論文 / 著書情報
Article / Book Information

Title	RI-DC: Rotation-Invariant Detection and Classification for Wheat Head Detection
Author	Takeru Ito, Kuniaki Uto, Koichi Shinoda
Journal/Book name	IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium Proceedings, , , pp. 5750-5753
Pub. date	2022, 7
DOI	https://doi.org/10.1109/IGARSS46834.2022.9883405
Copyright	(c)2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Note	This file is author (final) version.

RI-DC: ROTATION-INVARIANT DETECTION AND CLASSIFICATION FOR WHEAT HEAD DETECTION

Takeru Ito, Kuniaki Uto, Koichi Shinoda

Tokyo Institute of Technology

ABSTRACT

We propose a novel automatic detection method of wheat heads from images taken from above wheat fields. The automatic detection of heads is useful for predicting yields. For this purpose, deep learning based methods has proved to be effective recently, but they are not robust against the variations of head directions. To tackle this problem, we utilize a two-step approach which first carries out object detection for augmented test images rotated in many directions, and then classifies the detected objects by using a classifier trained with many rotated images. It was evaluated by using GWHD dataset and proved to be effective.

***Index Terms*—** rotation-invariant object detection, wheat head detection

1. INTRODUCTION

Wheat is the most cultivated cereal crop in the world. Usually, the amount of its yield can be found only after its harvest. If we can precisely predict its yield before its harvest, we can deal with climate changes, unstable market demands, and various wheat diseases. Until now, yield is estimated by manually counting the number of wheat heads of a small portion of fields, which is costly and not precise. Recently, a drone has been often used to take images from the sky. Object detection using deep learning from images (e.g. [1]) has been proved to be effective. There have been several studies which use these techniques for automatically detecting wheat heads [2–4].

Wheat heads have various directions in such images taken. Most deep learning based object detection methods are not designed to deal with these variations. If we can make them robust against the rotation, its performance would much improve.

In this paper, we propose Rotation-Invariant Detection and Classification (RI-DC) for wheat head detection. RI-DC consists of two steps. First, to achieve rotation-invariant prediction, we augment test images by rotating them (test-time augmentation). Unlike the conventional method [5] that uses rotations with 90-degree interval, it can use rotations with any intervals. We call this procedure Rotation-Invariant Detection (RI-D). While RI-D detects objects overlooked by previous methods, i,e., has high recall rate, it tends to generate many false positives. To reduce them, as the second step, we use a Rotation-Invariant Classification (RI-C) that adapts a rotation-invariant classifier [6] to identify whether each detected object is a head or not.

2. RELATED WORKS

2.1. Wheat Head Detection

Object detection is a computer vision task that locates objects in an image and identifies their classes. As deep learning has evolved, numerous object detection methods have been proposed [1] [7] and researchers have begun to use them for wheat head detection [2–4]. Hasan $et\ al.$ [3] demonstrated the potential of Faster-RCNN [1] for the task. Gong $et\ al.$ [4] adapted YOLO [8] to achieve real-time wheat head detection. Unlike these previous researches that apply off-the-shelf object detection networks, we develop rotation-invariant object detection to detect wheat heads with various directions more precisely.

2.2. Rotation-invariant Detection

Although there have been several rotation-invariant detection methods [9, 10], they used images labeled with angles for training, which needs costly annotation. In contrast to them, we explore an approach to augmenting images at testing stage, Test Time Augmentation [5], which predicts outputs from multiple augmented copies of an image in the test set and then ensembles the prediction results. With regard to object detection, augmentation on a test image is usually a compilation of detection results from images rotated with 90-degree intervals. In contrast, we use images rotated with any intervals.

Several rotation-invariant image recognition methods have been proposed. Worrall $et\ al.$ [11] used filters from a family of circular harmonics to obtain rotation-invariance. Patrick $et\ al.$ [12] rotated convolution filters and back-rotates the corresponding feature map. Some methods [12, 13] labeled rotated samples as positive and feed those samples into a network independently to train it. However, when appearance of rotated positive sample is similar to that of negative, its performance degrades. On the other hand, TI-Pooling [6] inputs an original and rotated samples to a CNN network at once. Rotation-invariant features are achieved by max-pooling of features from those samples. In TI-Pooling, the information from the other samples of the same image are utilized to identify a positive sample as positive even when it looks like negative.

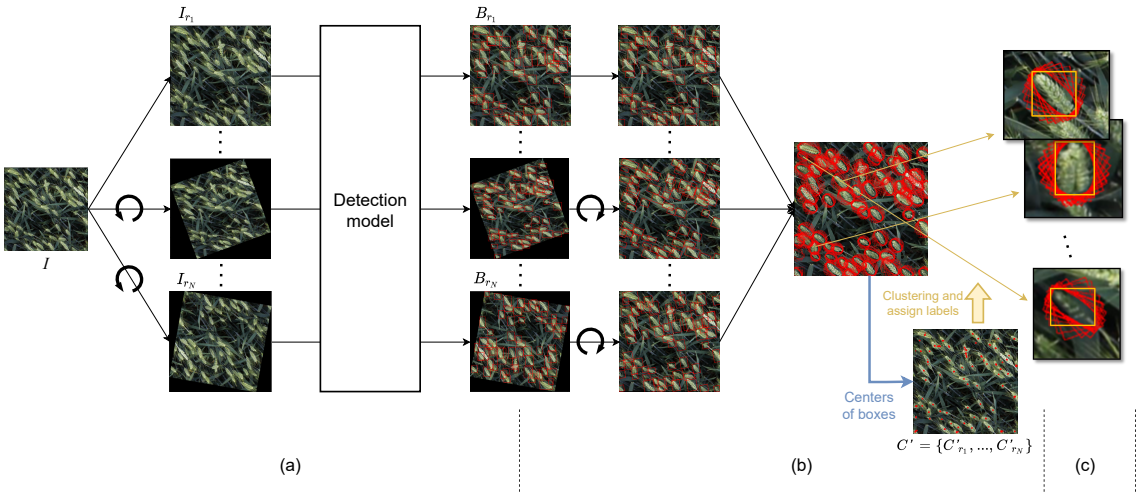

Fig. 1. The overview of RI-D. (a) An original image is duplicated by rotating it and object detection is applied for each of them. (b) All detected images are rotated back to 0 degree, and then a mean-shift algorithm is used for grouping bounding boxes. (c) The smallest rectangle that encloses the product set of all bounding boxes is defined as the final prediction.

3. PROPOSAL METHOD

The proposed RI-DC consists of two components: (3.1) Rotation-Invariant Detection (Figure 1) and (3.2) Rotation-Invariant Classification (Figure 2).

3.1. Rotation-Invariant Detection (RI-D)

The key assumption of the proposed rotation-invariant detection, RI-D, is that a single model can realize rotation-invariance by rotating an input image at the inference stage. RI-D consists of the following three steps (Figure 1(a)(b)(c)).

a). An original image is duplicated and rotated to generate a set of N images $\boldsymbol{I} = \{I_r | r = r_1, r_2, ..., r_N\}$. Here, $r_i = 360° \times (i-1)/N, (i = 1, ..., N)$. I_r represents an image rotated around its center by r degrees. The size of each rotated image is adjusted to that of the original image, and its pixels not overlapped with the original image are filled with zeros. $\boldsymbol{I}$ is then input to a detection model and the model predicts bounding boxes(bboxes), each with a detection score which represents how likely the corresponding box contains the target object. We select 100 bboxes with high detection scores for each image. We select this number because the average number of heads in an image is around 60. Then, we apply Non Maximum Suppression (NMS) [14] to reduce the number of boxes in I_r to n_r by eliminating redundant boxes. Finally we obtain $\boldsymbol{B} = \{\boldsymbol{B}_r | r = r_1, ..., r_N\}$ where $\boldsymbol{B}_r$ represents a prediction result for the I_r, which consists of n_r predicted boxes $\boldsymbol{B}_r = \{b_{r,j} | j = 1, ..., n_r\}$. Each $b_{r,j}$ is with detection score $s_{r,j}$.

b). To ensemble boxes in $\boldsymbol{B}$ to make final predictions, we estimate which boxes in $\boldsymbol{B}$'s with different r include the same object. Let a center coordinate of $b_{r,j}$ be $c_{r,j}$. Then image I_r is rotated around its center by $-r$ degrees to be reset to 0-degree image coordinate. For each $c_{r,j}$, $c'_{r,j}$ denotes the center coordinate after rotating back to 0 degree. Then we cluster a set of all centers $C' = \{c'_{r,j} | r = r_1, ..., r_N, j = 1, ..., n_r\}$ over all the angles over all the rotated images. We employ mean-shift algorithm [15], which is an algorithm that aims to discover the maximum of a probability density given by samples. For each $c'_{r,j}$, the local maximum of the density of c's within a bandwidth, which is a predefined parameter, is calculated iteratively until convergence. The centers converging to the same maximum belong to the same cluster.

c). For each cluster, we define the smallest upright rectangle that encloses the product set of all boxes as the predicted box, and its detection score is calculated by averaging scores of bboxes in the cluster.

RI-D feeds an image into the detection network N times. We do not have to limit the number N. As a detector, we select EfficientDet [7] for its balanced computational efficiency and

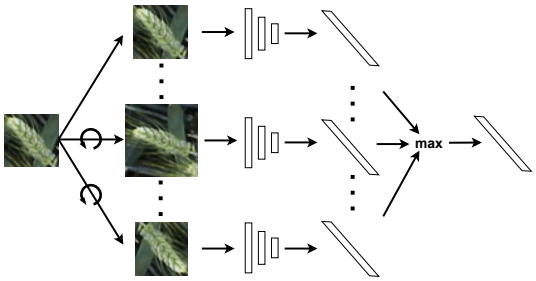

Fig. 2. The overview of RI-C. We apply TI-Pooling [6] as the classifier. Copied and rotated boxes are fed into a CNN network. Rotation-invariant features are achieved by max-pooling.

accuracy.

3.2. Rotation-Invariant Classification (RI-C)

RI-D generates predictions N times with different I_rs and may detect negative samples with high detection scores as positive in some I_r's even after NMS and clustering. Thus, RI-D tends to generate many false positives. To screen them, we add a subsequent classification network which classifies the predictions into wheat heads (true positive) and background (false positive).

We use TI-Pooling [6] for the classifier. In TI-Pooling, each bbox generated by the detection model is augmented with rotation and fed into the network all at once to extract rotation-invariant features (Figure 2). An image cropped with a predicted bbox that has a detection score s_b is input to the TI-Pooling network and the network outputs s_t. The final detection score s_a is calculated by

$$s_a = ws_b + (1 - w)s_t, \tag{1}$$

where w is a control parameter. We call this part as RI-C.

3.3. RI-DC

In RI-DC, RI-D and RI-C are concatenated in series. In the first step, RI-D detects most of the candidates with high recall, RI-C then screens the candidates.

4. EXPERIMENTS

4.1. Dataset

We use Global Wheat Head Detection dataset [16] to evaluate our method. It consists of high resolution RGB images of wheats with various sizes, growth stages and genotypes collected from around the world. The images were taken from a height in the range of 1.8 m to 3 m above the ground. We use images from source [usask_1, arvalis_1, inrae_1, arvalis_3, rres_1] as a training set and [ethz_1, arvalis_2] as a validation set. The training and validation sets contain 2422 and 951 images, respectively.

4.2. Implementation Details

To evaluate the detection performance, EfficientDet is trained as the baseline and the same model is used for RI-DC to validate effectiveness. In the training of EfficientDet, the images are resized to 512x512 with batch size 4, epoch 100 and learning rate 0.0002. In RI-DC, N is 18, which means the rotation interval is 20 degrees.

In RI-C training, positive and negative samples are fed into the classification after being resized to 32x32. Its training samples are predicted boxes from a detection model trained with GWHD training set. We select positive samples that have higher than 0.7 IoU with the ground truth, and negatives that have lower than 0.5. Cross Entropy Loss is used for the loss function with batch size 64, epoch 50 and learning rate 0.00001.

We run inference on a NVIDIA Titan Xp GPU that has 12GB memories and Mean Average Precision (mAP) whose IoU thresholds is 0.5 (mAP_{50}) is used as evaluation metrics.

Table 1. Comparison with baseline

method	$\text{mAP}_{50}(\%)$
EfficientDet [7]	86.0
RI-D	87.9
RI-C	87.4
RI-DC	**88.4**

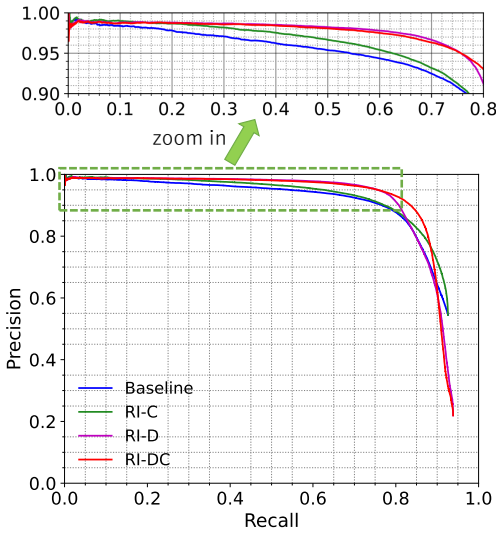

Fig. 3. PR curve comparison of the proposed methods with the baseline.

We select the optimal bandwidth used in mean-shift, 10 pixel, by our preliminary experiment and $w = 0.5$ used in (1).

4.3. Experiment Results

Table 1 shows mAP comparison between the baseline, RI-D, RI-C and RI-DC. Figure 3 shows their Precision and Recall curves (PR curve). RI-D and RI-C have higher mAP_{50} than the baseline (86.0%) by 1.9 point and 1.4 point, respectively. Furthermore, their combination is higher by 2.4 point. These results demonstrate that our proposed method is effective.

We further conducted ablation studies with different $N = [1, 2, 4, 9, 18]$. For instance, $N = 1$ means $r_1 = 0$ and $N = 2$ means $r_1 = 0, r_2 = 180$. Figure 4 shows PR curves for different Ns. As the value N increases, the performance improves. Particularly, improvements can be seen in between $N = 4$, $N = 9$ and $N = 18$. On the other hand, a downside of our method is that the larger N is, the longer it takes to infer a single image. In other words, there is a trade-off between the inference speed and accuracy.

4.4. Qualitative evaluation

To evaluate our proposed method qualitatively, we visualize some examples of detection results in Figure 5. In the upper row, RI-D detects heads that are not detected by the baseline. In the lower row, a large predicted box is falsely generated by RI-D but properly eliminated by the subsequent RI-C. It means RI-C correctly identifies false positive.

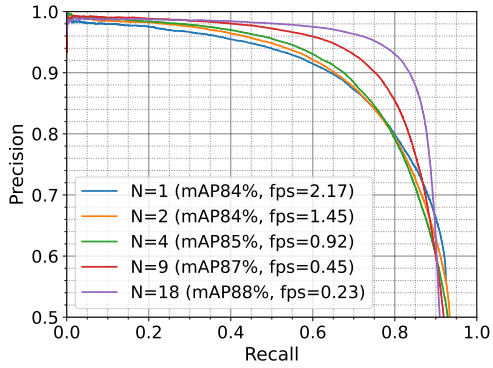

Fig. 4. Comparison of mAP and inference speed when changing N.

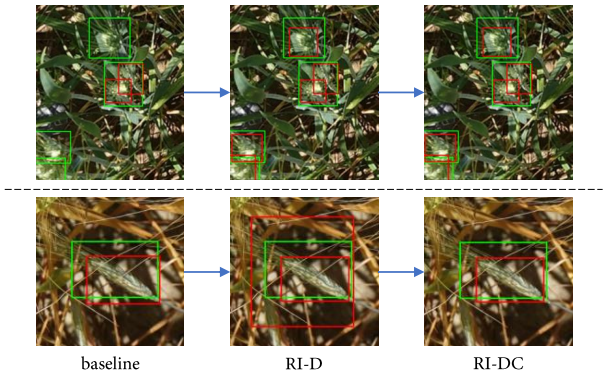

Fig. 5. Examples of detection. Green box and red box represent the ground truth and the predicted box, respectively. From left to right, (1) baseline, (2) after RI-D and (3) after RI-DC.

5. CONCLUSION

In this paper, we proposed RI-DC, which is composed of rotation-invariant object detection and subsequent rotation-invariant classification. Compared to the baseline, RI-D and RI-C improve mAP by 1.9 point and 1.4 point respectively, and RI-DC achieves the highest improvement by 2.4 point. From these results, we conclude that our proposed rotation-invariant model is effective for the automatic wheat head detection. For further study, we will explore the way to speed up inference.

6. REFERENCES

[1] Shaoqing Ren *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, Cambridge, MA, USA, 2015, NIPS'15, p. 91–99, MIT Press.

[2] Fares Fourati *et al.*, "Wheat head detection using deep, semi-supervised and ensemble learning," *Canadian Journal of Remote Sensing*, vol. 47, no. 2, pp. 198–208, Mar 2021.

[3] Md Hasan *et al.*, "Detection and analysis of wheat spikes using convolutional neural networks," *Plant Methods*, vol. 14, 11 2018.

[4] Bo Gong *et al.*, "Real-time detection for wheat head applying deep neural network," *Sensors (Basel, Switzerland)*, vol. 21, 2021.

[5] Ángela Casado-García *et al.*, "Ensemble methods for object detection," in *ECAI*, 2020.

[6] D. Laptev *et al.*, "TI-POOLING: Transformation-invariant pooling for feature learning in convolutional neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 289–297, 2016.

[7] Mingxing Tan *et al.*, "Efficientdet: Scalable and efficient object detection," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10778–10787, 2020.

[8] Joseph Redmon *et al.*, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[9] Xue Yang *et al.*, "Learning high-precision bounding box for rotated object detection via kullback-leibler divergence," *ArXiv*, vol. abs/2106.01883, 2021.

[10] Xue Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8231–8240, 2019.

[11] Daniel E. Worrall *et al.*, "Harmonic Networks: Deep Translation and Rotation Equivariance," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7168–7177, 2017.

[12] Patrick Follmann *et al.*, "A rotationally-invariant convolution module by feature map back-rotation," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 784–792, 2018.

[13] Gong Cheng *et al.*, "Rifd-cnn: Rotation-invariant and fisher discriminative convolutional neural networks for object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2884–2893.

[14] Navaneeth Bodla *et al.*, "Soft-NMS — improving object detection with one line of code," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5562–5570.

[15] K. Fukunaga *et al.*, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 32–40, 1975.

[16] E. David *et al.*, "Global Wheat Head Detection (GWHD) Dataset: A Large and Diverse Dataset of High-Resolution RGB-Labelled Images to Develop and Benchmark Wheat Head Detection Methods," *Plant Phenomics*, vol. 2020, 2020.