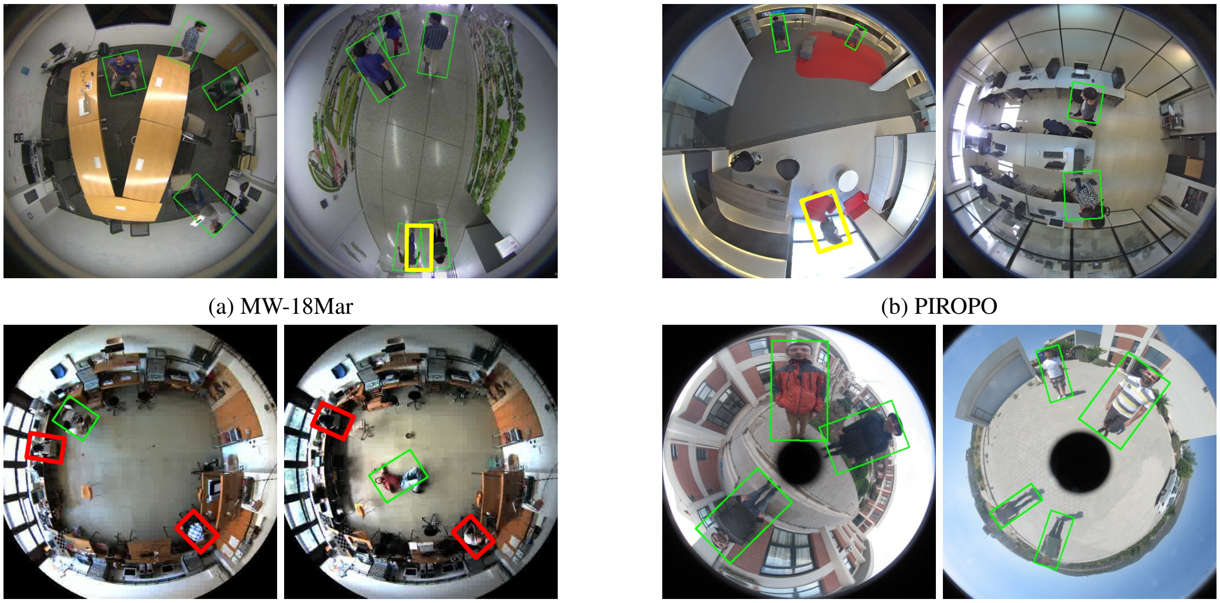
In Bomni, the LAMRs of all types are relatively high. We annotated all the pedestrians in the frames, and as a result, about two thirds of the ground truth labels are for sitting people. This means that the appearance of sitting people in omnidirectional images is hard to reproduce from perspective images.

Figure 5 shows a few of the detection results of the benchmarks. The green, yellow, and red boxes are true positives, false positives, and false negatives, respectively. In Fig. 5a, the area between two people is falsely detected. This type of false positive tends to occur in crowded areas. The yellow box in Fig. 5b is also a false positive, the size of which is not correctly predicted. When a part of the body is occluded, the size tends to be wrong. In Bomni, there are a lot of false negatives due to the appearance of sitting people, as explained above.

4.4. Comparison with previous method

To demonstrate the superiority of our method, we implement one baseline method and compared our method to it. The baseline is as follows:

Calibration This baseline is the original YOLOv2 with the camera calibration described in [34]. In this baseline, first, omnidirectional images are transformed into perspective images using calibrated camera parameters as described in [16]. YOLOv2 detects pedestrians in the transformed perspective images, and then the predicted bounding boxes are inverse transformed to the omnidirectional images. The final results are circumscribed horizontal bounding boxes of the inverse transformed bounding boxes. Normal NMS is replaced with soft NMS using Gaussian smoothing. Note that YOLOv2 is finetuned using only data labeled as people



(a) MW-18Mar

(b) PIROPO

(c) Bomni

(d) CVRG

Figure 5: Detection results of four benchmarks. Green boxes are true positives, yellow boxes are false positives, and red boxes are false negatives.

in COCO, the same as with the proposed detector.

For PIROPO and Bomni, provided calibration parameters are used for transforming the images into perspective images. For MW-18Mar and CVRG, calibration parameters are not provided, so we estimate them from the sizes of the people in the images. For fair comparison, we show the results of the proposed detector finetuned only on COCO.

Figure 6 shows the comparison results. The numbers at the end of legends refer to the size of the input image. Even though our method does not require any camera calibration, it outperforms the baseline method. The performance of our method increases when the size of the input image becomes large, while the performance of the baseline method degrades even though the size becomes large on MW-18Mar, Bomni, and CVRG, as shown in Figs. 6a, 6c, and 6d. This means that the detection performance is saturated at a certain point due to the interpolation artifacts of image transformation. In terms of detection speed, our method has an advantage because the baseline method has to carry out detection in overlapped multiple perspective images transformed from one omnidirectional image.

4.5. Test-time data augmentation

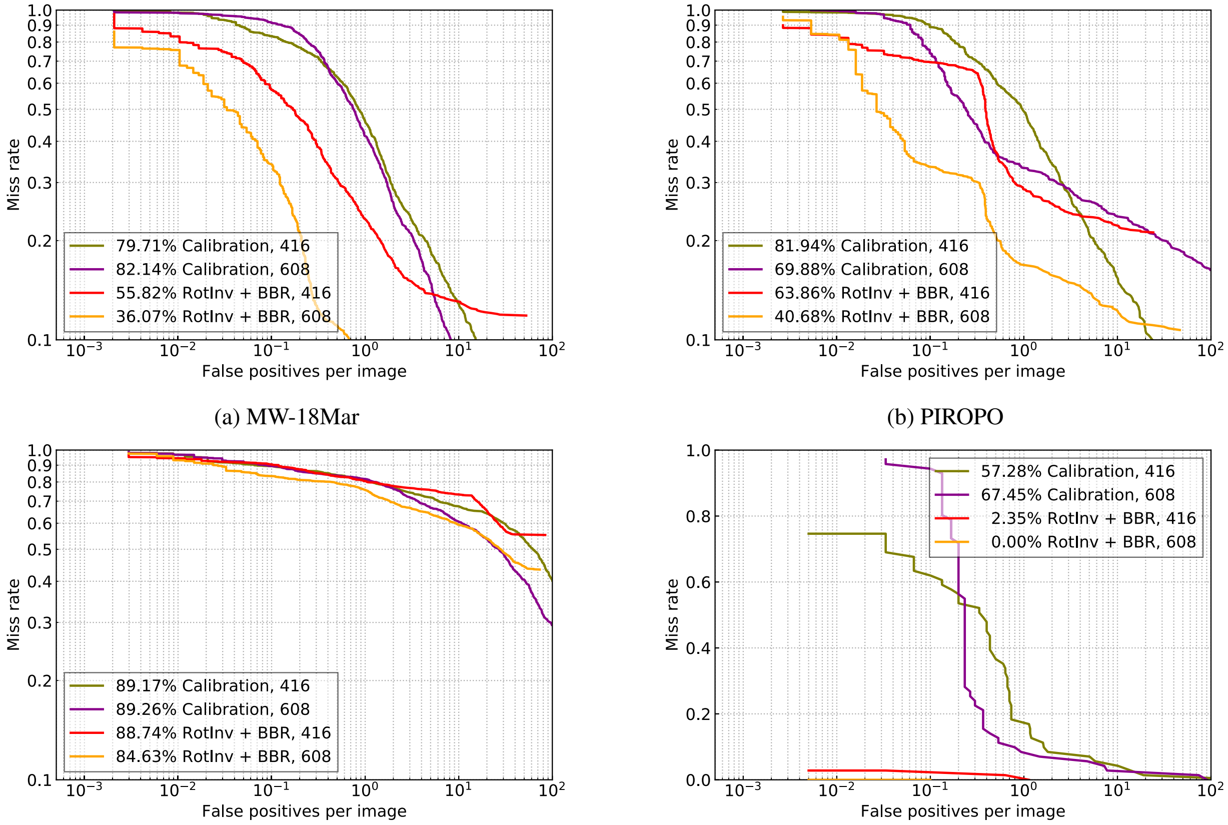
Since the proposed detector can detect pedestrians for any angles, the detection performance can be improved not only by horizontal flipping (HF) but by other test-time augmentation methods, e.g. vertical flipping (VF) and rotation. To test the effect of the augmentation, we augmented input images by combining HF, VF, and rotation of 90 degrees (Rot90). The LAMRs are shown in Table 2. When the number of the augmented images increases, the performance improves except for CVRG. As the rotation angle for the augmentation is not limited to 90 degrees, the performance is possibly improved by augmenting images with the rotation of other angles. In addition to changing the size of the input image, this augmentation also offers an easy tradeoff between speed and accuracy.

Table 2: LAMRs for various test-time augmentation.

		Dataset			
Augmentation	#Input	MW-18Mar	PIROPO	Bomni	CVRG
HF	2	50.86	56.97	89.62	**1.41**
HF + VF	4	49.84	54.01	89.60	4.69
HF + VF + Rot90	8	**47.98**	**49.61**	**88.74**	5.32

4.6. Position dependency

For deeper understanding of the proposed detector, we make an analysis on the relationship between LAMR and the pedestrian's position in the images using MW-18Mar dataset. To remove the dataset bias on the number of pedestrians and the detection difficulty for each angle, we rotate each image at 5 degree intervals and calculate LAMRs for all the rotated images. We used horizontal flipping for test-time data augmentation in this analysis. Figure 7a shows LAMRs on each distance from the image center. The results show that it is difficult to detect pedestrians at the center or



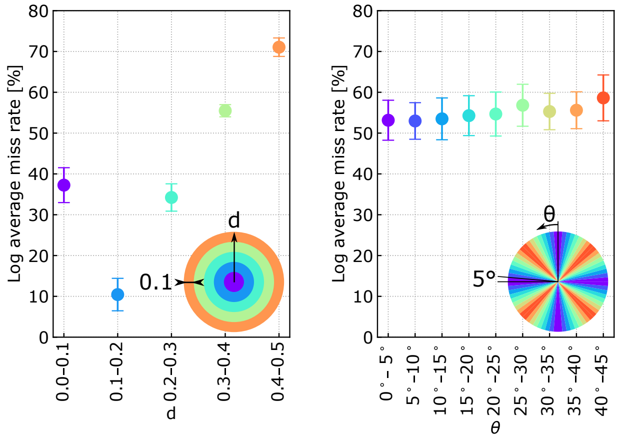
(a) MW-18Mar

(b) PIROPO

(c) Bomni

(d) CVRG

Figure 6: Comparison results of four benchmarks. Note that miss rate in (d) CVRG is linear scaled, while that in others is log scaled. Percentages in the legends express the LAMRs. The numbers at the end of legends are the size of the input image. Calibration: original YOLOv2 applied to transformed perspective images. RotInv + BBR: proposed method.



(a) LAMRs on each distance

(b) LAMRs on each angle

Figure 7: LAMRs on each position for MW-18Mar dataset. The bars in each graph express standard errors. The marker colors denote the positions in the circle.

the edge of the field of vision. This means it is difficult to detect pedestrians taken from their overhead or at tiny scales. Performance improvements of detecting pedestrians at such conditions are left to a future work. Figure 7b shows LAMRs on each angle. The results show that the detection difficulty does not depend on the angles of pedestrians; we conclude that the proposed detector successfully detects rotated pedestrians.

5. Conclusion

In this paper, we have proposed a rotation invariant training method for omnidirectional pedestrian detection. Our method deals with the lack of omnidirectional images to train CNNs by utilizing large-scale perspective image datasets. We also proposed BBR for grouping bounding boxes. BBR works well for the proposed detector that outputs multiple bounding boxes of significantly different sizes to one pedestrian. Experiments on four public benchmarks show that the proposed detector can achieve state-of-the-art performance in omnidirectional pedestrian detection.

References

[1] Mirror worlds challenge. https://icat.vt.edu/mirrorworlds/challenge/index.html.

[2] People in indoor rooms with perspective and omnidirectional cameras (PIROPO) database. https://sites.google.com/site/piropodatabase/.

[3] Y. Bastanlar, A. Temizel, Y. Yardimci, and P. Sturm. Multi-view structure-from-motion for hybrid camera scenarios. *Image and Vision Computing*, 30(8):557–572, 2012.

[4] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? In *ECCV*, pages 613–627, 2014.

[5] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE TPAMI*, 17(8):790–799, 1995.

[6] A.-T. Chiang and Y. Wang. Human detection in fish-eye images using HOG-based detectors over rotated windows. In *ICMEW*, pages 1–6, 2014.

[7] I. Cinaroglu and Y. Bastanlar. A direct approach for human detection with catadioptric omnidirectional cameras. In *SIU*, pages 2275–2279, 2014.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[9] M. Demirkus, L. Wang, M. Eschey, H. Kaestle, and F. Galasso. People detection in fish-eye top-views. In *VISIGRAPP*, pages 141–148, 2017.

[10] B. E. Demiröz, İ. Ari, O. Eroğlu, A. A. Salah, and L. Akarun. Feature-based tracking on a multi-omnidirectional camera dataset. In *ISCCSP*, pages 1–5, 2012.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[12] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE TPAMI*, 36(8):1532–1545, 2014.

[13] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, pages 91.1–91.11, 2009.

[14] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE TPAMI*, 34(4):743–761, 2012.

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

[16] M. Findeisen, L. Meinel, M. Heß, A. Apitzsch, and G. Hirtz. A fast approach for omnidirectional surveillance with multiple virtual perspective views. In *EUROCON*, pages 1578–1585, 2013.

[17] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE TIT*, 21(1):32–40, 1975.

[18] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, pages 3354–3361, 2012.

[19] R. Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015.

[20] A. Haque, A. Alahi, and L. Fei-Fei. Recurrent attention models for depth-based person identification. In *CVPR*, pages 1229–1238, 2016.

[21] J. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians. In *CVPR*, pages 4073–4082, 2015.

[22] O. Krams and N. Kiryati. People detection in top-view fisheye imaging. In *AVSS*, pages 1–6, 2017.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014.

[24] L. Liu, Z. Pan, and B. Lei. Learning a rotation invariant detector with rotatable bounding box. *arXiv preprint arXiv:1711.09405*, 2017.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, pages 21–37, 2016.

[26] Y. Liu and L. Jin. Deep matching prior network: Toward tighter multi-oriented text detection. In *CVPR*, pages 3454–3461, 2017.

[27] L. Meinel, C. Wiede, M. Findeisen, A. Apitzsch, and G. Hirtz. Virtual perspective views for real-time people detection using an omnidirectional camera. In *IST*, pages 312–315, 2014.

[28] W. Nam, P. Dollár, and J. H. Han. Local decorrelation for improved pedestrian detection. In *NIPS*, pages 424–432, 2014.

[29] T. B. Nguyen, S.-T. Chung, S. Cho, et al. Real-time human detection under omni-directional camera based on CNN with unified detection and agmm for visual surveillance. *Journal of Korea Multimedia Society*, 19(8):1345–1360, 2016.

[30] J. Noh, S. Lee, B. Kim, and G. Kim. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In *CVPR*, pages 966–974, 2018.

[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[32] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In *CVPR*, pages 6517–6525, 2017.

[33] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[34] R. Seidel, A. Apitzsch, and G. Hirtz. Omnidetector: With neural networks to bounding boxes. *arXiv preprint arXiv:1805.08503*, 2018.

[35] L. Zhang, L. Lin, X. Liang, and K. He. Is faster R-CNN doing well for pedestrian detection? In *ECCV*, pages 443–457, 2016.

[36] S. Zhang, C. Bauckhage, and A. B. Cremers. Informed Haar-like features improve pedestrian detection. In *CVPR*, pages 947–954, 2014.

[37] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? In *CVPR*, pages 1259–1267, 2016.

[38] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. Towards reaching human performance in pedestrian detection. *IEEE TPAMI*, 40(4):973–986, 2018.

[39] S. Zhang, R. Benenson, and B. Schiele. CityPersons: A diverse dataset for pedestrian detection. In *CVPR*, pages 3213–3221, 2017.

[40] S. Zhang, R. Benenson, B. Schiele, et al. Filtered channel features for pedestrian detection. In *CVPR*, pages 1751–1760, 2015.

[41] S. Zhang, J. Yang, and B. Schiele. Occluded pedestrian detection through guided attention in CNNs. In *CVPR*, pages 6995–7003, 2018.