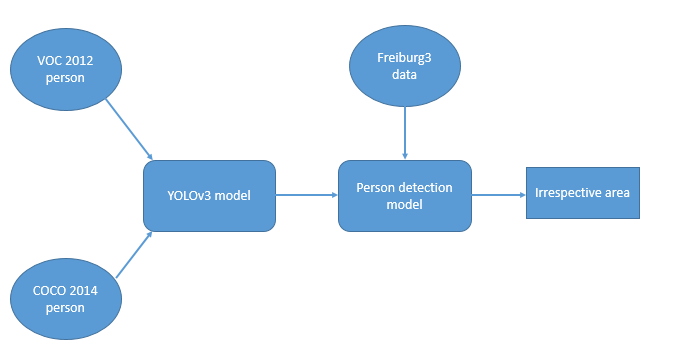
1. Object Detection

In order to ignore the influence of the person’s movement, we design a detection model to detect the person in the dynamic environment.

Such model is based on the YOLOv3 and is trained by the VOC 2012 and COCO2014 dataset. We choose the person data with label to train the model and our model have a good performance in the detect person in Freiburg3 data.



1.1 Train dataset

1.1.1 PASCAL Visual Object Classes 2012

The PASCAL Visual Object Classes 2012 dataset also called VOC is a well known dataset for image classification, detection and segmentation. It has 20 object classes and 9993 picture with labels.



Figure1, the example of VOC 2012 data[1]

The object labels in VOC include person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train,bottle, chair, dining table, potted plant, sofa, tv/monitor. What we use is the person pictures in VOC2012. The number of trained person data is 5717.



Figure 2, the example of person data in VOC.

In order to train the data in the YOLOv3 framework, we should change the label annotation into the way YOLOv3 can access.



Figure 3, the original format of VOC data

The original format of VOC data labels is ‘xml’, we extract the useful bounding box coordinate and labels and stored them in the ‘txt’ files.

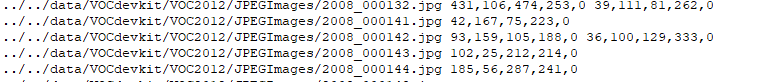


Figure 4, the changed label which the YOLOv3 model will receive

1.1.2 Microsoft Common Objects in Context (COCO ) 2014

The Microsoft Common Objects in Context dataset is also a widely used dataset for image classification, detection and segmentation. It has 91 object classes and 328000 picture with labels.



Figure 5, the coco 2014 example[2]

The original format of label data is ‘JSON’, just as what we do in VOC data , we also change the format of the label data.

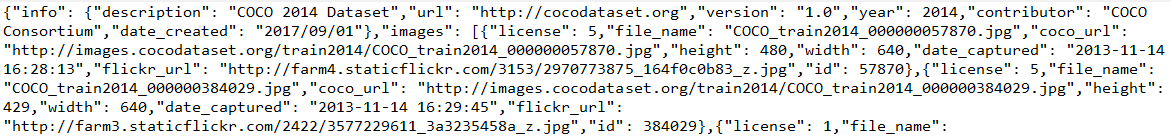


Figure 6, the original label data in COCO 2014.

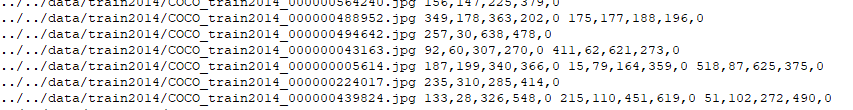


Figure 7, changed label data YOLOv3 will access.

* 1. YOLOv3

YOLOv3 is a usefull network in object detection. It draws on the practice of residual network and sets up fast links between some layers.

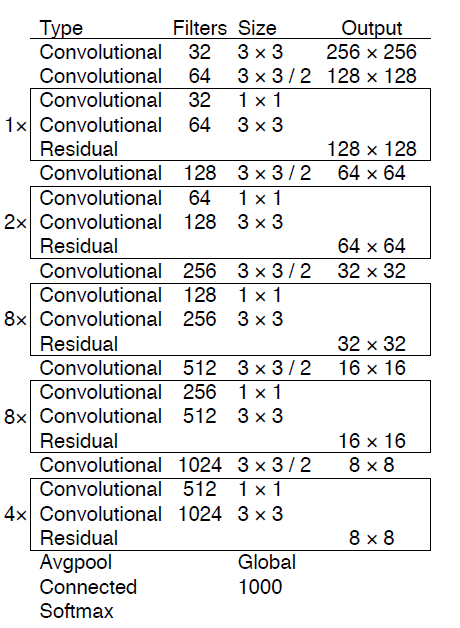


Figure 8, the YOLOv3 network architecture[3].

It is based on the darknet-53 network, which has 53 convolutional layers and 252 layers.

The original loss function can be defined as:

++

Where means the center coordinate of the object ( x and y) loss between labeled data and predicted data.

means the bounding box weight and height loss.

means whether the picture has the detected object.

means whether the label is right.

In our project, we only have one label –person, so our loss function can be define as:

+

The original anchor boxes length and width are:

(10,13), (16,30), (33,23), (30,61), (62,45), (59,119), (116,90), (156,198), (373,326). Which are not suitable for our data very well.

We use the k-means algorithm to get the suitable anchor boxes:

(6,11), (13,25), (21,57), (29,36), (42,77), (63,137), (110,204), (187,310), (360,393)

1.3 Experiment result

We use the Tesla P100 GPU to train our model.

We use the frozen method to accelerate the speed and unfrozen way to get better performance.

The batch size of frozen method is 256 and the batch size of unfrozen way is 20.

After 6 epochs frozen training and 3 unfrozen training, the loss become stable.

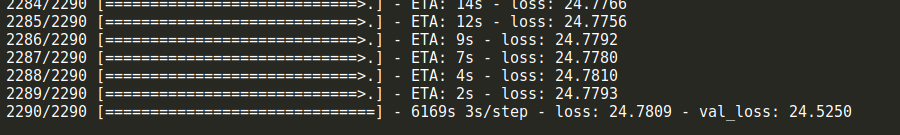


Figure 9, the training loss of final epoch.

Then we use the trained model to detect the person in the common SLAM problem dataset Freiburg3. The image id and the predicted person label can be seen in the following picture:

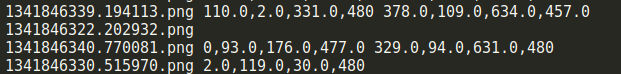


Figure 10, example of detection result.

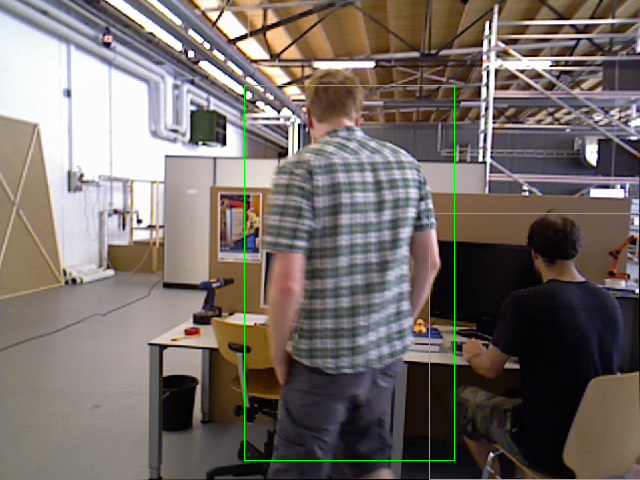
 

Figure 11, the detection result draws on the Freiburg3 data

Reference

[1] <http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html>

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ra-manan, P. Doll ́ar, and C. Zitnick. Microsoft COCO: Common Objects in Context. In ECCV, 2014

[3] J R, Redmon, A Farhadi. An Incremental Improvement. arXiv, 2018