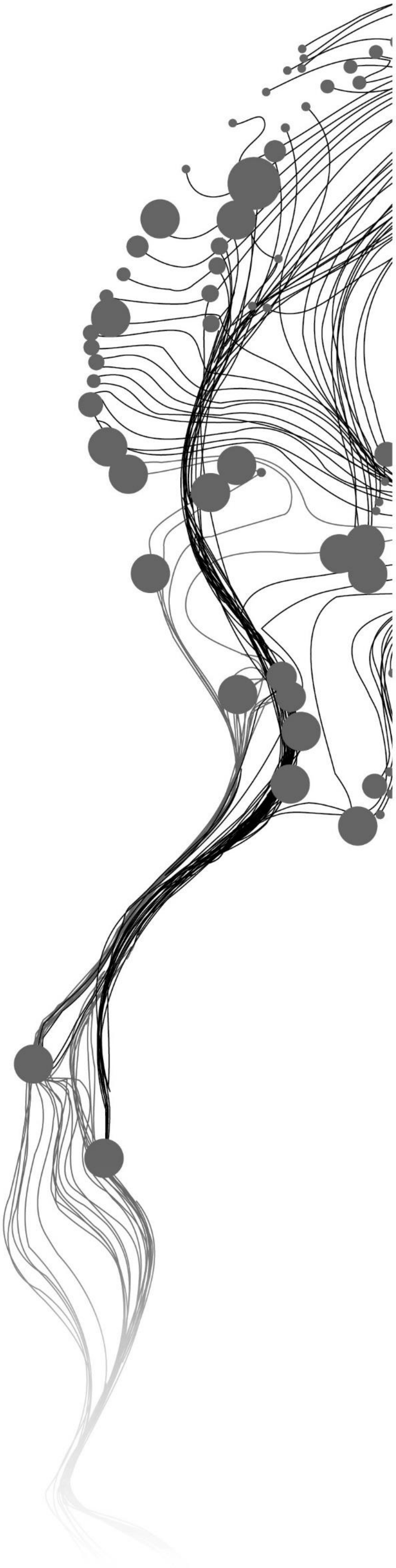
Human Detection in a Sequence of Thermal Images using Deep Learning

XINRAN WANG
FEBRARY, 2019

SUPERVISORS:
Dr, S. Hosseinyalamdary
Dr, F.C. Nex



Human Detection in a Sequence of Thermal Images using Deep Learning

XINRAN WANG
Enschede, The Netherlands, FEBRARY, 2019

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Dr, S. Hosseinyalamdary
Dr, F.C. Nex

THESIS ASSESSMENT BOARD:
Prof. Dr. ir. M.G. Vosselman (Chair)
Dr.M.Koeve, University of Twente, ITC-PGM
etc

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Human detection technology plays an irreplaceable role in many areas, such as search and rescue (SAR), autonomous driving, and surveillance, in recent years. Human detection is a still challenging task because, for the group of people, each individual has his unique appearance and.body shape. At the same time, humans can make thousands of gestures.

Compared with the traditional method, the deep learning neural network has the advantages of shorter computing time, higher accuracy and easier operation. Therefore, deep learning method has been widely used in object detection. The current state of art in human detection is RetinaNet. It is a robust one-stage object detector (Lin, Goyal, Girshick, He, & Piotr Dollar, 2018). This approach proposed a new function of loss to address the imbalance between foreground and background classes.

The temporal component of video provides additional and significant clues as compared to the static image. In this paper, the temporal relationship of the images is utilized to improve the accuracy of human detection. Compared to using only an image, the accuracy of human detection is higher when a sequence of images is applied.

The dataset used in this thesis is from KAIST. The dataset is disappointing because it has suffered a lot of occlusion and wrong annotations. In addition, the image resolution is low, and the l distance between people and the camera makes the outline of people very vague. That will have a negative impact on the result.

In this study, three temporally-consistent convolutional neural networks and a basic convolutional neural network have been used to do human detection. And the results show that all temporally-consistent CNNs perform better than the basic CNN. The continuous later CNN has the best performance that the accuracy of human detection is 21.4% higher than that of the basic model.

Keywords
human detection, temporal consistency, deep learning, thermal images

ACKNOWLEDGEMENTS

Here, I want to thank my first supervisor Dr. S.Hosseinyalamdary give my enthusiasm to encourage in my paper, intelligent ideas, support, guidance, and valuable discussions. He gave me many constructive Suggestions and criticisms and made me think harder and deeper in the paper. I can't finish my paper without their help.

I would like to thank Dr, F.C. Nex and all other teachers in the GFM program for their inspiring lectures and exercises that helped me learn a lot in geoinformatics.

Thank you very much for helping my classmate, Ms. QIAO Ren, when I encountered difficulties in the paper. I would also like to thank my ITC friends for their encouragement and positive attitude when I was depressed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Background information

Human detection is a useful tool in many research fields. This technology provides a solid technical foundation for these problems and guarantees their development. The importance of human detection in autonomous driving, Post-disaster rescue, automated surveillance, military and robotics services has become significant (Gajjar, Gurnani, & Khandhediya, 2017).

1.1.1. Post-disaster rescue

Unmanned aerial vehicles (UAV) can be automatically operated in dynamic and sophisticated operating environments (Doherty & Rudol, 2007). One particularly important application is in the searching assistance after the occurrence of a disaster.

After the disaster, a device equipped with a human detector could help the rescue team find out where the survivors are (Doherty & Rudol, 2007). If the specific location of the trapped people can be detected and reported to the search and rescue personnel, it will greatly enhance the efficiency of the rescue team. According to the information, rescuers can plan the most reasonable rescue plan. This allows the rescue personnel to work more efficiently and search the disastrous regions without being physically present. It guarantees the safety of the rescue team in the searching operation. This will provide more time for rescue work and reduce the occurrence of other accidents and avoid unnecessary losses.

However, more reliable human detection algorithm can help to rescue the victims of a disaster or an accident.

1.1.2. Autonomous driving

The main purpose of developing autonomous systems is to replace human beings to complete tedious and boring daily tasks. This is one of the major challenges facing modern computer science. An example is the automated driving system that can help reduce deaths caused by traffic accidents (Geiger, Lenz, & Urtasun, 2012). Nowadays, many experts and scholars are devoted to the research and development of autonomous driving cars.

In autonomous driving, human detection technology ensures the safety of both drivers and pedestrians (Ballas, Larochelle, & Courville, 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.

1.1.3. Automated surveillance

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. This is very helpful for the maintenance of public order.

With the rising crime rate in the world today, people are also aware of the need to detect abnormal activities (Gowsikhaa, Abirami, & Baskaran, 2014). Through human detection, some abnormal phenomena can be found in time to maintain public safety.

1.1.4. In military

In the military, the 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 at night or dark place. Thermal detectors can even go through thin walls or other shielding to detect human beings. The higher the accuracy of human detection, then there will be more initiatives in military campaigns.

No matter which application is mentioned, how to obtain high precision human detection algorithm is the primary consideration. The increased precision achieved in the field of human detection will play a huge role in other applications that rely on human detection technology. So how to improve the accuracy of human detection is a problem worth studying.

1.2. Problem statement

1.2.1. 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.

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 only pay attention to 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, as shown in Figure 1.1.

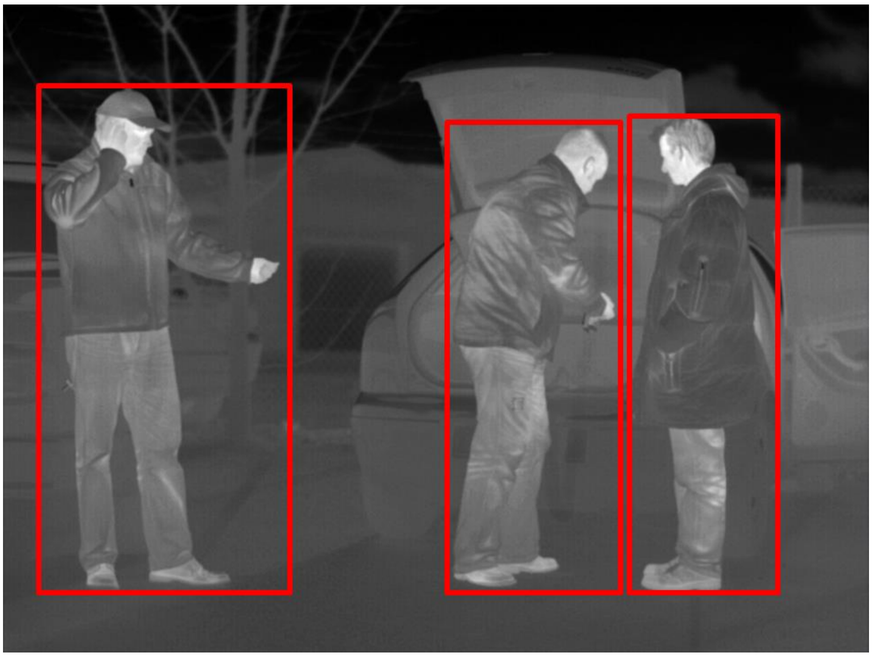


Figure 1.1 The task of human detection

1.2.2. Challenges in human detection

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, blurred 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 1.2. Human with different skin colors[1] (left), different body shapes[2] (middle) and different postures[3] (right).

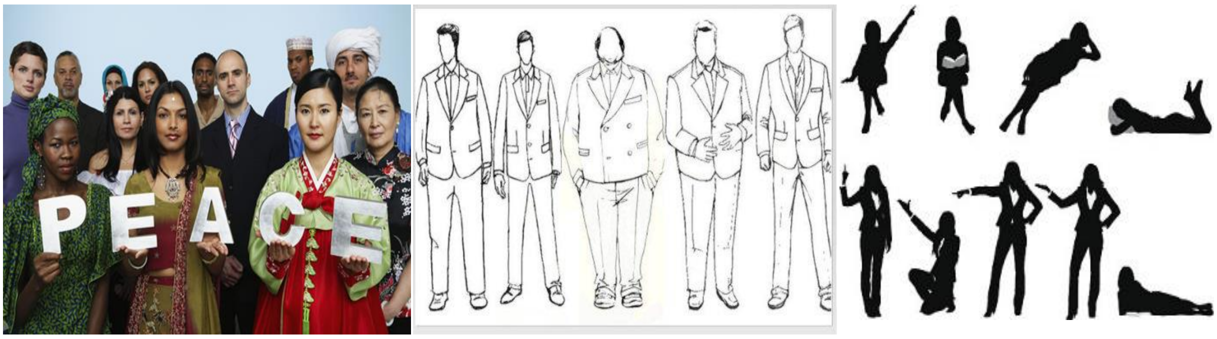


Figure 1.2 Human with different skin color(left), different body shape (middle) and different postures (right)

While the computer can only use graphic information, so different clothing will make it more difficult for computers to recognize a human, as shown in Figure 1.3. Human with various clothes (right and left-bottom)[4] and the different view of a people captured by a camera (left-top)[5].

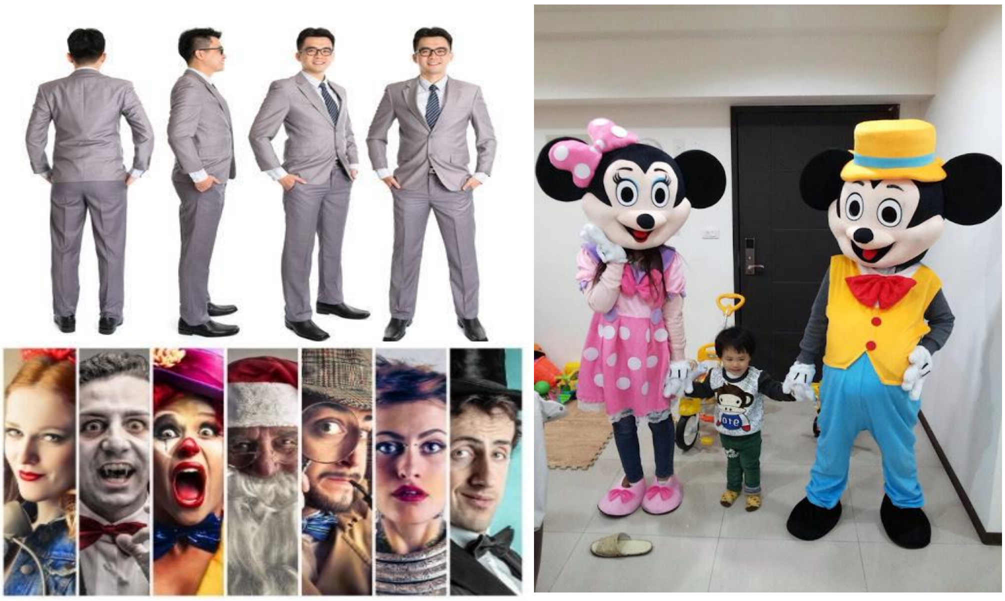


Figure 1.3 Human with various clothes (right and left-bottom) and the different view of a people captured by a camera (left-top)

[1] http://baike.chinaso.com/wiki/pic-view-781842-234165.html
[2] http://blog.xiaojunche.com/post/97.html
[3] https://www.vcg.com/creative/811803218
[4] http://werlovewoman.blogspot.com/2015/08/blog-post_80.html
http://www.52112.com/pic/325058.html
[5] https://www.123rf.com/

As well as camera capture at various views to the human body (shown in Figure 1.3 left top). This can cause humans to be obscured by other people or objects (Figure 1.4shows). The contours of people far from the camera are blurred, which can also affect the accuracy of human detection. People who are occluded by other people[6] and by other objects[7].

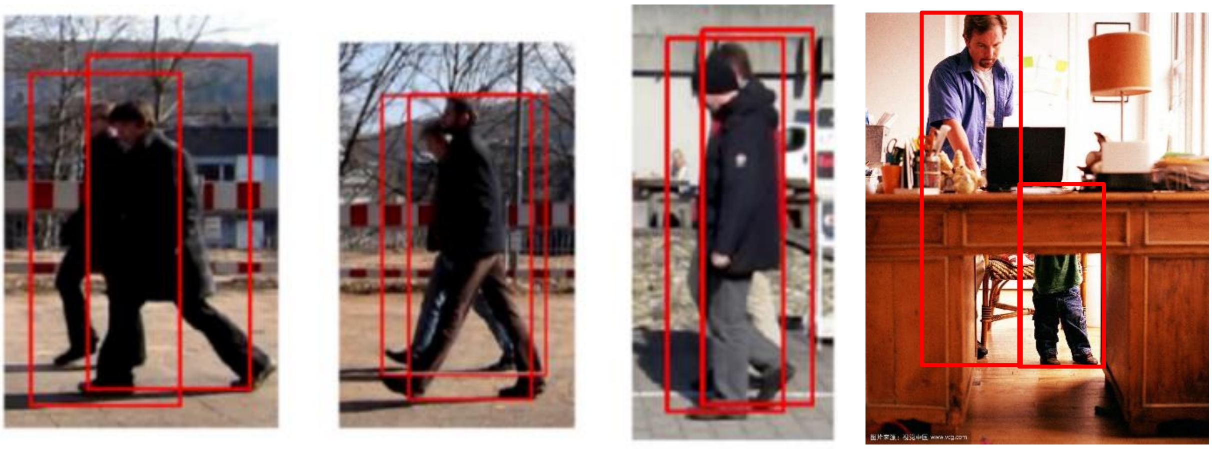


Figure 1.4 People who are occluded by other people and by other objects

1.2.3. The characteristics of thermal and visual images

Visual images and thermal images are the two major information sources used in human detection researches (Fan, Xu, Zhang, & Chen, 2008).

Visual images are more widely used than thermal images. Visual image refers to an image synthesized using RGB channels. In the visual images, the human detection has been extensively studied and has achieved good results. Most human detection tasks are still based on visual images (Hwang, Park, Kim, Choi, & Kweon, 2015). 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. Visual images are affected by poor lighting condition. Because 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 the representation of the amount of infrared energy emitted, transmitted and reflected by the object in terms of brightness (Correa, Hermosilla, Verschae, & Ruiz-del-Solar, 2012). The amount of radiation emitted by the object increases as the temperature of the object increases. Due to this, the thermal camera can measure the temperature of an object. The body temperature is different from the temperature of the environment. Thus, the thermal image can distinguish between human and other objects, particularly at night or in shadow. Advanced thermal detectors also have the ability to detect infrared radiation behind thin walls or other obstructions.

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). They provide limited performance in human detection. The closer the body temperature to the external environment temperature, the worse human detection accuracy in the thermal image. In desert areas, for example, thermal images perform poorly because the temperature of sand changes frequently and human detection

[6] https://ps.is.tuebingen.mpg.de/publications/tangijcv

[7] http://www.vcg.com/

can be a challenge. In addition, many other things that generate heat automatically may interfere with human detection.

Besides, the resolution of the thermal image is low, and it's 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. So, the thermal information is not as detailed as the visual images.

The differences of visual image and thermal image are shown in Figure 1.5.

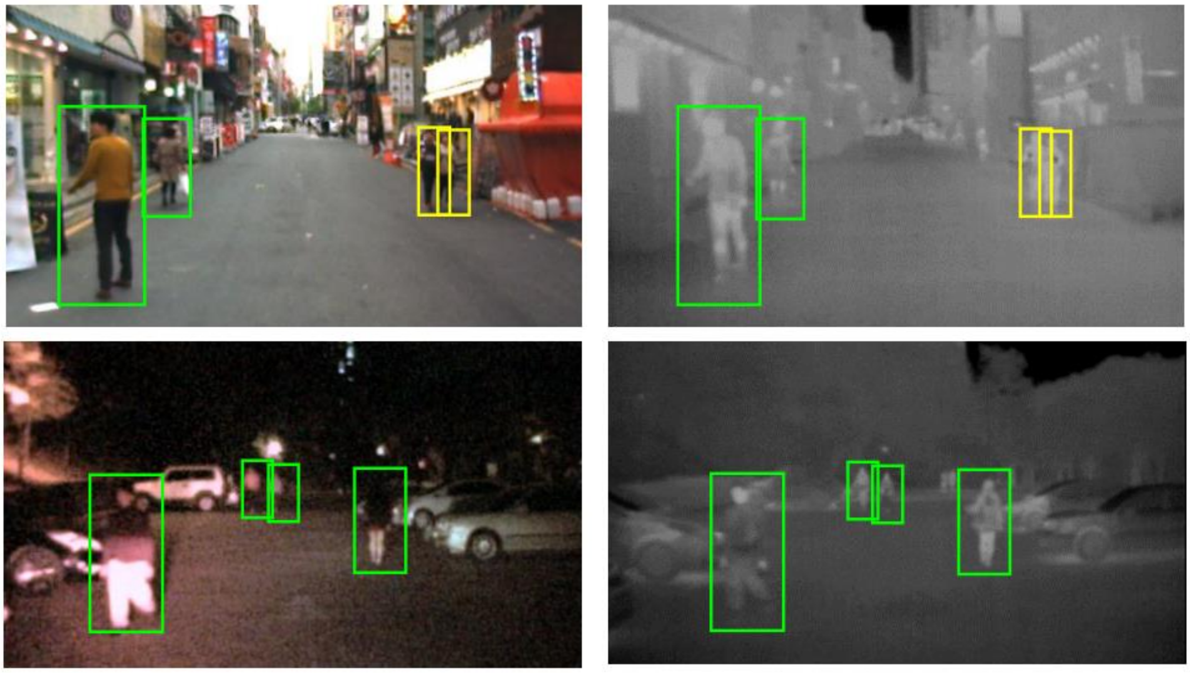


Figure 1.5 Pairs of visual images and thermal images (Hwang et al., 2015). The two images were taken during the day (top) and at night (bottom). The green bounding box shows the people who is not obscured. The green bounding box shows the people who is not obscured. The green bounding box shows the people who is obscured by other people.

1.2.4. Temporal consistency

Compared with the single image, video analysis provides more information for the identification of the task. It adds the time component and therefore, it improves the accuracy of human detection by adding movement information and trajectory (Ng et al., 2015).

The temporal component of video provides additional and significant clues as compared to the static image classification since many actions can be reliably identified based on motion information (Simonyan & Zisserman, 2014). Compared to using only an image, the accuracy of human detection is higher when a sequence of images is applied. In other words, we can use the temporal relationship of the images to improve accuracy of the human detection.

All objects are in constant movement. The occurrence of motion takes time. The time is added to the neural network while motion information is added. Nowadays, there are two main ways to add time information into convolutional neural networks. One is to process two or more frames simultaneously, while the other works with optical flow and corresponding frames at the same time.

The essence of the second method is to extract the motion information between different frames. The temporal information is converted into movement of detected objects and added to the convolutional neural network.

An example is shown in Figure 1.6. The images show that if a human being is identified in a series of continuous frames, it will be more confidently detected as human.

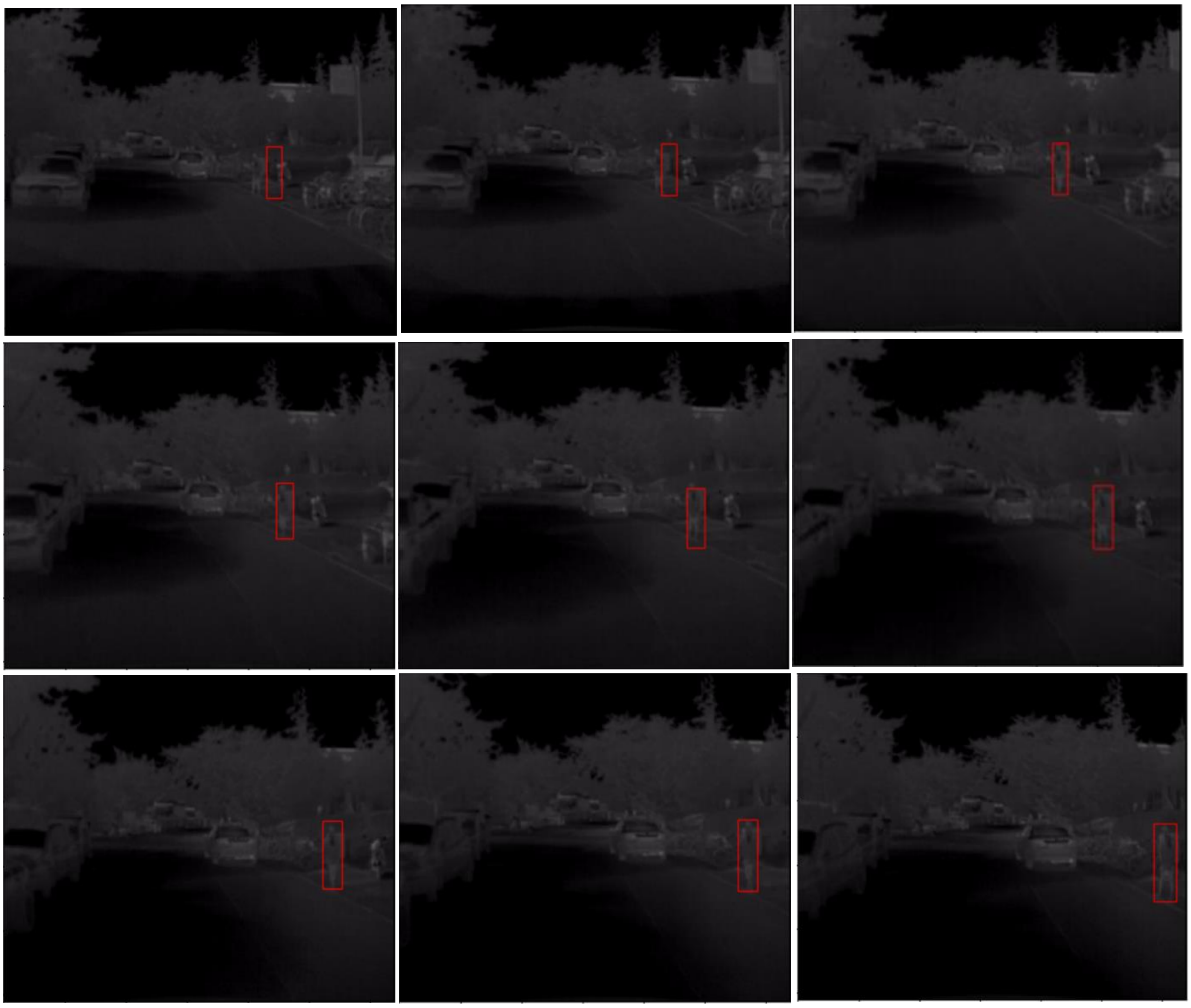


Figure 1.6 Human detection in a sequence of thermal images. The temporal consistency is applied to detect human in a sequence of thermal images.

1.3. Objective and research questions

1.3.1. Research objective

The main objective of this study is to apply deep learning method and temporal information to convolutional neural networks to do human detection with a sequence of thermal images.

1.3.2. Research questions

1. What is the state of art in human detection? How accurate is the state of art?
2. How much temporal CNN can improve human detection using the state of art?

3. What is the best temporal CNN architecture and how much it improves the state of the art?

1.3.3. Innovation aimed at

Many methods of deep learning have already been applied to object detection and have made remarkable achievements. Temporal information is of great importance here because it enforces temporal consistency among thermal images and improves human detection accuracy.

Here, I propose a novel idea to add temporal information to the thermal images and apply different types of Temporal convolutional neural network (T-CNN). Later, I am going to compare the results of these approaches to human detection.

The key contribution of this study is to find out which type of deep learning methods perform better in human detection by using a sequence of thermal images. This is the first time to use a sequence of thermal images to do human detection based on my own knowledge. We are the first group to use this innovative approach to do human detection.

1.4. Organization of this Dissertation

This thesis consists of eight chapters.

Chapter 1 describes the motivation and background information of this study, as well as the problems to be solved in this study.

Chapter 2 gives a brief review of scene interpretation methods in the current literature.

Chapter 3 focuses on datasets which have been used in this study.

Chapter 4 explains the architecture of methodology applied in this research.

Chapter 5 introduces some necessary implementation of this research.

Chapter 6 is a description of the experimental results and the analysis of the results.

Chapter 7 gives some discussion on this study. This chapter also describes some problems can be addressed in the future.

Chapter 8 is a short conclusion on this study and some recommendation.

Appendix shows some both good results and failed cases of this thesis.

2. LITERATURE REVIEW

This chapter briefly reviews the existing human detection approaches related to this research. In general, there are two types of human detection, one applies computer vision approaches and another utilizes the deep learning methods.

2.1. Traditional approches

One of the most famous computer vision approaches is Histogram of Oreinted Gradient (HOG). This main idea of HOG is feature extraction and object detection chain, as shown in Figure 2.1.

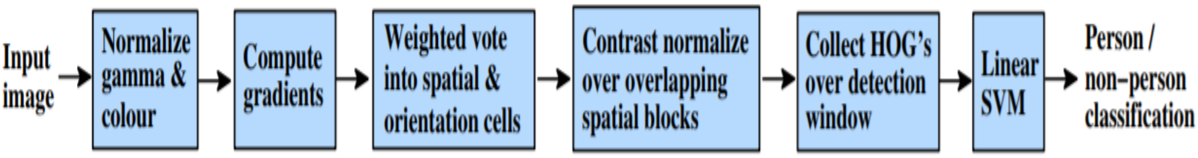

Input image
Normalize gamma & colour
Compute gradients
Weighted vote into spatial & orientation cells
Contrast normalize over overlapping spatial blocks
Collect HOG's over detection window
Linear SVM
Person / non-person classification

Figure 2.1 The main steps of HOG human detection(From (Dalal & Triggs, 2005))

In the detection window, detector is used to analyze interest points and draw histogram with gradient feature vector. Then all the histograms are 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 the detection of the human body in visual images, but the approach has been effectively applied to thermal images.

HOG takes visual images as input while THOG takes thermal images as input (Baek et al., 2017) In this study, they proposed an approach called thermal-position-intensity-histogram of oriented gradient (TPIHOG or TπHOG). TPIHOG or TπHOG improves pedestrian detection performance of HOG by incorporating temperature gradient information and its location as well as the temperature intensity of the night or dark environment.

Standard aggregated channel feature detector (ACF) is widely used as a basis algorithm on KAIST dataset (Dollar, Appel, Belongie, & Perona, 2014; Yang, Yan, Lei, & Stan Z. Li, 2014 ; Nam, Dollar, & Hee Han, 2014). Standard ACF uses color images as input. Standard ACF consists of 10 augmented channels, including color channels, gradient magnitude and gradient histograms (Hwang et al., 2015). This approach decreases computational costs substantially (Zhang, Bauckhage, & Cremers, 2014 ; Paisitkriangkrai, Shen, & Hengel, 2014)

Wang, Zhang, and Shen applied a new method based on the Shape Context Descriptor (SCD), using the Adaboost cascade classifier framework (Wang, Zhang, & Shen, 2010). This approach makes the thermal detector is not only in the night or dark environment 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, Cao, & Yang, 2019). 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.

2.2. Deep learning

The concept of deep learning stems from research on artificial neural networks. Multiple-hidden layer perceptron is one of the structures of deep learning (Lecun, Bengio, & Hinton, 2015). The main idea of depth learning is to incorporate low-level features into a more abstract level of advanced presentation attribute categories or features. The distributed characteristic representation of the data is found.
Deep learning is a kind of machine learning method based on data representation. Observations (such as images) can be expressed in a variety of ways. For example, a vector of the intensity value of each pixel, or more abstractly represented as a series of edges, areas of a particular shape, and so on (Ramachandran, Rajeev, Krishnan, & Subathra, 2015).
A basic CNN consists of two main processes: feature extraction and classification (Zhu et al., 2017). The purpose of feature extraction is to extract different parts of each object, such as human head, arms, and legs. Classification refers to calculating the degree of certainty of a person at this location. If the certainty is high, then the outcome is going to be one person in that position and vice versa. The basic principle of feature extraction is the detection of features from low to high levels. Low-level features include edges and colors. A high-level feature is an object, such as a cat, a tree, and a table.

Nowadays, more and more deep learning approaches are applied to human detection.
R-CNN (region-based convolutional neural network method) combine region proposals with CNNs (Girshick, Donahue, Darrell, Berkeley, & Malik, 2012) (Uijlings, Sande, Sande, & Smeulders, 2012). First, they applied high-capacity 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)(Hosang, Omran, Benenson, & Schiele, 2015 ; Li et al., 2017)). 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 deeper neural networks than R-CNN. F-CNN is faster than R-CNN and it can achieve higher accuracy in testing task (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 images and predict the offset value of the bounding box.

2.3. Temporal Convolutional Neural Network

Ng et al. propose two deep learning methods for handling long video classification (Ng et al., 2015). The first approach explores the various convolutional time property pool architectures and finds the design choices that need to be made to adapt to CNN for this task. The second way they used a cyclic neural network to model video as an ordered frame sequence.

Simonyan applied the motion information in two adjacent images to the deep convolutional network of video action recognition (Simonyan & Zisserman, 2014).

In action recognition, both Ng et al. and Simonyan & Zisserman suggested applying optical flow to better enforce temporal consistency.

In the detection of human abnormal activities, Zhou et al. proposed a method to detect and locate abnormal activities in the video sequence of crowded scenes (Zhou et al., 2016). The main novelty of this method lies in the coupling of anomaly detection and spatial-temporal CNN). By performing space-time convolution, the architecture captures features from the spatial and temporal dimensions while the appearance and motion information contained in the continuous frames is extracted. Experimental results Ballas and his colleagues proposed a method which successfully considered the local and global time structures of video to generate the description (Ballas et al., 2015). This method combines the representation of temporal and spatial 3D convolutional neural network (3D-CNN) to short-time dynamics. 3D CNN represents training through the video action identification task, giving a representation that is compatible with human action and behavior.

Karpathy studies multiple ways to extend CNN connectivity in the time domain to take advantage of local spatiotemporal information and proposes a multi-resolution, centrally structured architecture as a promising approach to accelerate training (Karpathy et al., 2014). Lea proposed temporal CNN which is a unified approach to capturing relationships in a hierarchical manner at low, medium and high time scales. Figure 2.2 shows the architecture of temporal CNN.

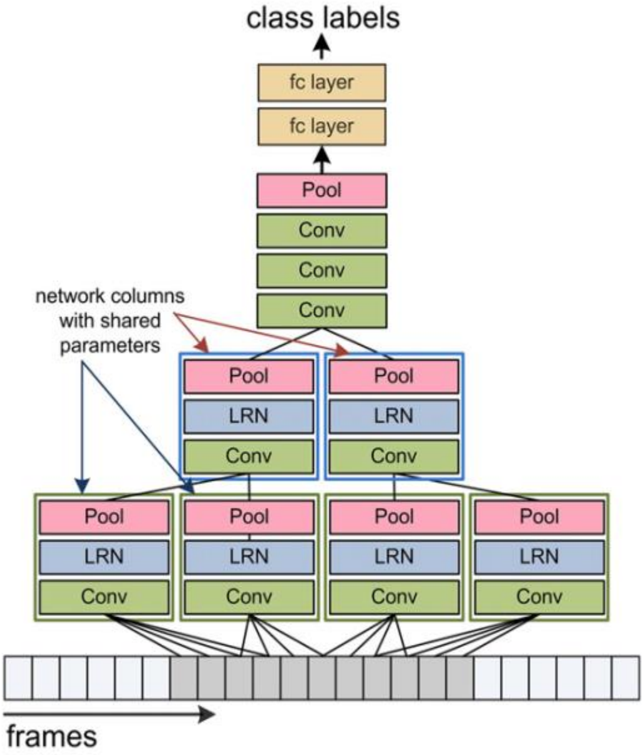


Figure 2.2 The architecture of CNNS. Denoted in yellow, blue, red and green are respectively fully connected layers, normalization(linear regression network), spatial-pooling and convolutional layers (Herath, Harandi, & Porikli, 2017).

Ji, Xu, Yang and Yu have developed a new 3D CNN motion recognition model for automatic recognition of human behavior in surveillance video (Ji, Xu, Yang, & Yu, 2013). The model is a model of the three-dimensional convolution of two dimensions from space and temporal, to get the movement information that is encoded in continuous frames. The developed model generates multiple channel information from the input frame, and the last feature represents the combination of information for all channels. They developed model was applied to the airport to monitor video identification of human behavior under the real environment, compared with the baseline method, this model has better performance.

König and his colleagues add temporal consistency among images in their work. In other words, they apply a sequence of multispectral images to detect people on the street. They also conclude that the halfway sensor fusion architecture is the best among different architecture. They show middle features are the best to integrate (König et al., n.d.).

3. DATASET

This chapter describes the details about the dataset. In this thesis, a total of three datasets were applied. They are COCO, KITTI and KAIST dataset.

As mentioned in the first chapter, human detection is still a challenging task. Because of the diversity of human clothing, posture, and appearance. Compared with other kinds of object detection, human detection needs a large number of samples to get an ideal result. The thermal image dataset is much smaller than the visual image dataset. The KAIST dataset is far from sufficient for human detection. To deal with this problem, we adopted the idea of "domain adaptation". Here, COCO and KITTI images are introduced to pre-train the model.

3.1. COCO

COCO is a large-scale dataset which can be used for object detection, segmentation, and captioning[8].

COCO dataset contains visual images which have 3 channels (RGB channels).

COCO dataset is an excellent object detection dataset not only because of its object segmentation and context recognition, and it also has the characteristics of super-pixel segmentation.

The COCO dataset has 330K images, of which more than 200K are labelled. There are 1.5 million object instances in this dataset, covering 80 object categories and 91 item categories. The most important of these is that 250,000 people with key points.

The ratio of training images to validation images is approximately two to one.

In this thesis, the COCO images related to humans such as pedestrians and cyclists can be used to do human detection.

3.2. KITTI

KITTI dataset is from The KITTI Vision Benchmark Suite. This is a project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago. This project aims at making full use of the autonomous driving platform Annie way to develop novel challenging real-world computer vision benchmarks[9].

KITTI dataset was captured while driving in Karlsruhe, rural areas and highways. Each image can see up to 15 cars and 30 pedestrians.

The KITTI dataset used in this thesis is from Object Tracking Evaluation 2012. The benchmark consists of eight different classes. Training sequences and test sequences are 21 and 29, respectively[10].

3.3. KAIST

A set of thermal images used in this thesis is from the website: KAIST (Korea Advanced Institute of Science and Technology) Multispectral Pedestrian Detection Benchmark. The KAIST Multispectral Pedestrian Dataset consists of 95000 colour-thermal pairs. They are captured by a vehicle which carries a

[8] http://cocodataset.org/#home

[9] http://www.cvlibs.net/datasets/kitti/

[10] http://www.cvlibs.net/datasets/kitti/eval_tracking.php

colour camera and a thermal camera. These images are paired as shown in Figure 3.1. The images are captured during day and night time. The dimension of these images is 640*512 pixels. Both horizontal and vertical resolutions are 96 dpi.

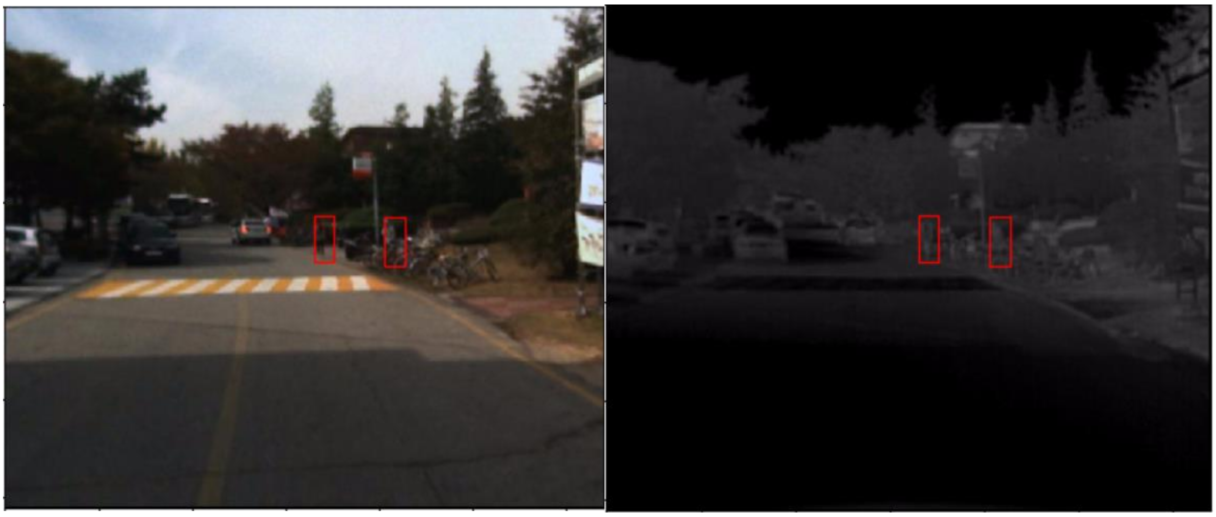


Figure 3.1 The comparison of human detection in visual images(left) and thermal images(right)

In this thesis, I only use thermal images part. The dataset used a long-wave infrared (7.5~13μm, also known as the thermal band) camera. The data is available online[11].

The dataset has 95328 images in total. The training part contains 50187 images, and testing part has 45141 images. The proportion of data used for training and testing is about 57% and 43% individually. This ratio is reasonable in model training. In order to make the training model more stable, 10 percent of the training data was used for validation. Each image has an annotation file corresponding to it. All human in images been labelled by annotation files. This file records the number and coordinate of bounding box in the images.

However, there are a lot of errors in annotation files, especially in testing part. So I searched for one fixed testing annotations provided by Ms Jingjing Liu[12]. The new annotations file fixes the previous part of the testing dataset. This improved testing annotation file is only one twentieth of the original test part. The improved testing annotation files are much better than the original annotation files. But there are still problems in this folder. There are also many uncertainties about the labelled ground truth. Through my own visual inspection, there are still a lot of problems with this improved annotation folder. Such as missing labels, wrong labels, and ambiguous labels, etc. Finally, the number of datasets used in this article is shown in Table 3-1.

[11] https://onedrive.live.com/?authkey=%21ADG6wuQeYqCroBI&id=1570430EADF56512%21624&cid=1570430EADF56512

[12] https://li-chengyang.github.io/home/MSDS-RCNN

Table 3-1 The number and allocation of datasets

Dataset	Training part	Validation part	Testing part
Day Campus	Set 00 15748 frames	Set 00 1750 frames	Set 06 648 frames
Day Road	Set 01 7232 frames	Set 01 803 frames	Set 07 406 frames
Day Downtown	Set 02 7080 frames	Set 02 786 frames	Set 08 401 images
Night Campus	Set 03 6001 frames	Set 03 667 frames	Set 09 175 frames
Night Road	Set 04 7480 frames	Set 04 720 frames	Set 10 444 frames
Night Downtown	Set 05 2628 frames	Set 05 292 frames	Set 11 178 frames
Total	**45169 images**	**5019 images**	**2252 images**

4. METHODOLOGY

This chapter describes the methodology of the whole thesis. The overview of the methodology is shown in the flowchart (Figure 4.1).

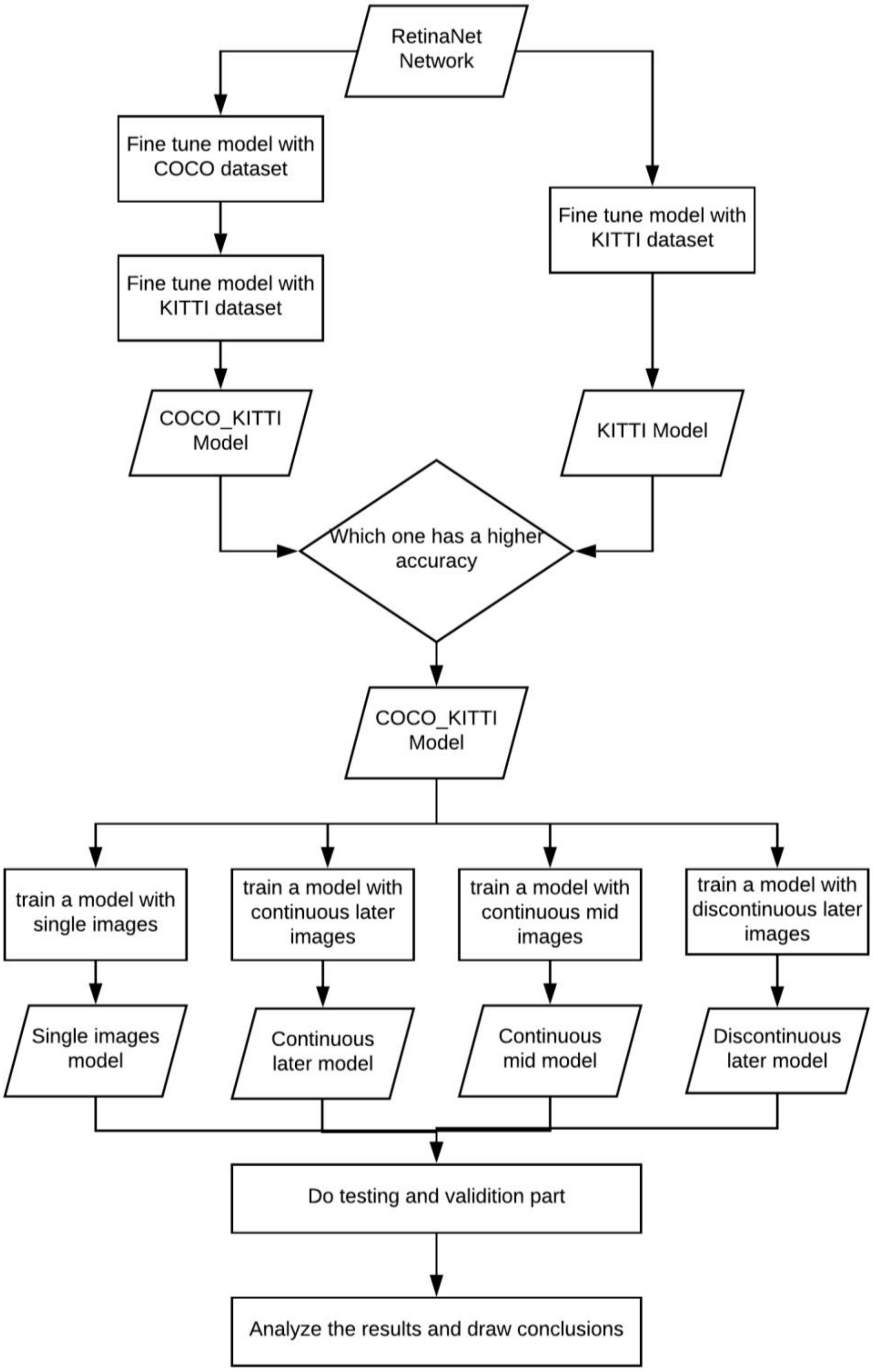

RetinaNet Network
Fine tune model with COCO dataset
Fine tune model with KITTI dataset
Fine tune model with KITTI dataset
COCO_KITTI Model
KITTI Model
Which one has a higher accuracy
COCO_KITTI Model
train a model with single images
train a model with continuous later images
train a model with continuous mid images
train a model with discontinuous later images
Single images model
Continuous later model
Continuous mid model
Discontinuous later model
Do testing and validition part
Analyze the results and draw conclusions

Figure 4.1 The architecture of thesis

4.1. Data Preparation

To study the effect of temporal consistency on human detection, this research set up four CNNs in total. The role of temporal information in human detection is studied by controlling the way in how time information is added.

When images are static, processing video requires modelling their dynamic time structure and then integrating this information appropriately into the natural language description model (Ballas et al., 2015). This requires the preparation of datasets for training, validation, and testing.

The process of data preparation is shown in the flowchart (Figure 4.2).

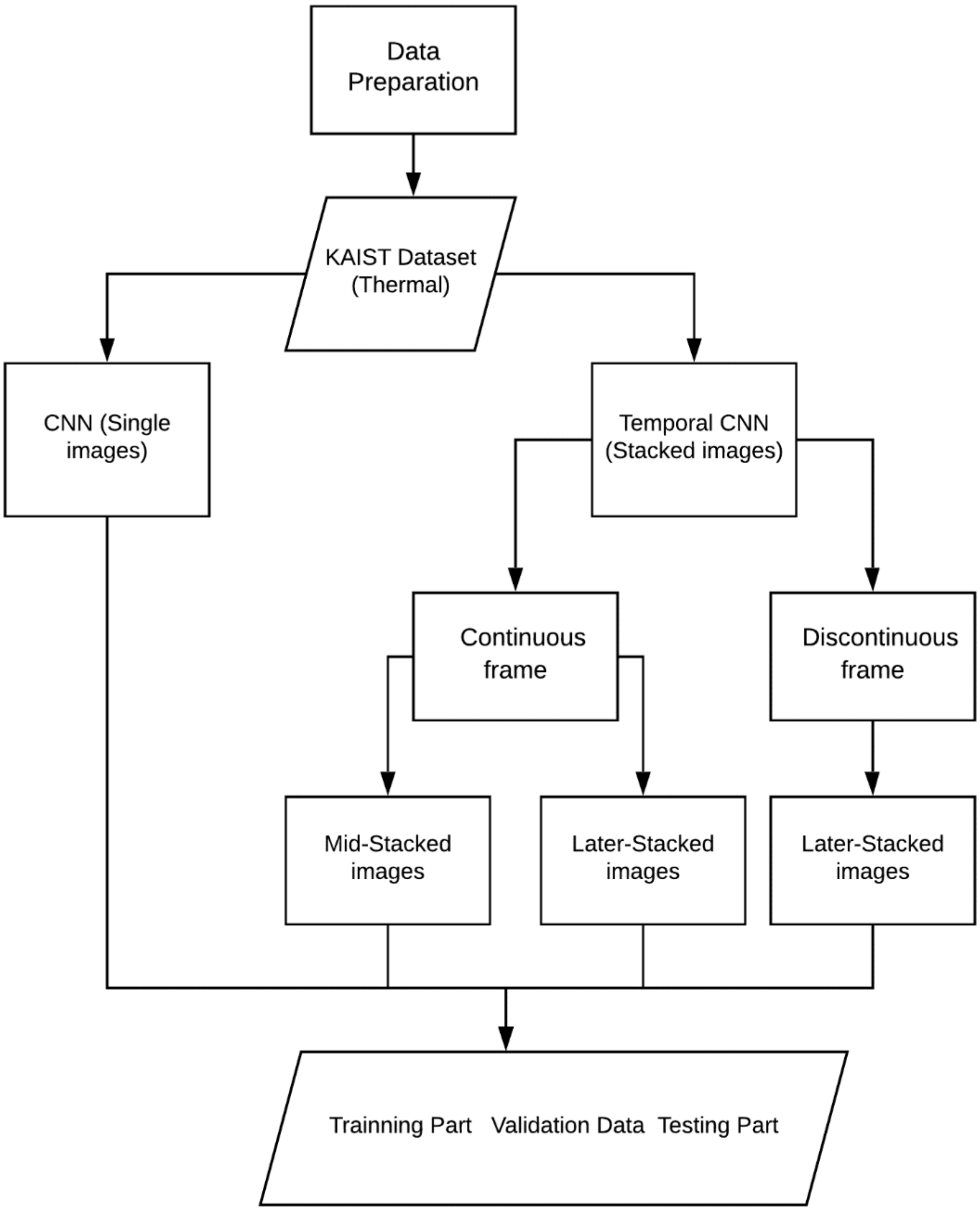


Figure 4.2 The architecture of Data preparation

Corresponding to four different types of networks, there are four corresponding methods for data processing. They are single images dataset, continuous later-core image set, continuous mid-core image set, and discontinuous later-core image set. Here stacked images mean three images are stacked to one image.

4.1.1. Continuous Frames

A continuous network uses adjacent frames to train the network without skipping any frames. In this case, the amount of calculation will be large, and the training time of the neural network will be long.

Basic CNN (Single image): For basic convolutional neural network, images are used separately to train the network and do validation and testing part. This means that the model is trained with single images. Only one image is processed at a time. Temporal information is not added to the neural network in this case. This network is just a basic convolutional neural network.

Temporal CNN (Stacked images): To add temporal information into neural network, three frames are stacked into one image to train the neural network. There are two ways to stack images, later core T-CNN and middle core T-CNN. The principles of these two methods are shown in Figure 4.3. Then these stacked images are used to train temporal convolutional neural networks. The time span between the two adjacent frames is 0.1 seconds.

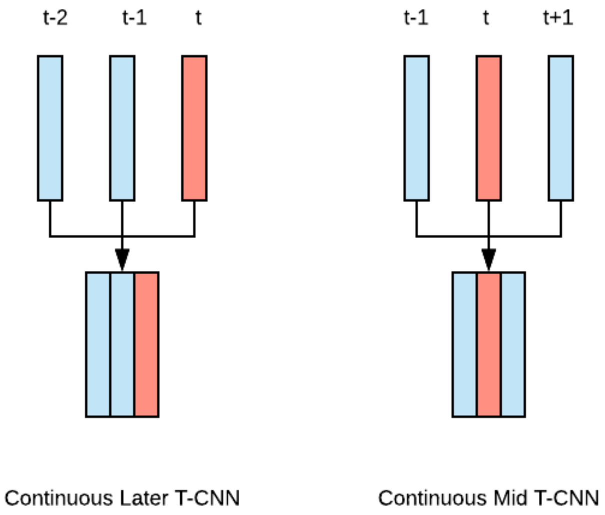


Figure 4.3 Continuous Later T-CNN (left) and Continuous Mid T-CNN (right). (The red rectangle is key frame and the blue rectangles are attached frame)

Here is an example from the dataset (shown in Figure 4.4). From left to right, they are I001364, I001365, and I001366. They are three adjacent images from Set00 V000. The coloured rectangle represents the position and coordinates of the person in the images from the annotation file corresponding to them individually.

Three images are stacked one image. The later-core images mean the stacked images (left bottom) using the later image's annotation (the blue rectangle). The mid-core images mean the stacked images (right bottom) using the middle image's annotation (the red rectangle).

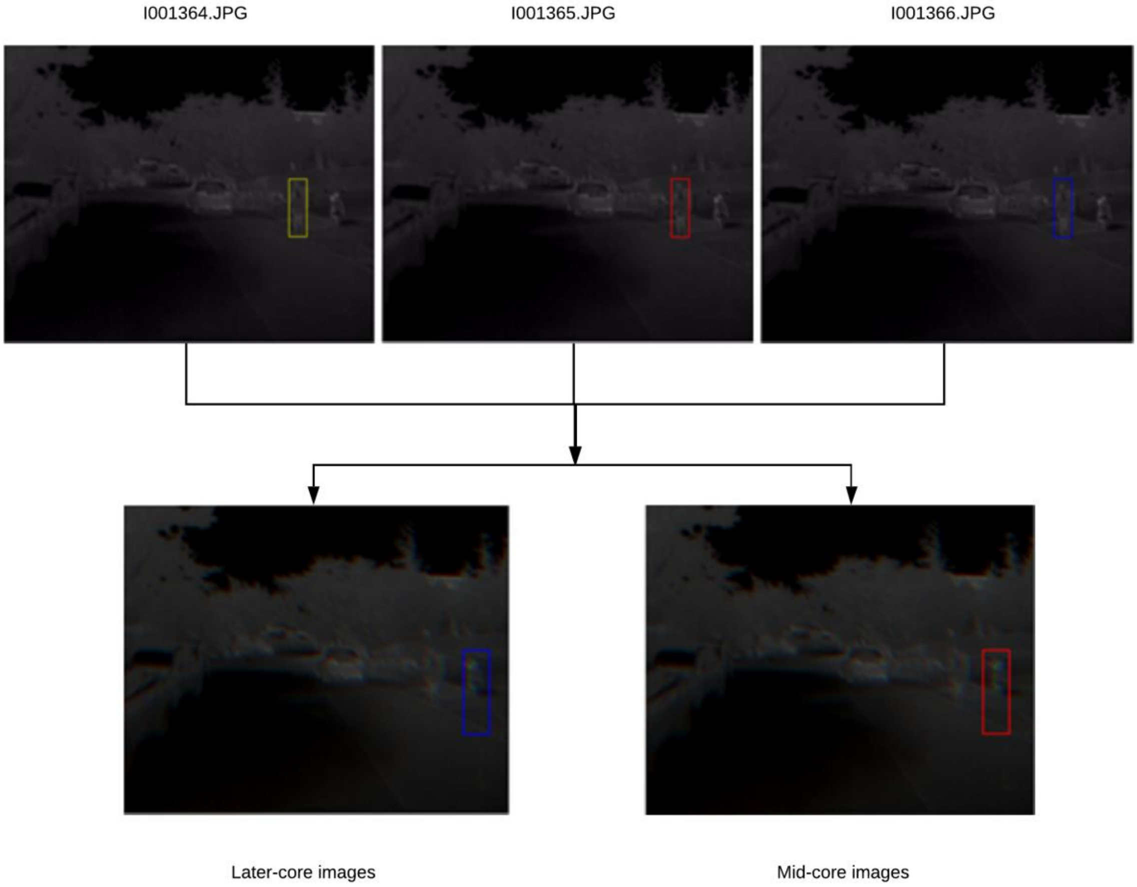


Figure 4.4 The example of later-core images and mid-core images

4.1.2. Discontinuous Frames

Temporal CNN (discontinuous stacked images): Due to the large amount of data, some frames are removed when stacking images. Because the time interval between each frame is very short (0.1 second), the movement of human in the adjacent two frames is not obvious. This way may lose some information but save plenty of time. And the difference between each frame will become obvious. Therefore, this dataset is used to study the influence of the input of a sequence of thermal images with intervals on the accuracy of human detection (shown in Figure 4.5).

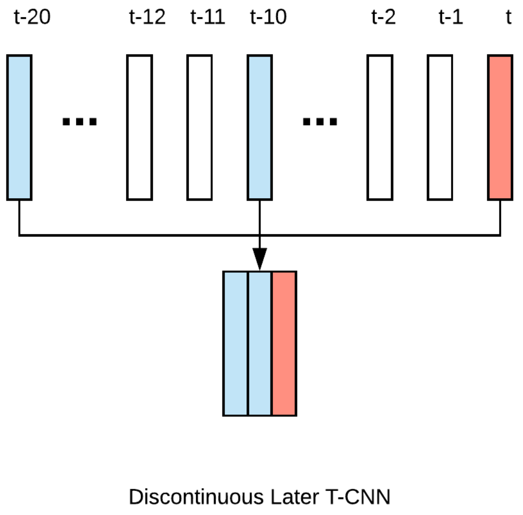


Figure 4.5 Discontinues Later T-CNN (The red rectangle is key frame; the blue rectangles are attached frames and white rectangle is skipped frames)

Here is an example from the dataset (shown in Figure 4.6). From left to right, they are I001346, I001356, and I001366. They are three images from Set00 V000. The coloured rectangle represents the position and coordinates of the person in the images from the annotation file corresponding to them individually. So this kind of image is called discontinuous image. The later-core images mean the stacked images using the later image's annotation.

In this study, nine frames are removed between every left two frames. This means the time span between the two adjacent frames is 0.1 seconds. The time interval between two frames in discontinuous T-CNN is 1 second.

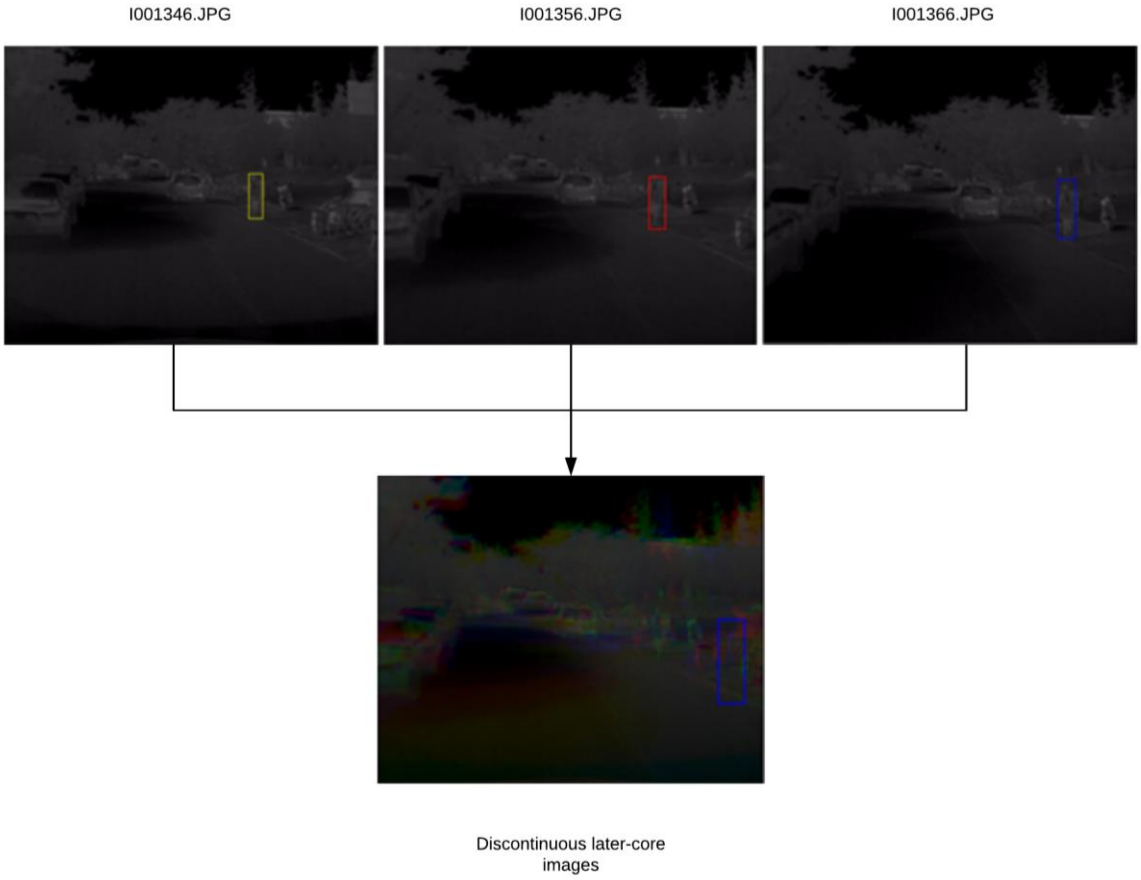


Figure 4.6 An example of discontinuous later-core image

4.2. Training prat and validation part

Unlike other stationary objects, humans move all the time. The essence of adding temporal information to the convolutional neural network is to add motion information to the model in order to improve the accuracy of human detection. In this study, I will compare the results of the basic CNN and temporally-consistent CNN for human detection. It is then concluded that there is a conclusion about that effect of temporal consistency on the accuracy of human detection.

4.2.1. Contrast network

Four different types of models were trained in order to investigate whether the time information would help improve the accuracy of human detection. One is trained with single images and the other three are temporally consistence neural network. This means temporal information has been added to the neural network. And stacked images are used instead of singles images to train the neural network.

These four networks are respectively trained by the dataset of single images, the dataset of the continuous stacked image with the middle image as the core, the dataset of the continuous stacked image with the later image as the core, and the dataset of the discontinuous stacked image with the later image as the core. The method of data preparation is introduced in the previous subsection.

Basic CNN means the model is trained by using single images. No temporal information has been added to this network. Temporally-consistent networks are trained with stacked images, which can add corresponding temporal information to the networks. In this thesis, I trained three different temporal consistency neural networks to study the effects of different types of time information on human detection.

Two temporal consistency networks are used to compare the effects of time series on human detection accuracy. In order to ensure the consistency of other variables, the two CNNs were trained with continuous images. The difference is that one network is at the core of the later image and the other the middle image is the core. For the later-core network, there are two front periods of time have been added to the network. For the other network, it takes advantage of a period of front time information and a period of behind time information.

Another control group studied the effect of time step on the accuracy of human detection. This may because the time interval between each frame is very short, it is 0.1 second, the movement of human in the adjacent two frames is not obvious. Perhaps we can remove some of the frames in the middle before the image is stacked. In this way, a lot of computation can be reduced, but some information will also be lost. In this thesis, continuous later-core CNN and discontinuous later-core CNN are utilized to find out the influence from time step on human detection.

4.2.2. Validation

Validation part is done with the training of the model. Validation dataset is ten percent from training dataset. Because the KAIST dataset does not have independent validation data. In order to make the model training more complete and to monitor the overfitting problem, I extracted the last ten percent of each folder of the training dataset used for validation.

4.3. Testing part

The testing assignment was done on each of the above trained four models. Then the same testing dataset, the improved data annotation files are given by Liu, has been used to evaluate the output. The mean steps of testing a model are shown in Figure 4.7.

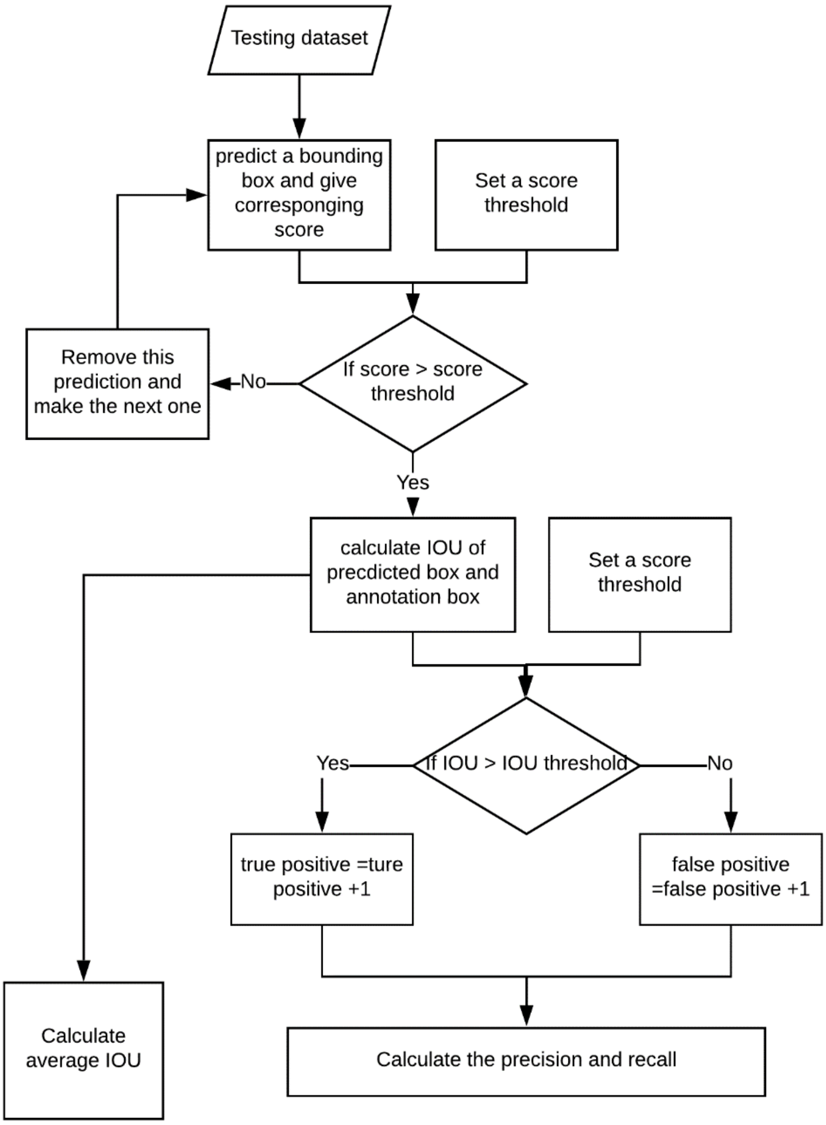

Testing dataset
predict a bounding box and give corresponging score
Set a score threshold
Remove this prediction and make the next one
No
If score > score threshold
Yes
calculate IOU of precdicted box and annotation box
Set a score threshold
Yes
If IOU > IOU threshold
No
true positive =ture positive +1
false positive =false positive +1
Calculate average IOU
Calculate the precision and recall

Figure 4.7 The mean steps of testing part

A parameter needs to be set during the testing process, score threshold. “score” is a value which represents the level of confidence that the retinal network identifies an object as a human. Compared with the ground truth, each bounding box predicted by RetinaNet network has corresponding confidence, which is called as "score". Score threshold is used to judge whether the prediction of this bounding box to be trusted. If the score given by the model of the bounding box is higher than the score threshold, the bounding box is accepted. Otherwise, the bounding box will not be considered correct. 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. A bounding box with a score higher than the score threshold will be recorded. Thus, each model has 9 results corresponding to 9 different score thresholds.

On the accuracy of the model evaluation mainly based on the classification accuracy, localization accuracy, and computational complexity in three aspects. The way to evaluate classification and positioning accuracy is mean average precision(mAP) (Han, Zhang, Cheng, Liu, & Xu, 2018).

In this paper, the average accuracy theory is used to find the best results for each model among different score thresholds. Recall, precision, F1 score and average Intersection of Union (IOU) will be the four criteria to evaluate the outcome of a human detection model.

Precision tells you how many of the detected objects were correct. It is a measure of completeness (Powers, 2007). Recall, also known as sensitivity, tells you how many of the objects that should have been detected were actually selected. It is a measure of exactness (Powers, 2007). F value, on the other hand, is the evaluation index integrating these two indicators and is used to comprehensively reflect the overall index. After calculating the value of precision and recall, F1 value can be obtained, which is the harmonic average of the two value(Sasaki, 2007).

Intersection over Union (IoU) is equal to overlap region divided by union region. IoU evaluates the geometric relation between labeled bounding box and predicted bounding box (Han et al., 2018). Equation 4.1, 4.2, 4.3 and 4.4 shows the formula for precision, recall, IoU calculation, and F1 score. According to the calculation formula, the output of true positives adds false positives is equal to the number of detected people. The output of true positives adds false negatives is equal to the number of labeled people. Figure 4.8 shows the schematic of IOU.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{4.1}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{4.2}$$

$$Intersection\ of\ Union = \frac{Intersection}{Union} \tag{4.3}$$

$$\mathrm{F1} = 2 * \frac{precision\ \cdot recall}{precision + recall} \tag{4.4}$$

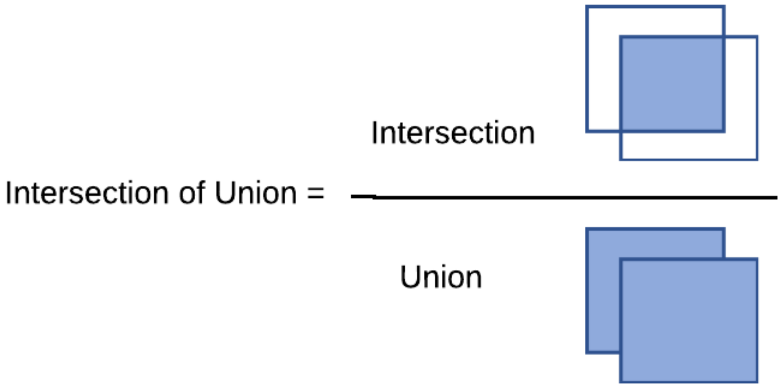


Figure 4.8 The formula for intersection of union

Average precision computes the average value of precision with respect to recall (Han et al., 2018). The higher the mAP is, the higher the accuracy is.

Mean average precision calculate the average of APs, when there are multiple queries. The calculation method of MAP is shown in Equation 4.5 and Equation 4.6.

$$\text{Avep} = \sum_{i=1}^{n} Precision_i * Recall_i \qquad 4.5$$

$$\text{MAP} = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \qquad 4.6$$

where Q is the number of queries.

5. IMPLEMENTATION

In this chapter, some implementation of this research has been introduced. There are three main aspects: the architecture of RetinaNet, fine-tuning model and validation results.

5.1. The architecture of RetinaNet

All the models are trained using RetinaNet network. RetinaNet, introduced by Lin et al, achieves a higher accuracy compared with the aforementioned neural networks on COCO dataset (Lin et al., 2018). Among all the deep learning approaches, RetinaNet gives the highest accuracy of human detection.

The current state of art in object detection is RetinaNet. It is a robust one stage object detector (Lin et al., 2018). RetinaNet is composed of a backbone network and two subnetworks. One of the subnetworks is used for classification, called classification subnet. The second subnet called box regression subnet which performs convolution bounding box regression. RetinaNet has a specific loss function, which can be used to address imbalance between foreground and background classes during training. Figure 5.1 shows the architecture of one-stage RetinaNet network.

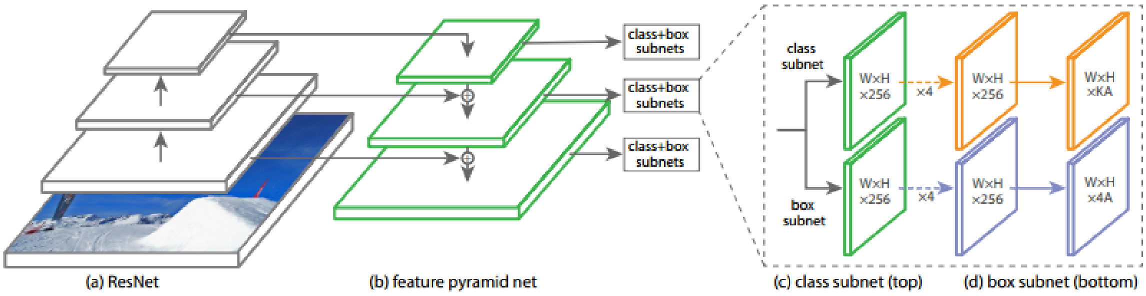


Figure 5.1 The architecture of one-stage RetinaNet network uses a Feature Pyramid Network (FPN) backbone on top of a feedforward ResNet architecture (Lin et al., 2018). (a) shows multi-scale convolutional feature pyramid. (b) represents feature pyramid net. (c) and (d) represent classification subnet and box regression subnet individually.

ResNet-101 is the backbone of this research. The default weight of RetinaNet network is trained by ImageNet. Table 5-1 shows main parameters of RetinaNet.

Table 5-1 Main parameters of Retinanet

Backbone	ResNet-101-Feature Pyramid Network (FPN) backbone
Learning rate	0.01(initial), divided by 10 at 60k and in 80k iterations
Weight decay	0.0001
Momentum	0.9
Batch size	1 for fine tuning model 4 for basic CNN and temporally-consistent CNN
Epochs	50
Steps	10000

5.2. Fine-tuning model

5.2.1. Two pretrained model

In order to get better results and solve the overfitting problem, a fine-tuning model was used. I prepared two kinds of fine-tuning models. One of them is based on a pretrained COCO model and then trained on KITTI data. The COCO model is found from GitHub. Another one only uses KITTI training data to train a model. The accuracy of these two models are shown in Table 5-2. The COCO-KITTI model has a higher accuracy than KITTI model, so in this study, COCO-KITTI model was chosen as fine-tuning model to train four models mentioned above.

Table 5-2 The comparison of COCO_KITTI model and KITTI model

COCO_KITTI Model	KITTI Model
Pedestrian with average precision: 0.4436 Cyclist with average precision: 0.4563	Pedestrian with average precision:0.3432 Cyclist with average precision: 0.3910

5.2.2. Whether to use a fine-tuning model

It is considered that both COCO and KITTI dataset are colour images. And in this thesis, I will use thermal images to do human detection. The different data types may have a negative effect on the model. Because of this, I trained another model only using the KAIST thermal dataset. This means this is a network without fine-tuning model.

I used RetinaNet evaluation code to evaluation these two models. The result can be seen in Table 5-3.

Table 5-3 The evaluation results of COCO_KITTI model and KITTI model

COCO_KITTI_KAIST(thermal) Model	KAIST(thermal) Model
average precision: 0.0149	average precision: 0.0140

I used tensorboard to read the log files and got the loss information. The graph of loss, classification loss, and regression loss can be seen. Classification loss is used to show whether the classification of human beings is correct or not. Regression loss is used to evaluate the accuracy of the coordinates of the bounding box. Shown as Figure 5.2 (COCO, KITTI, and KAIST thermal model) and Figure 5.3 (KAIST thermal model).

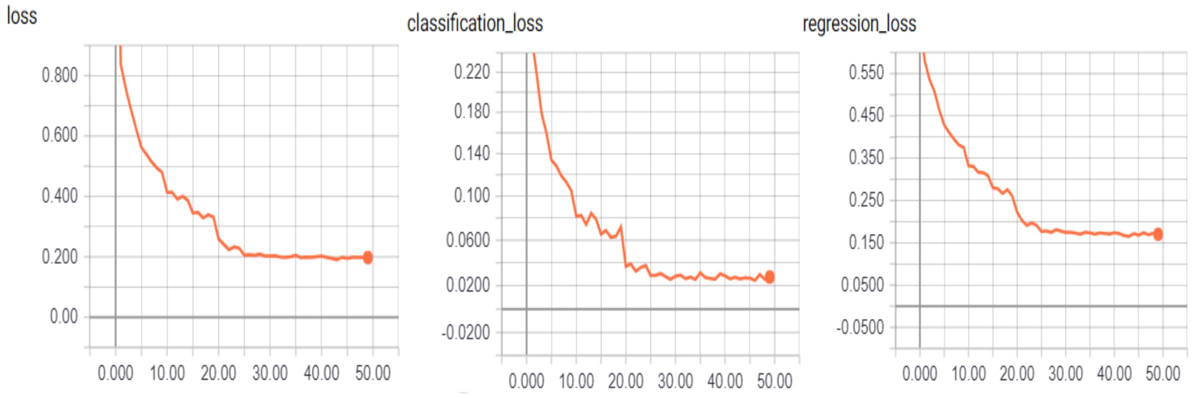


Figure 5.2 The loss(left) and classification_loss(middle) and regression_loss(right) of COCO, KITTI and KAIST thermal model

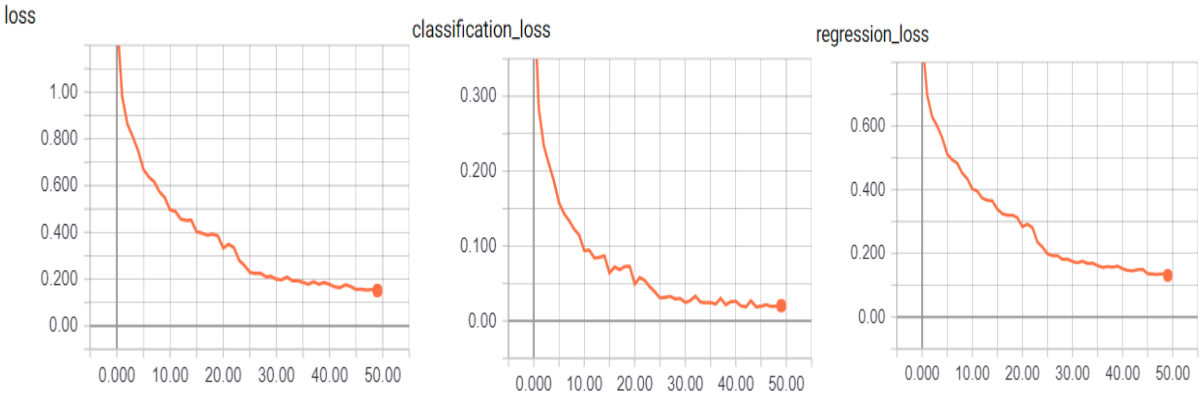


Figure 5.3 The loss(left) and classification_loss(middle) and regression_loss(right) of KAIST thermal model

The model with pretrained model can get a lower loss. These two models' regression losses are not satisfying. This may because this task is done on a challenging dataset. For this thesis, some errors occur in annotation files which lead to the bad regression losses.

We can see from these pictures that there are some fluctuations in the graph. This because the default batch size is small. The default value is one. Therefore, the models are easy to reach a local maximum or minimum. This problem can be solved by increasing batch size. But it's going to increase the computation as well. The requirements for hardware will also become more stringent.

I tested these two models using testing dataset. Here I use mAP to evaluate the results. I calculated the Mean IoU, precision and recall. The results are shown in Table 5-4.

Table 5-4 The testing results of two model

COCO_KITTI Model	KITTI Model
Mean IoU is 0.4556 Precision is 0.1085 recall is 0.0975	Mean IoU is 0.4494 Precision is 0.0904 recall is 0.0818

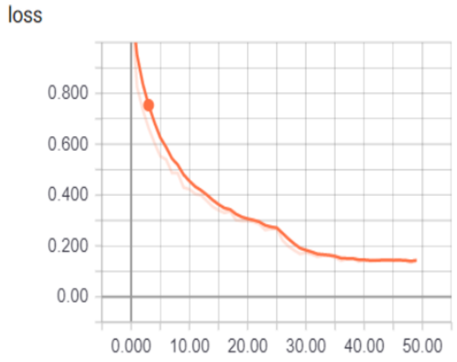
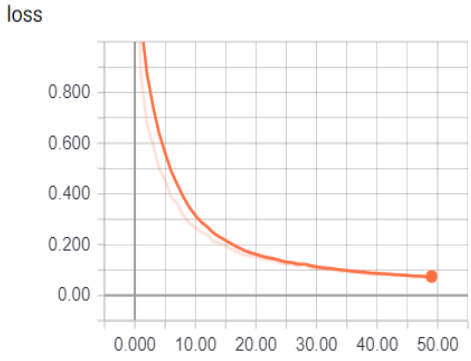
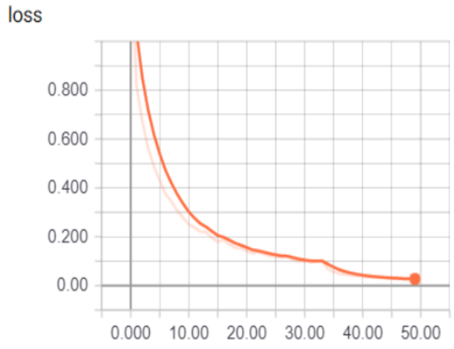
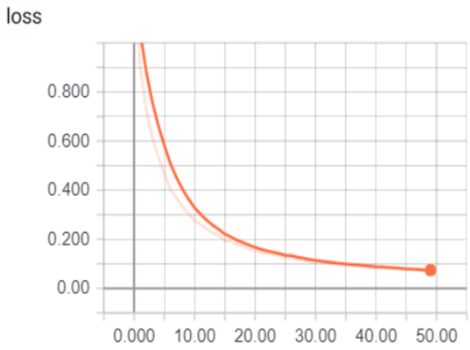
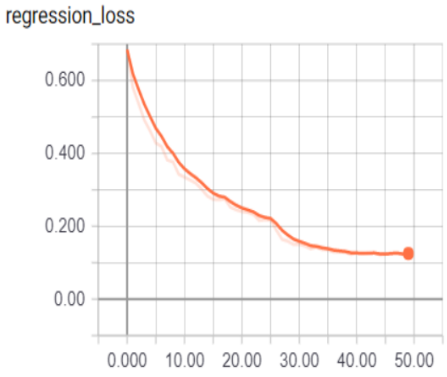
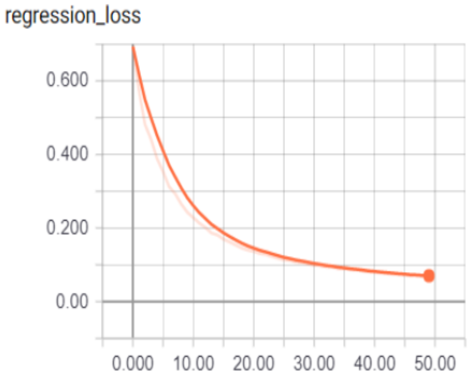
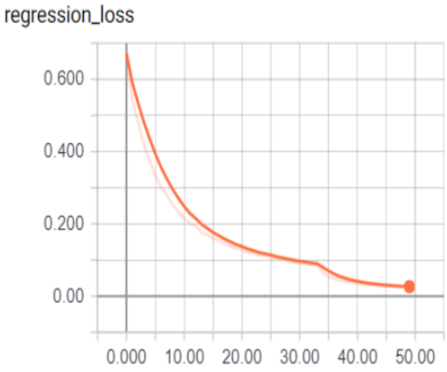
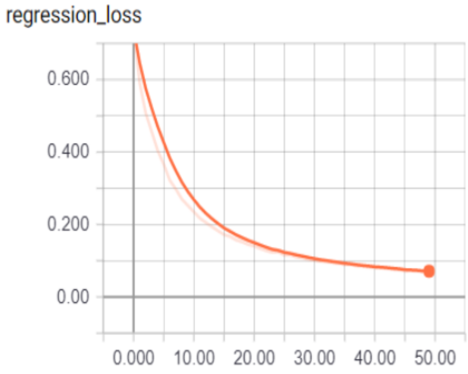
As can be seen from the above results, the COCO, KITTI and KAIST thermal model has a lower loss and a higher testing accuracy. Here, the testing dataset is the original dataset instead of the improved one. So the output is very disappointing. This shows that the fine-tuning model performs better in overfitting than the model without fine-tuning.

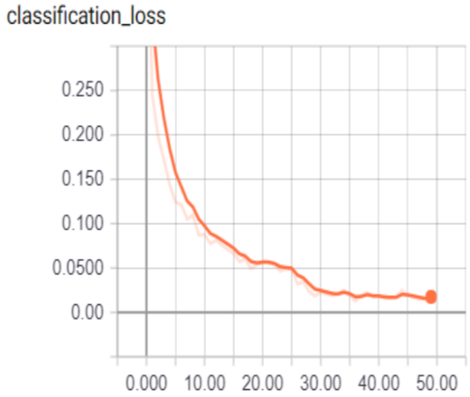
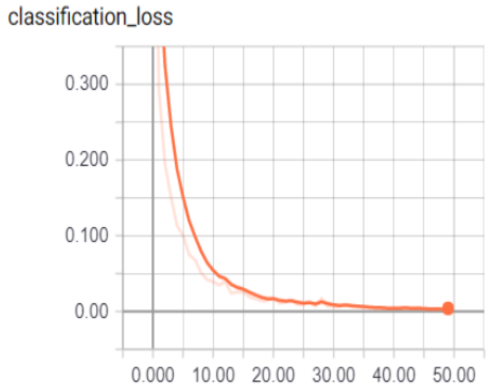
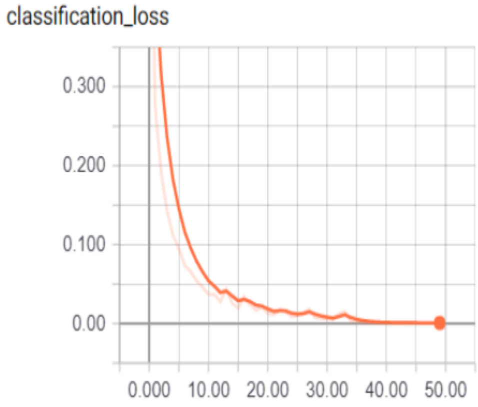
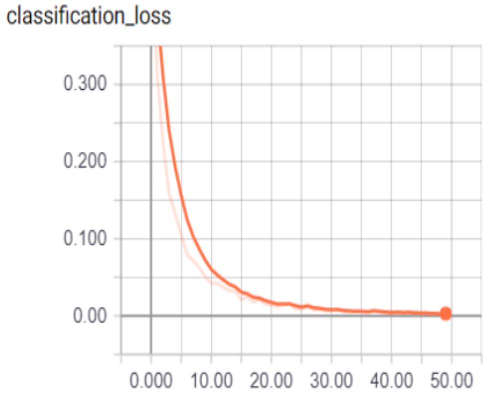
The COCO, KITTI and KAIST thermal model is chosen as fine-tuning model to train other four networks.

5.3. Validation results

The results of evaluation for these four models mentioned above are shown Table 5-5. Here I use four different curves given by tensorboard. There are graphs of loss, regression loss, and classification loss. By increasing the batch size, the curves are smoother than the models I got before. How much is the optimal value of batch size, is under the condition of the current hardware I cannot.

Table 5-5 The validation results of four models

Loss	Single images  	Continuous later 
	Continuous mid  	Discontinuous later 
Regression loss	Single images  	Continuous later 
	Continuous mid  	Continuous later 

Classification loss	Single images  	Continuous later 
	Continuous mid  	Discontinuous later 

6. RESULTS

This chapter is a demonstration and analysis of results. In this thesis, the IOU threshold is fixed to 0.7. The recall and precision of each score are represented on the graph (Figure 6.1). Precision decreases with the increase of recall. Go through each point and draw two lines parallel to the axis individually. The area of the rectangle between these two lines and the axes can be used to compare which score threshold is the most appropriate. To achieve a balance between recall and precision, for each model score threshold with the largest product of recall and precision is considered as the “best”. The testing results of each model are shown in Table 6-1.

Table 6-1 The precision, recall and mean IOU on each score for four models

Single images Basic CNN									
Score threshold	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
precision	0,3123	0,3988	0,4421	0,4722	0,4983	0,5182	0,5370	0,5594	0,5991
recall	0,3512	0,3353	0,3194	0,3070	0,2996	0,2837	0,2712	0,2626	0,2490
mean IOU	0,5520	0,6089	0,6362	0,6552	0,6695	0,6813	0,6913	0,7057	0,7197
Continuous TCNN_later									
Score threshold	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
precision	0,4719	0,5016	0,5223	0,5358	0,5496	0,5646	0,5767	0,5877	0,6037
recall	0,3487	0,3380	0,3320	0,3262	0,3230	0,3189	0,3142	0,3073	0,2994
mean IOU	0,6419	0,6593	0,6692	0,6763	0,6832	0,6894	0,6954	0,7016	0,7089
Continuous TCNN_mid									
Score threshold	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
precision	0,4671	0,4867	0,4980	0,5102	0,5188	0,5268	0,5357	0,5451	0,5663
recall	0,3485	0,3398	0,3366	0,3350	0,3323	0,3285	0,3267	0,3184	0,3013
mean IOU	0,6414	0,6527	0,6589	0,6655	0,6690	0,6725	0,6804	0,6848	0,6933
Discontinuous TCNN_later									
Score threshold	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
precision	0,4771	0,5063	0,5214	0,5330	0,5430	0,5548	0,5778	0,5778	0,5849
recall	0,3487	0,3100	0,3053	0,2967	0,2938	0,2918	0,2909	0,2870	0,2750
mean IOU	0,6489	0,6661	0,6737	0,6790	0,6836	0,6885	0,6946	0,6982	0,7027

6.1. Single images model against temporally-consistent model

The results show that the convolutional neural networks which added temporal information all have a better performance than the neural network trained using single images (as shown in Figure 6.1). The biggest point in each line shows the best performance with the maximum rectangle area of each model (blue dot at 0.5, green and red dot at 0.7, yellow dot at 0.8). The recall-precision lines of temporal CNN (green, yellow and red) are obviously above the recall-precision line of single images neural network. This means that the accuracy of human detection using temporally-consistent network has been greatly improved.

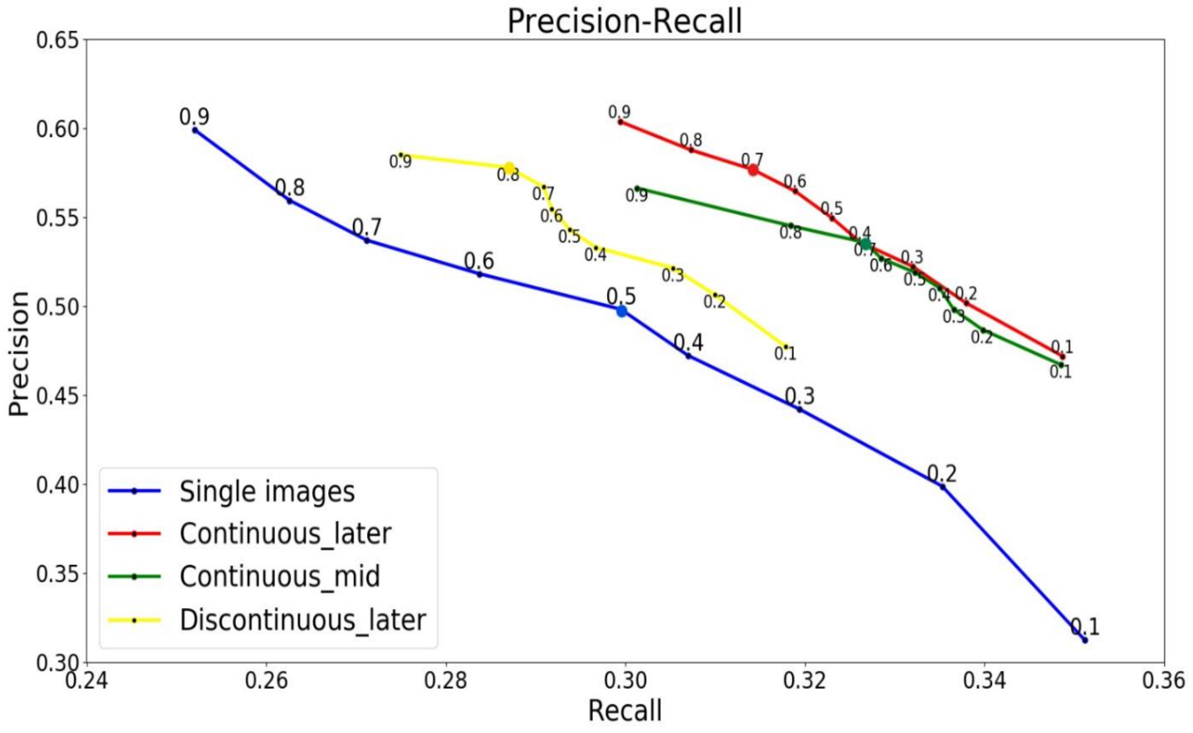


Figure 6.1 Precision-recall graphs of 4 models

The areas of rectangle of each model on every score threshold can be seen in Table 6-2. The highlights are the maximum value of precision and recall products of each model. The corresponding score is considered as the "best" score threshold. By comparing the "best" results of these models, we can find that continuous later neural network is better than the other three models. The largest area of the rectangle (score threshold on 0.7 continuous later network) is 21.4 percent larger than the smallest (score threshold on 0.5 single images network).

According to this table, the "best" performance of the temporally-consistent network will occur with a higher score threshold. This also means that the quality of the detected human is also higher than that of the single images network. A prediction of the bounding box of the higher the score, then we will have more confidence to say this object can be considered as a human. We can draw a conclusion that the network which added temporal information has a great improvement on the accuracy of human detection than the network only trained with single images.

Table 6-2 Precision*recall of four models on different score threshold

Precision * Recall	Score								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Single images	0.1097	0.1337	0.1412	0.1450	0.1493	0.1470	0.1456	0.1467	0.1491
Continuous TCNN_later	0.1646	0.1695	0.1734	0.1748	0.1775	0.1800	0.1812	0.1807	0.1806
Continuous TCNN_mid	0.1628	0.1654	0.1676	0.1709	0.1724	0.1731	0.1750	0.1736	0.1706
Discontinuous TCNN_later	0,1517	0,1570	0,1592	0,1581	0,1593	0,1619	0,1649	0,1658	0,1608

In order to add temporal information into the neural network, two methods of stacking images have been come up with. These two methods have been mentioned in section 5.1.

By comparing these two precision-recall lines, as seen in Figure 6.2, the network which uses later-core stacked images performances better than the network which uses mid-core stacked images. The red line is basically floating above the green line, except for the very little bit of overlap. It can be seen from Table 6-2 that the "best" performance of the two models are all with the score threshold equal to 0.7. When it comes to the maximum area of the rectangle, the later-core images model is three percent larger than mid-core images model.

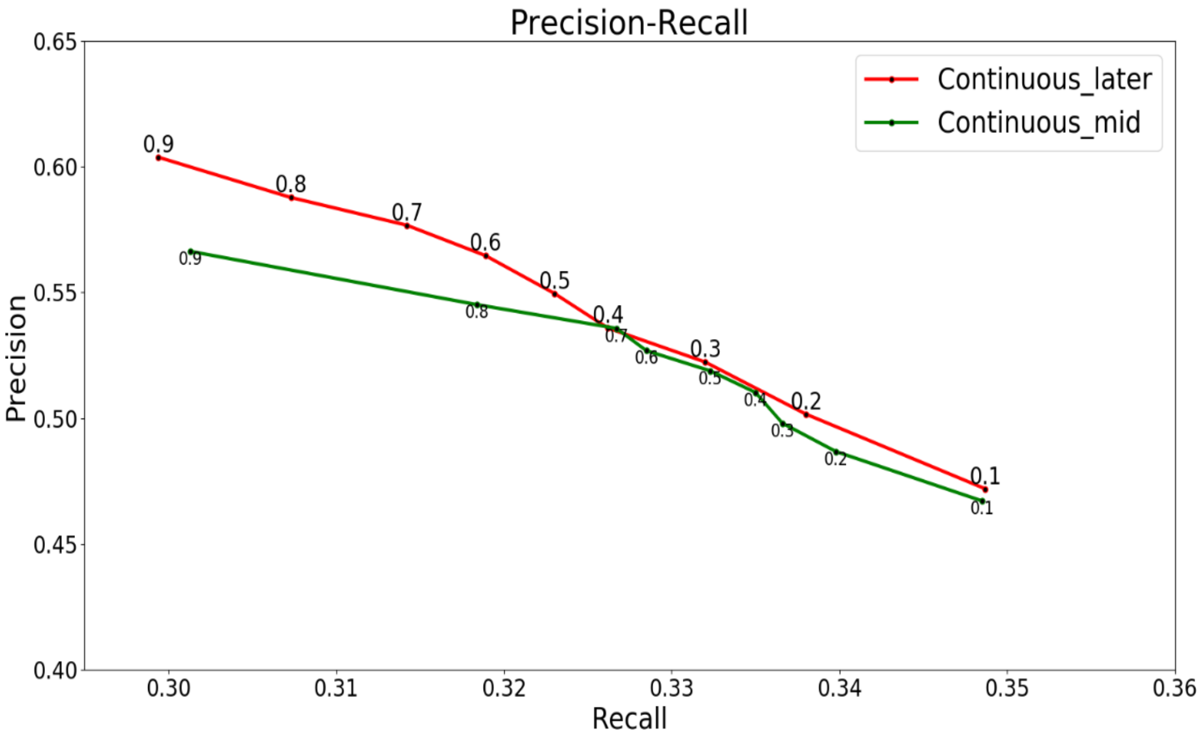


Figure 6.2 Precision-recall graphs of continuous later(red) network and continuous mid network(green)

This may because other things being equal, the network with a later frame as the core can use more information than the network with a middle frame as the core. This can be seen more intuitively in Figure 6.3. The yellow line is the time information between the two frames.

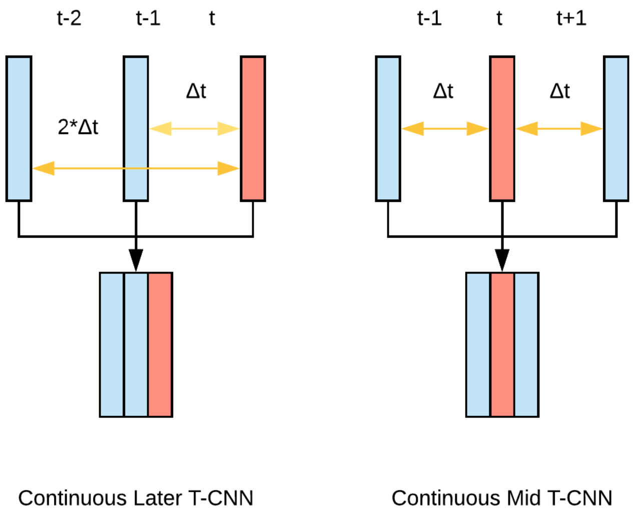


Figure 6.3 The temporal information contained in Continuous Later T-CNN (left) and Continuous Mid T-CNN (right). (The red rectangle is key frame and the blue rectangles are attached frame). The yellow line is the time information between the two frames.

In the continuous later T-CNN (left), the time information contained in an image is 3* Δt in total. But for continuous mid T-CNN (right), the temporal information one image has is only 2* Δt, which is less than the previous one. Here, Δt represents the time interval between two frames who are next to each other (Δt = 0.1s).

On the other hand, the way data is obtained also has an effect on this phenomenon. The camera keeps moving forward as the photo is taken. So that nearly all persons in the photograph appear in a fixed manner.

The camera is moving away from humans, and the distance between them is getting smaller and smaller. Eventually the person moves behind the camera and disappears into the image. The size of the human contour is slowly becoming larger and then suddenly disappear. This leads to a situation where the middle frame is the core and the person cannot be detected in the next frame. This will also lead to reduced precision of the model.

6.2. Continuous later model against discontinuous later model

The precision-recall graphs of these two models are shown in Figure 6.4. The red line as a whole in the top right of the yellow line. It can be seen from this that the model of training with continuous stacked image has a better performance than the model which trained while skipping several frames.
This is because after skipping servals frames, a lot of computation is reduced, the human motion information between each frame of the image will be more obvious. But there are downsides to this approach. The temporal information in the omitted frame is also removed.
In this thesis, all images are captured by the vehicle camera. The vehicle is moving continuously while collecting images. This makes the speed of relative motion of the person and the camera is high. In one second's time interval, the person will step out of the range of the camera.
Such a situation may arise. At frame t1, a person stands at point A. Then he walks to point B during the time interval between frame t1 and frame t2. Then he returns to point A at frame t3. If frame t2 has been removed, the movement information of the person will also lose. For this reason, the discontinuous later network has a lower accuracy of human detection.

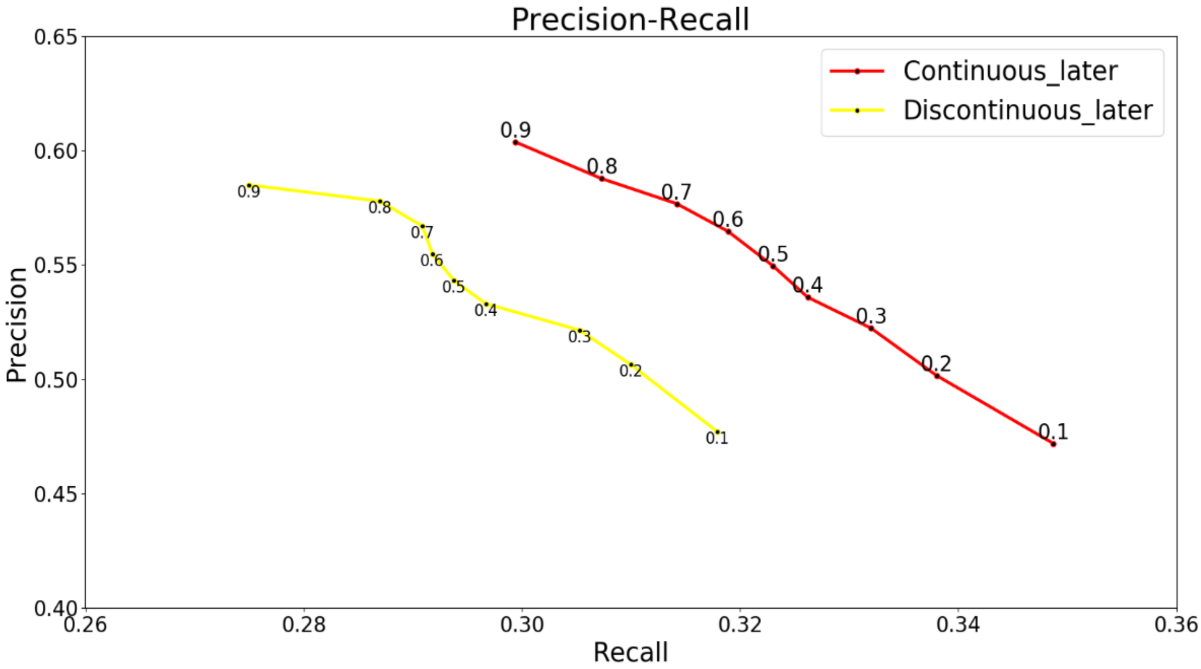


Figure 6.4 Precision-recall graphs of continuous later(red) network and discontinuous later network(yellow)

6.3. F1 score

The results for F1 Score is shown in Table 6-3. The highlight is the highest F1 score value among this model. We can see that score threshold with the highest F1 score is different from the score threshold with highest precision*recall value. However, the F1score result still indicates that the temporally-consistent network performs better than the basic network in human detection.

The F1 score-score threshold graphs of four models are shown in Figure 6.5. From this figure, we can see that F1 score expresses the same information as the precision-recall diagram. All lines for temporal CNNs are above the line of basic CNN. The continuous later model is the best one.

Recall reflects the model's ability to detect positive samples. The higher the recall is, the better the model can recognize positive samples. Precision reflects the ability to tell negative samples apart. F1-score is a combination of these two criteria. The higher the F1-score is, the more robust the model is.

Table 6-3 F1 score of four models on different score threshold

F1 Score	Score Threshold								
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
Single images	0,3306	0,3643	0,3709	0,3721	0,3742	0,3667	0,3604	0,3574	0,3518
Continuous TCNN_later	0,3992	0,4002	0,4017	0,4044	0,4051	0,4047	0,4059	0,402	0,3933
Continuous TCNN_mid	0,4011	0,4039	0,4060	0,4055	0,4069	0,4076	0,4068	0,4036	0,4003
Discontinuous TCNN_later	0,3816	0,3845	0,3851	0,3812	0,3813	0,3824	0,3845	0,3835	0,3741

It can be seen from this graph that all graphs for temporally-consistent CNN are nearly parallel to the horizontal axis. This indicates that the networks containing time information play a positive role in the stability of the human detection model under the extreme value of score threshold that is too large or too small.

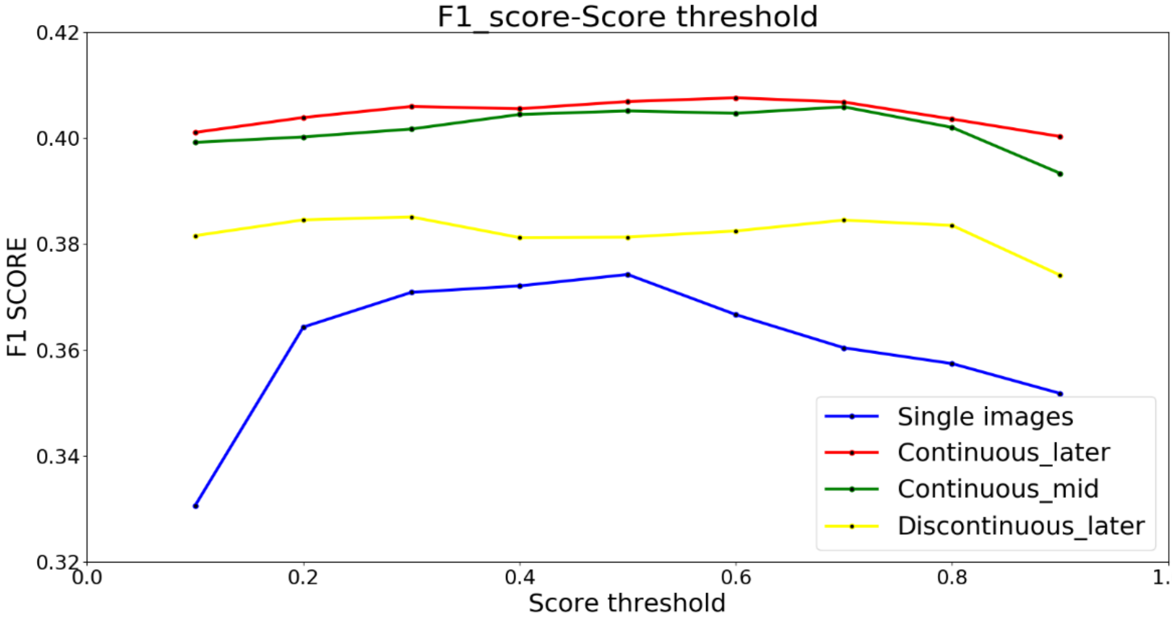


Figure 6.5 The F1 score- score threshold graph of four models.

6.4. Average IOU

Each score threshold has a corresponding average IOU. Figure 6.6 shows the relationship between average IOU and corresponding score threshold for each model.
The average IOU will grow with the score threshold. This is because, with the higher the score threshold, we can become more confident to say the detected human is correct. As you can see from this graph, the basic model grew faster than the others. This also shows that the networks with temporal information are more stable than the ordinary network.

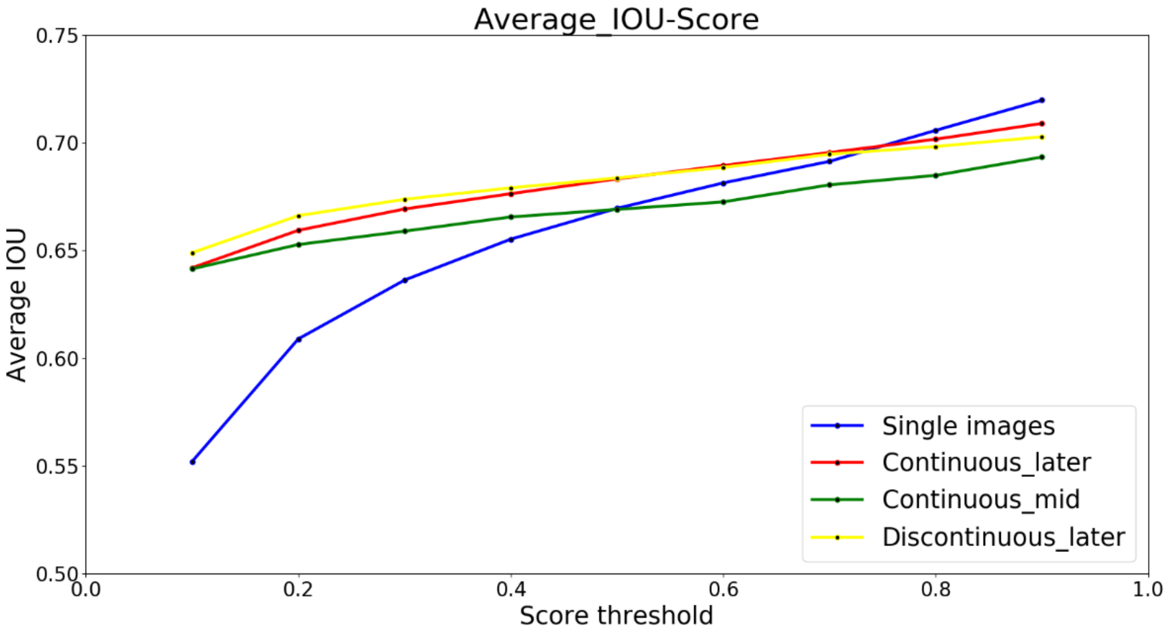


Figure 6.6 Average IOU-Score threshold graphs of 4 models

Table 6-4 shows the value of precision, recall and average IOU of the "best" score of each model.

Table 6-4 The best performance of each model

Name of a model	IOU threshold	Score threshold	Recall of the best performance	Precision of the best performance	Average IOU
Single images	0.7	0.5	0.2995	0.4982	0.6695
Continuous mid	0.7	0.7	0.3267	0.5357	0.6804
Continuous later	0.7	0.7	0.3142	0.5767	0.6954
Discontinuous later	0.7	0.8	0.2870	0.5778	0.6982

7. DISCUSSION

This chapter discusses the advantages and disadvantages of this study. This discussion focuses on the methodology, hardware, and dataset.

In this study, I successfully proved that the convolutional neural network with temporal information can greatly improve the accuracy of human detection. By comparing different temporally-consistent networks, the best way to add time information to the convolutional neural network is obtained.

7.1. Methodology

In this thesis, temporal information has been successfully added to convolutional neural networks. And it proves that temporal information has a positive effect on human detection. Because the human detection accuracy of all temporally-consistent networks is higher than that of single images networks. This also achieved the expected goal of the study.

Since RetinaNet network can only process the images of three channels, only three images are selected to be stacked in the temporal convolutional neural network training task. We cannot study whether human detection accuracy will be improved if four or more frames are processed at the same time. The more frames are processed at the same time, the richer the time information will be. At the same time, the amount of calculation will be greatly increased. Unfortunately, the relationship between these two variables was not found in this study.

In a discontinuous temporal convolutional neural network, ten images were skipped. This makes the time interval between the remaining two frames to be 1 second. Depending on the number of frames removed, the results will vary. How to balance the quantity of computation with the quality of human detection is a problem for future research.

In this thesis, we do not attempt to combine the optical flow and images to train the model. It cannot be determined which method is better. In future studies, more ways to incorporate time information will need to be developed. Make the time efficiency is higher.

7.2. Hardware

Some parameters cannot be adjusted due to hardware limitations.

Normally, it takes about one week to train a RetinaNet model which has 50epochs, and each epoch has 10000steps by ThinkPad P51. However, due to the limitations of time and hardware devices, many parameters cannot be adjusted. An optimal model may not be obtained.

The overfitting problem can be solved by adding batch size. However, increasing batch size will greatly increase the amount of calculation. This increases the time required to train the model. When the batch size is larger than 4, my supervisor's computer which equipped with GPU will also crash down.

The limitation of hardware equipment affects not only the time of model training but also the quality of the model.

7.3. Dataset

The dataset used in this thesis is from KAIST (Korea Advanced Institute of Science and Technology) Multispectral Pedestrian Detection Benchmark. Unfortunately, there is a lot of disappointment in this dataset. Either the images themselves or the annotation files that correspond to them. First of all, the

resolution of the whole dataset is very low, and the contour of human body is not clear. This will greatly affect the accuracy of human detection. In other words, this is a very challenging dataset.
All examples given here are from the "best" model, the continuous later model with score threshold equal to 0.7.

7.3.1. Occlusion

Occlusion can be serious in areas where pedestrians are concentrated. In this case, the geometric features of the human body would not be obvious. This may have some bad implications for computer recognition.

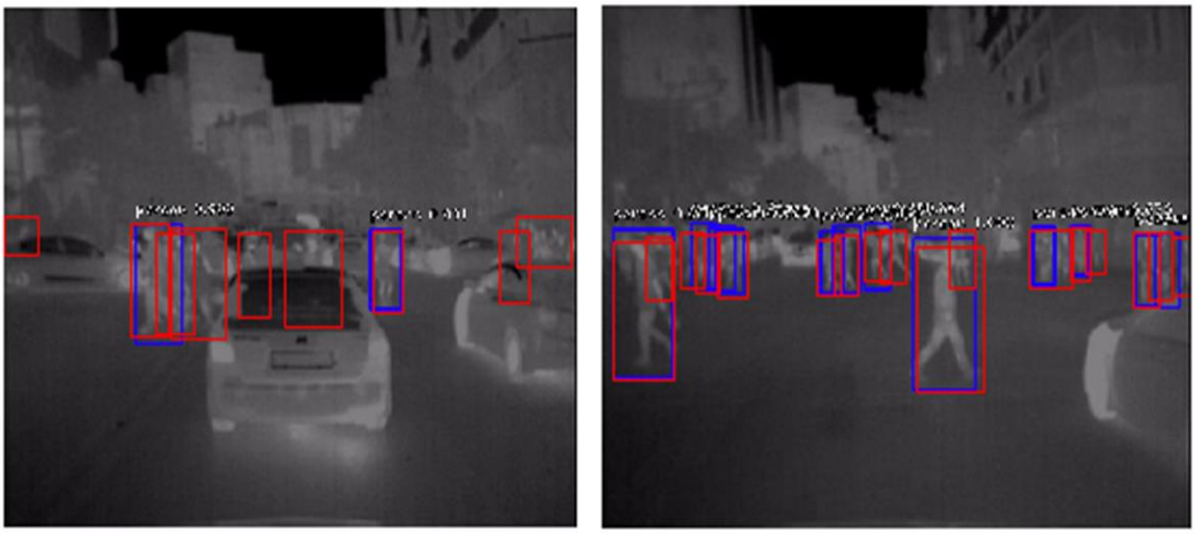


Figure 7.1 The left image is from Set08 V002 I00459.jpg and the right image is from Set08 V000 I02739.jpg

Figure 7.1(left) shows that people who are occluded by the cars were failed to be detected. On a crowded street, many people were obscured by other people, shown in Figure 7.1(right). Such phenomena can have an impact on the accuracy of human detection.

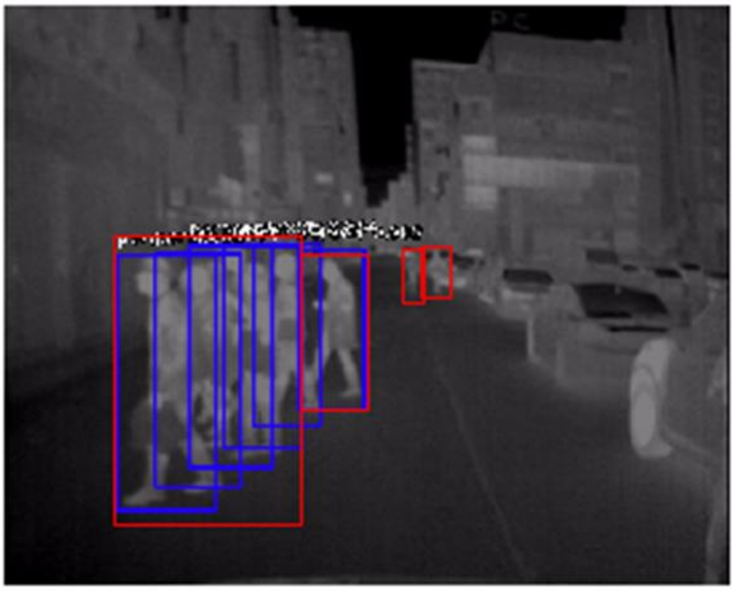


Figure 7.2 The image is from Set08 V000 I01939.jpg

In the face of such situation models also have a good performance. Figure 7.2 shows the model nearly detect all people even though there's a lot of overlap between them.

7.3.2. Distance

The models do not perform well for people who are far away from the camera. Because people are very small and blurry. It is difficult to tell them apart even by visual interpretation. Thus, this leads to low recall values of all models. From Figure 7.3 we can see that the people who stand far from the camera are hardly detected.

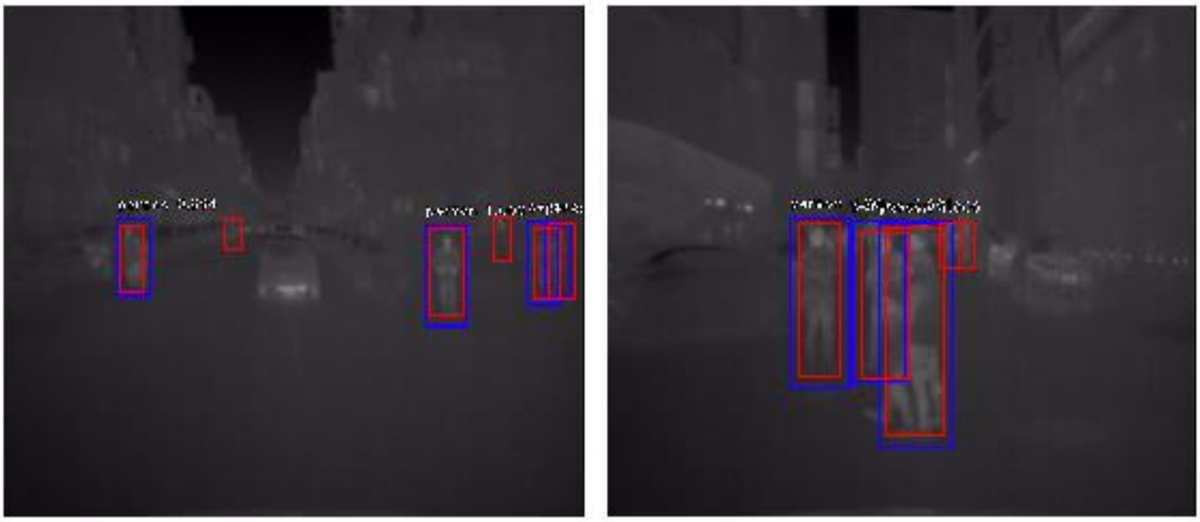


Figure 7.3 The left image is from Set11 V001 I00159.jpg and the right image is from Set11 V000 I01359.jpg

7.3.3. Annotations

Some errors occur on annotations. Most of the testing dataset suffered from this problem. Similar errors were found in training dataset. The most prominent problem is that the annotation files do not match the ground truth.

Even when we found the improved annotation files from Ms. Liu, the problem has not been completely solved. Because Liu's improved annotation files contain only one-twentieth of the original testing dataset. There was no improvement in training dataset. In addition, there are still many uncertain or wrong annotations in the improved dataset.

But the good thing is, even with all these problems. I still validated the proposed research objectives. Temporal information has a good effect on the accuracy of human detection.

7.3.4. Four major types of annotations.

The blue rectangle is the model detected people. the red rectangle is annotations. The yellow rectangle is used to highlight the person (Based on my visual interpretation.).

- Problem 1: Here is a person but annotation doesn't give the bounding box (Figure 7.4).

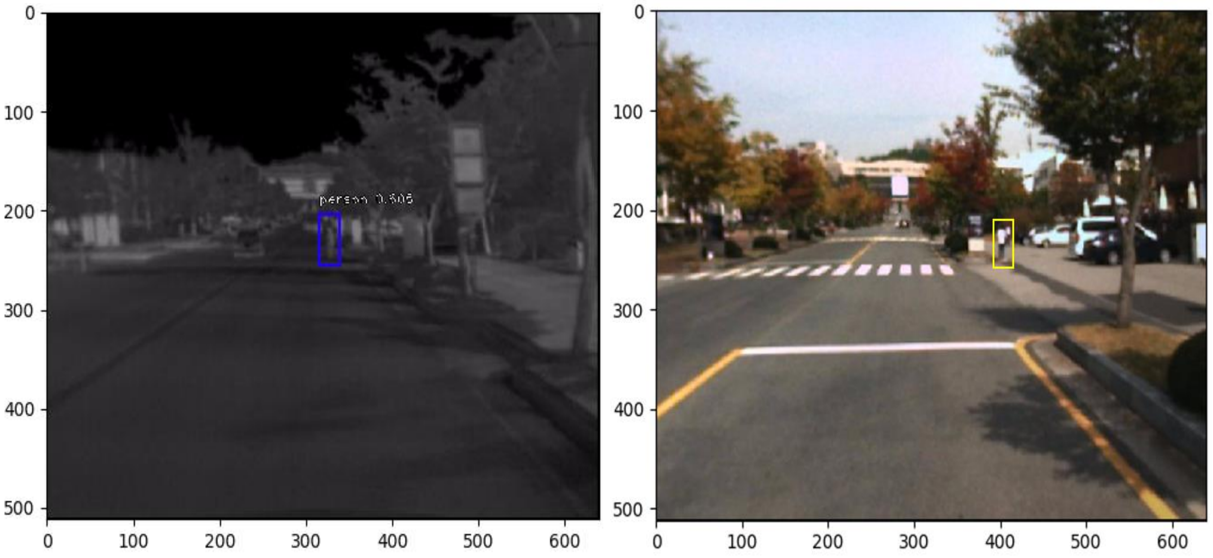


Figure 7.4 The pare of images is from Set06 V000 I00514.jpg

- Problem 2: There is no person, but annotations give a bounding box (Figure 7.5)

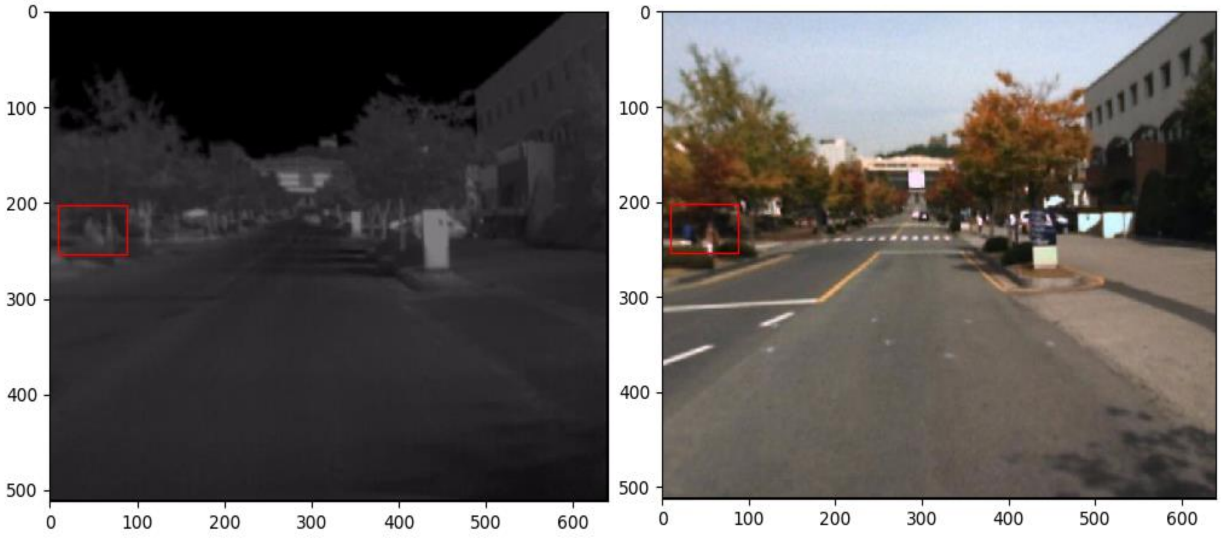


Figure 7.5 The pare of images is from Set06 V000 I00631.jpg

- Problem 3 The annotations give a bounding box which contains a group of people rather than a single person. This phenomenon is shown in Figure 7.6.

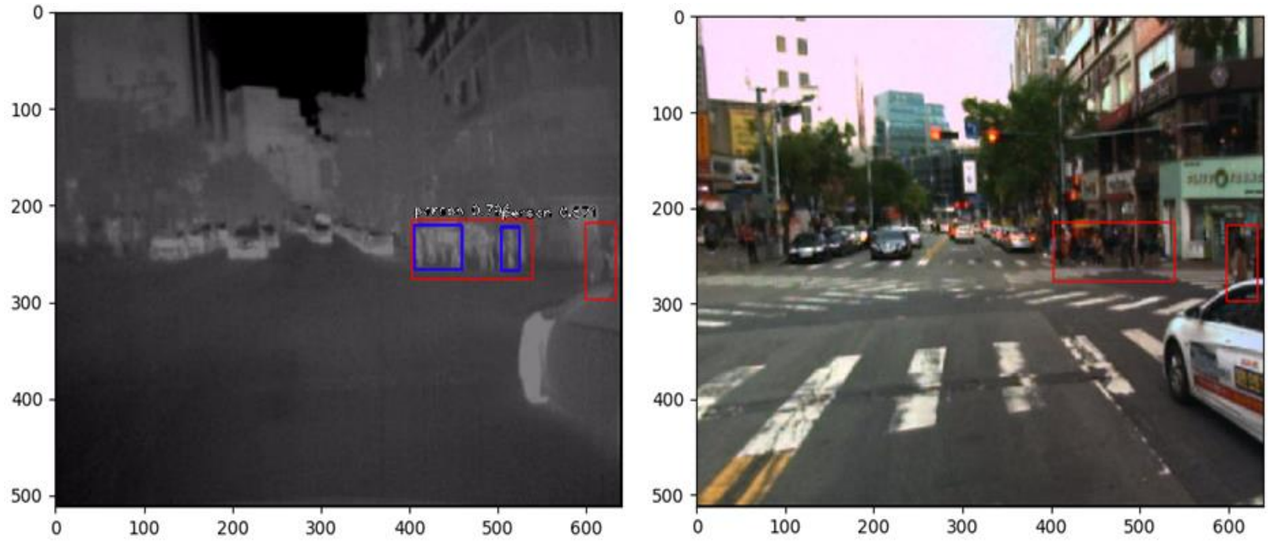


Figure 7.6 The pare of images is from Set08 V001 I01959.jpg

- Problem 4 Sometimes the annotations give a bounding box larger than the real person (the yellow one) shown in Figure 7.7

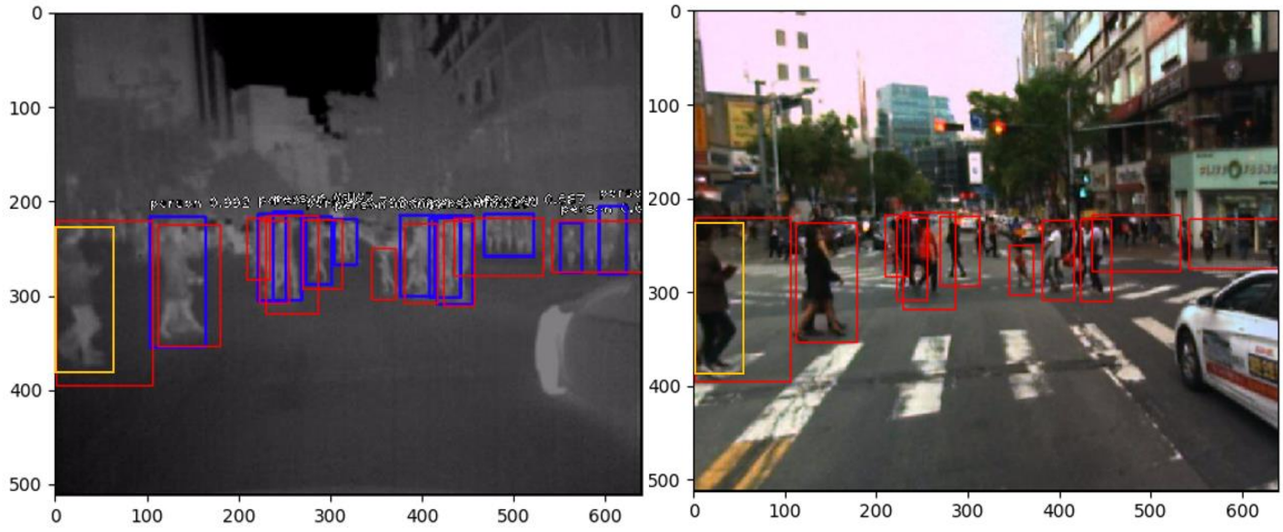


Figure 7.7 The pare of images is from Set08 V001 I02498.jpg

8. CONCLUSION AND RECOMMENDATION

8.1. Conclusion

In this thesis, we discussed the role of temporal consistency in human detection. This study trained three different types of temporally-consistency convolutional neural networks. By comparing the testing results of different models, we know that the human detection accuracy of the convolutional neural network with temporal information is better than that of the basic convolutional neural network. The accuracy of the three models continuous later model, continuous mid model and discontinuous for human detection is 21.4%,17.2%, and 11.1% higher than that of the basic model, respectively.

By analyzing different temporally-consistency convolutional neural networks, the following results can be obtained:

- The model that trained with continuous frames is better than models that skip some frames between two adjacent frames.
- In the continuous T-CNN model, the model with the later frame as the core is better than the model with the middle frame as the core.

8.2. Answers to research questions

1. What is the state of art in human detection? How accurate is the state of art?
   The current state of art in object detection is RetinaNet. RetinaNet has a special loss function to solve imbalance between foreground and background classes. Among all the deep learning approaches used in human detection, RetinaNet gives the highest accuracy. According to the results, the basic CNN achieve its "best" performance with score threshold is equal to 0.5. The precision, recall, and average IOU is 0.4982, 0.2995 and 0.6695 respectively under the circumstance.

2. How much temporal CNN can improve human detection using the state of art?
   Compared with the basic CNN network, temporal CNN has greatly improved the accuracy of human detection. The accuracy of the three models: continuous later model, continuous mid model and discontinuous for human detection is 21.4%,17.2%, and 11.1% higher than that of the basic model, respectively (based on precision*recall).

3. What is the best temporal CNN architecture and how much it improves the state of the art?
   In this thesis, the best temporal CNN architecture is continuous later-core network. The accuracy of human detection with this model is 21.4 percent higher than basic model (based on precision*recall). The model has the "best" result on 0.7 score threshold. The precision, recall, and average IOU is 0.5767, 0.3142 and 0.6954 respectively.

8.3. Recommendation

- For the following research, the research results can be more convincing by adjusting the parameters. For example, the number of frames skipped between two frames, and the number of frames processed together. As well as the value of epoch, steps and batch size.
- Other neural networks such as RNN can also be used to add time information for human detection.

- Optical flow and images can be processed together for human detection in the future. To achieve higher accuracy and faster speed.
- Human detection can be done by using a better-quality dataset, or based on cognitive semantics.

LIST OF REFERENCE

REFERENCES

Baek, J., Hong, S., Kim, J., & Kim, E. (2017). Efficient Pedestrian Detection at Nighttime Using a Thermal Camera. *Sensors*, *17*(8), 1850. https://doi.org/10.3390/s17081850

Ballas, N., Larochelle, H., & Courville, A. (2015). Describing Videos by Exploiting Temporal Structure, 4507–4515. https://doi.org/10.1109/ICCV.2015.512

Correa, M., Hermosilla, G., Verschae, R., & Ruiz-del-Solar, J. (2012). Human Detection and Identification by Robots Using Thermal and Visual Information in Domestic Environments. *Journal of Intelligent & Robotic Systems*, *66*(1–2), 223–243. https://doi.org/10.1007/s10846-011-9612-2

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, *I*, 886–893. https://doi.org/10.1109/CVPR.2005.177

Doherty, P., & Rudol, P. (2007). A UAV Search and Rescue Scenario with Human Body Detection and Geolocalization. In *AI 2007: Advances in Artificial Intelligence* (pp. 1–13). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-76928-6_1

Dollar, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast Feature Pyramids for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(8), 1532–1545. https://doi.org/10.1109/TPAMI.2014.2300479

Fan, X., Xu, L., Zhang, X., & Chen, L. (2008). 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* (pp. 139–143). IEEE. https://doi.org/10.1109/ICNC.2008.315

Gajjar, V., Gurnani, A., & Khandhediya, Y. (2017). Human Detection and Tracking for Video Surveillance A Cognitive Science Approach. *ArXiv:1709.00726v1 [Cs]*, 2805–2809. https://doi.org/10.1109/ICCVW.2017.330

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? the KITTI vision benchmark suite. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074

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

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

Gowsikhaa, D., Abirami, S., & Baskaran, R. (2014). Automated human behavior analysis from surveillance videos: a survey. *Artificial Intelligence Review*, *42*(4), 747–765. https://doi.org/10.1007/s10462-012-9341-3

Gowsikhaa D, Manjunath, & Abirami S. (2012). Suspicious Human Activity Detection from Surveillance Videos. *(IJIDCS) International Journal on Internet and Distributed Computing Systems*, *2*(2), 141–149.

Guan, D., Cao, Y., Yang, J., Cao, Y., & Yang, M. Y. (2019). Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, *50*(November 2018), 148–157. https://doi.org/10.1016/j.inffus.2018.11.017

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

Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. *Image and Vision Computing*, *60*, 4–21. https://doi.org/10.1016/j.imavis.2017.01.010

Hosang, J., Omran, M., Benenson, R., & Schiele, B. (2015). Taking a Deeper Look at Pedestrians. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4073–4082.

Hwang, S., Park, J., Kim, N., Choi, Y., & Kweon, I. S. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07–12–June*, 1037–1045. https://doi.org/10.1109/CVPR.2015.7298706

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D Convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231. https://doi.org/10.1109/TPAMI.2012.59

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Li, F. F. (2014). Large-scale video classification with convolutional neural networks. *Proc. IEEE CVPR*, 1725–1732. https://doi.org/10.1109/CVPR.2014.223

Kim, J. H., Hong, H. G., & Park, K. R. (2017). Convolutional Neural Network-Based Human Detection in Nighttime Images Using Visible Light Camera Sensors. *Passaro VMN, Ed. Sensors (Basel, Switzerland)*, *17*(5), 1065. https://doi.org/10.3390/s17051065

König, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., & Teutsch, M. (n.d.). *Fully Convolutional Region Proposal Networks for Multispectral Person Detection*.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, J., Liang, X., Shen, S., Xu, T., Feng, J., & Yan, S. (2017). Scale-aware Fast R-CNN for Pedestrian Detection. *IEEE Transactions on Multimedia*, 1–10. https://doi.org/10.1109/TMM.2017.2759508

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

Moore, D. (2003). A real-world system for human motion detection and tracking.

Nam, W., Dollar, P., & Hee Han, J. (2014). Local Decorrelation for Improved Pedestrian Detection. *ARXIV*, 1–9.

Ng, J. Y. H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07–12–June*, 4694–4702. https://doi.org/10.1109/CVPR.2015.7299101

Paisitkriangkrai, S., Shen, C., & Hengel, A. Van Den. (2014). Strengthening the Effectiveness of Pedestrian Detection with Spatially Pooled Features, 546–561.

Powers, D. M. W. (2007). Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation.

Ramachandran, R., Rajeev, D. C., Krishnan, S. G., & Subathra, P. (2015). Deep learning – An overview. *International Journal of Applied Engineering Research*, *10*(10), 25433–25448. https://doi.org/10.1016/j.neunet.2014.09.003

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

Sasaki, Y. (2007). F-measure.pdf, 1–5.

Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos, 1–9. https://doi.org/10.1017/CBO9781107415324.004

Uijlings, J. R. ., Sande, K. E. . Van De, Sande, T., & Smeulders, A. W. M. (2012). Selective Search for

Object Recognition. *International Journal of Computer Vision*, *104*(2), 154–171. https://doi.org/10.1007/s11263-013-0620-5

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

Yang, B., Yan, J., Lei, Z., & Stan Z. Li. (2014). Aggregate Channel Features for Multi-view Face Detection. *International Joint Conference on Biometrics*.

Zhang, S., Bauckhage, C., & Cremers, A. B. (2014). Informed Haar-like Features Improve Pedestrian Detection. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 947–954.

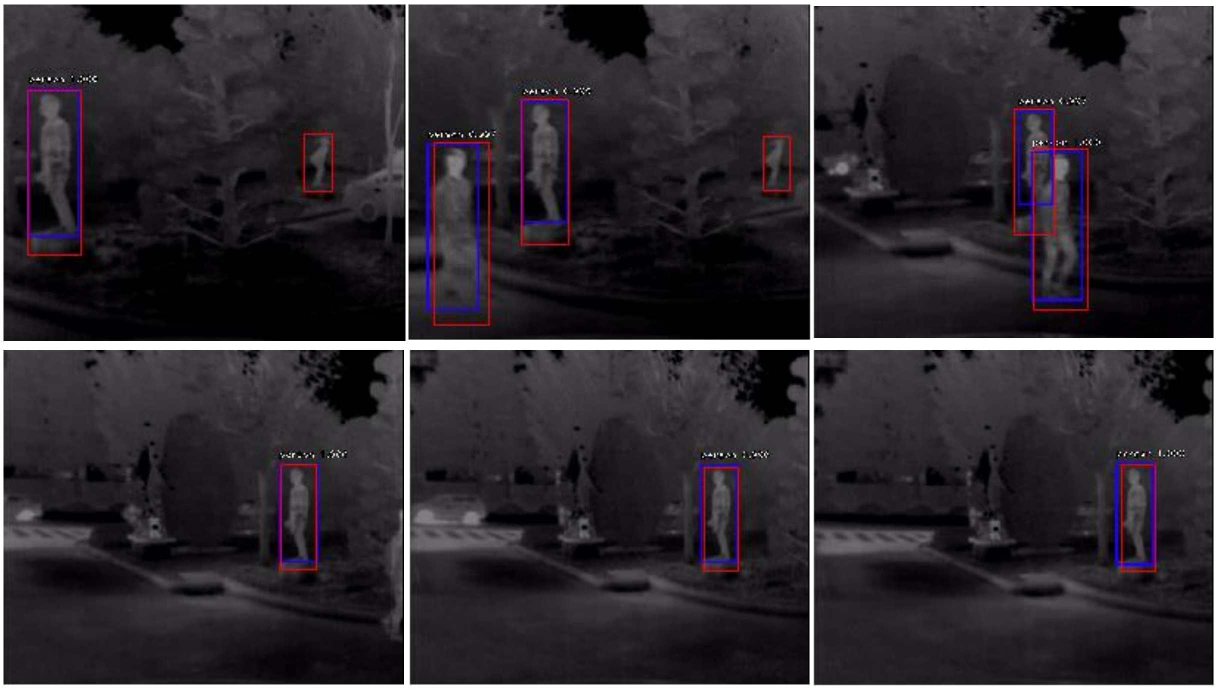
Zhou, S., Shen, W., Zeng, D., Fang, M., Wei, Y., & Zhang, Z. (2016). Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication*, *47*, 358–368. https://doi.org/10.1016/j.image.2016.06.007

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017, December). Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*. https://doi.org/10.1109/MGRS.2017.2762307

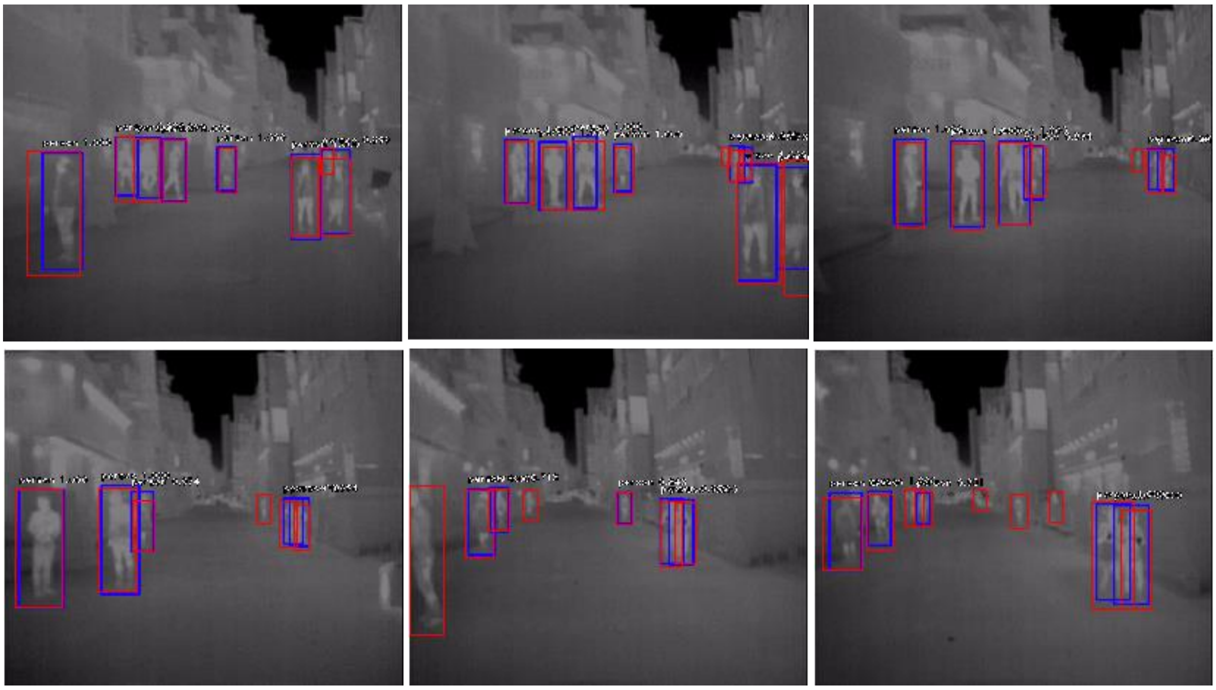
APPENDIX 1

Appendix 1 shows some successful stories about this thesis. All examples given here is from the "best" model, the continuous later model with score threshold equal to 0.7. In this chapter, the red rectangle is the bounding box given by annotation files. The blue rectangle shows the bounding box predicted by the model.

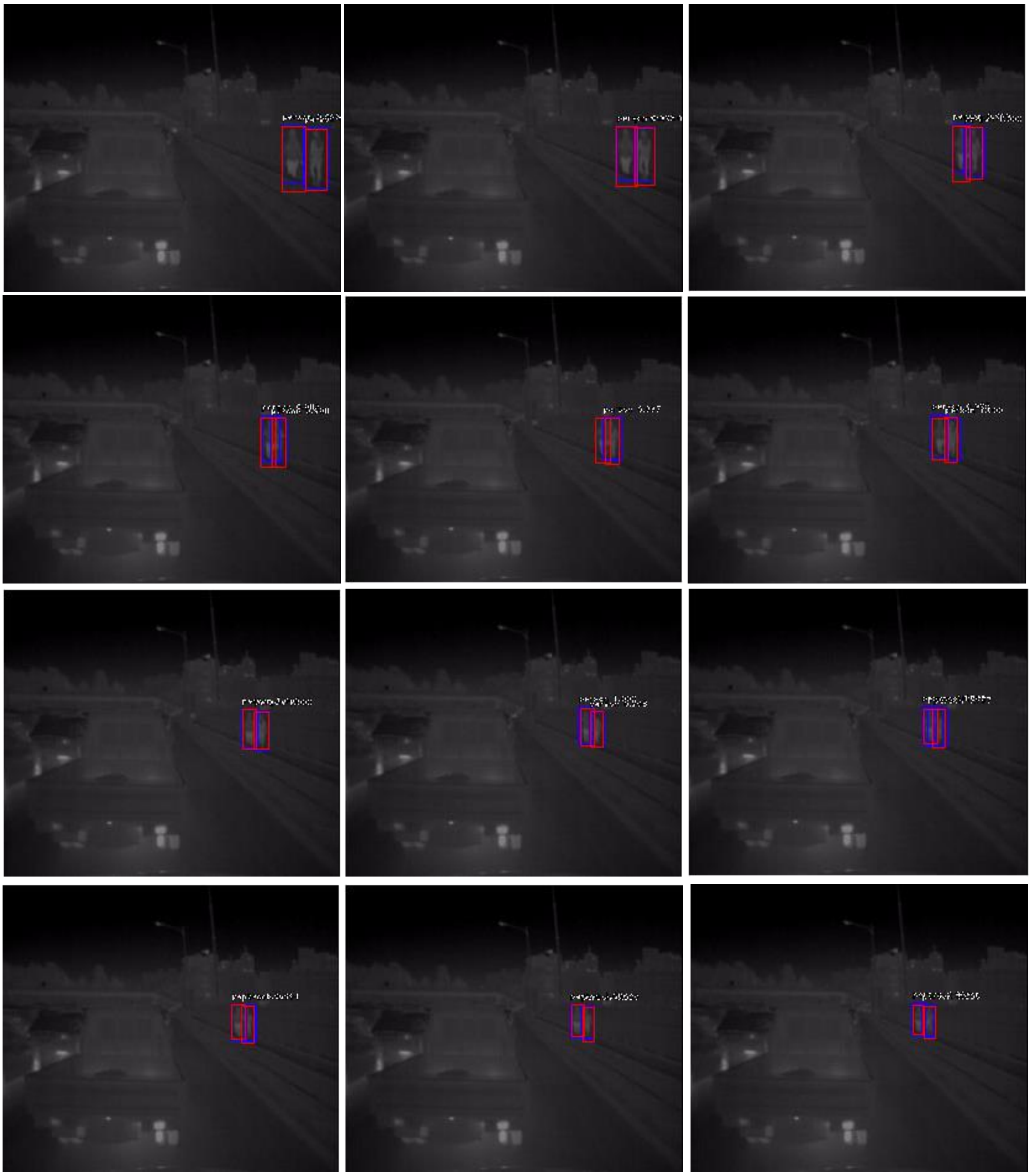
A series of test results which came from Set06 V004 (Day Campus).



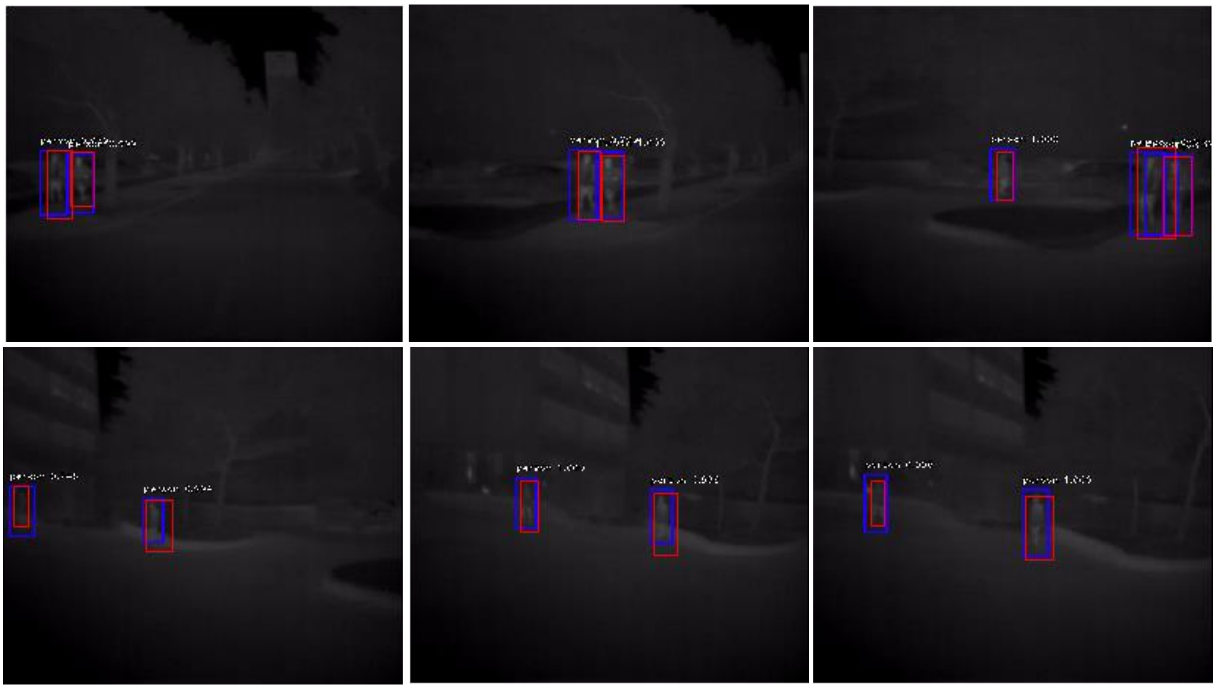
A series of test results which came from Set08 V004 (Day Downtown).



A series of test results which came from Set10 V000 (Night Road).



A series of test results which came from Set09 V000 (Night Campus).



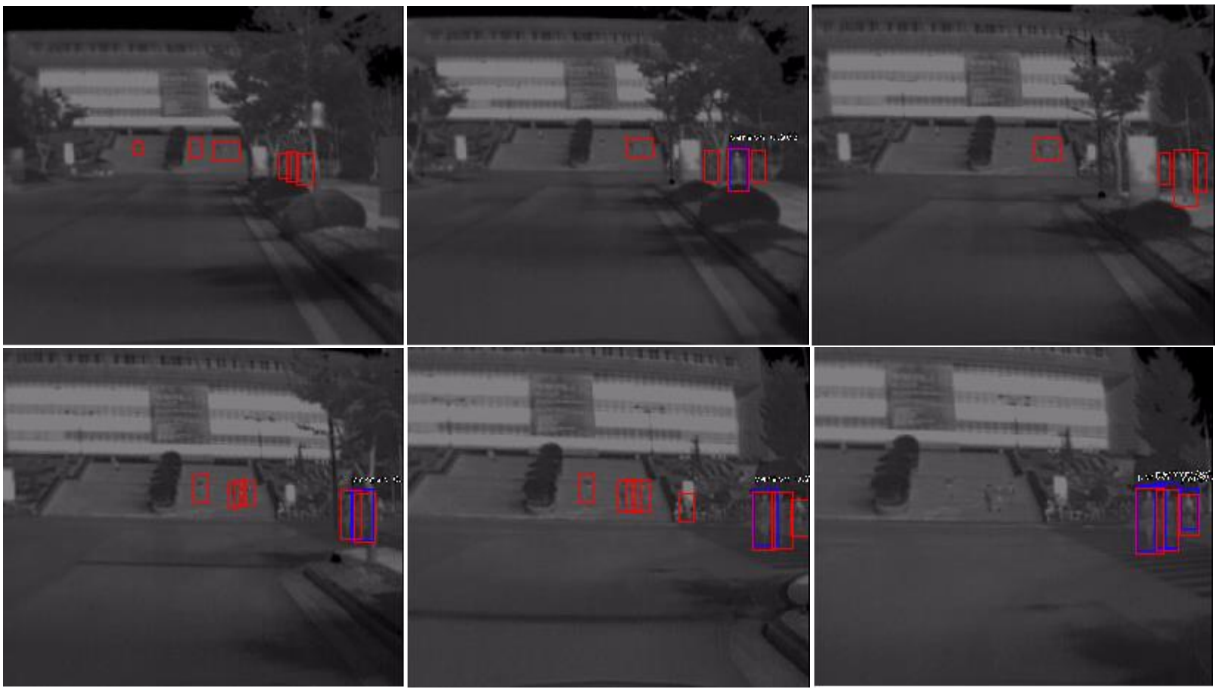
APPENDIX 2

Appendix 2 shows some failed cases about this thesis. All examples given here are from the “best” model, the continuous later model with the score threshold equal to 0.7. The red rectangle is the bounding box given by annotation files. The blue rectangle shows the bounding box predicted by the model.

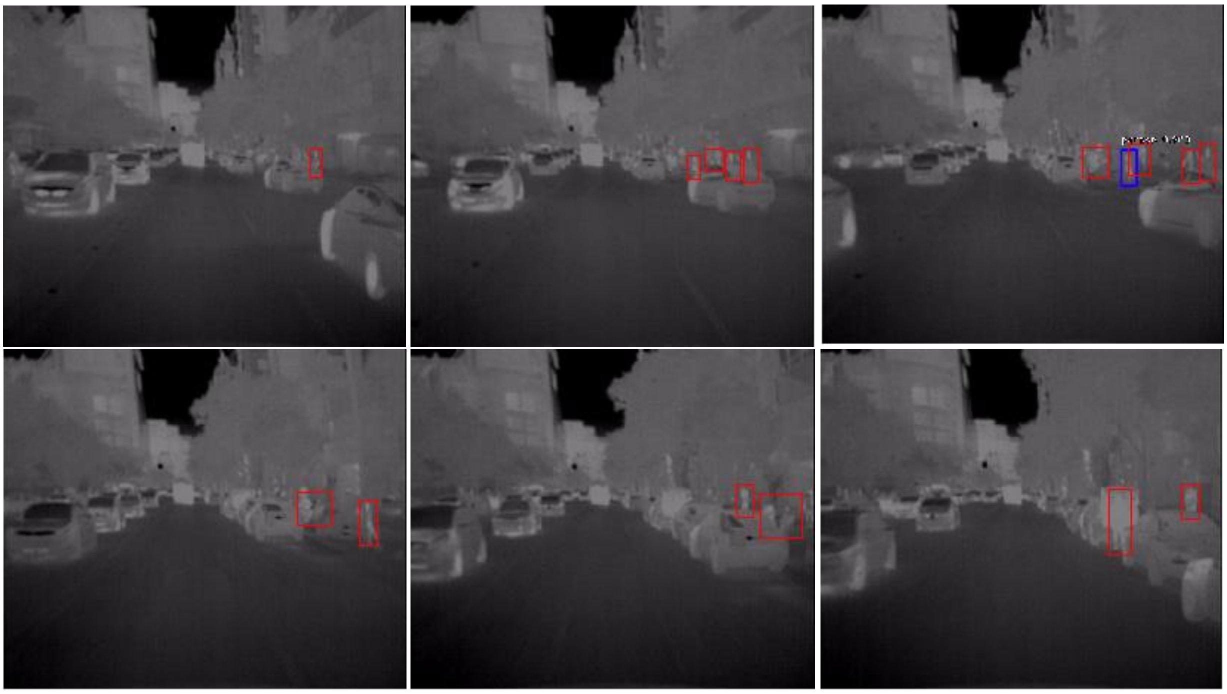
The first series of test results which came from Set06 V001 (Day Campus). The second series of test results which came from Set08 V004 (Day Downtown). The causes of failure of the two groups of pictures have much in common.

The people in these images are far away from the camera, and the thermal image resolution is very low. These two groups of images were taken during the day, when temperatures were higher, so the contour of the human body is not very clear. The people on the right side of the image is occluded by trees or cars. All these reasons resulting in poor detection results.

- Set06 V001 (Day Campus)

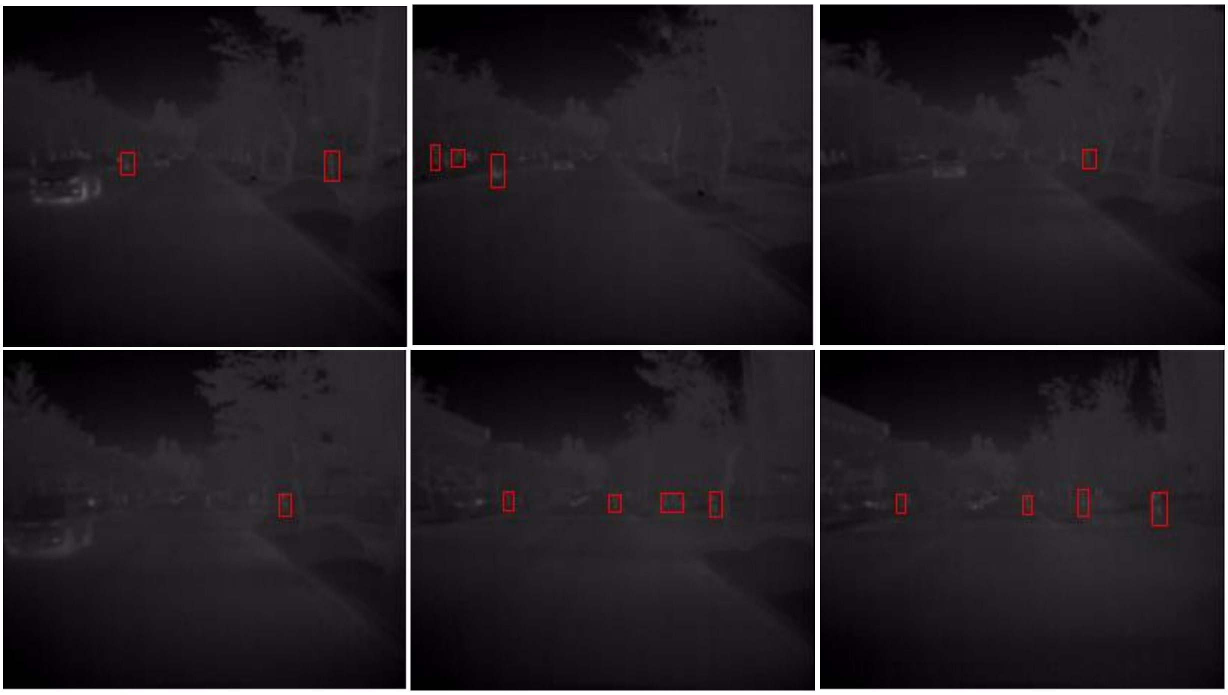


- Set08 V004 (Day Downtown)



This group of images is taken from Set09 V000 (Night Downtown). In these images, the distance between the person and the camera is large. It is difficult to identify human contour even visual interpreting.

- Set09 V000 (Night Downtown)



The series of images is taken at night in downtown. People are standing together and covering each other. The group of people in these images are all occluded by a car. Even the bounding boxes given by annotations (red rectangles) are not accurate marked the location of the people.

- Set11 V000 (Night Downtown)