

The Integration of Thermal and RGB Images For Human Detection Using Deep Learning

Qiao Ren
Geo-Information Science and Earth
Observation
University of Twente
Enschede, the Netherlands
oreolinda20130828@gmail.com

Siavash Hosseinyalamdary
Geo-Information Science and Earth
Observation
University of Twente
Enschede, the Netherlands
s.hosseinyalamdary@utwente.nl

Xinran Wang
Geo-Information Science and Earth
Observation
University of Twente
Enschede, the Netherlands
x.wang-2@student.utwente.nl

Abstract—Detection of human is important in a wide range of applications nowadays. This research aims at finding out the best architecture with the approach of convolutional neural network applied in human detection by integrating thermal and visual images. The convolutional neural network RetinaNet is a state-of-art approach. This research implements two multi-sensor fusion models and four single sensor models, based on RetinaNet. Multi-sensor fusion models are early-fusion and late-fusion. The early fusion model integrates thermal and visual images on pixel level. The late fusion model integrates the predictions that are provided by single-sensor models. The result shows that late fusion model has significantly improves the performance of single sensor models. The improvement made by late fusion model is 7.1 %, 3.8%, 11.1% and 8.1%, compared with non-finetuned thermal model kaistT, finetuned thermal model cocokittikaistT, non-finetuned RGB model kaist RGB and finetuned RGB model cocokittikaistRGB respectively. Early fusion model does make improvement. Therefore, it is concluded that late-fusion model with RetinaNet is the best approach in human detection.

Keywords—multi sensor integration, deep learning, human detection, thermal images, RGB images

I. INTRODUCTION

Human detection is essential for various applications (Liu, Zhang, Wang, & Metaxas, 2016). It is important in disasters management, the autonomous driving systems, automated surveillance and human-robotics interaction [2].

Human detection is a challenging task [1]. Because, people have different ages, genders, body shapes, positions, appearances. In addition, human can be captured from different views. Moreover, human may be occluded, which makes it difficult to detect human.

The datasets mainly used in human detection are visual images and thermal images [3]. Visual images are represented in RGB channel [4]. Thermal images are visual displays of the amount of infrared energy emitted, transmitted, and reflected by an object (Correa, Hermosilla, Verschae, & Ruiz-del-Solar, 2012). As the temperature of an object increases, the amount of radiation emitted by the object increases.

Both visual images and thermal images have drawbacks [6]. The drawback of visual images is that they are sensitive to illumination changes. Therefore, they are easily underexposed or overexposed in the sudden changes of illumination. In addition, they require sufficient illumination. Consequently, the quality of visual images deteriorates when the illumination is insufficient, such as at night, at dusk, in the shadow regions, and in foggy weather.

The disadvantage of thermal images is that, when the difference between the temperature of human and non-human

background is small, it is difficult for thermal images to detect human. This situation can happen in the daytime during hot summer. Besides, the resolution of thermal images is low [3], which causes difficulties in human detection.

Sensor fusion is applied in human detection, in order to overcome the drawback of individual sensors [7]. Sensor fusion in human detection means to integrate features generated by two different sensors, RGB camera and thermal camera[7]. These two types of sensors provide complementary detection decisions [8]. When illumination is sufficient, such as in a daytime, it is easy to detect human by visual images. When illumination is insufficient, such as during the night or dusk, it is easy to detect human by thermal images. The combination of visual and thermal images provides a robust algorithm and gives a higher accuracy in human detection, compared with using a single type of image [9].

The approach to implement sensor fusion is decided to be deep learning. Because deep learning is a highly efficient method, compared with other traditional methods like HOG (histogram of gradient) [10]. Traditional methods use a dictionary to store all the human in the training data [9]. It has two drawbacks [1]. Firstly, its speed is slow, because it compares with all the objects in the dictionary with the unknown object in the image. Secondly, if a human in an image does not exist in the dictionary, then the traditional approach is not able to detect it. In contrast, deep learning is able to capture the essence of human features.

Currently, RetinaNet is the best approach in human detection [11]. Its accuracy surpasses all the deep learning approaches. The innovation of RetinaNet is that it solves two problems [11]. Firstly, it applies a special function, called focal loss function to solve the class imbalance problem. Secondly, it has a unique architecture Pyramid Feature Network. This network solves the problem that high resolution images contributes to weak features while low resolution images contribute to strong features. This problem is not solved by other deep learning networks [11]. Therefore, RetinaNet is the best approach in human detection.

II. RESEARCH IDENTIFICATION

A. Research Objectives

The main objective of this study is to find out the best architecture with the approach of convolutional neural network (CNN) applied in human detection. The architecture takes thermal and visual images as input and it detects the human in these images. We aim to test different architectures and provide the most accurate architecture in human detection.

B. Research Questions

The main objective is achieved by answering the following research questions:

Research Question 1: What is the best human detection approach?

Research Question 2: How to integrate multiple sensors in the state of art?

Research Question 3: How much does multi-sensor approach improve the single sensor approach?

C. Innovation

RetinaNet is the recent state of art in object detection. It is robust due to its special loss function [11]. Previous studies use Fast RCNN, Faster RCNN and ACF+T+THOG in human detection. The state-of-the-art approach surpasses the accuracy of those approaches [11].

Based on my knowledge, it will be the first time to adapt RetinaNet to multi-sensor integration on human detection. This is the first time that the RetinaNet is incorporated with different architectures of sensor fusion: early fusion level and late fusion.

III. DATASETS

Three datasets have been used in used in this research: Kaist dataset, Coco dataset and Kitti dataset.

A. Kaist dataset

Data used in this study is provided by KAIST (Korea Advanced Institute of Science and Technology). The KAIST Multispectral Pedestrian Dataset contains 95,000 pairs of color and thermal images. The images are captured by a vehicle which carries a color camera and a thermal camera. The images are captured during day and night time. All the images have the same dimension 640 pixels * 512 pixels and the same resolution 96 dpi * 96 dpi. In this research, 53293 annotated persons are used in training dataset. 2742 annotated persons are used in testing dataset. The testing dataset uses the corrected annotations which are provided by Liu [1].

B. Coco dataset

Coco dataset¹ has 91 object classes². Person is one of the 91 classes. Coco provides segmentation masks for every object instance. Coco dataset has 330K images. 250,000 people are annotated.

C. Kitti dataset

Kitti dataset has 8 object classes: pedestrian, cyclist, car, van, truck, sitter, tram and misc. Images are captured by two cameras: a color and a grayscale camera.

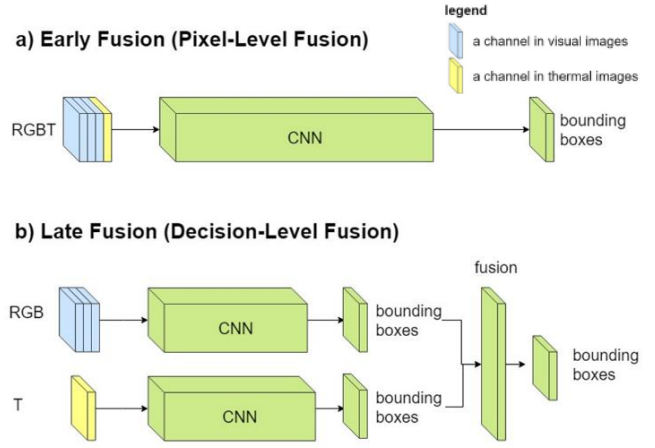


Fig. 1 architecture of early fusion and late fusion

IV. METHODOLOGY

A. Methodology Overview

This research implemented multi-sensor-fusion models and single sensor models. Two models are multi-sensor-fusion models: early fusion model and late fusion model. Four models are single-sensor models. This research compares multi-sensor-fusion models with single-sensor models. Single sensor models are reference models. They are trained on either thermal or RGB images. Early fusion means to integrate images from different sensors on pixel level and then train the CNN (Fig. 1 a). Late fusion means to integrate the decisions made by two subnetworks (Fig. 1 b). One subnetwork has been trained on thermal images. The other subnetwork has been trained on RGB images.

Output of a model are bounding boxes and corresponding scores. A Bounding box is a rectangle with the smallest perimeter within which the whole human lies. Each bounding box has a corresponding score. Score is the degree of certainty to which there is a human on a certain location. The range of a score is from 0 to 1.

B. Assumptions

- 1) It is assumed that the time of capturing visual images and thermal images has been synchronized.
- 2) It is assumed that the location of visual camera and thermal camera are at the same location.
- 3) It is assumed that the labeled ground truth is perfect and correct. This means that all the rectangular bounding box of labeled human are in correct position and with correct size.
- 4) It is assumed that, if only a part of human present in an image, such as a finger, a hand or a foot, then it is unnecessary to detect it as a human. Because it is incomplete.

C. Single Sensor Models

Single sensor models are the models which are trained with either thermal or RGB images. There are two purposes to use single sensor models in this research. Firstly, they are used to compare with multi-sensor fusion models. Secondly, the best single sensor models are chosen as the subnetworks of late-fusion model.

¹ <https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/>

² <https://tech.amikelive.com/node-718/what-object-categories-labels-are-in-coco-dataset/>

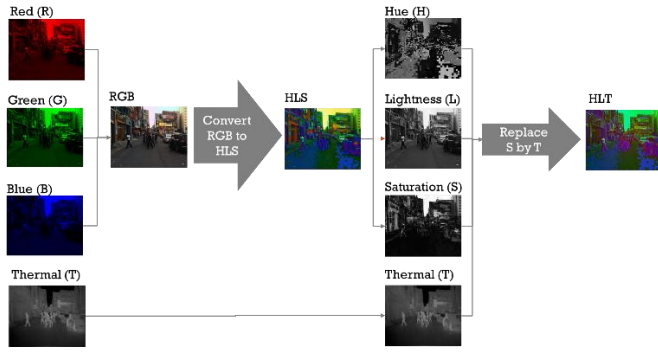


Fig. 2 converting RGBT to HLT in early fusion

An important approach applied in single sensor models is fine tuning. Fine-tuning is the process in which parameters of a model is adjusted very precisely in order to fit with certain observations. Fine tuning requires a model 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.

The datasets used for finetuning are coco dataset and kitti dataset. The default weights in RetinaNet are the weights that have been trained by ImageNet dataset. Therefore, four models have been generated:

- Model kaistT: a model trained with thermal images. The weights were initialized by default.
- Model kaistRGB: a model trained with visual images. The weights were initialized by default.
- Model cocokittikaistT: a model trained with thermal images. The model has been finetuned on Coco and Kitti dataset before training on Kaist dataset.
- Model cocokittikaistRGB: a model trained with visual images. The model has been finetuned on Coco and Kitti dataset before training on Kaist dataset.

D. Early Fusion Architecture

Early fusion means to integrate thermal and visual images on pixel level before training CNN. In this research, the proposed method is to convert RGBT to HLT (Hue, Luminance and Thermal) (Fig. 2). Firstly, RGB is converted to HSL. Then the saturation band (S) in HSL is replaced by thermal band (T). The removed band is saturation, because saturation is the least informative band in HSL. The reason that RGBT must be converted to a 3-band image is that the maximum amount of input bands in RetinaNet is three. It is impossible to feed a 4-band image (R+G+B+T). HLT solves this problem.

E. Late Fusion Architecture

Late fusion aims at integrating the decisions which has been provided by thermal model and RGB model. Late fusion contains two steps.

Step 1: use single-sensor models as subnetworks. The sub-networks of late fusion are the best model on thermal images and the best model on RGB images. In this research, the model CocokittikaistT with score threshold 0.2 has the

best performance on thermal images. The model Cocokittikaist RGB with score threshold 0.2 has the best performance on visual images.

Step 2: decision integration. This step integrates the bounding boxes predicted by the two subnetworks, thermal model and RGB model. The following two cases need to be analyzed.

In one type of situation (Fig. 3), there is no overlap between a T bounding box and an RGB bounding box. In this case, both of bounding boxes are generated as the final output. Because thermal and RGB models extract different features. A person who is detected by thermal model may not be detected by RGB model and vice versa. Generating bounding boxes from both models leverages the advantages of both models.

In the other type of situation (Fig. 4), there is an overlap between a T bounding box and an RGB bounding box. Then the question is whether both should be the output or only one of them should be the output. The algorithm to solve this problem is the following. If the overlap between two boxes is small, then the two bounding boxes are interpreted as detecting different persons. If the overlap is large, then the two bounding boxes are interpreted as detecting the same person. Because it is very likely to happen that there is a small difference between the locations which are detected by two different models. Overlap is defined as IOU (intersection over union) between thermal and RGB bounding boxes. The way to distinguish whether the overlap is large or small is to set a threshold of IOU. This threshold is called as merge threshold. If IOU is smaller than the merge threshold, then both bounding boxes should be generated in the output. If IOU is larger than the merge threshold, then only one bounding box which has larger score should be generated. The bounding box with larger score is more trustworthy. Therefore, this bounding box is the output.

It is necessary to set a reasonable merge threshold. Because if the merge threshold is too large, then two bounding boxes which detect the same person will be interpreted detecting different persons. The amount of false positive will increase. If the merge threshold is too small, then the person who has been detected by one of the models will be missed out. False negative will increase. Fig. 5 visualizes the output bounding boxes generated by different merge threshold. Different values of merge threshold cause different output bounding boxes (white bounding boxes). When merge threshold is 0.8, redundant bounding boxes are generated.

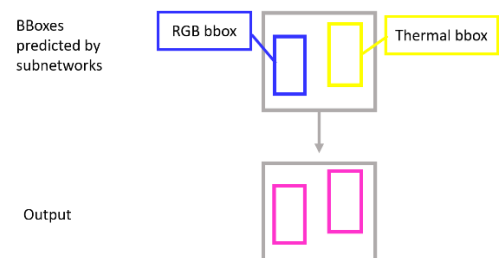


Fig. 3 algorithm which integrates RGB and Thermal bounding boxes, in case there is no overlap between the two bounding boxes

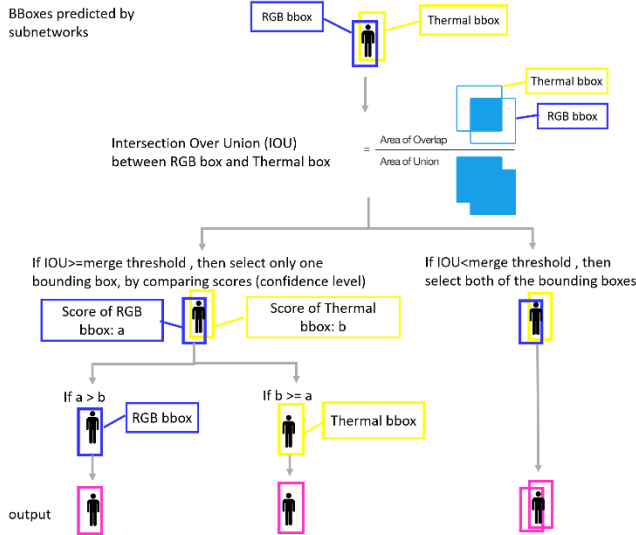


Fig. 4 algorithm which integrates RGB and Thermal bounding boxes, in case there is an overlap between the two bounding boxes

The method to optimize the merge threshold is to evaluate the precision and recall. Late fusion model has been run with different candidate merge threshold. The best merge threshold is the one which provides the largest F1 score. The candidate merge thresholds are 0.5, 0.6, 0.7, 0.8 and 0.9. It has been found that the best merge threshold is 0.7 (Fig. 6). When merge threshold is 0.7, F1 score reaches its maximum, which is 0.3332. Therefore, 0.7 is set as the best merge threshold.

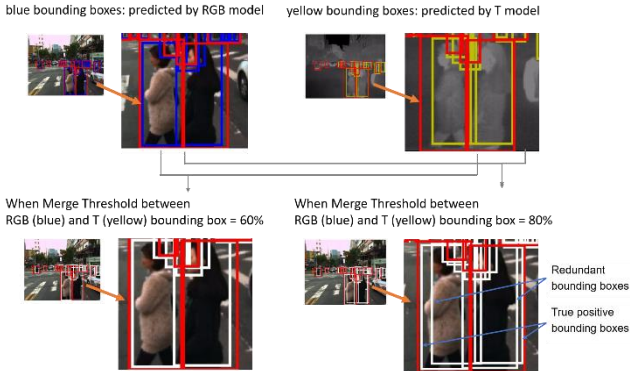


Fig. 5 The output bounding boxes are different when merge threshold has different value. In this examples, two merge thresholds are tested: 60% and 80%. White bounding box is the output of late fusion. Red bounding boxes are annotations.

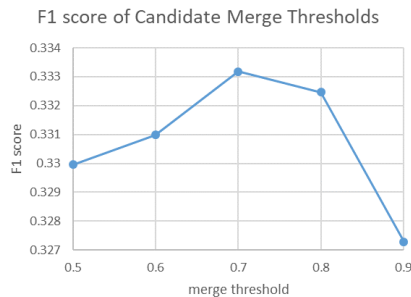


Fig. 6 F1 score of candidate merge thresholds

F. Quality Assessment

Quality assessment is based on precision, recall, F1 score, miss rate and false positive per images [12]. The formulas in quality assessment are the following equations.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$F1\ score = 2 \frac{precision * recall}{precision + recall} \quad (3)$$

$$Miss\ Rate = \frac{false\ negative}{false\ negative + true\ positive} \quad (4)$$

$$False\ Positive\ Per\ Image = \frac{false\ positive}{total\ amount\ of\ images} \quad (5)$$

The terms in formulas are explained in this paragraph. True positive (TP) is the number of persons who are correctly predicted by a model. False positive (FP) is the number of non-humans who are incorrectly predicted as human. False negative (FN) is the number of persons who are incorrectly predicted as non-human. Precision is defined as, among all the detected objects, how many detections are correct [13]. Recall tells how many of the objects that should have been predicted are predicted. It is a measure of exactness [13]. IOU is intersection over union between two bounding boxes. It measures to what degree that a predicted bounding box and an annotated bounding box are overlap (Fig. 7).

The criteria to define a true positive is to set a threshold of IOU. The threshold is called as IOU threshold or true positive threshold. In this research, the true positive threshold is set as 0.7. This means that if IOU is larger than this threshold, then this predicted bounding box is a true positive. Otherwise, it is a false positive.

A model has different precision and recall, under different score threshold. A score threshold is used to filter out the bounding boxes which has low confidence level. As shown in Fig. 8, when the score threshold is set too low (Fig. 8 a), false positive would increase. When the score threshold is too high (Fig. 8 c), a lot of human are not detected. A low score threshold causes recall to be high and precision to be low. A high score threshold causes precision to be high and recall being low. In this research, each model is tested with different score thresholds: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9.

The approach to find out the best performance of a model is to find out the largest F1 score. F1 score is the harmonic average between precision and recall. F1 score measures the balance between precision and recall. F1 score is 0 in the worst case. F1 score is 1 in the best case. For each model.

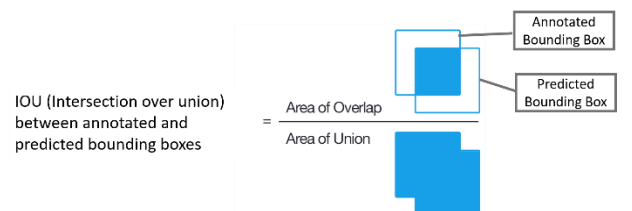


Fig. 7 IOU between annotated and predicted bounding boxes

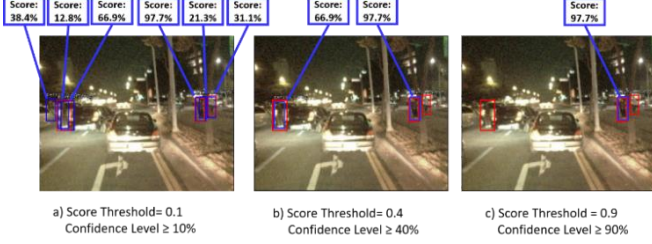


Fig. 8 predicted bounding boxes caused by different score threshold

V. RESULTS

Quantitative results are shown in Fig. 12, Fig. 13, Fig. 14 and Table 1. Fig. 12 shows the precision and recall of single-sensor models and multi-sensor-fusion models. Six lines corresponding to six models. Each line has nine dots which represents the precision and recall under a certain score threshold. The model which is the closest to the optimal point (precision is 100% and recall is 100%) is the best model. Fig. 13 shows the F1 score of each model. The largest F1 score of each model is represented by the big dot. The model which has the largest F1 score is the best model. Fig. 14 shows the miss rate of six models. The model which has the lowest miss rate is the best model. Table 1 Shows the best performance of each model. The examples of bounding boxes predicted by six different models are shown in Fig. 15.

Late fusion has the best performance. F1 score of late fusion (0.360) is the highest among the six models. Comparing the F1 score of the best performance of each model, it has been found that late fusion surpasses the performance of kaistT by 7.1%, cocokittikaistT by 3.8%, kaistRGB by 11.1%, cocokittikaistRGB by 8.1%. The miss rate of late fusion (0.67) is the lowest among all the six models. The miss rate of late fusion is 11% lower than kaistT, 5% lower than cocokittikaistT, 15% lower than kaistRGB and 11% lower than cocokittikaistRGB. The reason that late fusion has the best performance is that late fusion successfully integrates the decision from RGB model and T model. When a person is detected by one of the single sensor models, but not detected by the other one (Fig. 9), late fusion makes a compensational decision. Late fusion model leverages the advantage of single sensor models and remedies the deficiencies of single sensor models. Late fusion model improves the detection in challenging situations: when a person is in low illumination, in a shadow, is far away from the camera or is occluded.

Early fusion has the worst performance. F1 score of early fusion is the lowest among the six models. It has three reasons. Firstly, information loss happened due to the conversion from RGBT to HLT. The band saturation in HSL has been removed. This causes information loss in visual images. Secondly, information loss happened in the process converting from RGB to HLS. The resolution of band hue decreases.

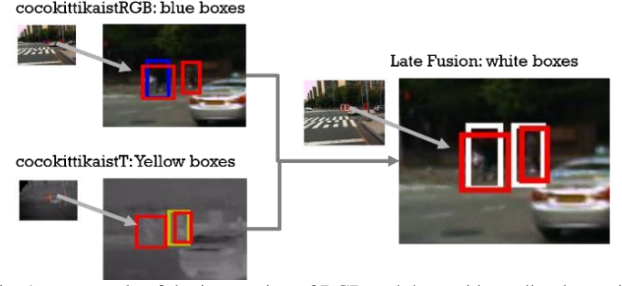


Fig. 9 an example of the integration of RGB and thermal bounding boxes in late fusion

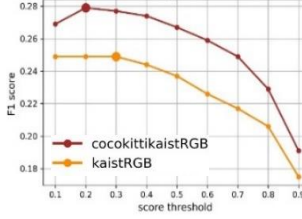


Fig. 10 F1 score of kaistRGB and cocokittikaistRGB

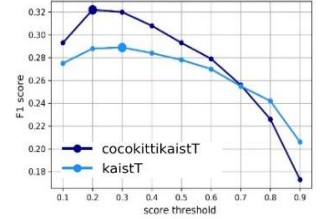


Fig. 11 F1 score of kaistT and cocokittikaistT

Fine-tuned models have better performance than non-finetuned models, as shown in Fig. 10 and Fig. 11. Because weights have been adjusted when the models are trained by Coco and Kitti data. Weights have been precisely adjusted to their optimal values.

Models trained on thermal images have better performance than trained on RGB images. Because RGB images are sensitive to illumination. When it is under-exposure or over-exposure, it is difficult for RGB images to detect a person. Thermal images do not suffer from the illumination problem.

The result of this research is not comparable with the result of other literatures. Because, firstly, the threshold of IOU used for counting true positive is not declared in other researches. Different threshold of IOU result in different precision and different recall. Secondly, in this research, the testing images include all the person instances which are non-occluded, partially-occluded, heavily occluded, close-to-camera and far-away-from-camera. However, the other researches use a subset of kaist dataset to do testing. They use a “reasonable subset” which means “the images in which human are at least 50% non-occluded” [14], [15].

TABLE I. BEST PERFORMANCE OF EACH MODEL

Model name	ST ³	Recall (%)	Precision (%)	F1 score (%)	Miss Rate	FP per img ⁴
kaistT	0.3	22.43	40.48	0.289	0.78	0.62
cocokittikaistT	0.2	27.74	38.49	0.322	0.72	0.84
kaist RGB	0.3	17.68	42.08	0.249	0.82	0.46
cocokittikaist RGB	0.2	21.58	39.62	0.279	0.78	0.62
Late fusion	0.3	33.42	38.91	0.360	0.67	0.99
Early fusion	0.1	11.24	41.35	0.177	0.89	0.30

Table 1 the best performance of each model

³ Score threshold

⁴ False positive per image

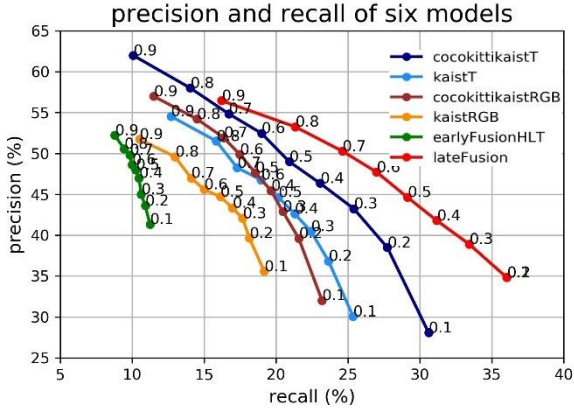


Fig. 12 precision and recall of six models

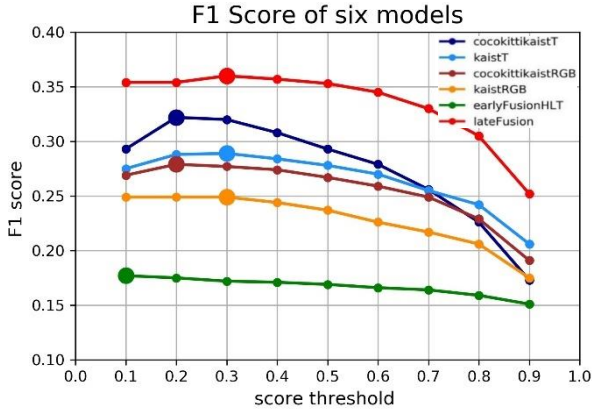


Fig. 13 F1 score of six models

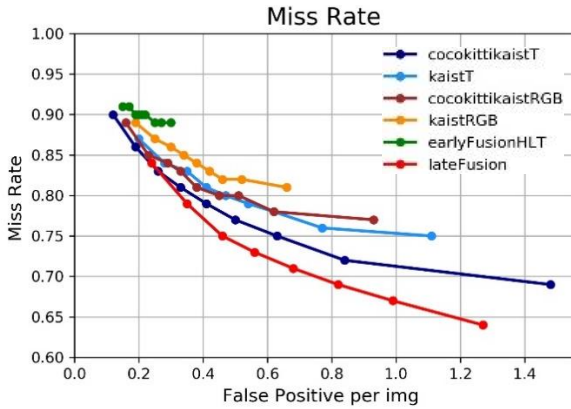


Fig. 14 miss rate of six models

VI. DISCUSSION

Two fusion models in this research have both advantages and disadvantages.

A. Early Fusion

The advantage of early fusion model is that the weights in RetinaNet have been adjusted. RetinaNet learns the features from both RGB and thermal images. However, the drawback is that information loss occurs when RGBT is converted to HLT. Because the band saturation has been removed.



Fig. 15 examples of bounding boxes predicted by six models

B. Late Fusion

The advantage of late fusion model is that the bounding boxes predicted by single-sensor models are all taken into its consideration. Late fusion model remedies the deficiencies of single sensor models. No information is lost. The limitation of late fusion is that the precision and recall of late fusion model is constrained by single sensor models. If a person is not detected by any single sensor model, then it will not be detected in late fusion model.

C. Challenging Dataset

The kaist dataset is very challenging. Firstly, RGB and Thermal camera are not perfectly located on the same location. Secondly, RGB and Thermal camera do not perfectly capture images at the same time. These two reasons make it not easy to integrate RGB and T bounding boxes in late fusion. Thirdly, annotations are imperfect. Some annotated bounding boxes are too much bigger than the size of persons. Several non-human objects are annotated as human.

It is challenging to detect human, in some situations. Table 2 shows the cases which has an influence on recall. The cases include when a person is far away from the camera, or a person is occluded, or only a part of a person is captured in an image, or the resolution is low, or it is under-exposure, or it is over-exposure.

In the other situations, persons have been correctly detected but does not counted into true positive. Table 3 shows the cases which has an influence on precision. One case is that persons are detected individually but an annotated bounding box contains multiple persons. The other case is that the size of annotated bounding box is larger than a real person.

Therefore, the challenging kaist dataset has an influence on the performance of models.

TABLE II. CHALLENGING SITUATIONS THAT INFLUENCES RECALL



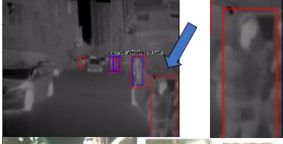




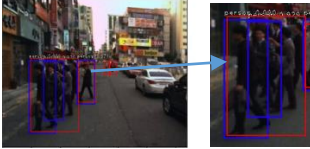
Far distance	
occlusion	
Half of a person	
Low resolution	
Under-exposure	
Over-exposure	
A non-human object is annotated as human	

Table 2 challenging situations that influences recall

TABLE III. CHALLENGING SITUATIONS THAT INFLUENCES PRECISION

1) an annotated bounding box contains multiple persons



2) size of annotated bounding box is larger than a person



Table 3 challenging situations that influences precision

VII. FUTURE SUGGESTION

A. Suggestions on Early Fusion

The suggestion to solve information loss problem in early fusion is to modify the code of RetinaNet in order to make it accept four bands. It requires a deep understanding of RetinaNet. It is a complicated and time-demanding work. Because RetinaNet is very complex.

B. Suggestions on Late Fusion

In the future, it is suggested to integrate thermal and RGB images in the internal layers of RetinaNet, such as fully convolutional layers. This approach is different from integrating the bounding boxes. It will be interesting to

compare the two different architectures of late fusion and to find out which architecture is better.

VIII. CONCLUSION

This research aims at finding out the best architecture with the approach of convolutional neural network (CNN) applied in human detection. The architecture takes thermal and visual images as input and it detects the human in these images. RetinaNet is the state-of-art approach in this research.

Six deep learning models have been implemented. Four reference models are single-sensor models. They are trained by either thermal or visual images. Two models are multi-sensor fusion models. They are early fusion model and late fusion model. Early fusion model integrates thermal and RGB images on a pixel level. This is implemented by converting RGB and thermal images to HLT (Hue, Luminance, Thermal). Late fusion model integrates thermal and RGB images on a decision level. Two single sensor models, *cocokittikaistT* and *cocokittikaistRGB*, are chosen as the subnetworks of late fusion model. The integration is implemented by incorporating the bounding boxes.

It is concluded that late fusion model surpasses the performance of all single-sensor models. Early fusion does not make improvement on single-sensor models. Furthermore, it has been observed that fine-tuned models have better performance than non-finetuned models. Models trained on thermal images have better performance than trained on RGB images.

REFERENCES

- [1] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, 'Multispectral Deep Neural Networks for Pedestrian Detection', pp. 1–13, 2016.
- [2] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, 'Computer vision and deep learning techniques for pedestrian detection and tracking: A survey', *Neurocomputing*, vol. 300, pp. 17–33, Jul. 2018.
- [3] X. Fan, L. Xu, X. Zhang, and L. Chen, 'The Research and Application of Human Detection Based on Support Vector Machine Using in Intelligent Video Surveillance System', in *2008 Fourth International Conference on Natural Computation*, 2008, pp. 139–143.
- [4] J. L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, 'Transition-Aware Human Activity Recognition Using Smartphones', *Neurocomputing*, vol. 171, pp. 754–767, 2016.
- [5] M. Correa, G. Hermosilla, R. Verschae, and J. Ruizdel-Solar, 'Human Detection and Identification by Robots Using Thermal and Visual Information in Domestic Environments', *J. Intell. Robot. Syst.*, vol. 66, no. 1–2, pp. 223–243, 2012.
- [6] J. H. Kim, H. G. Hong, and K. R. Park, 'Convolutional Neural Network-Based Human

Detection in Nighttime Images Using Visible Light Camera Sensors’, *Passaro VMN, ed. Sensors (Basel, Switzerland)*, vol. 17, no. 5, p. 1065, May 2017.

- [7] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, ‘Multimodal Machine Learning: A Survey and Taxonomy’, *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–20, 2017.
- [8] P. Afsar, P. Cortez, and H. Santos, ‘Automatic visual detection of human behavior: A review from 2000 to 2014’, *Expert Syst. Appl.*, vol. 42, no. 20, pp. 6935–6956, Nov. 2015.
- [9] J. Wagner, V. Fischer, M. Herman, and S. Behnke, ‘Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks’, *ESANN 2016 proceedings, Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn.*, pp. 27–29, 2016.
- [10] X. X. Zhu *et al.*, ‘Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources’, *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.
- [11] T. Lin, P. Goyal, R. Girshick, K. He, and Piotr Dollar, ‘Focal Loss for Dense Object Detection’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.
- [12] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, ‘Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection’, *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, 2018.
- [13] D. M. W. Powers, ‘Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation’, 2007.
- [14] V. Fischer, M. Herman, S. Behnke, and A. I. Systems, ‘Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks’, *Conf. 24th Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn.*, no. April, pp. 27–29, 2016.
- [15] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Ying Yang, ‘Fusion of Multispectral Data Through Illumination-aware Deep Neural Networks for Pedestrian Detection’, *ARXIV*, 2018.