INSTANCE SEGMENTATION OF CROWD DETECTION IN THE CAMERA IMAGES

Saziye Ozge Atik (1)*, Cengizhan Ipbuker (2)

[1] Gebze Technical University, Cayirova Campus, Kocaeli, 41400, Turkey
[2] Istanbul Technical University, Ayazaga Campus, 34467, Turkey
Email: soatik@gtu.edu.tr; buker@itu.edu.tr

KEYWORDS: Instance segmentation, Object detection, Deep learning, public camera, UAV

ABSTRACT

Artificial Intelligence (AI) is the new era in Remote Sensing (RS) applications like many other research areas. Convolutional Neural Network (CNN) structures are widely used in supervised and unsupervised classification within deep learning methods. One of the main fields of these deep learning methods is grouped under semantic segmentation applications. Instance segmentation according to the classes that are detected in the images is a type of them. In this study, in the images and real-time frames different deep learning methods are conducted and person class is detected. As data, public camera images, and Unmanned Aerial Vehicles (UAV) images are used. Each object that detected is segmented and the results are shown quantitatively. In the experiments, Mask-RCNN and Yolact++ architectures are used with selected backbones such as ResNet. Also, the time durations of each model's applications are calculated. As total time consuming for each frame, Yolact++ is faster, but the scores are yielded better in the Mask-R CNN model in the experiments. Security, target detection, and metropolitan city vision systems and many other industries are using such as crowd and person detecting applications. It is also expected to increase by the time more and more shortly, as well.

1. INTRODUCTION

Computer vision approaches on the images and frames such as UAV or public CCTV surveillance cameras can aim for several purposes. Object detection, segmentation, and classification can count as main topics. Semantic segmentation, instance segmentation, and panoptic segmentation are under the group of segmentation title. Instance segmentation examples also vary with the visualization like the black-white result of the images, to be with bounding boxes or scores. In this study, the different sources of data (UAV and public camera images) are used for detecting people inside. Detecting and masking phases are conducted on the processes on the images and video frames as well. Mask-R CNN has yielded a better score than Yolact ++ as scores. On the other hand, as time process the Yolact++ approach is faster.

2. RELATED WORKS

Object detection is one of the common applications all-around computer vision tasks. Approaches are using two-stage detectors (example: RCNN family, SSD) or single-stage detectors (YOLO family) with or without anchor boxes. The models can be pre-trained with commonly shared datasets such as Pascal VOC, COCO (Common Object in Context), or Open Images Challenge. COCO dataset is a challenging one with 80 classes and has over 1.5 million object instances. At the end of the studies, several evolution metrics can be conducted to the results.

Pedestrian detection, or person class detection there are many datasets such as Caltech Pedestrians [1], Wider Challenge - Pedestrian Detection [2], and so on.

Additionally, to person detection, masking the objects and segmented classes is another task in the studies. Instance segmentation or semantic segmentation on the images is based on these approaches. Thian Z. et al. made a study about conditional convolutions for Instance Segmentation and they proposed a new instance segmentation framework, named CondInst alternative to Mask-R CNN [3]. Another architecture is Cascade R-CNN, a multi-stage object detection architecture is used for instance segmentation that is studied by Cai Z. and Vasconcelos N. [4].

3. EXPERIMENTS

In the study, two different datasets are used for instance segmentation person detection and instance segmentation. The first dataset is VisDrone includes UAV images [5]. The second dataset is the Oxford Town Centre dataset has video that is obtained by the public camera [6]. One of the architecture for instance segmentation is Yolact++, Bolya D et al. performed an algorithm for real-time instance segmentation [7]. Also for the second method, another popular approach is chosen as Mask R-CNN framework [8].

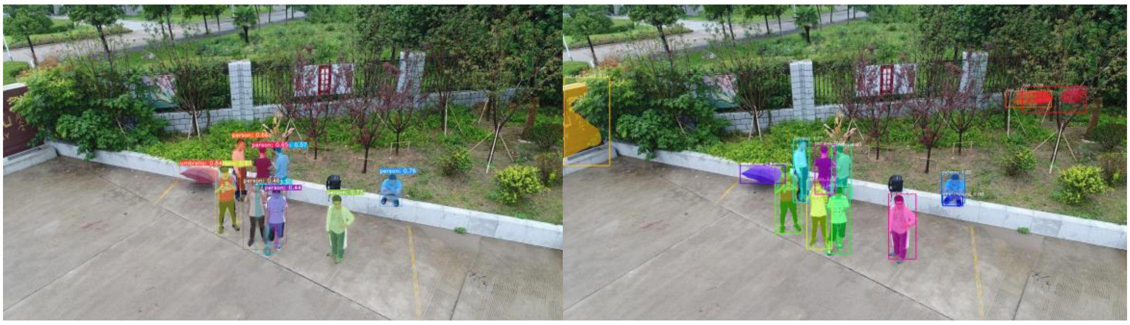

Figure 1 VisDrone image sample 1 Right Yolact ++ result, Left Mask-R CNN result

In this image scores are come as Yolact++ mean is 0.60 and Mask-R CNN is 1.00. Also Yolact ++ has one person false positive.

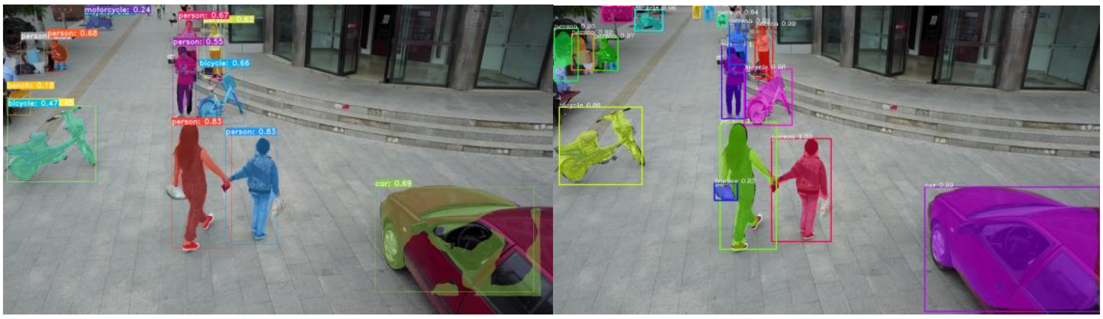

Figure 2 VisDrone image sample 2 Right Yolact ++ result, Left Mask-R CNN result

In the sample 2 results are Yolact++ is 71,25 and Mask-R CNN is 93,9. In this sample Yolact ++ yielded four False Negative results as person class.

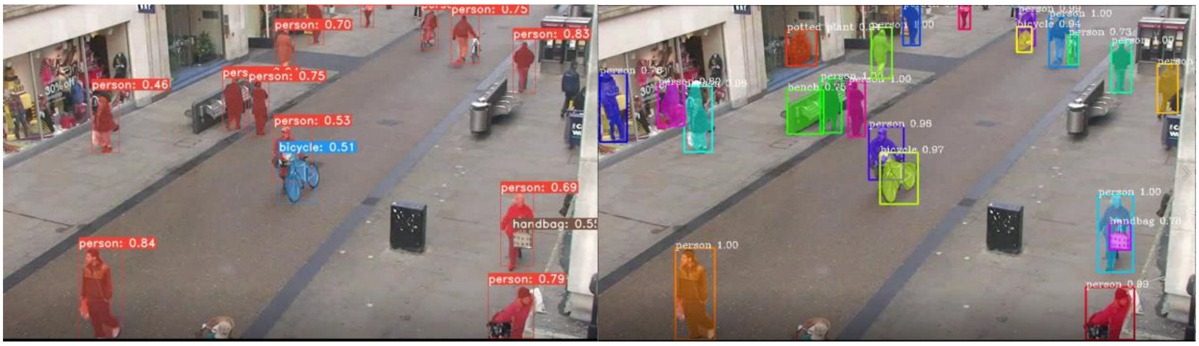

Figure 3 Town Centre video sample frame Right Yolact ++ result, Left Mask-R CNN result

In the video, person class is detected and scores are shown with bounding boxes and instance segmentation is done for each frame. Yolact++ in the frame all person classes are shown as the same color code in the frames. The mean of the scores is calculated as 0,67 in Yolact++ architecture. However, in this model score calculation method every 20 frames scores are used not all scores in the video. More than 300 people are detected. As a second model, Mask R-CNN pre-trained weights are used with the pixellib library. The mean of the person detection is obtained as 0,96 with this model. The scores are given at the end of the algorithm itself, the mean is calculated only. Besides, the total time for the video process is 451 frames in 10446.9 seconds. The models are both processed for 20 frames per second (fps).

Mean	Image 1	Image 2	Video
Yolact ++	0,60	0,71	0,67
Mask R CNN	1,00	0,94	0,96

Table 1 Scores of the person classes in the UAV Images and Public Camera Video

Each model is used pre-trained weights that are applied to COCO challenge dataset. Additionally, in image 2, Mask R CNN detected a false positive about another class: car. However, ın general for person class. Mask R CNN scores are higher.

4. CONCLUSIONS

One of the computer vision techniques for identifying and locating the objects in an image or video is object detection. Object detection can be grouped under object localization and object classification. In this study for object classification and defining the localization of the object is the chosen person class. In the images and video, each person is aimed to detect and each detection is aimed to segmented on the frames and images. Two different methods are applied. Mask-R CNN score is higher than Yolact++. In future studies, density maps can be shown with hot spot analysis according to the person counts on the images. Crowd detection applications are getting wider by time, also successes in this field are getting higher throughout the years tremendously. Computer vision has still a big part of this kind of application in the near past. Real-time applications are getting more importance due to planning, security, surveillance aims, and so on.

REFERENCES

[1] http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/ date retrieved 15.10.2020

[2] https://wider-challenge.org/2019.html date retrieved 15.10.2020

[3] Tian, Z., Shen, C., & Chen, H. (2020). Conditional Convolutions for Instance Segmentation. arXiv preprint arXiv:2003.05664.

[4] Cai, Z., & Vasconcelos, N. (2019). Cascade R-CNN: high quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[5] Visdrone dataset, Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Qinghua Hu, Haibin Ling. Vision Meets Drones: Past, Present and Future, arXiv:2001.06303 (2020). Dataset URL= http://aiskyeye.com/, date retrieved 10.09.2020

[6] Oxford Town Centre Dataset, Harvey, Adam. LaPlace, Jules. , MegaPixels.cc: Origins, Ethics, and Privacy Implications of Publicly Available Face Recognition Image Datasets, 2019, URL= https://megapixels.cc, date retrieved 10.09.2020

[7] Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019). Yolact++: Better real-time instance segmentation. arXiv preprint arXiv:1912.06218.
[8] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross ´ Girshick, “Mask r-cnn,” in The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980– 2988.