Instance Segmentation on Omni-directional Images Based on Mask R-CNN

Mengcheng Song[1]　Junichi Hara[2]　Hiroshi Watanabe[1,2]

[1] Graduate School of Fundamental Science and Engineering, Waseda University
[2] Global Information and Telecommunication Institute, Waseda University

1. Introduction

Instance segmentation is of utmost importance for many applications of computer vision such as complex visual understanding, augmented reality, etc. With the appearance of portable omni-directional (OD) camera like Ricoh Theta [1], the on-line OD images is increasing dramatically, which will make instance segmentation and scene understanding of OD images become one of important studies in computer vision area. Mask R-CNN [2] is the state of the art for both objects detection and instance segmentation on conventional perspective images. However, the non-linear distortion invariance on OD image cannot be achieved because of the property of CNN which is designed for rectangular windows. In this paper, a system for simultaneously detecting objects and producing pixel-wise masks on OD images in an end-to-end manner is proposed. Instead of directly applying Mask R-CNN to the raw equirectangular images, we extend Mask R-CNN to tackle the problem of non-linear distortion. The results show that our proposed system can improve recall and precision on object detection. Moreover, it can produce satisfactory high-quality masks subjectively.

2. Proposed method

A diagram of proposed system is shown in Fig.1. Our method takes an equirectangular image as input and maps it to six cube faces using the pre-processing steps explained in 2.1. The entire cubic image is directly sent through Mask R-CNN which is described in 2.2. The output of Mask R-CNN is then remapped to equirectangular image using the post-processing steps mentioned in 2.1.

2.1 Pre- and post-processing

In pre-processing, we first project equirectangular image (also known as longitude-latitude image) on unit-sphere surface with normalized spherical coordinates. Then we apply cubic mapping to obtain a cubic image as shown in Fig.3(a), which will be fed into Mask R-CNN at inference phase. Cubic mapping has been used in computer graphics for a long period and recently used in preprocessing for compression of equirectangular images [3]. It can project the OD image to six patches same as perspective images. Thus, the non-linear distortion can be addressed by pre-processing steps.

After the masks, bounding boxes, classes, confidences of input cubic image are produced by Mask R-CNN. We draw masked cubic image, then remap it to equirectangular image with same size as input image. By post-preprocessing, we can view the OD image at arbitrary viewpoints as shown in Fig 4.

2.2 Mask R-CNN

Researches based on Faster R-CNN [4] and Fully Convolutional Network (FCN) [5] have achieved great advances for object detection and semantic segmentation respectively compared with traditional manual feature extractor. Mask R-CNN, proposed by Facebook Research recently, extends the object detection of Faster R-CNN by just adding a branch of FCN to parallelly provide pixel-wise segmentation. The backbone architecture combining ResNeXt-101 with FPN [6] achieves strong scale-invariant feature extraction. Then, regions of interest produced by RPN mapped on extracted pyramid feature maps are passed through 3 branches for classification, bounding box regression and binary mask prediction respectively. In addition, the improvement from ROIPooling to ROIAlign can produce more accurate bounding boxes and masks. As a result, we use Mask R-CNN for instance segmentation, a model [7] pre-trained on MS COCO is used in our experiments.

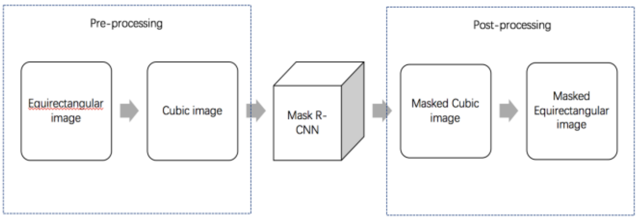


Fig. 1 Proposed system

3. Results

3.1 Image source

Ten equirectangular images with resolution of 2160x1080 pixels captured by Ricoh Theta S at different positions and viewpoints in our laboratory are used as input OD images. One of them (5th) is shown in Fig.2(a).

3.2 Instance segmentation

We first directly feed the raw equirectangular images into Mask R-CNN as baseline experiments for comparison. One of the output masked images (5th) is shown in Fig.2(b).

We implement pre-processing steps for obtaining 10 cubic images, one of them (5th) is shown in Fig.3(a). The corresponding masked cubic images is shown in Fig.3(b). We compare the number of detected objects and correct classified objects between baseline method and our proposed method. As shown in Table.1, with the pre- and post-processing steps, our proposed method showed higher performance on both recall and precision compared with direct approach.

4. Conclusion

In this paper, we proposed a system for instance segmentation on omni-directional images based on Mask R-CNN. Future work may focus on collecting and building instance segmentation dataset of omni-directional images.

References

[1] https://theta360.com.

[2] He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. Mask r-cnn. arXiv preprint arXiv:1703.06870.

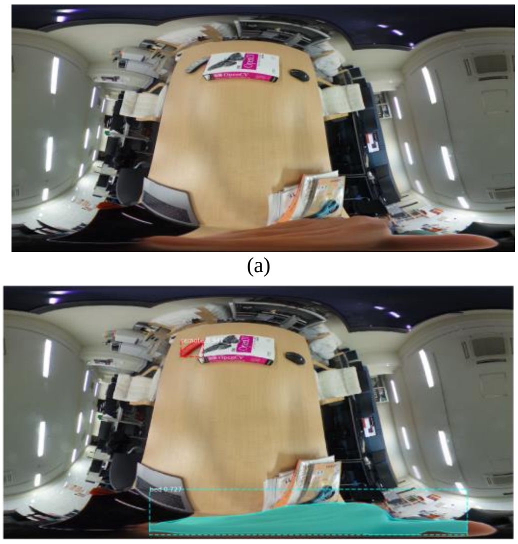
[3] https://code.facebook.com/posts/1638767863078802/under-the-hood-building-360-video/

[4] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

[5] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[6] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." arXiv preprint arXiv:1612.03144 (2016).

[7] https://github.com/matterport/Mask_RCNN

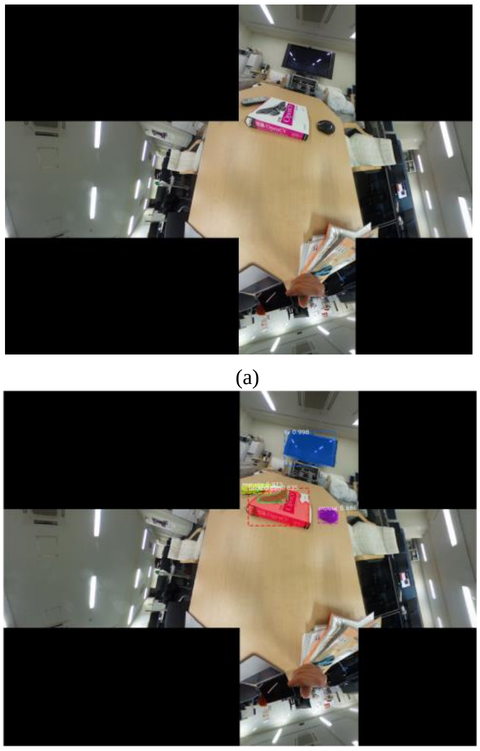


(a)

(b)

Fig. 2 Equirectangular image　a. Source image 5th,

b. Corresponding output masked image by direct approach



(a)

(b)

Fig. 3 Cubic image　a. Pre-processed Source image 5th,

b. Corresponding output masked image by proposed approach

Table.1 Performance comparison for box detection

Method	Baseline method	Proposed method
	Number of detected objects	Number of detected objects
Mean	3.1	5.5
Min	1	2
Max	5	13
	Number of correct classified objects	Number of correct classified objects
Mean	2.1	4.4
Min	1	2
Max	3	10

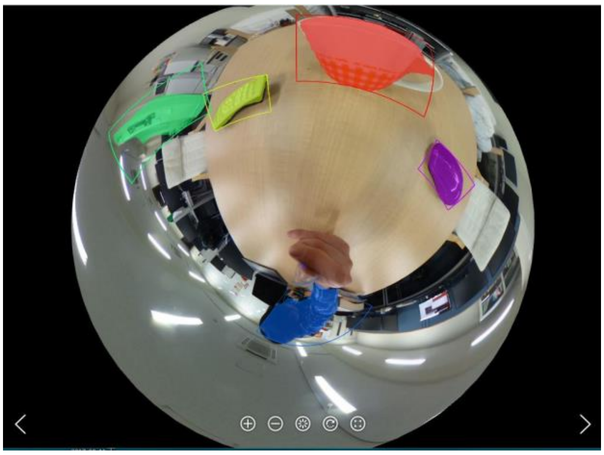


Fig. 4 Output of proposed system viewed at arbitrary viewpoint