The Integration of Thermal and Visual Images for Human Detection Using Deep Learning

GFM M.Sc. Research Proposal

Qiao REN

August 28, 2018

Contents

Figures 3

Tables 3

1 Motivation and Problem Statement 4

1.1 Motivation 4

1.2 Problem Statement 4

2. Research Identification 7

2.1 Research Objectives 7

2.2 Research Questions 7

2.3 Innovation Aimed At 7

2.4 Related Work 8

1) human detection using single sensor 8

2) human detection using multiple sensor 9

3. Project Set-up 11

3.1 Method Adopted 11

1) Model Training 12

2) Accuracy Evaluation 13

3.2 Plan of the project 15

3.3 Risks and contingencies 16

3.4. Expected Output 16

4. Preliminary Results 16

5. Resources Required 17

5.1 Information 17

5.2 Data 18

5.3 People 18

5.5 Finances 18

6. Supervision 18

6.1 Staff Already Consulted 18

6.2 Communication plan 18

7. Reference 19

Figures

Figure 1 (a) human with different skin colors (b) human with different genders and ages (c) fat and slim people (d) short and tall people (e) the front, back and side view of a people captured by a camera (f) different positions of human. (g) (h) (i) (j) show people's clothes and headwear are in different colors, styles and can even be exaggerated. 6
Figure 2 people who are occluded by other people 6
Figure 3 Image (a) and (b) are captured in daytime on the same scene(Hwang et al., 2015). Image (c) and (d) are captured at night time on the same scene(Hwang et al., 2015). (a) and (c) are visual images. (b) and (d) are thermal images. 6
Figure 4 a workflow of this research 11
Figure 5 the architecture of early fusion, half-way fusion and late fusion that will be tested in this research 12
Figure 6 different overlap cases between ground truth bounding box and predicted bounding box with corresponding IoU value 14
Figure 7 precision-recall curves of three fusion models (sensitivity is recall) 14
Figure 8 Two images in the KAIST dataset. Image (a) is a visual image. Image (b) is a thermal image. Image (a) and (b) were captured in the same location at the same time. Annotations are shown in red bounding box. In image (a) and (b), there are two annotated persons. The annotations are provided by KAIST dataset. 18

Tables

Table 1 an overview of the literature research 8
Table 2 An evaluation of accuracy and computation expenses. 15
Table 3 a schedule of my research phases 15
Table 4 the preliminary classification and localization results of Mask RCNN and RetinaNet applied on three pairs of example images in KAIST dataset, compared with ground truth annotation.. 17

1 Motivation and Problem Statement

1.1 Motivation

Human detection is essential for various applications (Liu, Zhang, Wang, & Metaxas, 2016). It is important in disasters management, the autonomous driving systems, automated surveillance and human-robotics interaction (Brunetti, Buongiorno, Trotta, & Bevilacqua, 2018):

1) when a disaster, such as earthquake and flooding, occurs the automatic human detection is able to aid the rescue work (Bharathi.V.S, 2005 ; Niels Gerlif, 2013). The location of victims can be detected and sent to the rescue team. An effective rescue scenario can be planned based on the location of the detected survivors. Without human detection, it is difficult for the rescue team to search for all the victims in the sophisticated disastrous regions.
2) In the autonomous driving systems, the human detection technique supports safety of the platform (Balani, Deshpande, Nair, & Rane, 2015). In order to ensure the safety of the pedestrians, the adjacent people should be detected, and their location should be estimated. If their distance is closer than a critical distance, the vehicle should slow down, detour, or halt.
3) Human detection plays a key role in automated surveillance (D, Manjunath, & Abirami, 2012 ; Moore, 2003). The access of human should be limited in certain areas, such as runways of airports museums in closed hours. In addition, automatic human detection algorithms can alert the owner of a house about the stranger intrusion.
4) In interaction with mobile robotics, robotics could provide better service to human customers if it detects the presence of human and his location (Moore, 2003). For example, in a seamless assistance system, the robot should be able to detect the location of the user and interact efficiently with him.

In conclusion, human detection is of importance in many applications and it is an ongoing research topic.

1.2 Problem Statement

This section aims at explaining the following questions. Why it is difficult to detect human? What are the drawbacks of visual and thermal images? Why it is necessary to integrate thermal and visual images for human detection? What are the challenges in sensor fusion? Why deep learning is better than other approaches in human detection?

Human detection is a challenging task (J. Liu et al., 2016). Because, people have different ages, genders, body shapes, positions, appearances. In addition, human can be captured from different views (Figure 1). Moreover, human may be occluded, which makes it difficult to detect human. (Figure 2)

Visual images and thermal images are the two major information sources used in human detection researches (Fan, Xu, Zhang, & Chen, 2008). Visual images are in RGB channel (Reyes-Ortiz, Oneto, Samà, Parra, & Anguita, 2016). Thermal images are visual displays of the amount of infrared energy emitted, transmitted, and reflected by an object (Correa, Hermosilla, Verschae, & Ruiz-del-Solar, 2012). As the temperature of an object increases, the amount of radiation emitted by the object increases. With the help of thermal images, warm objects like human become easily visible against the cool environment. An example of visual and thermal images in day and night is shown in Figure 3 (Hwang, Park, Kim, Choi, & Kweon, 2015).

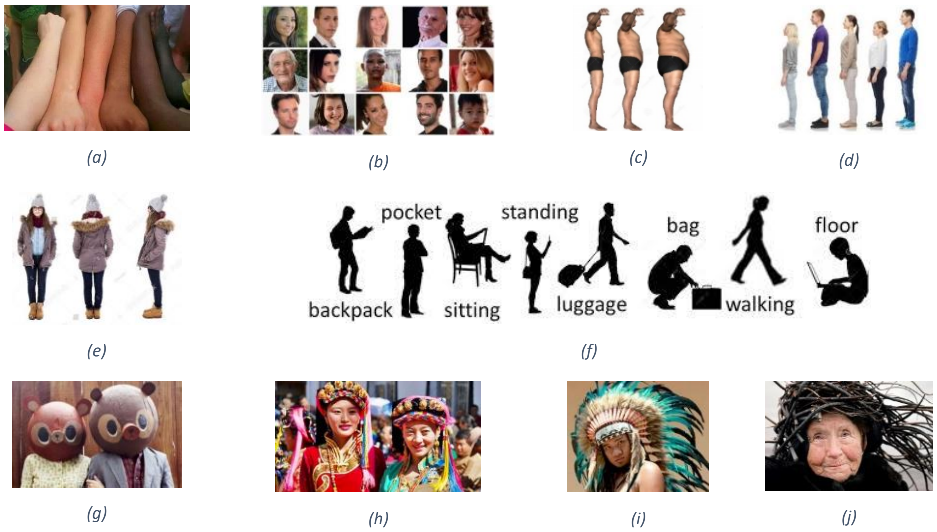
Both visual images and thermal images have drawbacks (Kim, Hong, & Park, 2017). The drawback of visual images is that they are sensitive to illumination changes. Therefore, they are easily underexposed or overexposed in the sudden changes of illumination. In addition, they require sufficient illumination. Consequently, the quality of visual images deteriorates when the illumination is insufficient, such as at night, at dusk, in the shadow regions, and in foggy weather.

The disadvantage of thermal images is that, when the temperature of the background is high, for instance above 36 Celsius degrees, human detection in thermal images can easily be disturbed by the non-human environment (Kim et al., 2017; Baek, Hong, Kim, & Kim, 2017). Because, the difference between the temperature of human and non-human is small. This situation can happen in the daytime during hot summer. Besides, the resolution of thermal images is low (Fan et al., 2008), which causes some difficulties in human detection.

Sensor fusion is applied in human detection, in order to overcome the drawback of individual sensors (Baltrušaitis, Ahuja, & Morency, 2017). Sensor fusion means to integrate data generated by two different sensors (Baltrušaitis et al., 2017). In human detection, sensor fusion means to integrate visual and thermal images. These two types of images provide complementary detection decisions (Afsar, Cortez, & Santos, 2015). When illumination is sufficient, such as in a daytime, it is easy to detect human by visual images. When illumination is insufficient, such as during the night or dusk, it is easy to detect human by thermal images. The combination of visual and thermal images provides a robust algorithm and gives a higher accuracy in human detection, compared with using a single type of image (Wagner, Fischer, Herman, & Behnke, 2016).

There are challenges in sensor fusion. Firstly, images extracted by different sensors (thermal and visual camera) may have different size and resolution. It is necessary to convert both images to the same size and resolution. Secondly, thermal and visual images have different properties. Thirdly, occlusion makes it difficult to detect human. Fourthly, in convolutional neural network, which sensor should have larger weight.

The approach to implement sensor fusion is decided to be deep learning. Because deep learning is a highly efficient method, compared with other traditional methods like HOG (histogram of gradient) (Zhu et al., 2017). Traditional methods use a dictionary to store all the human in the training data (Wagner et al., 2016). It has two drawbacks (J. Liu et al., 2016). Firstly, its speed is slow, because it compares with all the objects in the dictionary with the unknown object in the image. Secondly, if a human in an image does not exist in the dictionary, then the traditional approach is not able to detect it. In contrast, deep learning, such as convolutional neural network, is able to capture the core of human features.



(a) (b) (c) (d)

(e) (f)

(g) (h) (i) (j)

Figure 1 (a) human with different skin colors[1] (b) human with different genders and ages[2] (c) fat and slim people[3] (d) short and tall people[4] (e) the front, back and side view of a people captured by a camera[5] (f) different positions of human[6]. (g)[7] (h[8]) (i)[9] (j)[10] show people's clothes and headwear are in different colors, styles and can even be exaggerated.

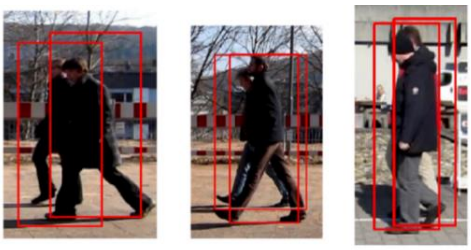
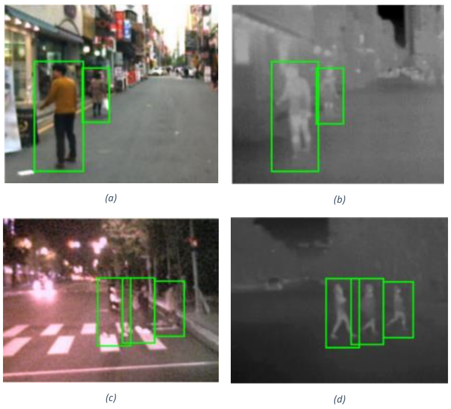


Figure 2 people who are occluded by other people[11]



(a) (b)

(c) (d)

Figure 3 Image (a) and (b) are captured in daytime on the same scene(Hwang et al., 2015). Image (c) and (d) are captured at night time on the same scene(Hwang et al., 2015). (a) and (c) are visual images. (b) and (d) are thermal images.

[1] https://www.quora.com/Is-the-variation-in-human-skin-color-another-example-of-evolutions-natural-selection

[2] https://www.123rf.com/photo_24176012_collage-of-many-different-human-faces.html

[3] https://pt.dreamstime.com/illustration/homem-gordo-e-magro.html

[4] https://www.sciencedaily.com/releases/2017/12/171205115936.html

[5] https://www.123rf.com/photo_24138754_businesswoman-front-back-side-view-isolated.html

[6] http://mbaservicesllc.com/statement-of-position-2015-its-time-to-hit-your-stride/

[7] https://qzhyxx.com/tupian/%E5%8D%A1%E9%80%9A%E5%A4%B4%E5%A5%97.html

[8] https://www.pinterest.ca/pin/332914597439335503/

[9] https://www.etsy.com/sg-en/listing/208818837/feather-headdress-indian-style-green

[10] https://www.kickstarter.com/projects/eyesasbigasplates/eyes-as-big-as-plates/posts/1636340

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

2. Research Identification

2.1 Research Objectives

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

2.2 Research Questions

Research questions cover two aspects: network architecture and sensor fusion.

Research questions on network architecture:

- What are the optimal values of the hyperparameters?
    - How many layers in total are needed in each fusion architecture? How many kernels we need in each layer?
    - How may training data and validation data are sufficient for training a CNN?
    - In activation layer, which function should be used?
    - In maximum pooling layer, how big should a window size be?
- What is the computation complexity of early, hallway and middle fusion architecture?
- What is the optimal threshold of the degree of certainty for human detection, in order to provide the highest accuracy?
- Comparing fine-tuning a pre-trained CNN and generating a CNN from scratch, which model provides a higher accuracy?
- If the number of training data is insufficient, comparing fine-tuning a pre-trained CNN and implementing data augmentation, which approach provides a higher accuracy?

Research questions on sensor fusion:

- Compare the accuracy of early, halfway and late fusion CNN architectures, which architecture provides the highest accuracy?
- Compare the computation expenses of early, halfway and late fusion CNN architectures, which architecture has the least computation complexity?

2.3 Innovation Aimed At

The current state of art in object detection is RetinaNet. It is a robust one-stage object detector (Lin, Goyal, Girshick, He, & Piotr Dollar, 2018). RetinaNet has a special loss function to solve imbalance between foreground and background classes. Previous studies use Fast RCNN, Faster RCNN and ACF+T+THOG in human detection. However, the state-of-the-art approach surpasses the accuracy of those approaches (Lin et al., 2018). Therefore, we anticipate that our human detection approach using RetinaNet and multiple sensors outperform the previous human detection approaches.

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

2.4 Related Work

Various methods have been applied on human detection. There are two ways to categorize human detection approaches. From the aspect of sensor, methods can be classified by using single sensor and using multiple sensor. From the aspect of the accuracy and computation expenses, methods can be grouped by traditional approaches and deep learning approaches. An overview of this literature research is shown in Table 1

Using single sensor	Traditional approaches		HOG
			THOG
			Standard ACF
	Deep learning approaches	Two stage	R-CNN
			Fast R-CNN
			Faster R-CNN
		One stage	RatinaNet
			SSD[12]
			YOLO[13]
Using multiple sensors	Traditional approaches		ACF+T
			ACF+T+TM+TO
			ACF+T+THOG
			ACF+C+T
	Deep learning approaches	Multimodal applications	Applications in object detection
			Applications in human detection
		Fusion types	Pixel-level fusion
			Feature-level fusion
			Decision level fusion

Table 1 an overview of the literature research

1) human detection using single sensor

a. Traditional approach

One traditional approach applied in human detection is Histogram of oriented gradients (HOG) (Dalal & Triggs, 2005). HOG extracts the distribution of local intensity gradients along all the possible edge directions, in order to detect the appearance and shape of human. HOG takes visual images as input while THOG takes thermal images as input. In (Baek et al., 2017), thermal-position-intensity-histogram of oriented gradient (TPIHOG or TπHOG) has been proposed. TPIHOG or TπHOG improves nighttime pedestrian detection performance of HOG by incorporating thermal gradient information and its locations and thermal intensities. Currently, 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).

[12] (Du, El-khamy, Lee, & Davis, 2017)

[13] (Redmon, Divvala, Girshick, & Farhadi, 2016)

b. Deep learning

A bunch of deep learning approaches have been designed for human detection, including R-CNN, Fast R-CNN, Faster RCNN and RatinaNet.

The novelty of R-CNN (region-based convolutional neural network method) is to apply a region proposal (Girshick, Donahue, Darrell, Malik, & Berkeley, 2013) (Uijlings, Sande, Sande, & Smeulders, 2012). R-CNN selects candidate object locations and then a convolutional neural network (CNN) is implemented on each of the candidate locations. Classification and localization are finally be computed as output. The drawback of R-CNN is that it is slow at training-time (Girshick et al., 2013). 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 (Girshick et al., 2013).

Compared with R-CNN, Fast R-CNN (Girshick, 2015 ; Hosang, Omran, Benenson, & Schiele, 2015 ; Li et al., 2017) improves training speed and detection accuracy. Li et al. (2016) proposed Fast R-CNN in pedestrian detection. Fast R-CNN processes the whole input image in CNN, generates a high-resolution feature map. Region proposal method is then implemented on the feature map. Fast R-CNN is faster than R-CNN during training time. However, computing region proposals still takes long time. Furthermore, Fast R-CNN is not applicable for real time detection, because the test time for each image is slow (Girshick, 2015).

Faster R-CNN incorporates Fast R-CNN network and a region proposal network (RPN) (Ren, He, Girshick, & Sun, 2015) (J. Liu et al., 2016). RPN is trained to produce high quality region proposals. So there is no need to do external independent region proposals, compared with R-CNN and Fast R-CNN (Hosang et al., 2015; Li et al., 2017). After RPN is trained, the process is the same as Fast R-CNN. The computation speed of Faster R-CNN is much faster than Fast R-CNN. Liu et al. (2016) applied Faster R-CNN on color image and visual image separately for human detection.

RetinaNet, introduced by Lin et al. (2018), achieves a higher accuracy compared with aforementioned neural networks on COCO dataset. RetinaNet, as a one stage detector, is different from previous two stage detectors. It consists a backbone network and two subnetworks, one for classification and one for generating bounding box. This research proposed a new loss function in order to address the imbalance between foreground and background classes. RetinaNet-101 is the benchmark of this research.

2) human detection using multiple sensor

a. traditional approaches

Many traditional multimodal approaches in human detection are extensions of the standard ACF, including ACF+T (Hwang et al., 2015) , ACF+T+TM+TO (Hwang et al., 2015), ACF+T+THOG (Hwang et al., 2015 ; Baek et al., 2017), and ACF+C+T (J. Liu et al., 2016 ; Wagner et al., 2016). ACF+T incorporates thermal intensity as an additional channel (Hwang et al., 2015). ACF+T+TM+TO adds three channels to Standard ACF: thermal intensity, normalized gradient magnitude of thermal images and the histogram of oriented gradients of thermal images. ACF+T+THOG (Hwang et al., 2015 ; Baek et al., 2017) incorporates a contrast enhanced version of the thermal images and HOG features of thermal image as channels. ACF+C+T detector used 10-channel aggregated features to fuse color and thermal images (J. Liu et al., 2016). Generally speaking, The extensions of ACF surpass the accuracy of standard ACF (Hwang et al., 2015). ACF+T+THOG gives the highest accuracy among these extension approaches (Hwang et al., 2015).

b. deep learning

Multi-sensor fusion approaches can be divided into three categories: pixel-level fusion, feature-level fusion and decision-level fusion (Zhu et al., 2017; Baltrušaitis, Ahuja, & Morency, 2017 ; Wagner et al., 2016). Pixel-level fusion means to integrate the information contained in multiple images of the same scene into one image (Y. Liu et al., 2018). Feature level fusion integrates features immediately after they are extracted and it learns the correlation and interactions between features (Baltrušaitis et al., 2017; Zhu et al., 2017). It is implemented on convolutional layers. Decision-level fusion integrates the outputs of multiple sub-neural-networks features (Baltrušaitis et al., 2017; Zhu et al., 2017). It is implemented on the last fully connected layers.

Multimodal fusion by deep neural network has been widely used in object detection. For instance, Images and texts are merged for visual questions answering (Visual QA) (Wu, Shen, Wang, Dick, & Van Den Hengel, 2016). Audios and videos are integrated for speech recognition (Mitra et al., 2016; Ngiam et al., 2011). Images and videos are fused for Action Recognition (Karpathy et al., 2014 ; Simonyan & Zisserman, 2014). The integration of RGB visual images and dense depth images (RGB-D) is used for 3D geometric reconstruction of static and dynamic scenes (Zollhöfer et al., 2018). Human pose detection and hand gesture recognition are two examples that apply RGB-D (Zollhöfer et al., 2018). Besides, multimodal fusion are also used in medical imaging, digital photography, remote sensing and video surveillance (Y. Liu et al., 2018).

Human detection is a sub-branch of object detection. Two sensors, visual and thermal detector, are integrated in human detection (Wagner et al., 2016). This is denoted by RGB-T. In previous researches, different deep neural networks are blended with various levels of fusion. Wagner et al. (2016) applied R-CNN in pixel level fusion and decision Level fusion. J. Liu et al. (2016) implemented Faster R-CNN on Feature level fusion and decision level fusion. This research suggests that halfway fusion on feature level gives the highest accuracy. Guan et al. (2018) introduced a deep neural network which is able to be aware of illumination so that it learns human-related feature under different illumination conditions (day and night).

Previous studies have shown that multi-sensor fusion gives higher accuracy than single-sensor applied in object detection. Deep learning approaches are faster and more robust than traditional approaches. Among all the deep learning approaches, RetinaNet gives the highest accuracy.

3. Project Set-up

3.1 Method Adopted

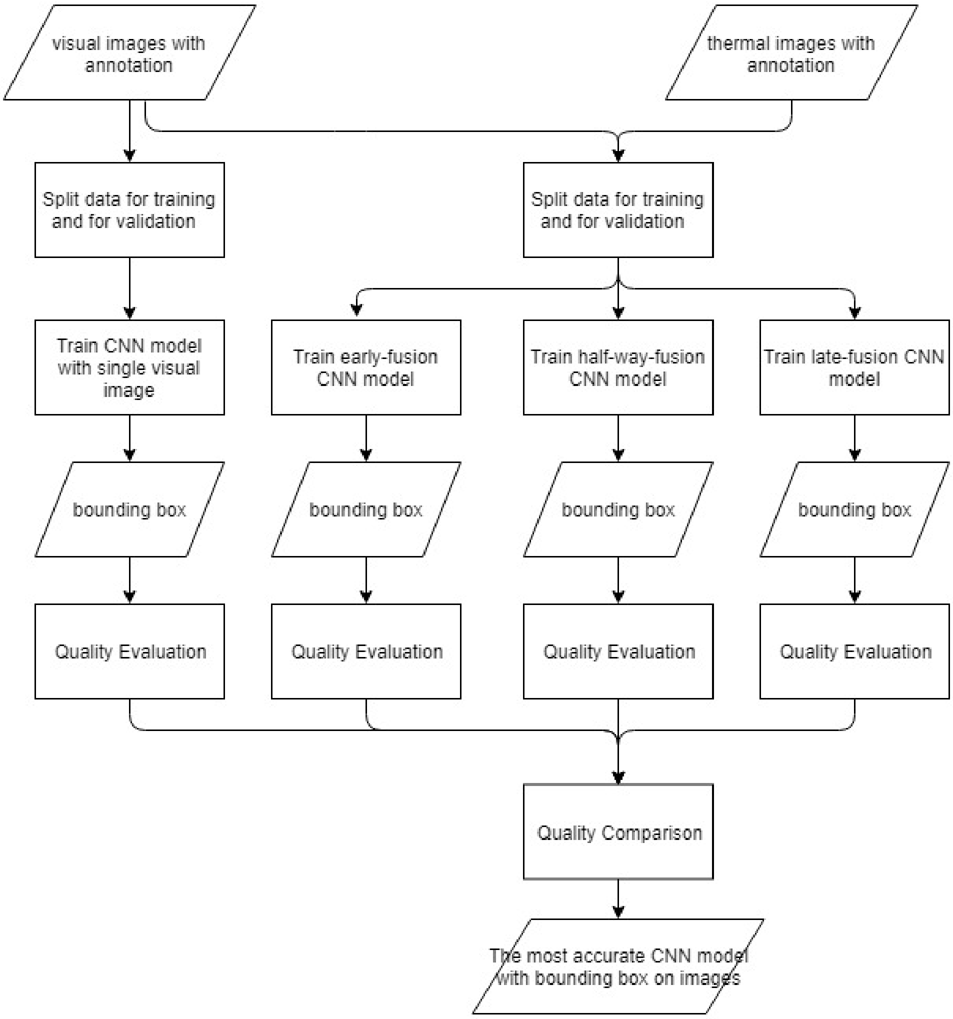

visual images with annotation
thermal images with annotation
Split data for training and for validation
Split data for training and for validation
Train CNN model with single visual image
Train early-fusion CNN model
Train half-way-fusion CNN model
Train late-fusion CNN model
bounding box
bounding box
bounding box
bounding box
Quality Evaluation
Quality Evaluation
Quality Evaluation
Quality Evaluation
Quality Comparison
The most accurate CNN model with bounding box on images

Figure 4 a workflow of this research

The workflow of this research is demonstrated in Figure 4. The proposed method is to use a convolutional neural network (CNN) to integrate thermal images and visual images and then use rectangular bounding boxes to label human in both images. The input of CNN is pairs of visual and thermal images. The visual and thermal images will train the CNN to learn the features of human. The output of the CNN will be the degree of certainty to which there is a human on a certain location. After that, if a human is detected with a high degree of certainty, a bounding box will be drawn on this location. A bounding box is a rectangle with the smallest perimeter within which the whole human lies. A bounding box is defined by three parameters: coordinates of the left top corner of the rectangle, a height and a width. Outcome will be the bounding boxes on both visual images and thermal images. Finally, the quality of the human detection will be evaluated. By comparing the quality of three fusion models, the model which provides the highest quality will be chosen as the best model. The next two sections will give a detailed explanation on model training and accuracy evaluation.

1) Model Training

There are three ways of fusion to build up a CNN model: early fusion, half-way fusion and late (Figure 5). Early, halfway and late fusion are based on pixel level, feature level and decision-level respectively. The early fusion (Figure 5(a)) means to firstly integrate thermal and visual images and later train the CNN. The late fusion (Figure 5(c)) means that, firstly, thermal and visual images are used to train the CNN separately, and the outcome of two CNNs are fused. The half-way fusion network (Figure 5(b)) is between these two types. Its fusion process happens after the two sub CNNs has been trained separately.

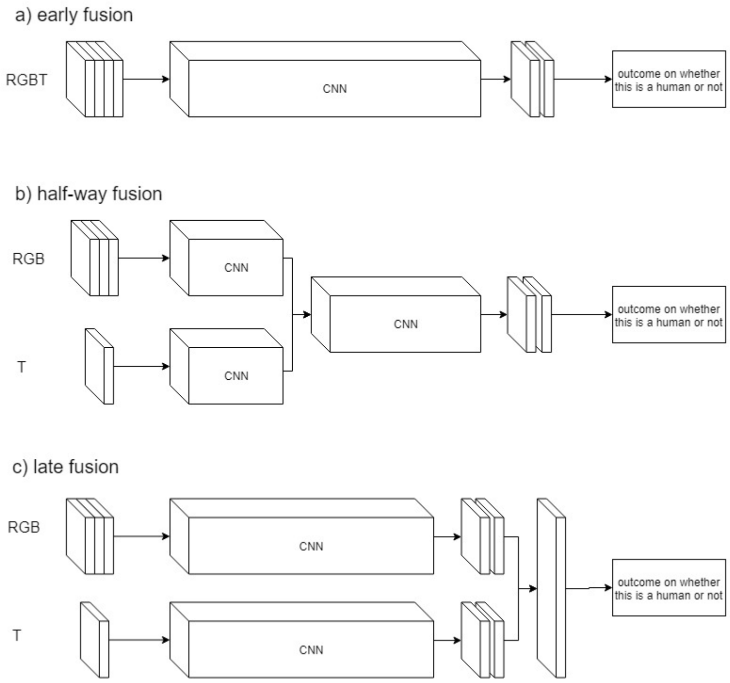

a) early fusion
RGBT
CNN
outcome on whether this is a human or not
b) half-way fusion
RGB
T
CNN
CNN
CNN
outcome on whether this is a human or not
c) late fusion
RGB
T
CNN
CNN
outcome on whether this is a human or not

Figure 5 the architecture of early fusion, half-way fusion and late fusion that will be tested in this research

A basic CNN consists of two main processes: feature extraction and classification (Zhu et al., 2017). Feature extraction aims at extracting the distinctive parts of each objects, such as the ears, the tail, the nose, the legs and other features of a dog. Classification means to compute the degree of certainty of whether there is a dog at this location or not. If the degree of certainty is high, then the outcome will be there is a dog at this location and vice versa. The basic principle of feature extraction is to detect features from low level to high level. Low level feature contains edges and colours. High level feature means an object, such as a flower, a human hand and a cloud.

Feature extraction is composed of three types of layers: **c**onvolutional layer (CONV)**,** rectifying-linear-units layer (RELU), and maximum pooling layer (POOL). Convolution layer calculates a dot product of a kernel and a local subset of the input image. Rectifying-linear-units layer is a non-linear function which removes negative values in the output images of convolutional layer. Maximum pooling layer decreases the size of the output images of convolutional layer without distorting them. The more amount of iterations there are, the finer the features will be extracted.

After feature extraction, classification will be implemented. In this process, all the neurons are fully connected to each other, therefore this layer is called "fully-Connected Layer (FC)". Each neuron will compute the class scores. For example, if there are 10 features in FC, then a column vector with length 10 will provide a class score and it will be used to categorize the object into one of the classes.

Parameters that need to be defined in this research are the following:

- Define kernels of each convolutional layer
- Define the rectifying-linear-unit layer: to what proximity the image should be shrinked.
- Define the maximum pooling layer: the size of the sliding window.
- Define how many layers in total are needed.
- In half-way fusion CNN, at which layer the fusion should start?

2) Accuracy Evaluation

The evaluation of the accuracy of each fusion model is based on three aspects: accuracy of classification, accuracy of localization and computation complexity. The method to evaluate the accuracy of classification and localization is mean average precision (mAP) (Han, Zhang, Cheng, Liu, & Xu, 2018). The method to evaluate the computation complexity is time measurement.

i. Mean Average Precision (mAP)

Three terms are necessary to compute mAP:

- Intersection over Union (IoU) = area of overlap/area of union. IoU evaluates the geometric relation between ground truth bounding box and predicted bounding box (Han et al., 2018). Figure 6 demonstrates different IoU values.

- Precision =TP/(TP+FP) [14] . Precision tells you how many of the selected objects were correct. It is a measure of completeness (Powers, 2007).

- Recall =TP/(TP+FN) [15]. Recall, also known as sensitivity, tells you how many of the objects that should have been selected were actually selected. It is a measure of exactness (Powers, 2007).

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.

$$AP = \sum_{i=1}^{n} Precision_i \cdot \Delta Recall_i$$

Equation 1equation of average precision [16]. Precision$_i$ is a percentage of correct items among first i recommendations. ΔRecall$_i$ equals 1/n if ith item is correct and 0 otherwise.

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

Equation 2 equation of mean average precision[17] . Q is the number of queries

[14] http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf

[15] http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf

[16] https://medium.com/@jonathan_hui/map-mean-average-precision-for-object-detection-45c121a31173

[17] https://www.quora.com/How-can-I-measure-the-accuracy-of-a-recommender-system

The definition of the above abbreviations are the following:

- True positive (TP): number of human who are correctly identified as human
- False positive (FP): number of non-human who are incorrectly identified as human
- True negative (TN): number of non-human who are correctly identified as non-human
- False negative (FN): number of human who are incorrectly identified as non-human

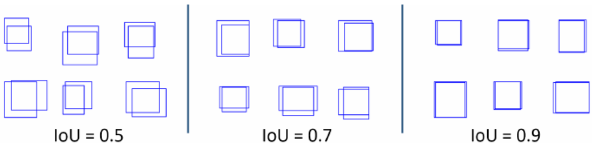


Figure 6 different overlap cases between ground truth bounding box and predicted bounding box with corresponding IoU value[18]

Precision recall curve (Han et al., 2018) will be used in my evaluation. In Figure 7, yellow, green and orange lines correspond to the precision-recall curve of different fusion architectures. They also need to be compared with the benchmark model. The line which are the closest to the green dot is the most accurate fusion model. Mean absolute will also be used in order to take true negative error into account (Han et al., 2018).

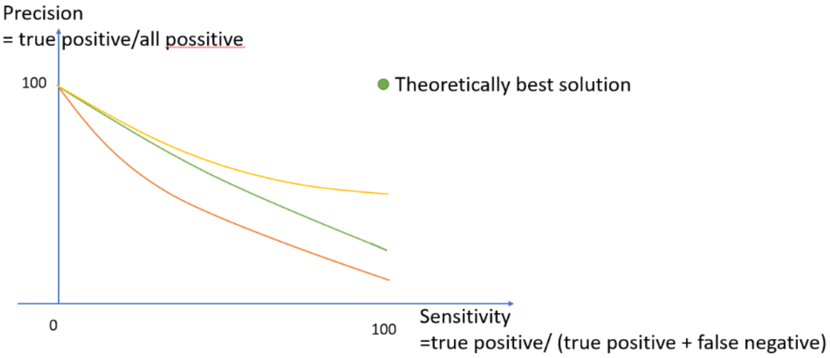


Figure 7 precision-recall curves of three fusion models (sensitivity is recall)

ii. Computation Expenses

The computation complexity of a fusion network during the training time should also be considered into quality evaluation. If the running time is too long, then it means that the time cost of the architecture is high.

To wrap up, the evaluation of accuracy and computation expenses is expected to be shown in Table 2. The optimal fusion architecture is the one with highest accuracy and with acceptable consumed time.

[18] https://www.researchgate.net/figure/An-illustration-of-random-bounding-boxes-with-Intersection-over-Union-IoU-of-05-07_fig6_319770284

		Early fusion CNN	Halfway fusion CNN	Late fusion Faster CNN
Average Training time per image (in seconds)				
Speedup				
mean average precision	IoU=a%			
	IoU=b%			
	IoU=c%			
	IoU=d%			

Table 2 An evaluation of accuracy and computation expenses. Speedup means how many times that one network is faster than the other network.

3.2 Plan of the project

The plan of my research is shown in Table 3.

	July				August				September				October				Norvermber				December				January				February				March			
	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61		
literature research	■	■	■	■	■	■	■	■	■	■	■	■																								
study the theory of "deep learning"	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■																				
download data, read data, install software	■	■	■	■																																
search and run CNN code for object detection				■	■	■	■	■																												
write proposal and prepare for presentation					■	■	■	■																												
data preprocessing							■	■	■																											
design the CNN architecture on early, halfway and late level of fusion							■	■	■																											
train the CNN on early, halfway and late level of fusion								early	■	■			halfway	■	■				late	■	■															
test the CNN on early, halfway and late level of fusion											early	■				halfway	■	■				late	■													
evaluate and compare the result of early, halfway, late fusion CNN, find the best trained CNN												■	■	■	■	■	■	■	■	■	■	■	■	■	■											
write thesis																			■	■	■	■	■	■	■	■	■	■	■	■	■					
prepare for MSc defences																															■	■	■	■		
MSc proposal submission 28 August 2018																																				
MSc proposal presentations 03 September 2018 through 07 September 2018																																				
MSc mid-term presentation 19 November 2018 through 23 November 2018																																				
MSc thesis submission 25 February,2019																																				
MSc defences 11 March 2019 to 15 March 2019																																				

Table 3 a schedule of my research phases

3.3 Risks and contingencies

- It is assumed that the time of capturing visual images and thermal images has been synchronized. If the visual images and thermal images are not captured simultaneously, then it is assumed that the time has been calibrated. Otherwise, it is out of the scope of this study.

- It is assumed that the location of visual camera and thermal camera are at the same location. If they are at different location, then it is assumed that the location has been calibrated.

- It is assumed that the labeled ground truth is perfect and correct. This means that all the rectangular bounding box of labeled human are in correct position and with correct size. There is no object which is not a human but labeled as a human. There is no object which actually is a human but not labeled.

- It is assumed that if only a part of human present in an image, such as a finger, a hand or a foot, it is unnecessary to detect it as a human. Because it is incomplete.

3.4. Expected Output

The most important expected output will be the code of different fusion models based on the state of art approach. The code will probably be uploaded to Github. An explanation of the code will be attached. The other output will be a thesis and several presentations.

4. Preliminary Results

Two deep neural networks, Mask RCNN and RetinaNet, have been used to generate preliminary results. These two networks have different architectures. Mask RCNN is a two-stage approach (He, Gkioxari, Dollar, & Girshick, 2017). It is based on Faster RCNN with additionally object masks. It generates three outputs: classification, bounding box and mask. In classification, there are 82 classes in total. Person is one of the classes. Its loss function is a sum of loss from classification, bounding box and mask. Its code is available on this website[19].

Different from Mask RCNN, RetinaNet is a one-stage approach (Lin et al., 2018). It generates classification results and bounding boxes only. There are 80 classes in total. Person is one of the classes. Its loss function is specifically designed to address class imbalance (Lin et al., 2018). Its code is available on this website[20].

The reason that these two networks are chosen is the following. Mask RCNN is good at human poses estimation and person key point detection (He et al., 2017). RetinaNet surpasses the accuracy of all existing state-of-art two stage detectors (Lin et al., 2018).

The implementation of Mask RCNN and RetinaNet can be seen as a demo of this research. The implementation process of the two networks are similar. I created two virtual environments. One for Mask RCNN and one for RetinaNet. I installed TensorFlow, Keras, pip packages, Jupiter, Visual C++ and pycocotools in python. I cloned the repository of the two networks. These two networks have been trained on the same dataset: COCO dataset[21]. COCO is a large amount of visual image dataset

[19] https://github.com/matterport/Mask_RCNN
[20] https://github.com/fizyr/keras-retinanet
[21] http://cocodataset.org/#home

for object detection and segmentation[22]. Two networks were tested on the same dataset: KAIST dataset. I selected the three pairs of images from KAIST dataset. Each pair of images consist of a visual and a thermal image. By running the code, the pre-trained networks and pre-trained weights are loaded. After that, Mask RCNN and RetinaNet were applied on the selected KAIST images for testing. The preliminary results of Mask RCNN and RetinaNet are shown in Table 4.

The preliminary result shows that Mask RCNN and RetinaNet detect human sometimes correctly but not always. There are cases that annotated human cannot be detected. We cannot draw a conclusion on which network has higher accuracy than the other. Because firstly, the amount of testing images is limited. Secondly, the two networks were trained on COCO dataset and not KAIST dataset. COCO dataset only contains visual images and does not contain thermal images. Therefore, in general, it is more difficult to detect human in thermal images than in visual images. This test is a warm-up of this research. It gives an intuition on how the process and what the outcome would look like. Solid process of training, validation and testing will be done in the following months.

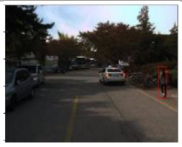
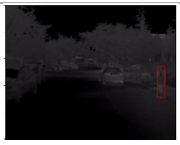
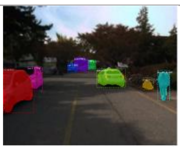
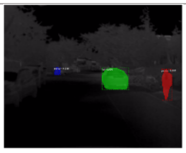
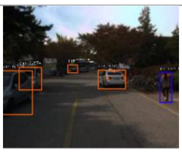
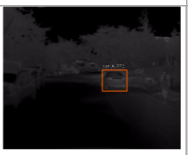
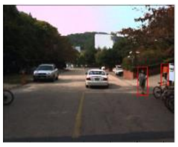
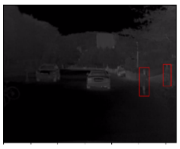
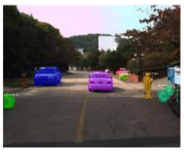
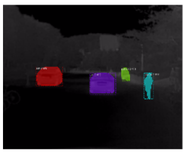
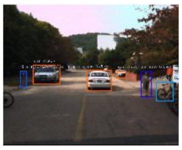
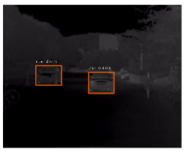
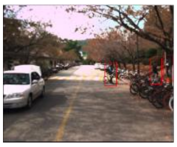
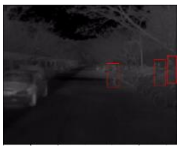
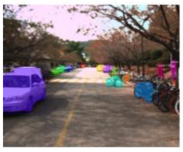
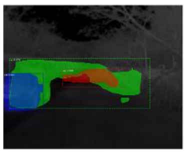
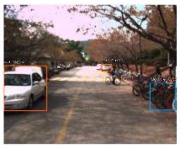
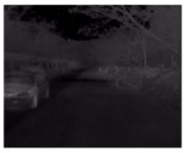
Number of annotated persons	Ground truth annotation		Mask RCNN		RetinaNet	
	visual	Thermal	Visual	thermal	visual	thermal
one person						
two persons						
three persons						

Table 4 the preliminary classification and localization results of Mask RCNN and RetinaNet applied on three pairs of example images in KAIST dataset, compared with ground truth annotation. In each pair of images, thermal and visual images are captured at the same location and at the same time. The three selected pairs of images have one, two and three annotated persons respectively. In ground truth annotation, red bounding boxes are annotated persons. In Mask RCNN and RetinaNet, the detected persons are labeled as "person" with the possibility that this is a person. The detected persons are masked in different colors in Mask RCNN and are surrounded by dark blue bounding boxes in RetinaNet.

5. Resources Required

5.1 Information

The information sources that I need are the following:

1)Training data and testing data: They all come from KAIST dataset. Section 4.2 gives a detailed description of the data.

2)Benchmark: RatinaNet (Lin et al., 2018)

3)My laptop and a server

[22] https://github.com/cocodataset/cocoapi

5.2 Data

Data used in this study is provided by KAIST (Korea Advanced Institute of Science and Technology) which is a research university in South Korea. The KAIST Multispectral Pedestrian Dataset consists of 95,000 colour-thermal pairs. They are captured by a vehicle which carries a colour camera and a thermal camera. The images are captured during day and night time. As an example, a pair of visual and thermal images with annotation is shown in Figure 8. The visual images are in size of about 200 kb while the thermal images are in size of about 77 kb. All the images have the same dimension 640 pixels * 512 pixels and the same resolution 96 dpi * 96 dpi. Human in all the pairs have been labelled. In total, there are 1182 pedestrians who have been annotated. The data is available online[23].

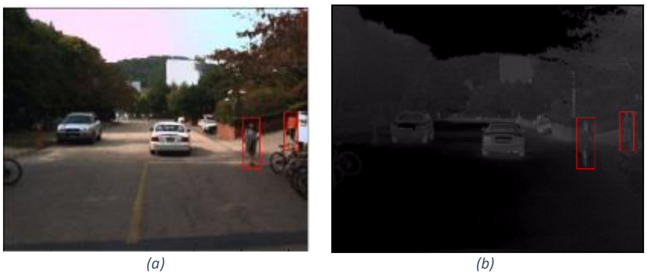


Figure 8 Two images in the KAIST dataset. Image (a) is a visual image. Image (b) is a thermal image. Image (a) and (b) were captured in the same location at the same time. Annotations are shown in red bounding box. In image (a) and (b), there are two annotated persons. The annotations are provided by KAIST dataset.

5.3 People

Besides my first and second supervisors, no other people are needed.

5.4 Software and hardware

Python, tensor flow, keras, server in ITC

5.5 Finances

No financial resource is needed.

6. Supervision

6.1 Staff Already Consulted

I have been consulting Dr. S. Hosseinyalamdary (Siavash) in the department of Earth Observation Science. He is my first supervisor. He guides me on learning convolutional neural network, reading literatures, understanding the core of this project, and writing a proposal. His guidance is very helpful. I have learned a lot during the discussion sessions with him.

My second supervisor is Dr. Francesco Nex.

6.2 Communication plan

I will have one or two meetings per week with my first supervisor. I will show him what I have achieved and what my questions are. My plan of communication is based on Table 3.

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

7. Reference

Afsar, P., Cortez, P., & Santos, H. (2015). Automatic visual detection of human behavior: A review from 2000 to 2014. *Expert Systems with Applications*, *42*(20), 6935–6956. https://doi.org/10.1016/J.ESWA.2015.05.023

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

Balani, K., Deshpande, S., Nair, R., & Rane, V. (2015). Human detection for autonomous vehicles. *2015 IEEE International Transportation Electrification Conference (ITEC)*, 1–8. https://doi.org/10.1109/ITEC-India.2015.7386891

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2017). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20. https://doi.org/10.1109/TPAMI.2018.2798607

Bharathi.V.S, G. (2005). Alive Human Detection in Disaster Zones using Manually Controlled Robots. *International Journal of Innovative Research in Computer and Communication Engineering*, *3*(2), 11–17. Retrieved from www.ijircce.com

Brunetti, A., Buongiorno, D., Trotta, G. F., & Bevilacqua, V. (2018). Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing*, *300*, 17–33. https://doi.org/10.1016/J.NEUCOM.2018.01.092

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

D, G., Manjunath, & Abirami, S. (2012). Suspicious Human Activity Detection from Surveillance Videos. *(IJIDCS) International Journal on Internet and Distributed Computing Systems*, *2*(3). Retrieved from https://pdfs.semanticscholar.org/c3bc/90003193ad9c1973a3529b551ab8857ad589.pdf

Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. *Conference on Computer Vision and Pattern Recognition (CVPR)*, *1*, 886–893. https://doi.org/10.1109/CVPR.2005.177

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

Du, X., El-khamy, M., Lee, J., & Davis, L. (2017). Fused DNN : A deep neural network fusion approach to fast and robust pedestrian detection. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 953–961. https://doi.org/10.1109/WACV.2017.111

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

Girshick, R. (2015). Fast R-CNN. *The IEEE International Conference on Computer Vision (ICCV)*, 1440–1448. Retrieved from https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Girshick_Fast_R-CNN_ICCV_2015_paper.html

Girshick, R., Donahue, J., Darrell, T., Malik, J., & Berkeley, U. C. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *ARXIV*. Retrieved from http://arxiv.org/abs/1311.2524

Guan, D., Cao, Y., Yang, J., & Yang, M. Y. (2018). Fusion of Multispectral Data Through Illumination-aware Deep Neural Networks for Pedestrian Detection. *ARXIV*. Retrieved from https://arxiv.org/pdf/1802.09972.pdf

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

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. *ARXIV*. Retrieved from http://arxiv.org/abs/1703.06870

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. Retrieved from http://arxiv.org/abs/1501.05790

Hwang, S., Park, J., Kim, N., Choi, Y., & Kweon, I. S. (2015). Multispectral Pedestrian Detection : Benchmark Dataset and Baseline. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1037–1045. https://doi.org/10.1109/CVPR.2015.7298706

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale Video Classification with Convolutional Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 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

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

Liu, J., Zhang, S., Wang, S., & Metaxas, D. N. (2016). Multispectral Deep Neural Networks for Pedestrian Detection, 1–13. Retrieved from http://arxiv.org/abs/1611.02644

Liu, Y., Chen, X., Wang, Z., Wang, Z. J., Ward, R. K., & Wang, X. (2018). Deep learning for pixel-level image fusion : Recent advances and future prospects. *Information Fusion*, *42*, 158–173. https://doi.org/10.1016/j.inffus.2017.10.007

Mitra, V., Vanhout, J., Wang, W., Bartels, C., Franco, H., Vergyri, D., … Morgan, N. (2016). Fusion Strategies for Robust Speech Recognition and Keyword Spotting for Channel- and Noise-Degraded Speech. *INTERSPEECH 2016*, 3683–3687. https://doi.org/10.21437/Interspeech.2016-279

Moore, D. (2003). A real-world system for human motion detection and tracking. Retrieved from http://www.vision.caltech.edu/~dmoore/dmoore-final-thesis.pdf

Nam, W., Dollar, P., & Hee Han, J. (2014). Local Decorrelation for Improved Pedestrian Detection. *ARXIV*, 1–9. Retrieved from https://papers.nips.cc/paper/5419-local-decorrelation-for-improved-pedestrian-detection.pdf

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal Deep Learning, 1–9. Retrieved from https://people.csail.mit.edu/khosla/papers/icml2011_ngiam.pdf

Niels Gerlif, M. (2013). Visual Detection of Humans in a Disaster Scenario. Retrieved from es.aau.dk

Paisitkriangkrai, S., Shen, C., & Hengel, A. Van Den. (2014). Strengthening the Effectiveness of Pedestrian Detection with Spatially Pooled Features, 546–561. Retrieved from http://arxiv.org/abs/1407.0786

Powers, D. M. W. (2007). Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation. Retrieved from http://david.wardpowers.info/BM/index.htm.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once : Unified , Real-Time Object Detection. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. Retrieved from https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks. *ARXIV*, 1–14. Retrieved from http://arxiv.org/abs/1506.01497

Reyes-Ortiz, J. L., Oneto, L., Samà, A., Parra, X., & Anguita, D. (2016). Transition-Aware Human Activity Recognition Using Smartphones. *Neurocomputing*, *171*, 754–767. https://doi.org/10.1016/j.neucom.2015.07.085

Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. *ARXIV*, 1–9. Retrieved from http://arxiv.org/abs/1406.2199

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

Wagner, J., Fischer, V., Herman, M., & Behnke, S. (2016). Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks. *ESANN 2016 Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 27–29. Retrieved from http://www.i6doc.com/en/

Wu, Q., Shen, C., Wang, P., Dick, A., & Van Den Hengel, A. (2016). Image Captioning and Visual Question Answering Based on Attributes and External Knowledge. *ARXIV*. Retrieved from https://arxiv.org/pdf/1603.02814.pdf

Yang, B., Yan, J., Lei, Z., & Stan Z. Li. (2014). Aggregate Channel Features for Multi-view Face Detection. *International Joint Conference on Biometrics*. Retrieved from http://arxiv.org/abs/1407.4023

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. Retrieved from https://www.cv-foundation.org/openaccess/content_cvpr_2014/html/Zhang_Informed_Haar-like_Features_2014_CVPR_paper.html

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

Zollhöfer, M., Stotko, P., Görlitz, A., Theobalt, C., Nießner, M., Klein, R., & Kolb, A. (2018). State of the Art on 3D Reconstruction with RGB-D Cameras. *Computer Graphics Forum*, *37*(2), 625–652. https://doi.org/10.1111/cgf.13386