Paper number ITS-XXXX

Human Detection Using RetinaNet Applied on Visual and Thermal Single-Sensor Images

Qiao Ren[1*], Siavash Hosseinyalamdary[2*], Xinran Wang[3]

1. University of Twente, the Netherlands email: oreolinda20130828@gmail.com

2. University of Twente, the Netherlands email: s.hosseinyalamdary@utwente.nl

3. University of Twente, the Netherlands email: x.wang-2@student.utwente.nl

Abstract

This research aims at comparing the performance of four models to perform human detection using a deep learning approach RetinaNet. Four models include two non-finetuned models, which are named as model kaistT and kaistRGB, and two finetuned models, which are named as model cocokittikaistT and cocokittikaistRGB. The result shows that the best model of training on the thermal kaist dataset is model cocokittikaistT. The best model of training on the visual kaist dataset is the model cocokittikaistRGB.

Keywords:

finetuning, human detection, single sensor

1.Introduction

1.Background information

Human detection has been playing an increasingly important role in many fields in recent years. The technology is widely used in autonomous driving, Post-disaster rescue,automated surveillance, military and robotics services (Gajjar, Gurnani, & Khandhediya, 2017).

After the disaster, a device equipped with a human detector could help the rescue team find out where the survivors are(Doherty & Rudol, 2007). According to these information rescuers can plan the most reasonable rescue plan. It guarantees the safety of rescue team in the searching operation. This will provide more time for rescue work.

In autonomous driving, human detection technology ensures the safety of both drivers and pedestrians(Balani, Deshpande, Nair, & Rane, 2015). Human detection systems detect pedestrians adjacent to autonomous driving cars and get their specific locations. So autonomous vehicles can avoid collisions in this way. However, more reliable human detection algorithm can help to rescue the victims of a disaster or an accident.

Human detection plays a key role in automated surveillance(D, Manjunath, & Abirami, 2012 ; Moore, 2003). Human detection technology can help to monitor some suspicious activities. Such as limitations for human activity in certain areas. The runway of the airport is not allowed to walk randomly by people other than the staff. Private houses do not wish to be disturbed by others. In military, human detection device can help to monitor the enemy's action. Especially the thermal detector, can help the army to obtain position and quantity of the enemy.

2. What is human detection

The object test is a computer technology that has to do with the visual and image processing of the machine, which is used to detect an instance of some kind of semantic object, like a person, a building or a car, in digital imaging and video[1].

Object detection is the task of recognizing the existence of predefined object types and estimating their position in images. This task includes identifying the existence of objects and drawing a bounding box of each object.

Human body detection is a sub-problem of object detection, and we are only interested in the existing human body in the images. It can be divided into two parts: we need to determine if there are people in that image, and we want to get their corresponding coordinates in the image.

3. Problem statement

So far, many scholars and scientists have invested a lot of efforts in human detection and made some achievements. Human detection is still a very challenging problem. It can be affected by occlusions, messy backgrounds and poor visibility at night.

Human detection is still a challenging task because of the different appearances and postures of each person (Gajjar et al., 2017), as shown in Figure 1. While, the computer can only use graphic information, so different clothing will make it more difficult for computers to recognition a human, as shown in Figure 2. As well as camera capture at various views to the human body (shown in Figure 3). This can cause humans to be obscured by other people or objects (Figure 4 shows).

[1] https://en.wikipedia.org/wiki/Object_detection

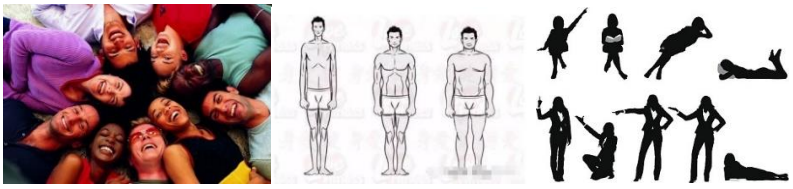


Figure 1: human with different skin colors[2] (left), different body shapes[3](middle) and different postures[4](right)

Figure 2： human with various clothes[5] and the different view of a people captured by a camera[6]

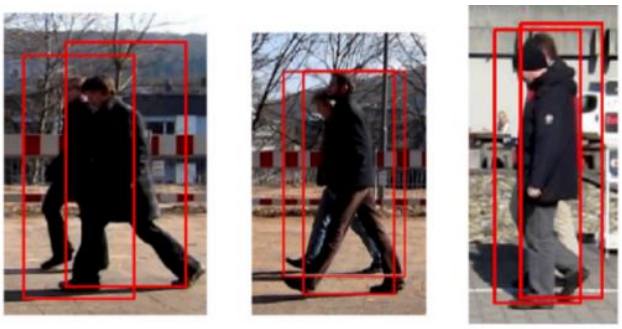


Figure 3: people who are occluded by other people[7]

Visual images and thermal images are the two major information sources used in human detection researches (Fan, Xu, Zhang, & Chen, 2008). Visual images are in RGB channel (Reyes-Ortiz, Oneto, Samà, Parra, & Anguita, 2016). Nowadays, most human detection tasks are still based on the visual data(Hwang, Park, Kim, Choi, & Kweon, 2015). Visual images are in RGB channel (Reyes-Ortiz, Oneto, Samà, Parra, & Anguita, 2016). And some achievements have been made.

The visual image has the disadvantage of being sensitive to light changes. As a result, they are vulnerable to insufficient exposure or excessive exposure during a sudden change in illumination. Besides, they need plenty of light. Therefore, when light is insufficient, such as in the night, at dusk and shadow area, visual image quality drops.

Thermal images are visual displays of infrared energy emitted, transmitted, and reflected by objects (Correa, Hermosilla, Verschae, & Ruiz-del-Solar, 2012). The amount of radiation emitted by the object increases as the temperature of the object increases. The thermal camera measures the temperature of an object. The body temperature is different from the temperature of the environment. Thus, the thermal image is used to distinguish human from other objects, particularly at night or in shadow. We can see the differences of visual image and thermal image in Figure 5.

[2] https://www.taringa.net/posts/salud-bienestar/14426457/Enterate-por-que-nos-reimos.html
[3] http://www.sohu.com/a/201618131_508479
[4] https://www.vcg.com/creative/811803218
[5] https://www.etsy.com/sg-en/listing/208818837/feather-headdress-indian-style-medium
https://item.jd.com/11064904661.html
http://www.52112.com/pic/325058.html
[6] https://www.123rf.com/
[7] https://ps.is.tuebingen.mpg.de/publications/tangijcv

Figure 5: Image (a) and (b) are captured in daytime on the same scene(Hwang et al., 2015). Image (c) and (d) are captured at night time on the same scene(Hwang et al., 2015). (a) and (c) are visual images. (b) and (d) are thermal images.

The disadvantage of thermal images is that the thermal detector is susceptible to non-human factors when the outside temperature is high (Kim et al., 2017; Baek, Hong, Kim, & Kim, 2017). Besides, the resolution of the thermal image is low. It is very difficult to identify the distant human body in a thermal image (Fan et al., 2008). Thermal light sources may also have an effect on the quality of the image at night.

Both visual images and thermal images have drawbacks (Kim, Hong, & Park, 2017). Visual camera is able to perceive the environment if the lighting condition is suitable. The visual images have poor quality at nights and under-shadow regions. In contrast, thermal images sense the human body temperature at night. Therefore, it can be applied to detect human body. Thermal images have lower resolution and thermal information is not as detailed as the visual images.

4. Research Objective

The research objective of this research is to find out the best model among the finetuned and non-finetuned models applied in human detection. The input is thermal and visual dataset respectively. The output is bounding boxes predicted by the model. The bounding boxes visualize the location detected by the models. The finetuned and non-finetuned models are based on an innovative convolutional neural network RetinaNet (Lin, Goyal, Girshick, He, & Piotr Dollar, 2018).

5. Innovation At

The current state of art in object detection is RetinaNet. It is a robust one-stage object detector (Lin, Goyal, Girshick, He, & Piotr Dollar, 2018). The innovation of RetinaNet is to apply a special loss function in order to solve the imbalance problem between foreground and background classes. Previous studies use Fast RCNN, Faster RCNN and ACF+T+THOG in human detection. However, the state-of-the-art approach surpasses the accuracy of those approaches (Lin et al., 2018). Therefore, we decided to use RetinaNet as the approach in thi research.

Based on my knowledge, it will be the first time to train RetinaNet on human detection. This is also the first time that the RetinaNet is incorporated with different levels of sensor fusion: pixel level, feature level, decision level.

2. Literature Research

One of the most famous computer vision approaches is Histogram of Oreinted Gradient (HOG). In the detection window, detector is used to analyse interest points and draw histogram with gradient feature vector. Then all the histograms is taken into the aggregation in different layers in order to detect the object instance. The prediction of the object category will be given. (Dalal & Triggs, 2005) Dalal and Triggs apply this approach to detection of the human body in visual images, but the approach has been effectively applied to thermal images.

Wang, Zhang and Shen applied a new method which is method is based on the Shape Context Descriptor (SCD) with the Adaboost cascade classifier framework(Wang, Zhang, & Shen, 2010). This makes the thermal detector is not only in the night performance is remarkable, and has some robustness to the change of light during the day. The experimental results show that the shape of the enhanced classification background characteristics of thermal images of the human body detection has significant improvement.

Guan et al proposed a network to distinguish between daylight and night time and propose different networks for the thermal and visual images in daylight and at night time (Guan, Cao, Yang, & Yang, 2018). This research combines the features extracted by two image sensors, firstly estimates the illumination, and then corrects the coefficients of the day and night network on the basis of estimated illumination.

Nowadays, more and more deep learning approaches are applied to the human detection.

R-CNN (region-based convolutional neural network method) combine region proposals with CNNs(Girshick, Donahue, Darrell, Berkeley, & Malik, 2012).First, they applied high-capacity convolutional neural networks (CNNs) to bottom-up regional proposals to locate and segment objects. Then, when marked as insufficient training data, the supplementary task is pre-trained with supervision, and then a domain specific fine adjustment is performed, which can significantly improve performance. The drawback of R-CNN is that it is slow at training-time. Because it needs to run full process of CNN for watch region proposal. The other drawback is that CNN features are not updated in response to regressors.

Soon afterwards, a Fast Region-based Convolutional Network method (Fast R-CNN) for object detection (Fast r-cnn) was proposed(Girshick, 2015). Compared with R-CNN, Fast R-CNN not only improves the training and testing speed, but also improves the detection accuracy. Region proposal method is then implemented on the feature map. Fast R-CNN trains the very deep VGG16 network 9× faster than R-CNN, is 213× faster at test-time, and achieves a higher mAP on PASCAL VOC 2012(Girshick, 2015).

Ren et al. came up with an object detection algorithm that eliminates the selective search algorithm and allows the network learn the region proposal (Ren, He, Girshick, & Sun, 2015). This method called Faster R-CNN.Similar to fast R-CNN, the image is provided as input to the convolution network, which provides the convolution feature graph. Instead of using a selective search algorithm on the feature map to identify regional Suggestions, it's using individual networks to predict regional Suggestions. Then use the RoI pool layer to predict regional suggest refactoring, the layer is used for classifying suggest area of the image, and predict the offset value of the bounding box.

3. Data Explanation

Data used in this study is provided by KAIST (Korea Advanced Institute of Science and Technology) which is a research university in South Korea. The KAIST Multispectral Pedestrian Dataset are captured by a vehicle which carries a colour camera and a thermal camera. The images are captured during day and night time. As an example, a pair of visual

and thermal images with annotation is shown in Figure 6. Human in all the pairs have been labelled. The data is available online[8].

(a) (b)

Figure 6 Two images in the KAIST dataset. Image (a) is a visual image. Image (b) is a thermal image. Image (a) and (b) were captured in the same location at the same time. Annotations are shown in red bounding box. In image (a) and (b), there are two annotated persons. The annotations are provided by KAIST dataset.

4. Methodology

The overview of the methodology is shown in the flowchart (Figure 7). In total, four models are generated. For each of the single sensor kaist dataset, two models are generated, which are a non-finetuned model and a finetuned model. The initial default weights of retinanet are the weights that have been trained by the dataset ImageNet. There are two branches of training. One of the branches is to train the retinanet with default weights directly on kaist training data. The other branch is to finetune the weights first and train the retinanet on kaist training data in the next step. The image dataset used for fine tuning are kitti and coco dataset. It will be justified in the next section. The output of the two branches are four models: kaistT, kaistRGB, cocokittikaistT, cocokittikaistRGB. The models kaistT and kaistRGB are the non-finetuned models. The models cocokittikaistT and cocokittikaistRGB are the finetuned models. After that, all the four models will be tested. Because the precision and recall of each model vary under different conditions of score threshold and iou threshold, it is necessary to find out the best score threshold which provide the best performance of a model. The "best" is defined by the optimal tradeoff between precision and recall. After the best model has been found, a comparison will be implemented. Among the non-finetuned and finetuned models, the best model that generate the highest average IOU (intersection over union) will be found out. The following sections will explain the methodology in detail.

[8] https://sites.google.com/site/pedestrianbenchmark/home

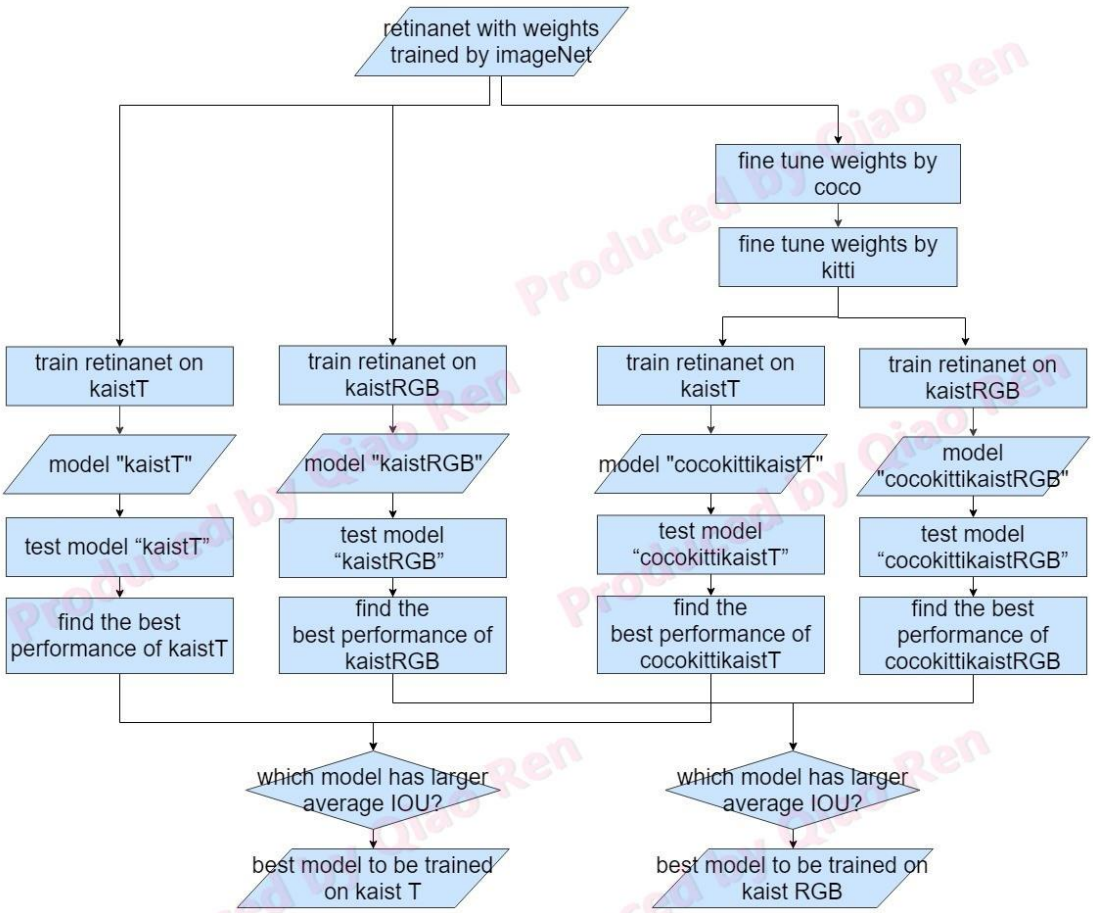


Figure 7 overview of four models in this research

Step1 prepare data set
There are two different channels: visual (RGB) and thermal (T). Each channel has two datasets: training dataset and testing dataset.

- kaistT dataset contains images and annotations of kaistT
- kaistRGB dataset contains images and annotations of kaistRGB

Step 2 train retinanet with and without finetuning
Fine-tuning means to learn features from a broad domain in order to help learning features from a specific domain. The advantage of fine-tuning is to speed up the training process and to overcome small dataset size.

Figure 8 shows the relation between ImageNet, coco, kitti and kaist dataset. The dataset represented by the large circle contains the classes of the dataset represented by the small circle. The fine-tuning process goes from broad domain to specific domain.

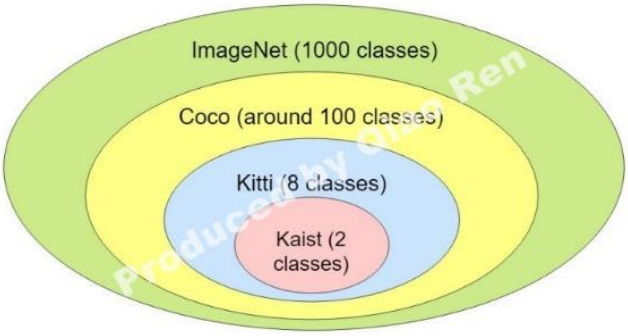


Figure 8 kitti dataset has 8 classes: car, van, truck, pedestrian, person_sitting, cyclist, tram, don't care. Kaist dataset has 2 classes: human and non-human.

In this research, model cocokitti and model kitti have both been trained (Figure 9). The reason that cocokitti has been decided as finetuning dataset, instead of kitti dataset is that the precision of model cocokitti is higher than the precision of kitti (Table 1).

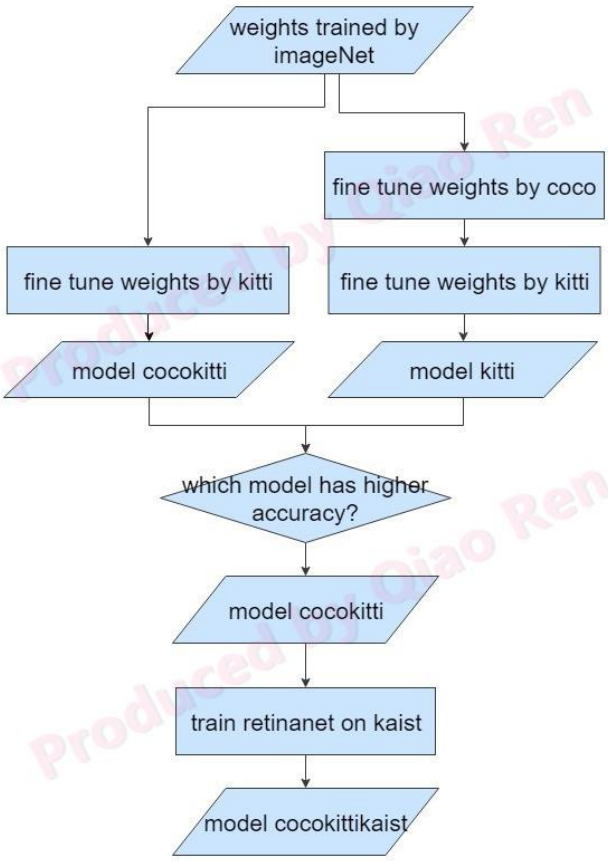


Figure 9 a flow chart to find the best model among the model cocokitti and the model kitti

	cocokitti	kitti
Average precision of pedestrian	0.44	0.34
Average precision of cyclist	0.46	0.39

Table 1 average precision of the model cocokitti and the model kitti

Step 3 test models

The flowchart of testing a model has been shown in Figure 10. This testing process is implemented on each of the above mentioned four models. The input is a retinanet model, a bunch of kaist testing images and a score threshold. A score means to which level the retinanet is confident with detecting an object as a human. Each bounding box predicted by retinanet has a corresponding confidence level which is also called as "score". A score threshold is used to determine whether or not a predicted bounding box is trustworthy. If the score of a bounding box is higher than the score threshold, then this bounding box is accepted. Otherwise, the bounding box is discarded. The score threshold in this research is set as 9 values: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. The final output is the bounding boxes which has scores higher than the score threshold. Therefore, each model has nine outcomes corresponding to nine different score thresholds.

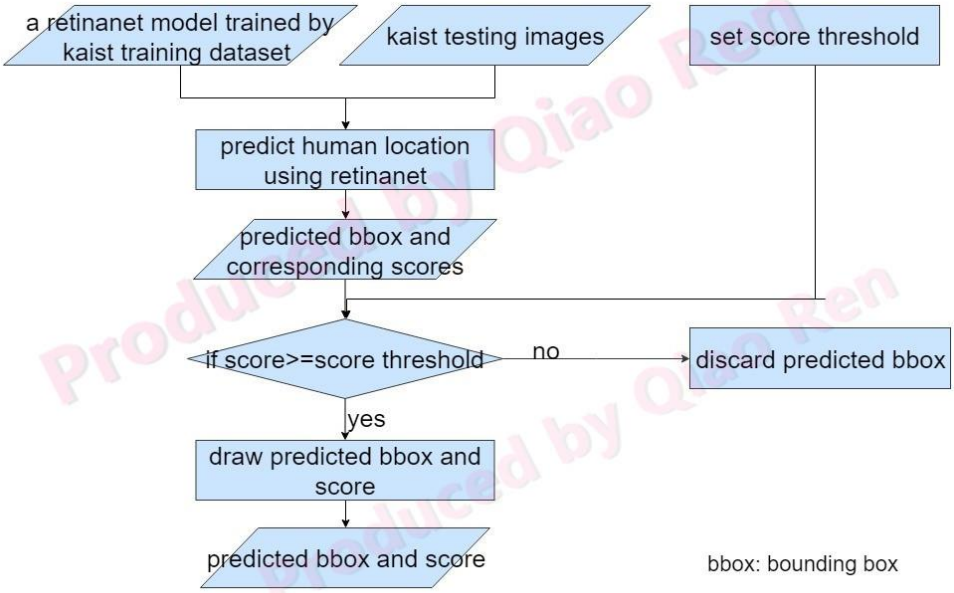


Figure 10 a flowchart of how to test a model

Step 4 find the best model by evaluating recall, precision and average IOU

Recall, precision and average IOU are used to evaluate the outcome of an object-detection model (Han, Zhang, Cheng, Liu, & Xu, 2018). As shown in Figure 11, the input are the bounding boxes predicted by RetinaNet, bounding boxes from annotation (ground truth) and a threshold of IOU (intersection of union). Firstly, the predicted bounding boxes are matched with annotated bounding boxes. This matching process is implemented in the following way. IOU of each single predicted bounding box and all the annotated bounding boxes are calculated. This is based on the formula (Formula 1). Among all the IOU, the largest IOU is seen as the IOU of this prediction bounding box and its corresponding annotated bounding box. Then this IOU is compared with the threshold of IOU. The IOU threshold provides a requirement that the predicted bounding box is seen as the same annotated bounding box. So it is a correct prediction. If an IOU is larger than the IOU threshold, then the predicted bounding box is true positive. Otherwise, it is false positive. Based on the accumulated number of true positive in the whole testing dataset, the precision and recall are calculated. There is a specific pair of value, precision and recall, under a certain threshold of IOU and a certain threshold of score. Based on the theory (Han et al., 2018), precision decreases with the increase of recall. Because there is a trade-off

between precision and recall. This research assumes that the score threshold has larger influence than the iou threshold. The test focuses on how various score threshold influence the tradeoff between recall and precision. This is visualized in a recall-precision graph. The iou threshold is fixed which is 0.7. At the same time, the average iou of all the predicted bounding boxes are calculated.

$$\text{IOU} = \frac{\text{area of overlap}}{\text{area of union}} \quad [9]$$

Formula 1 Formula of intersection over union

In order to find out the score threshold that provides the best performance of a model, a formula is applied. For each pair of precision and recall, the multiplication of precision and recall is calculated. The model that gives the maximum result of multiplication is the best model. Ultimately, the best performance will be found out for each model. The output will be the best performance of kaistT, the best performance of kaistRGB, the best performance of cocokittikaistT, the best performance of cocokittikaistRGB.

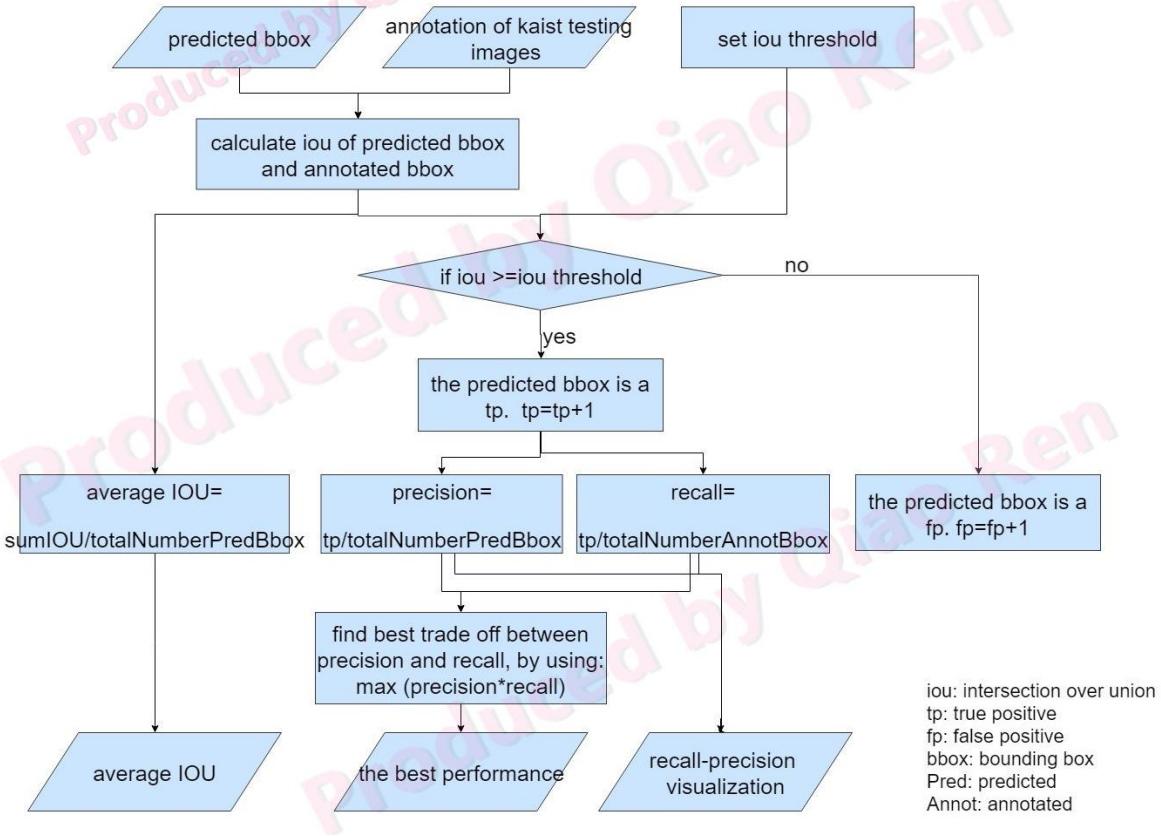


Figure 11 flow chart of how to calculate average IOU, precision and recall of a model.

5. Results

The results of the four models, which are kaistT, kaistRGB, cocokittikaistT and cocokittikaistRGB, are illustrated in this section. An example of the bounding box and score predicted by the four models are shown in Figure 12. These results are under the score threshold

[9] https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

0.5 and iou threshold 0.7. Blue rectangle box represents the predicted bounding box. Red rectangle box represents the annotated bounding box.

Figure 13 shows the classification loss and regression loss of each model. Classification loss measures whether the objects are classified correctly or not. Regression loss measures whether the coordinates of predicted bounding boxes are correct or not. Classification and regression losses are recorded during the training process. There are 50 epoches in the training process. From the images, all four models have been converged.

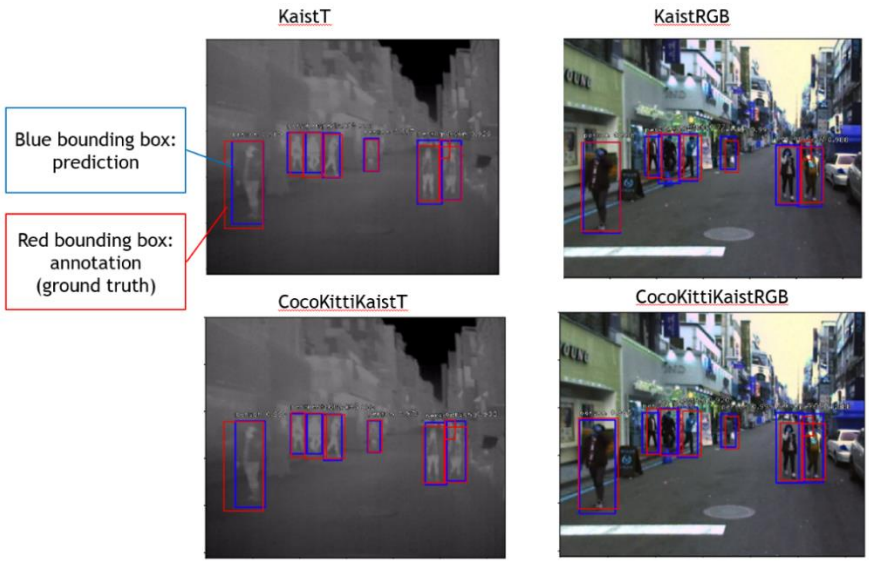


Figure 12 an example of bounding boxes (in red) predicted by four different models (kaistT, kaistRGB, cocokittikaistT, cocokittikaistRGB)

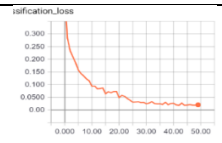
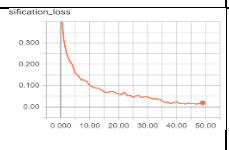
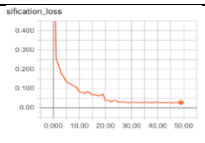
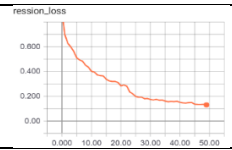
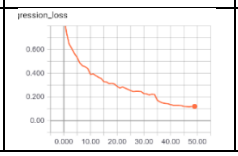
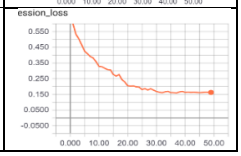
	KaistT	KaistRGB	cocokittikaistT	cocokittikaistRGB
Classification loss				
Regression loss				

Figure 13 the classification loss and regression loss of four models (kaistT, kaistRGB, cocokittikaistT, cocokittikaistRGB)

Figure 14 shows the recall-precision graph of the four models respectively. Each graph contains 9 dots, corresponding to 9 different score threshold: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. Figure 14 shows the trade-off of precision and recall of four models respectively. The trade-off is calculated by the formula 2.

$$Rectangle\ Area = precision * recall$$

Formula 2 a formula used for evaluating the tradeoff between precision and recall

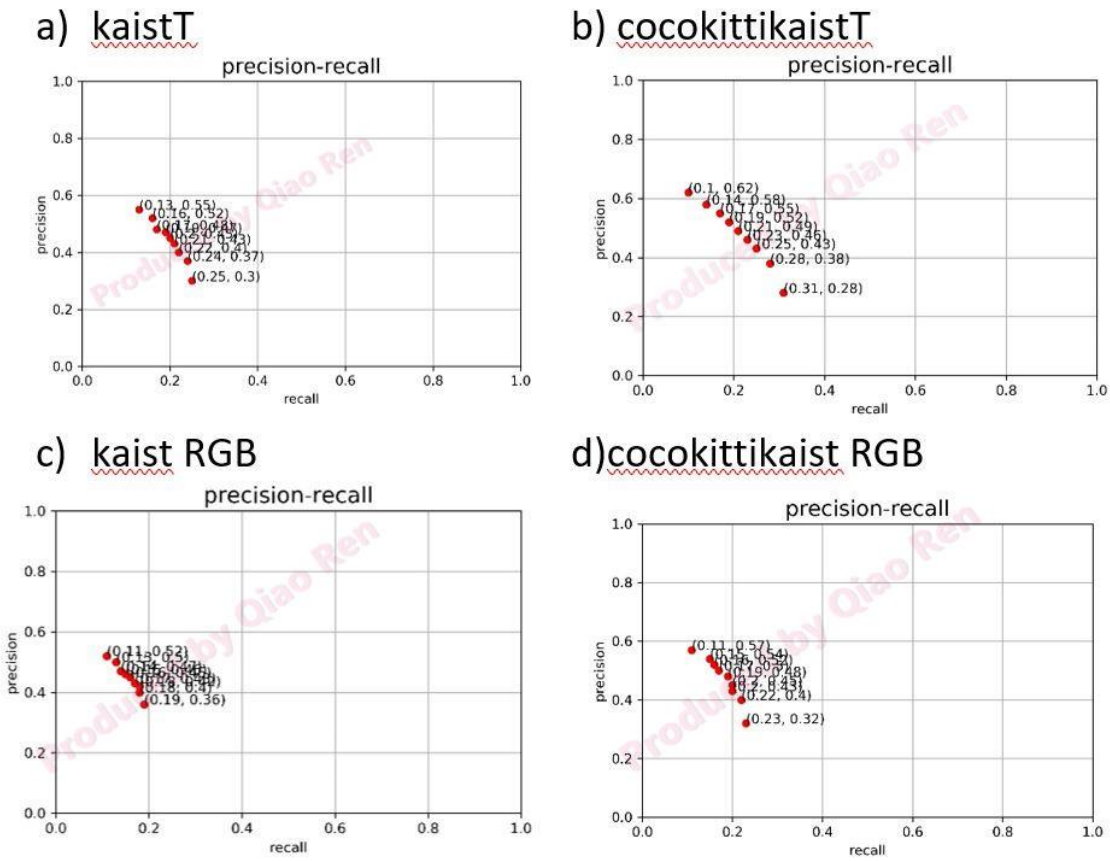


Figure 14 precision-recall graph of four models respectively

Figure 16 visualizes a comparison among the recall-precision graphs of the four models. The actual output are dots. The line that connects the dots are drew manually, in order to enhance the visualization. The big dot in each line shows the best performance with the maximum rectangle area of each model. Figure 17 visualize the comparison of average IOU of all four models. Each score threshold has a corresponding average IOU.

The maximum rectangle value, which is recall*precision, of each model is shown in table 2. In table 2, each row gives the best recall-precision performance of each model. The average IOU is written on the last column. Based on this table, the multiplication result of recall and precision indicates the best model for each dataset. Therefore, it is concluded that the best model of dataset kaistT is model cocokittikaistT (a finetuned model). Because recall*precision of cocokittikaistT (0.1099) is larger than the value of kaistT (0.0908). The best model of the dataset kaistRGB is the model cocokittikaistRGB (a finetuned model). Because recall*precision of cocokittikaistRGB (0.0893) is larger than the value of kaistRGB (0.0744).

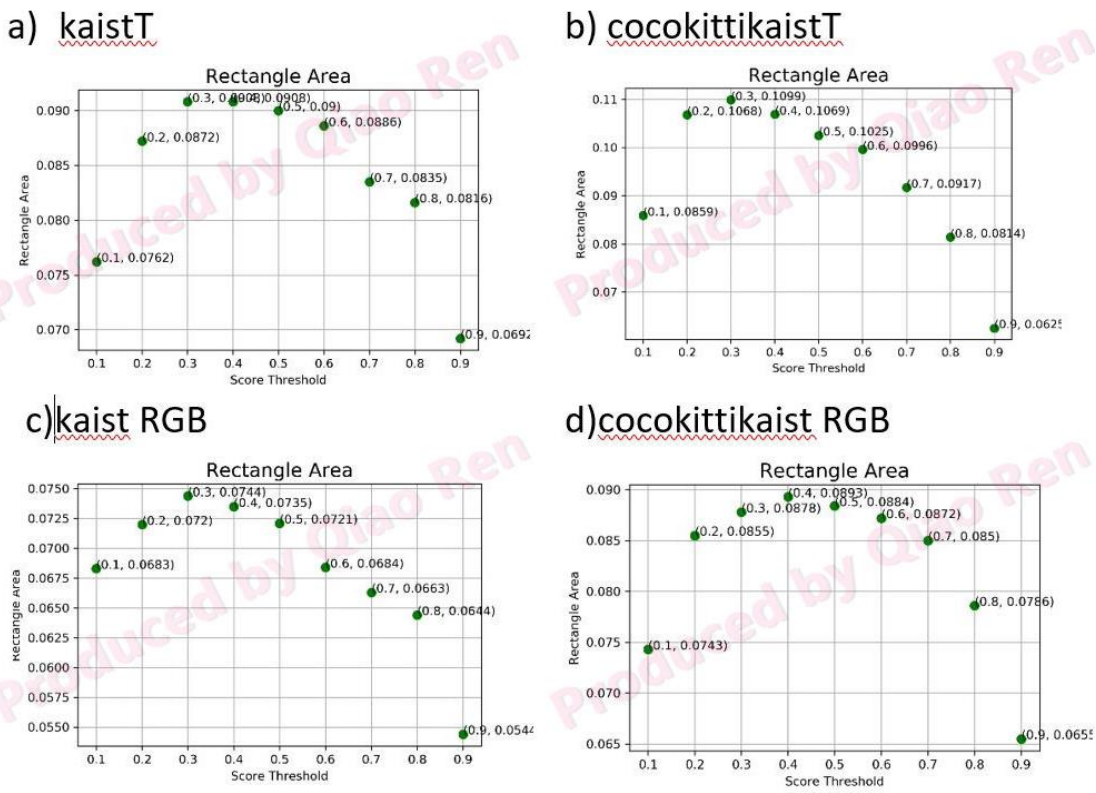


Figure 15 trade-off of precision and recall of four models respectively

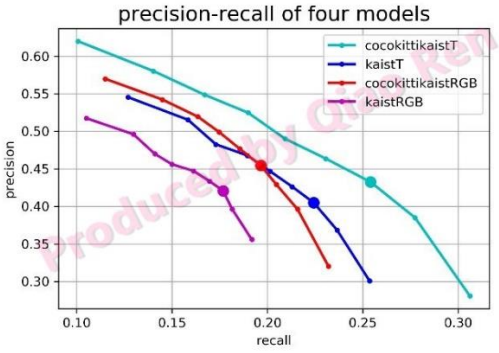


Figure 16 Precision-recall of four models

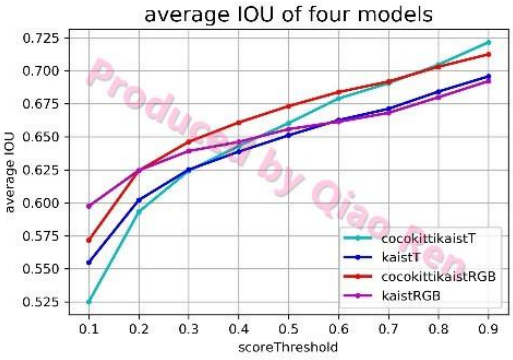


Figure 17 Average IOU of four models

Name of a model	Iou threshold	Score threshold	Recall * Precision	Average IOU
kaistT	0.7	0.3	0.0908	0.6251
cocokittikaistT	0.7	0.3	0.1099	0.6243
Kaist RGB	0.7	0.3	0.0744	0.6392
Cocokittikaist RGB	0.7	0.4	0.0893	0.6608

Table 2 The best recall-precision performance of each single-channel model (by calculating the max area of rectangle: precision*recall)

6. Discussion

There are several points need to be discussed in this research.

1) Overfitting occurs on all four models. Overfitting has an influence on the result of test. Overfitting problem is found based on the fact that Kaist T convergence value is smaller than cocokittikaistT convergence value, while precision-recall graph shows that Kaist T has lower precision. This implies that KaistT has a problem of overfitting.

2) Annotated bounding box dataset has problems. Four types of errors occur in the annotation. These errors are found by Xinran Wang. The examples of each type of error is illustrated in the appendix. “Problem 1 here is a person but annotation doesn’t give the bounding box. Problem 2 there is no person but annotations give a bounding box. Problem 3 the annotations give a bounding box which contains a group of people rather than a single person. Problem 4 sometimes the annotations give a bounding box larger than the real person.” The most frequently occurring problem is problem 3: a group of people are annotated by one bounding box. It causes two problems to the training model. Firstly, the model learns inaccurate features of human. Secondly, the test result of the model deviates from its true result, especially when a model detects individual persons and ground truth annotates a group of people.

3) Only one value of iou threshold is used in testing, which is 0.7. 0.7 is a good choice of iou. However, it is better to implement the test with different value of iou threshold, for example 0.6, 0.8 and 0.9.

7. Conclusion

It is concluded that finetuned models, which are cocokittikaistT and cocokittikaistRGB, has a better performance than the non-finetuned models, which are kaistT and kaistRGB. Therefore, it is suggested to use the finetuning approach cocokitti on kasit dataset to implement human detection.

References

1.Girshick, R. (2015). Full-Text. *IEEE International Conference on Computer Vision (ICCV 2015)*, 1440–1448. https://doi.org/10.1109/iccv.2015.169

2.Girshick, R., Donahue, J., Darrell, T., Berkeley, U. C., & Malik, J. (2012). Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2–9. https://doi.org/10.1109/CVPR.2014.81

3.Han, J., Zhang, D., Cheng, G., Liu, N., & Xu, D. (2018). Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection. *IEEE Signal Processing Magazine*, *35*(1), 84–100. https://doi.org/10.1109/MSP.2017.2749125

4. Lin, T., Goyal, P., Girshick, R., He, K., & Piotr Dollar. (2018). Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2018.2858826

5.Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks. *ARXIV*, 1–14.

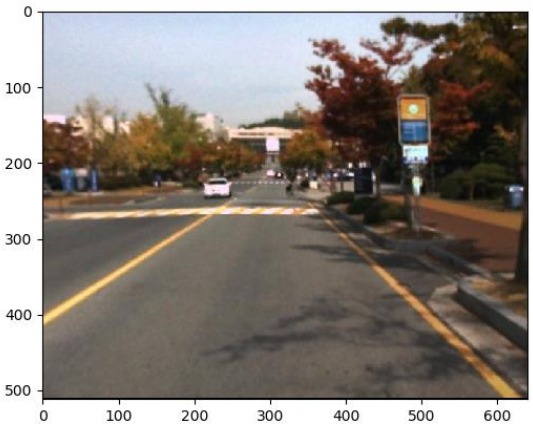
6.Wang, W., Zhang, J., & Shen, C. (2010). Improved human detection and classification in thermal images. *Proceedings - International Conference on Image Processing, ICIP*, 2313–2316. https://doi.org/10.1109/ICIP.2010.5649946

9. Appendix

Four problems in annotation (written by Xinran Wang)

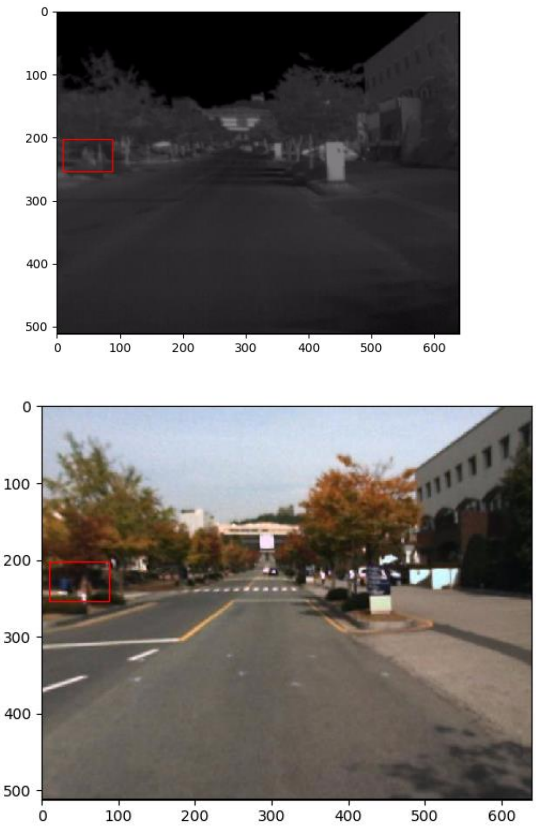
The blue rectangle is model detected people. the red rectangle is annotations. The yellow rectangle is used to show the person

Problem 1 here is a person but annotation doesn't give the bounding box.

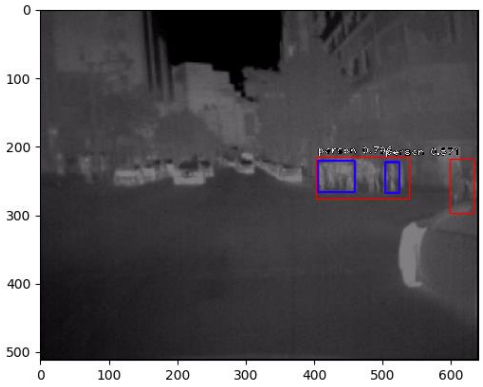


The blue rectangle is model detected people. the red rectangle is annotations. The yellow rectangle is used to show the person

Problem 2 there is no person but annotations give a bounding box



Problem 3 the annotations give a bounding box which contains a group of people rather than a single person.



Problem 4 sometimes the annotations give a bounding box larger than the real person which are represented by the yellow bounding boxes.