15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

# Refining Faster-RCNN for Accurate Object Detection

Myung-Cheol Roh
Kakao corp.
242, Cheomdan-ro, Jeju-si, Jeju-do, Korea
joshua.roh@kakaocorp.com

Ju-young Lee
Kakao corp.
242, Cheomdan-ro, Jeju-si, Jeju-do, Korea
michael.lee@kakaocorp.com

## Abstract

*Object detector with region proposal networks such as Fast/Faster R-CNN [1, 2] have shown the state-of-the art performance on several benchmarks. However, they have limited success for detecting small objects. We argue the limitation is related to insufficient performance of Fast R-CNN block in Faster R-CNN. In this paper, we propose a refining block for Fast R-CNN. We further merge the block and Faster R-CNN into a single network (RF-RCNN). The RF-RCNN was applied on plate and human detection in RoadView image that consists of high resolution street images (over 30M pixels). As a result, the RF-RCNN showed great improvement over the Faster-RCNN.*

Figure 1: In RoadView service, human faces and license plates should be blurred in order to protect privacy before the images are opened in public.

## 1 Introduction

Object detection is a challenging problem in computer vision. Classical object detection based on sliding windows, multi-scale and/or cascade approaches have shown successful results on several applications. Recently, new approaches based on deep learning outperformed the classical methods in various applications. Region-based convolutional network (RCNN) is one of the state art detection methods [4]. The RCNN proposed an efficient method that utilizes a feature map which is an output from a convolutional network. Around same time, many approaches based on feature map utilization was proposed [1, 2, 8, 9, 10, 11] They have achieved the state-of-the art performance. The first row in Table. 1 shows plate and human detection result of the Faster-RCNN. As shown in the Table 1, they have limitation for detecting relatively small objects in images where large (human) and small objects (plate) are presented mixed together.

To overcome the problem, we propose to use a refining stage for classification and bounding box regression blocks. The proposed method is implemented based on the Faster-RCNN architecture, however, it can be utilized with all the other CNN (Convolutional Nueral Network) based object detection architectures. We applied our method on *RoadView* images which Figure. 1 shows an example image. [1] As a result, we obtain more than 1.6 times improvement over Faster-RCNN for license plates detection while the performance of human detection is preserved.

## 2 Previous work

Classical object detection based on sliding windows and cascade method achieved fast and reasonable accuracy on several applications such as face detection and human detection. They also showed successful results on high resolution images for detecting faces and license plates [3]. However, they have limitations on accuracy and expanding multi-class. The deep neural network object detector overcome the limitations and show significant results on several benchmarks such as ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [5]. The success is due to considerable two components: an region proposal block and a classifier block.

**Region proposal block** Selective search [6] is a good partner with an earlier object detector. The RCNN is the one of the successful methods using the selective search. The RCNN is very slow because it requires CNN feature map calculation for every proposals obtained from the selective search. To overcome the problem, fast-RCNN was proposed [1]. The fast-RCNN proposed a ROI pooling method that utilizes a feature map for all proposals without re-calculating them. However, the proposal and RCNN blocks cannot be trained simultaneously. Recently, a region proposal network based method which provides an end-to-end training both the proposal and RCNN blocks was proposed [2]. It has shown the best performance in many applications and databases including ILSVRC and Microsoft COCO [7].

**Classifier block in object detection pipeline** The classifier block is composed of bounding box regressor and object classifier [1, 2]. They take candidate boxes from the output of region proposal block as inputs. A classifier classifies the candidate boxes and regressor finds precise positions and sizes of them If an input candidate box is small or far from true bounding box, their outputs are imprecise. GBD-Net [8] proposed a method that overcomes the case where an input candidate box is smaller than true bounding box. They won on an object detection track in ILSVRC 2016. Multi-box [9], YOLO [10], and SSD [11]

[1] The RoadView images are street-level scene images as google street-view [3]. Faces and license plates on images should be detected and blurred due to privacy issue. In the images, large objects and small objects are presented mixed together.

also achieved good performances for the speed and accuracy. Concurrent with our work, RefineNet [13] iteratively re-used the output from the Fast R-CNN. They did not demonstrate to multi-class and multi-size objects simultaneously. In addition, they did not present analysis with respect to the refinement gain.

## 3 Refining Faster-RCNN (RF-RCNN)

### 3.1 Faster-RCNN

The conventional Faster-RCNN is illustrated in Figure. 2a. It is composed of 5 main parts: a deep fully convolutional network, region proposal network, ROI pooling and fully connected networks, bounding box regressor, and classifier. Through the deep fully convolutional and region proposal networks, a lot of object candidates are proposed and the candidate regions (proposals) are normalized through ROI Pooling layer. Then, fully connected layers extract good features to conduct classification and regression. Most of the proposals are not always fit to target objects and the regressor tries to find a best fits. However, the classification is conducted on the proposals rather than on the best fits. In practical use, the size of target objects are various: for an example, car, human, and license plate. If sizes of target objects are various, many of proposals are larger so that it contains small objects as well as large objects. In this case, classifier should deal with not only the variations of the object region itself but also the variations of the outside region of the objects. It is not efficient for both classifier and regressor and produces false positives and false negatives in test. In order to cope with this problem, we introduce a refining stage.

### 3.2 Refining layer

The proposed RF-RCNN is composed of Faster-RCNN and two refinement steps: the refinement steps are classification refinement and regression refinement steps. As it is shown in the Figure. 2b, the refinement steps are structured nearly as same as the classification and regression steps in Faster-RCNN except that an additional fully connected layer is put before the regressor.

Classification refinement takes convolutional feature maps and ROIs from Faster-RCNN block as inputs. Note that the ROIs from Faster-RCNN block are outputs of the regressor in the Faster-RCNN. The ROIs are given either to fit closer to target objects than the proposal regions or to contains background. Thus, it is natural to expect better classification performance.

Regression refinement moves the ROIs toward target objects more precisely. Bounding boxes placed just around a target object after regression refinement are gathered and merged into one. In order to merge them, non-maximal suppression is applied as in Faster-RCNN.

### 3.3 Training

The RF-RCNN is trained through 3 stages: faster-rcnn block, classification refinement block, and regression refinement block training stages. Faster-RCNN is trained as same as it is in [2]. Given a Faster-RCNN model classification and regression refining networks are trained. In the refinement stages, the output regions of Faster-RCNN are taken as inputs to regression refining network.

During training classification refinenment block, regressor is also trained without fully connected layer. Soft-max and L1 loss functions are used for classification and regression refining networks.

After training classification refinement stages, a regressor refining network (two fully connected layers) is trained.

In the refinement network training stages, jittering technique is applied to the ROIs. Several bounding boxes are randomly sampled around the ROIs and they are fed into the refining stages. Then, the position variations are learned and it is expected to increase localization accuracy in testing. Given a ROI centered at $(cx_0, cy_0)$ with width $w_0$ and height $h_0$, a sample bounding box centered at $(cx_1, cy_1)$ is sampled as follows:

$$\begin{aligned} cx_1 &= \Delta x + cx_0 \\ cy_1 &= \Delta y + cy_0 \\ w_1 &= \Delta s + w_0 \\ h_1 &= \Delta s + h_0. \end{aligned} \quad (1)$$

$\Delta x$, $\Delta y$, $\Delta s$ are uniformly distributed random numbers in $[-r/k, r/k]$ where $r$ is $min(w_0, h_0)$ and $k$ is a constant. In the classification refining network, a few samples, 5, are used and in the regression refining network many samples, 100, are used so that all the bounding boxes are aligned toward target objects.

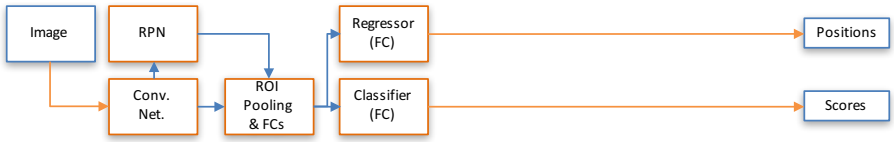
## 4 Experiment

We tested the proposed RF-RCNN on 900 road-view images, which are of 8000×4000 pixels. An input image is divided into several patches of 800×800 pixels with overlaps and each of the patches is fed into the RF-RCNN.

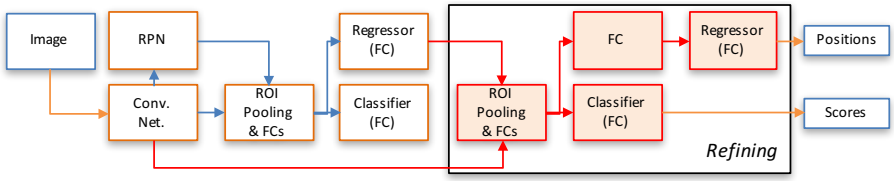
Figure 3 shows detection results using RF-RCNN and Faster-RCNN. In order to obtain the result image, thresholds were set at the point where both methods produce the same recall rate. First row, license plate detection results, shows that the proposed RF-RCNN detected correct one and no false positives. Meanwhile, Faster-RCNN gave several false positive results. Second row, human detection, shows that RF-RCNN detected one false positive (a dummy in red shirts) and missed one person. Faster-RCNN missed two people and detected no false positive.

Figures 4b and 4c shows ROC (Receiver Operating Characteristic) curves. '2nd layer' represents result after applying 'ROI Pooling and FCs' and 'classifier' in the Figure 2b. '3rd layer' represents result after applying all the refining steps after Faster-RCNN block in the Figure 2b, which is the proposed RF-RCNN.

As shown in Figure 4a overall performances of both the RF-RCNN and Refining-classification methods are increased. In Figure 4b the RF-RCNN and Refining-classification methods outperform Faster-RCNN greatly. Meanwhile, in Figure 4c the RF-RCNN and Refining-classification methods result in minor performance drop. During training refining stage jittering technique applied. The variations of the jittering were given by the Eq. 1, large variations of large

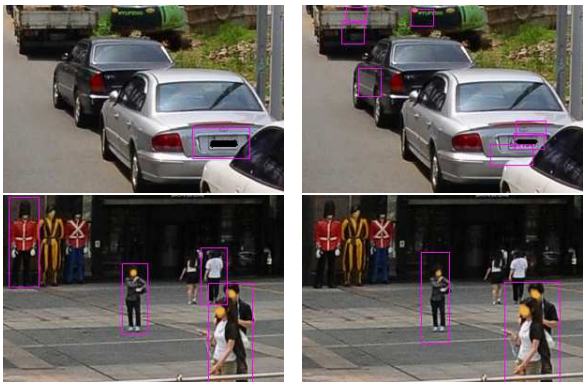



(a) Faster-RCNN




(b) Refining Faster-RCNN

Figure 2: Illustrations of the Faster-RCNN and the proposed Refining Faster-RCNN.



(a) RF-RCNN (b) Faster-RCNN

Figure 3: License plate (First row) and human (Second row) detection results using RF-RCNN (left column) and Faster-RCNN (right column).

Table 1: Detection performance, $(1 - EER)$.

| | Overall | Plate | Human |
|---|---|---|---|
| Faster-RCNN | 0.61 | 0.43 | 0.79 |
| Intermediate layer | 0.67 | 0.58 | 0.75 |
| RF-RCNN | 0.75 | 0.70 | 0.78 |

objects were learned, and it affected human classification. In our experiment, the result of human detection remained almost same as the Faster-RCNN if we use very small variation (for an instance, let $k$ in the Eq. 1 be 100). In order to raise the performance, new strategy of variation determination rather than in the Eq. 1 is needed. However, the performance drop recovered after the 3rd layer, and moreover, the overall performance of human and plate detection showed better result than the Faster-RCNN as shown in the Figure 4a.

Table 1 shows $(1 - EER)$ of the detection results where the EER (Equal Error Rate) is a value at a threshold value that false acceptance and false rejection rates are equal [14]. The proposed method achieved 0.75 while Faster-RCNN achieved 0.61.

**Analysis** In order to analyze the effect of refining regression, we measure mean of normalized distances between ground-truth and predicted centers. The normalized distance is defined by Euclidean distance between ground-truth and predicted bounding boxes divided by $\max(W, H)$. The $W$ and $H$ are width and height of the ground-truth. Only true positive bounding boxes are measured. We picked thresholds that the numbers of true positives are nearly same in both methods. Table 2 shows that the RF-RCNN is more precise than Faster-RCNN. The averages are nearly same, but the proposed method goes beyond Faster-RCNN in plate detection while the former falls behind the latter. We presume that it is because we used same jittering parameters during training, which caused larger variations of human bounding boxes than those of plate bounding boxes. This make the classifier to be trained on slightly off-position for human data due to the large position variation. With small amount of jittering on human data, RF-RCNN is expected better result on human detection as well as on plate detection.

Table 2: Normalized mean distance between centers of detected bounding boxes and groundtruth.

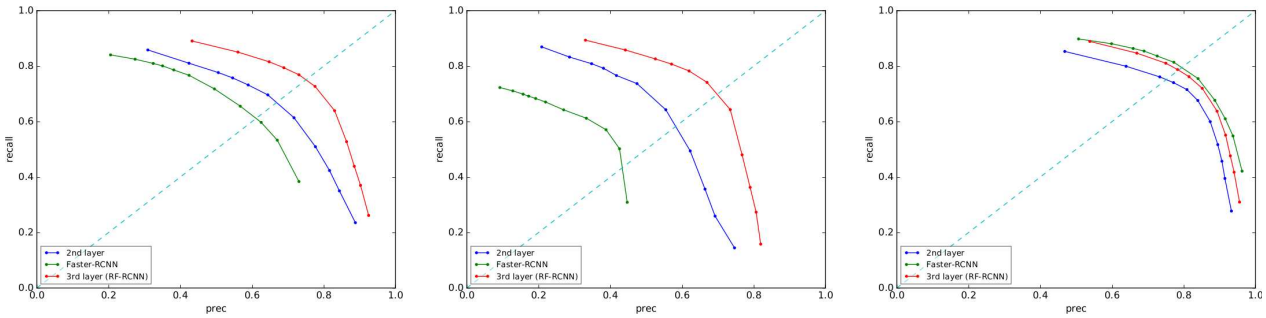
| | Overall | Plate | Human |
|---|---|---|---|
| Faster-RCNN | 0.40 | 0.46 | 0.37 |
| Intermediate layer | 0.46 | 0.42 | 0.48 |
| RF-RCNN | 0.39 | 0.36 | 0.41 |

Moreover, iterative regression can be used to obtain more accurate position of a target object as used in many of alignment approaches [12]. We leave it for our future work.

In this paper, we used same parameters in the refining stage as in the Faster-RCNN: the number of output in the fully connected layers. Optimizing the parameters could increase the performance.

## 5 Conclusion

We proposed RF-RCNN which refines classifier and regressor of Faster-RCNN. The proposed RF-RCNN is applied to human and license plate detection, and as a result, it outperforms Faster-RCNN on both detec-




(a) Overall detection (b) Plate detection (c) Human detection

Figure 4: ROC curves.

Figure 5: In RoadView service, human faces and license plates should be blurred in order to protect privacy before the images are opened in public.

tion performance and preciseness of detected bounding box. Since the proposed method showed promising performance on detection, we are going to extend the approach to detect and refine more object.

## References


[1] R. Girshick: Fast R-CNN. In: *Proc. ICCV.* (2015)

[2] S. Ren, et al.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *NIPS.* (2015)

[3] A. Frome, et.al.: Large-scale Privacy Protection in Google Street View. In: *Proc. ICCV.* (2009)

[4] R. Girshick, et.al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proc. CVPR.* (2014)

[5] O. Russakovsky, et al.: ImageNet Large Scale Visual Recognition Challenge. *IJCV.* (2015)

[6] J. Uijlings, et al.: Selective search for object recognition. *IJCV.* 104. (2013)

[7] T.Y. Lin, et al.: Microsoft COCO: Common Objects in Context. In: *Proc. ECCV.* (2014)

[8] X. Zeng, et al.: Gated Bi-directional CNN for Object Detection. In: *Proc. ECCV.* (2016)

[9] D. Erhan, et al.: Scalable Object Detection Using Deep Neural Networks. In: *Proc. CVPR.* (2014)

[10] J. Redmon, et al.: You Only Look Once: Unified, Real-Time Object Detection. In: *Proc. CVPR.* (2016)

[11] W. Liu, et al.: SSD: Single Shot MultiBox Detector. In: *Proc. ECCV.* (2016)

[12] S. Ren, et al.: Face Alignment at 3000 FPS via Regressing Local Binary Features. In: *Proc. CVPR.* (2014)

[13] R. Rajaram, et al.: RefineNet: Iterative refinement for accurate object localization. In: *Proc. ITSC.* (2016)

[14] https://en.wikipedia.org/wiki/Biometrics