Imaging Solution by Fusion of Thermal and Color Images using the Convolutional Neural Network

Author

Bushra Khalid

Fall 2016-MS (CE) 00000172621

Supervisor

Dr. Muhammad Usman Akram

COMPUTER ENGINEERING DEPARTMENT

COLLEGE OF ELECTRICAL AND MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

February, 2019

Imaging Solution by Fusion of Thermal and Color Images using the Convolutional Neural Network

Author

Bushra Khalid

Fall 2016-MS (CE) 00000172621

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Computer Engineering

Thesis Supervisor

Dr. Muhammad Usman Akram

Thesis Supervisor's Signature: ______________________________________

DEPARTMENT OF COMPUTER ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,

ISLAMABAD

FEBRUARY, 2019

Declaration

I certify that this research work titled "*Imaging Solution by a Fusion of Thermal and Color Images using the Convolutional Neural Network* " is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources, it has been properly acknowledged / referred.

Signature of Student

Bushra Khalid

Fall 2016-MS (CE) 00000172621

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. The thesis is also according to the format given by the university.

Signature of Student
Bushra Khalid
Fall 2016-MS (CE) 00000172621

Signature of Supervisor
Dr. Muhammad Usman Akram

Copyright Statement

Acknowledgements

All praise and glory to Almighty Allah (the most glorified, the highest) who gave me the courage, patience, knowledge and ability to carry out this work and to persevere and complete it satisfactorily. Undoubtedly, HE eased my way and without HIS blessings I can achieve nothing.

I would like to express my sincere gratitude to my advisor Dr. Muhammad Usman Akram for boosting my morale and for his continual assistance, motivation, dedication and invaluable guidance in my quest for knowledge. I am blessed to have such a co-operative advisor and kind mentor for my research.

Along with my advisor, I would like to acknowledge my entire thesis committee: Dr. Farhan Hussain, Dr. Arslan Shaukat and Dr. Usman Qayyum for their cooperation and prudent suggestions.

My acknowledgement would be incomplete without thanking the biggest source of my strength, my family. I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout every phase of my life and my loving siblings who were with me through my thick and thin.

Finally, I would like to express my gratitude to all my friends and the individuals who have encouraged and supported me through this entire period.

Dedicated to my exceptional parents: Mr. ***Khalid Hussain & Miss. Salma Shaheen****, and adored siblings whose tremendous support and cooperation led me to this accomplishment*

Abstract

In Computer vision, object detection and classification are active fields of research. Applications of object detection and classification includes a diverse range of fields such as surveillance, autonomous cars, robotic vision, search and rescue, driver assistance systems and military applications. Many intelligent systems are built by researchers to achieve the accuracy of human perception but could not quite achieve it yet. In the last couple of decades, Convolution Neural Network (CNN) emerged as the most active field of research. There are a number of applications of CNN, and its architectures are used for the improvement of accuracy and efficiency in various fields. In this research, we aim to use CNN in order to generate fusion of visible and thermal camera images to detect persons present in those images for a reliable surveillance application. There are various kinds of image fusion methods to achieve multi-sensor, multi-modal, multi-focus and multi-view image fusion. Our proposed methodology includes Encoder-Decoder architecture for fusion of visible and thermal images, ResNet-152 architecture for classification of images. KAIST multi-spectral dataset consisting of 95,000 visible and thermal images is used for training of CNNs. During experimentation, it is observed that fused architecture outperforms individual visible and thermal based architectures, where fused architecture gives 99.2% accuracy while visible gives 99.01% and thermal gives 98.98% accuracy. Images obtained from ResNet-152 are then fed into Mask-RCNN for localization of persons. Mask-RCNN uses ResNet-101 architecture for localization of objects. From the results it can be clearly seen that Fused model for object localization outperforms the Visible model and gives promising results for person detection for surveillance purposes. Our proposed localization module gives a miss rate of 5.25%, which is 5 percent better than previous best techniques proposed.

Table of Content:

List of Figures

List of Tables

Chapter 1 : INTRODUCTION

In the present age of technology, security of individuals is one of the most important concerns. Offices, schools, hospitals and organizations are provided with complete security measures to a avoid any kind of security breach. These security measures include security personals, Close Circuit Television cameras (CCTV) and surveillance. CCTV cameras can be used in a number of other applications as well, such as offense detection, public safety, crime prevention, quick emergency response, management and for the reduction of fear of crime in public [1]. Surveillance cameras can also be used for monitoring traffic flow and to keep an eye on the staff in regards to the complaints received [2].

While monitoring a vicinity there are many types of surveillance cameras to cater different weather and lighting conditions. During the day light visible cameras come in handy for clear observation while utilizing the bright light of the sun. Thermal or Long Wave Infrared (LWIR) cameras on the other hand becomes useful when there is a bad lighting condition, storm, fog or dark scene.

Visible cameras compliment with a source of information where thermal cameras identify the presence of persons, animals, weapons and other such things as having any form of heat signature. Heat signatures along with visible information becomes much more informative during surveillance. Intruders and robbers can be identified during night time and even if they have hidden themselves behind any solid object. Thermal cameras nullify the effect of occlusion.

In the war zone soldiers and vehicles such as tanks are camouflaged to disguise the opponents. For the detection of presence of foes in your secured vicinity LWIR cameras can be utilized. Heat signatures of camouflauged human beings and tanks can easily be detected and identified using LWIR camera. Surveillance cameras are proven to be very effective in terms of crime mitigation. A statistic report released by the Federal Bureau of Investigation (FBI), shows that in 2014 estimated 1165383 violent crimes and 8277829 property crimes were recorded [3]. From which 45% were caught on surveillance 65% of the them were judged correctly.

In this paper, we are proposing a solution for efficient surveillance by utilizing both visible and thermal modes of imaging. The technique we used for this purpose is convolutional neural network. First of all ResNet-152 architecture is used for the classification of images. Classification is done on the basis of features. Once the image is classified such that there is a person present in it, images from visible and thermal models are collected and fused using an encoder-decoder network. After fusion the fused images are sent into the localization, network for the segmentation and masking of persons present in the frames.

Object classification and detection is an active field of research since last couple of decades. There are many applications of object detection such as surveillance, autonomous vehicles and military purposes. Hand crafted features such as the Discrete Cosine Transform, Wavelet Transform and a few others are used for feature extraction and Principal Component Analysis (PCA) for classification [4]. In recent years CNN and deep learning architectures are used for efficient and accurate feature extraction as well as Classification and Segmentation. The need of thermal and visible image fusion arises when the content of a single image is not enough for identification of scenes and objects. When we want detailed classification, images obtained from different sources are merged or fused together. These fused images can be obtained from different sensors or they can be acquired at different time.

In our target approach we are using data sets consisting of registered Visible and Thermal or Long Wave Infrared (LWIR) images for training of the Convolution neural network. After training, rigorous tests are performed for validation of results.

A number of neural network techniques were proposed for image fusion in the past few years. In early fusion techniques, thermal and visible images are fused before giving them to the network as input. While in late fusion, images are fused inside the network after convolution layers and before Fully connected layers [5]. Another way of late fusion of images is to fuse the outputs of networks together. Other than these halfway fusion and score fusion techniques are also used [6]. Along side the techniques of fusion, choice of network architecture is also very important. According to ILSVR challenge results, ResNet [7] is by far the best architecture with a miss rate of 3.57 percent. In this thesis, we are using two encoders for encoding visible and thermal image features which are then fused and decoded back to fused image content. Fused image is then passed through the localization, network for localization of objects.

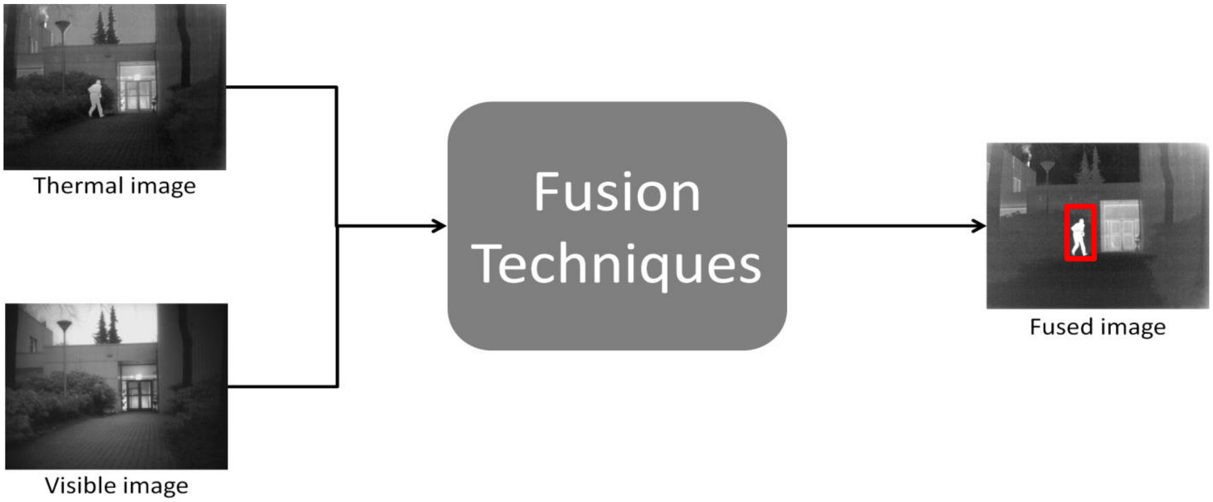


Figure 1.1: Fusion and detection flow diagram

1.1 Motivation

In the field of computer vision, object detection is still a crucial task and needs a lot of improvement to reach the level of human perception. The motivation behind this thesis was to utilize multiple modes of imaging to acquire a more informed image so that object detection becomes efficient. This system can be utilized in the war zone to detect stealth objects and can also be used in institutions for surveillance and security. Another prime purpose of this solution is to avoid accidents by providing clear visualization of the road to the driver even during night time and foggy weather conditions.

1.2 Problem Statement

To develop a better solution for image fusion so that object detection and scene identification becomes easier and achievable, mostly in stealth and almost invisible(dark) scenes.

1.3 Aims and Objectives

Major objectives of the research are as follows:

- It can be used in war zones to detect military camouflage and tanks which are made cold bodied. (Counter Stealth Imaging Technology).

- It can also be used for object detection in images or scenes where there is not enough brightness to identify things visually such as pedestrians at night.
- Can be used in vegetation for the detection of pesticides.

1.4 Contributions

- Comparison of recent developments in object detection and localization systems using a convolutional neural network.
- Fully automated system for fusion of visible and thermal images using a convolutional neural network.
- Fully automated system for classification and localization of objects from fused images for the purpose of safety and security.

1.5 Structure of Thesis

This work is structured as follows:

Chapter 2: States the literature review of related research for image fusion techniques and object detection techniques.

Chapter 3: Gives the details about the dataset used in this thesis.

Chapter 4: Consists of the proposed methodology in detail. It includes: feature level image fusion, decision level image fusion, object detection and localization.

Chapter 5: Experiments and results are discussed in detail with all desired figures and tables.

Chapter 6: Concludes the thesis and reveals the future scope of this research.

Chapter 2 : LITERATURE REVIEW

The Convolutional Neural Network has emerged as the best possible technique for detection and classification in the last couple of decades. After successful classification of ImageNet dataset, CNN was opted to classify PASCAL VOC and ILSVR2013 data set challenge. Ross Girshick et al. from UC Berkley accepted this challenge and proved that CNN is a far better approach than HOG and SIFT for classification as well as segmentation of images. They used region based convolutional neural network for object detection and named it R-CNN. Their technique involved: taking images as input, extracting regions of the data set, computation of features for each region using convolutional neural network and then the classification of a region using SVM [8].

2.1 CNN architectures

Convolutional Neural Networks are acting as a backbone solution for computer vision and machine learning tasks these days along with their growing popularity with every passing day. CNNs are widely admired and used due to their simple and understandable architecture as well as their potential to provide efficient solutions. First ever proposed architecture for Convolutional neural network was LeNet and it was published by LeCun et al. It takes an input image perform convolution on it using 5x5 filters with a stride of 1, then performs sub sampling after that some more convolutions and few pooling layers and then few fully connected layers. This architecture provided excellent results for digit recognition [9]. This architecture provided the basis for today's fast and efficient networks, but it could not provide better results on data sets due lack of availability of data on the internet and due to less computational power of GPUs present at that time.

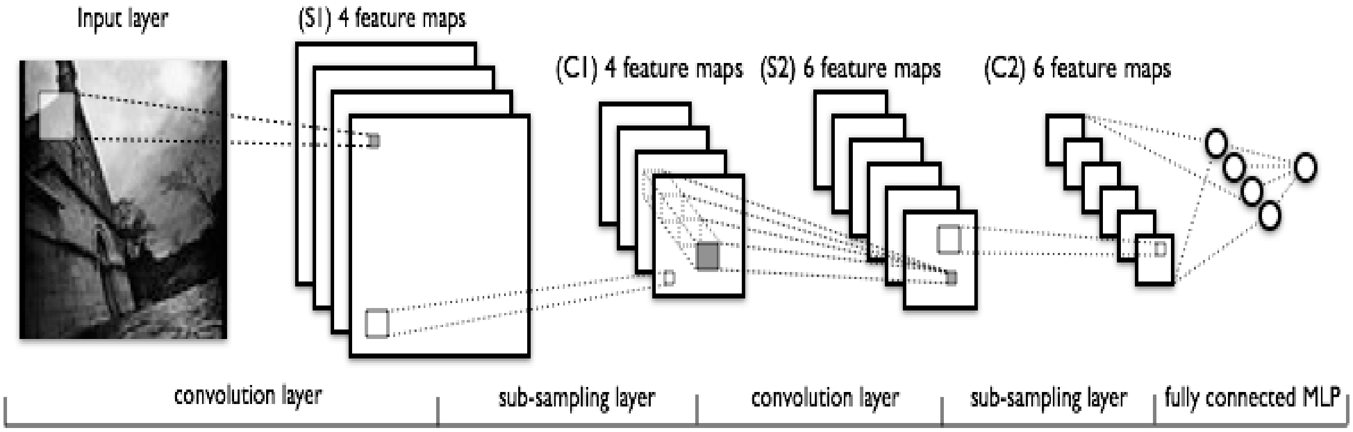


Figure 2.1: LeNet Architecture [10]

In 2012 Alex Krizhevsky et al. [11] proposed a CNN architecture called AlexNet. AlexNet won 2012 ImageNet classification challenge and beat up all other methods by a clear margin. It reduced error rate to 16.4%. Its architecture consists of five Convolutional layers, three Max pooling layers, two Normalization layers and three fully connected layers.

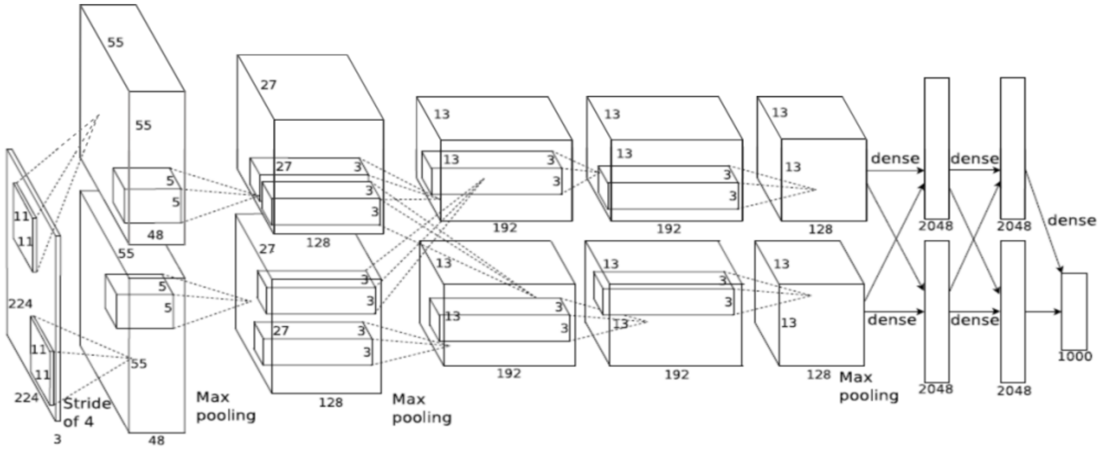


Figure 2.2: AlexNet Architecture [12]

With time CNNs became deeper by increasing layers, in AlexNet there were eight layers in total while in 2014 two networks emerged named as VGGNet [13] and GoogleNet. In VGGNet there were nineteen layers with smaller filters. Convolution layer filters were 3x3 with stride 1 and pad 1. Max pooling layers were 2x2 with stride 2. 7.3% error rate in ILSVR challenge.

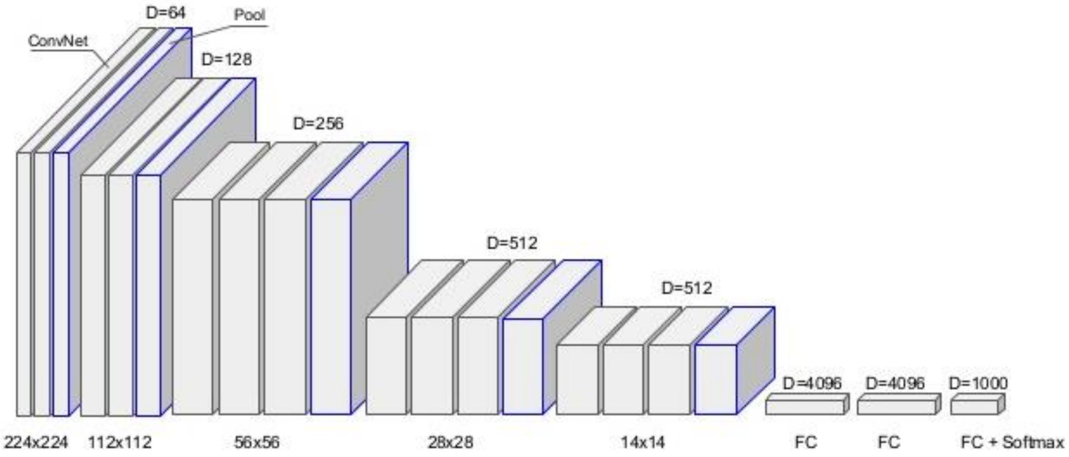


Figure 2.3: VGG-Net Architecture [14]

GoogleNet [15] won the classification challenge of ILSVR in 2014. In GoogleNet there were twenty two layers. GoogleNet used inception modules and is formed by placing multiple inception modules above each other. Error rate decreased to 6.7%. In the inception module, on each input coming from previous module multiple convolution are applied along with a pooling layer and results of all of them are then combined in a single layer which will then be fed to next module and in this way it goes on. In GoogleNet at the start of the network, we have stem network having convolution and pooling layers to start the network operations. Then we have inception modules placed above each other. After these inception modules we have output module also called classifier module for output classification. There are no fully connected layers in this architecture. Major problem of this architecture was computational complexity as each inception module will increase the depth of the output.

Figure 2.4: GoogleNet Architecture [16]

In 2015 ResNet [17] won ILSVR challenge it turned out to be the most dense architecture. ResNet has 152 layers in total. An Error rate of ResNet decreased to 3.57%. It is considered to be the best CNN architecture by far and it is used widely with small modifications here and there.

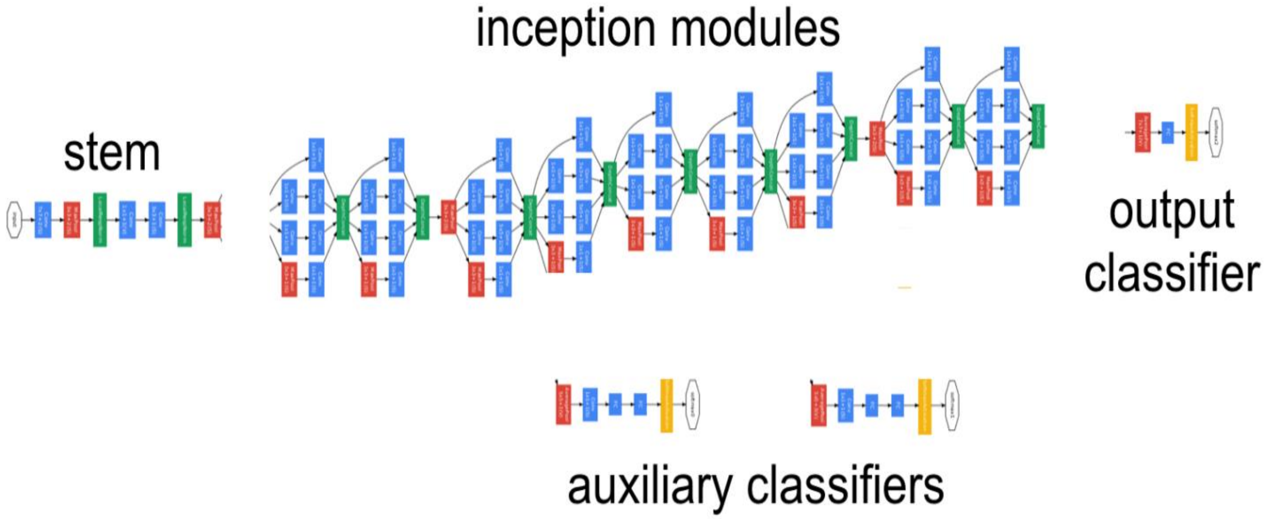


Figure 2.5: ResNet Residual block [18]

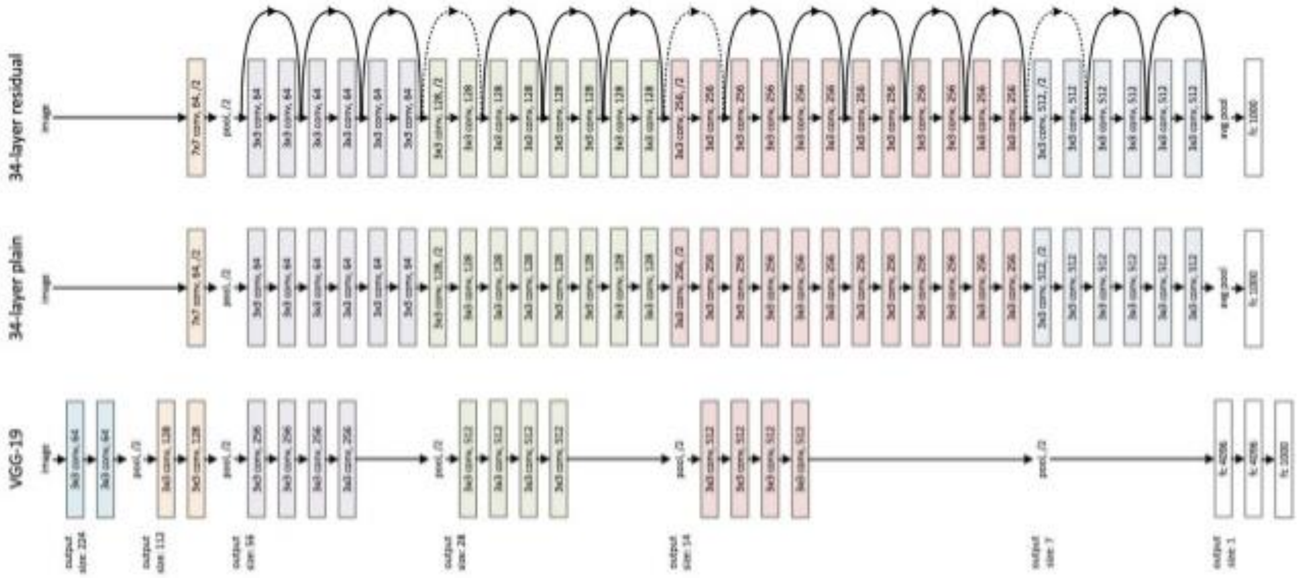


Figure 2.6: ResNet Architecture [19]

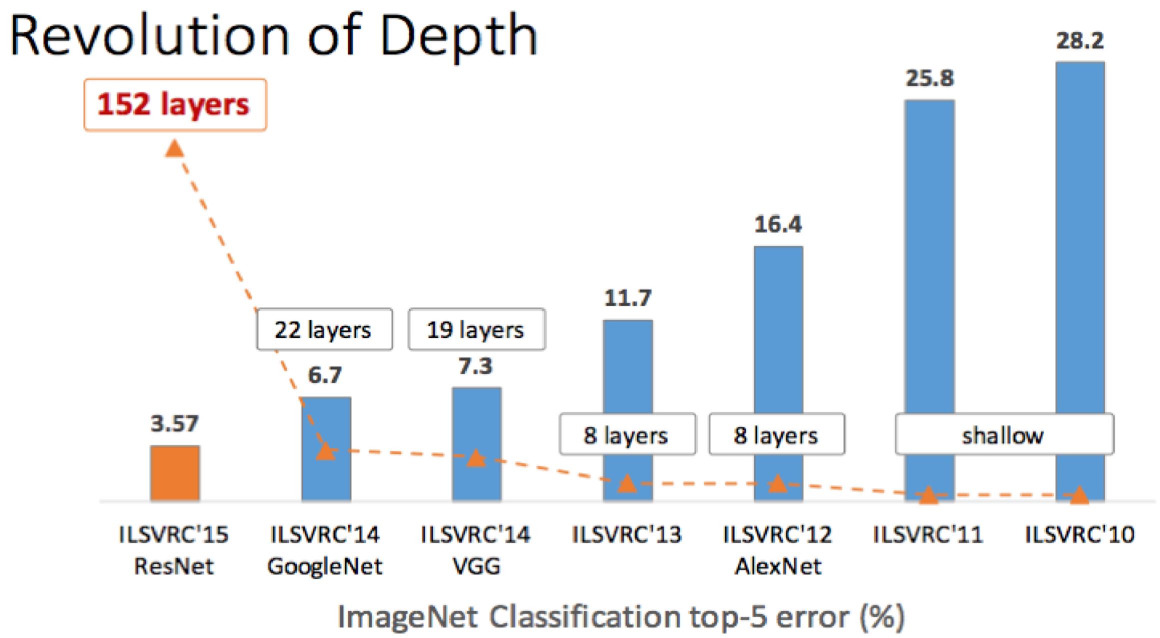


Figure 2.7: Comparison of Error rates of CNNs [20]

Table 2.1: Comparison of CNN Architectures

CNN	Conv.layers	MACCs (millions)	Parameters (millions)	Activations (millions)	ImageNet Top-5 error
AlexNet	5	1140	62.4	2.4	16.4%
VGGNet-16	16	15470	138.3	29.0	8.1%
GoogleNet	22	1600	7.0	10.4	6.7%
ResNet-50	50	3870	25.6	46.9	3.5%

2.2 Applications of CNN

Mathias Limmer et al. Proposed a useful technique for colorization of images, which is used to transfer RGB images to near infrared images using the convolutional neural network. Basic image colorization techniques comprise of three steps: first step is to segment images, the second is to assign a color palette for each region and in third step each palette is used to determine chrominance. During preprocessing step image pyramid is created than pyramid’s levels are assigned to the corresponding layer of the neural network. At the end of the neural network all layers are fused together to form on a fully connected layer of output. Size of filter changes after each pooling layer while the size of kernel remains same throughout the network. Once CNN’s result is obtained postprocessing is done to remove noise and ambiguities. For postprocessing, bilateral filters are used and once image is filtered it is compared with input image for high frequency augmentation to form a resultant colorized image [21].

CNN can be used in the field of medicine as well. In [22] Teresa Araujo et al. Proposed a convolutional neural network based technique which helps in the detection of cancerous cells in breast tissues. Deep neural networks are found to be more effective than other conventional techniques. CNN’s output provides certain features of the image which can then be combined with SVM later to get even better results.

Gustav Larsson et al. Proposed a technique based on the deep Convolutional neural network for colorization of grayscale images. As colored images provide much more information than black and white or grayscale images, it is of great consideration to convert already present grayscale datasets into colored datasets. The Convolutional neural networks can be used as a major and effective technique for this purpose as already trained networks produce efficient results. Proposed technique takes a grayscale image as input and produce colorized output by passing it through some Convolutional layers and couple of fully connected layers [23].

Another application of convolutional neural network is to check out the potential difference between colored and LiDAR data for object detection and classification. R. Niessner et al. Discussed this problem in detail and provided experimental conclusions about better techniques and effective approach. Three different approaches are taken into account for the variation of convolutional neural network. In the first method a Convolutional neural network is taken which is trained already and its output is given as an input into SVM for classification, In the second technique another pre trained neural network is taken and is used for the purpose of refinement of data within the network layers, and the third an last approach is to take a neural network which is not trained in advance and use it for the purpose of classification after training it. GRSS data provided by IEEE are used in the verification process. Upon evaluation it is concluded that LiDAR data provide better results as compared to RGB [24].

Image matting is a popular technique in imaging and video editing areas. This could be very useful when dealing with background alteration and film making, especially in animated or movies with special effects. To crop an object or number of objects from the foreground with such accuracy that on adding a virtual background it does not look unrealistic. Ning Xu et al. Proposed image matting technique based on neural networks, which also handles high level features and context of images. Their architecture consists of two neural networks, one of which takes the input image an its tri-map and uses encoder decoder mechanism to produce matte of the image. Second network is used for fine tuning of output of the first network [25].

In addition to above mentioned techniques convolutional neural network can be used in a number of other applications as well. In [26] Ashnil Kumar et al. Combined multiple Convolutional neural networks with ensemble data from different modalities which can be used for medical image classification and detection. In [27] Ronald Kemker et al. Used deep Convolutional neural network for the purpose of image segmentation. Multi spectral images are being segmented in this method and networks are trained and fine tuned over multi spectral datasets. Jiwen Lu et al. [28] proposed multi manifold deep metric learning for image classification. They utilized manifold models for classification of objects under certain circumstances, such as different lighting conditions and different angled images of the same scene. Simon Philipp Hohberg [29] did his thesis on the topic of wildfire smoke detection by using convolutional neural networks, which is a critical topic as detection of wildfire smoke at early stages can be very helpful in controlling the fire that's spreading in the area and hence saving a lot of resources, wildlife and human lives. In [30] Jin Kyu Kang et al. Proposed a fuzzy inference based Convolutional neural network, which takes RGB and infrared images as inputs and process them to find out which one is giving better results and in turn perform pedestrian detection. In [31] Natalia Neverova used convolutional neural network for human motion analysis. It takes inputs in different forms such as images, videos, audio and recorded voice and then combine the results of these modalities to give an estimation of human's emotional and physical state. Samer Hijazi et al. Used Convolutional neural network for image recognition [32].

2.3 FUSION

Image fusion is a technique of image processing through which useful information can be extracted from a combination of multiple images taken from different spectrums, foci, modes and views. Fusion can be done in multiple ways and it has different categories [33]. In Multiview fusion images are taken from different viewing angles or in different weather or lighting conditions and then they are fused together to gain useful information. In Multimodal fusion different sensors are used in the process of image acquisition like visual, thermal, laser and infrared sensors. Multimodal fusion can be done by pixel level fusion, object level fusion and through transformation as well. After acquisition of images from different sensors, these images are merged in such a way that the most prominent feature from each mode is utilized to form a final combined and more informative image than each individual. In Multitemporal fusion images of the same scene are taken at different times to detect changes. In Multifocus fusion images of the same scene with different focus regions are fused to get a well informed image.

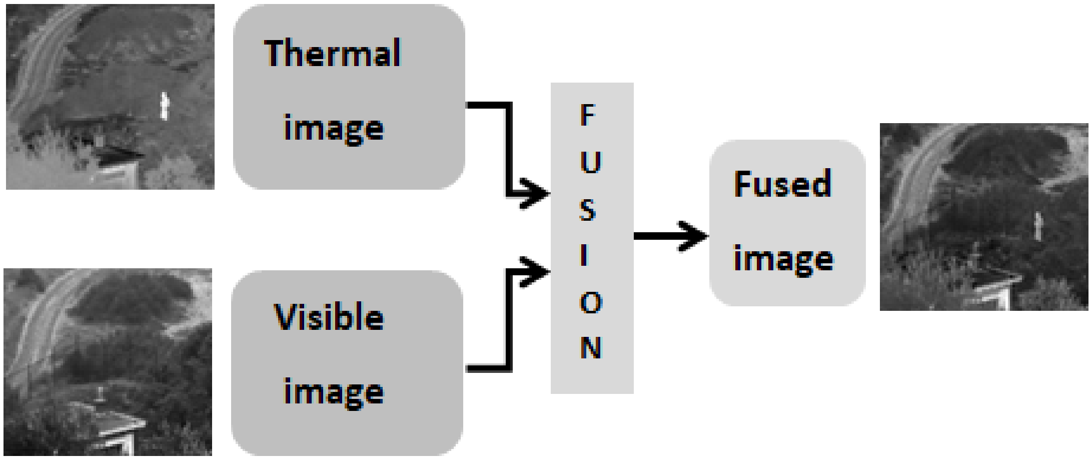


Figure 2.8: Fusion of images

2.3.1 Applications of Fusion

Image fusion has multiple applications and can be used in a variety of fields.

- Multi-spectral image fusion can be used in remote sensing applications
- Color fusion and multi-spectral fusion are used for enhanced night vision applications
- Image fusion is used in the colorization of images from grayscale to RGB colors
- In medical field image fusion is used for detection of tumors and problematic areas from images of different modes.
- Multi-source image fusion is used in GIS applications for monitoring
- Face detection, object detection, human expression, finding and hand pose estimation is also done widely using image fusion

Along with the above mentioned applications, fusion can be applied in a number of fields. In [34] Er-Yang Huan et al. Proposed a fusion technique which helps in combination of colored face images and its features to find out body constitution. This technique could be very helpful in medical field.

2.3.2 Fusion using hand crafted features

This paper discusses spatial and frequency domains of image fusion. S patial domain image fusion techniques involve Simple average, Minimum, Maximum, Max-Min, Simple Block Replace, Weighted Average, HIS, Brovey, PCA and Guided Filtering. Digging into brief detail of these techniques revealed that Simple Average fusion takes an average of corresponding pixels of images to be fused. Min and Max technique takes minimum or maximum of the corresponding pixels and form a new fused image. Max-Min takes an average of maximum and minimum from corresponding pixels. In Block pixel replace image is divided into bocks and average of each block is calculated, for the fusion maximum value of input average is taken. In weighted average different weights are assigned to each image and fusion is done on the basis of those weights. PCA transforms correlated elements of images into non correlated values also referred as principle components. IHS takes an intensity component from RGB image and merges it with infrared or thermal image, which shows exact boundaries and the presence of an object. Frequency domain image fusion is divided into two further categories. One is pyramid decomposition and the other is discrete transform. In pyramid decomposition input images are transformed into subsequent levels by convolution of image with defined filter for each level. Once levels are formed, corresponding levels of input images are merged by adding pixel values to give a fused output. There are a number of discrete transforms which can be applied to obtain image fusion. DCT, WT, KWT, KHWT, SWT, combination of curve and SWT [35]. These all techniques have common basic steps which include separation of red, green and blue components of the image, the application of the transform, combining red, green and blue components to form a whole image again. Finally, take an average of processed images to form a fused output.

Muhammad Hanif and Usman Ali proposed an optimized data fusion technique for the face recognition purpose. Thermal and visual images can be fused together in a number of ways, one of them is to use wavelet transform. The Optimized method proposed in [36] is quite efficient in a way that it can also handle problem of light and expression variations. They have compared spatial and wavelet domain fusions. In wavelet domain further four methods are applied from which optimized DWT happened to give the best results. Equinox dataset is used for experimentation which is a large database of both face and normal day life images. After the fusion is done, Gabour filter is used for face detection by comparing fused data with images stored in database.

2.3.3 Fusion using CNN

Shuo Li Hsu et al. proposed pixel level fusion technique using artificial neural network and they named it ‘region based image fusion [37]. In this paper information of intensity from infrared image is added to visual image for better visualization. For this purpose RGB image is converted into HSI format and after replacing Intensity component with infrared intensity, H and S components are converted back into RGB format. To fuse infrared image information, fused parameters are replaced with intensity, which can be estimated using neural network. Watershed algorithm is used for calculation of peak and bottom values for the purpose of region based segmentation using histogram. After removing noise from infrared images, multi-level thresholding is done. After the segmentation artificial neural network is used for calculation of fused parameters. Four features: the average intensity of visual image, the average intensity of infrared image, the average intensity of the region in the segmented infrared image and visibility of the segmented infrared image are taken into account for training of artificial neural network.

Fused parameters are output of the network, which are adjusted by authors and back propagation is done in training.

Liuhao Ge et al. Proposed a novel technique for detection of hand signs from a depth image using the 3D Convolutional neural network. As two dimensional information is not necessarily enough for estimation of gestures that is why the proposed technique takes in a three dimensional depth image as input and then uses a 3D convolutional neural network for estimations and calculations. Data augmentation is performed on the training set before CNN to manage different sizes and volumes of various hands. The Network consists of three Convolutional layers with two pooling and one ReLU layer, respectively. Afterwards three fully connected layers are present with two dropouts and one ReLU layer respectively. This architecture performs estimation of hand signs in a single pass which makes it a powerful approach to solving this problem. Experiments and their results have clearly shown that proposed method outperforms state of the art data sets [38].

Habibollah Agh Atabay showed a comparative analysis of some convolutional neural networks for object classification of binary images. In the proposed method objects are recognized based on their binary shapes and an input of 32x32 pixel is taken. The Proposed architecture of CNN is named as BS-CNN and is tested on Animal and Mpeg7 datasets. BS-CNN results are then compared with LeNet, AlexNet and MNIST-CNN and by performing augmentation of six folds on BS-CNN it was quite clear that the proposed method outperformed previous methods considerably [39].

Sachin et al. Proposed a technique for face detection from images captured from different views using the deep convolutional neural network. A lot of work has already been done on the face recognition problem, but most of it requires a face landmark and training of networks in every possible orientation which makes these methods more complex and computationally extensive. The proposed method is named as Deep Dense Face Detector DDFD. Use of convolutional neural network makes DDFD easy and effective to use. Furthermore segmentation, SVM and bounding box elements are also eliminated from DDFD. DDFD also solves the problem of various face orientations and occlusion through fusion. DDFD is compared with DPM, Head Hunter, TSM, OpenCV and a few others which resulted in better performance [40].

Face detection has been a busy area of research for researchers in the past few decades. Face detection is quite important for a number of applications such as security, surveillance, identification and verification. A lot of methods have been devised to handle this problem area. A challenging factor comes into action when faces are not fully visible or when they have some sort of occlusion which makes them unidentified. To handle this issue Shimming Ge et al. proposed a solution based on convolutional neural network, which turned out to be quite effective as it handles three levels of occlusion such as weak, medium and heavy occlusion. They pointed out two major problems of masked or occluded face detection at the start of their research, one is the lack of an image dataset with masked faces and the second is the lack of facial feature visibility due to occlusion. These researchers then created their own dataset by gathering masked face images from search engines and databases and then arranged those images manually. The Proposed architecture is named as LLE-CNN, which consists of two sub networks and an embedded module followed by a verification module. Subnetwork CNNs carry out the processes of facial region extraction and noisy such as masked feature identification, VGG is then used in combination of these two and for formation of clear descriptors. At the end of this network a verification module is placed which will perform identification tasks. Experimental results showed that LLE-CNN outperformed previous methods by a considerable percentage [41].

Before multimodal fusion, multi-focus image fusion was introduced. The need of multi-focus image fusion arose when pictures taken from different angles gave different focuses on objects. In multi-focus image fusion whole scene can be viewed clearly by fusing clear objects from different perspectives. DWT based image fusion takes two images as input, decompose them and then integrate them together to form fused coefficients which are then taken inverse of to form a final fused image. But this technique did not perform exceptionally well. The proposed technique in [42] took neural network as a solution to this problem. Two input images of the same scenario with different focuses are divided into multiple blocks ad then sent as input to the neural network after feature extraction. If the output of neural network is greater than 0.5 then an image block of first input is selected and if it’s not the case then image block of second input is taken as fusion output. At this stage a consistency verification filter is applied and after that final fusion output is received.

Deep Convolutional neural networks can also be used for action recognition in videos. The Proposed architecture is divided into two streams, one for spatial contents and one for temporal contents [43]. The spatial ConvNet deals with static images which combine to form a video while the Temporal ConvNet deals with motion recognition from video frames. Both these sub networks are finally fused at the end of fully connected layers after their softmaxs are achieved. Spatial stream ConvNet consists of five Convolutional layers with 7x7x96 stride 2, 5x5x256 stride, 3x3x512 stride 1, 3x3x512 and 3x3x512 dimensions respectively. Two fully connected layers with 4096 and 2048 neurons respectively. Temporal stream ConvNet have layers with similar dimensions as above only difference is that temporal ConvNet does not contain normalization for the second Convolutional layer. Optical flow stacking is used to recognize motion within frames, which after experimentation turned out to be better than the previous frame stacking techniques.

The RGB Depth analysis technique can also be referred as a fusion technique which can be implemented using deep neural networks. Andreas Eitel et al. Proposed a method for RGB-D based object detection. This method outperforms others as it can handle occlusion and noise in objects which is an important factor in robotics. This technique consists of two parts. One part processes RGB and other one processes depth image separately. Each part is made of CNN which is trained using ImageNet dataset already. CaffeNet architecture is used for CNN. The Proposed method is computationally cheap and quite effective. It involves following steps: after getting depth image data it is normalized between 0-255. Then attained depths are colorized into three basic channels by applying jet color method. As the network is made for RGB images, so the given information of depth and edges will be enough for feature extraction [44].

Object detection is the center of attention in computer vision and it is one of the basic building blocks. A number of researchers have worked on object detection more specifically human detection to avoid collision events while driving autonomous cars or during the movement of robots. Till late 90s detection of objects was totally based on visual images and visual scenarios. But now this problem is shifting towards multi-spectral identification, i.e. object detection using images of different spectrums or modalities. Different techniques have been devised till date for fusion of visual and spectral images. Fusion can take place at pixel, feature and decision level. The authors of this paper claim that proposed method of fusion of RGB and thermal images is the first method of its kind to the best of their knowledge. They proposed two architectures of convolutional neural network and named them Early fusion and Late fusion. In Early fusion visual and thermal image are fused together by pixel based fusion before taking the image on network input. The basic architecture of CaffeNet is used with some changes. Four channels (RGBT) Red, Green, Blue and Thermal are taken as input. Five Convolutional layers are used with first layer having a stride of two and the rest of the layers having a stride of one, the number of filters is increased at each Convolutional layer to adjust thermal channel. In fully connected layers number of neurons are decreased and classification layer is bound to be a binary classification layer. Late fusion also comprises of CaffeNet. In Late fusion whole network is divided into smaller subnetworks. Each network takes one image as an input, be it a visual image or thermal image. Five convolutional layers, two fully connected, one fusion and one classification layer are used. Subnetwork which is dealing with RGB or visual image takes double the filter size used in subnetwork dealing with a thermal image. After fully connected layers of reduced neurons a fusion layer is introduced which is basically a fully connected layer and it concatenates the results from previous fully connected layers. After the fusion layer classification layer is placed. The KAIST data set is used for training and after comparing with early fusion techniques it is concluded that late fusion techniques are far better and more reliable than earlier fusion and pre trained Late fusion architecture outperforms state of the art ACF+T+THOG detector. R-CNN framework is used which generates detected images which are later entered as input to convolutional neural network [45].

Deep Neural Networks are used by Jingjing et al. In [46] and they carried out experimentation using three different techniques of fusion, i.e. Early fusion, Halfway Fusion and Late Fusion. Later RPN and ROI proposals are generated for improvement of results. The Vanilla ConvNet model is used for experiments. Results showed that Halfway fusion was the best approach for fusion among other two. Miss Rate was minimized by 11% when compared to Faster-RCNN. 3.5% improvement was noted on KAIST as compared to earlier techniques.

Daniel Koenig et al. Proposed a different technique based on Faster-RCNN. KAIST dataset is used for experiments. They took out the RPN part and showed through the results that RPN can be as efficient as Faster-RCNN and with a few modifications it outperformed Faster-RCNN. At the end of RPN they applied Binary Decision Tree BDT for improvement of results and to decrease miss rate. After applying BDT miss rate dropped down to 29.83% [42]. Cross-Modality Transfer CNN (CMT-CNN) is used in this paper for detection of pedestrians in KAIST dataset. As compared to other method tried so far CMT-CNN provided most efficient results by showing a Miss rate of 10.69% [47].

Table 2.2: Overview of different techniques using CNN

S. No.	Author	Year	Technique	Data set	Reported results
1	**Shuo-Li Huo et al.**	2009	IHS+CNN	Self generated data set	---------
2	**Liu Hao G et al.**	2017	D-TSDF+3D CNN	MSRA, NYU	Angle estimates
3	**Habibollah et al.**	2016	CNN+binary shape class.	Animal, Mpeg7	-------------
4	**Sachin et al.**	2015	DDFD	PASCAL, AFW, FDDB	91.79% AP
5	**Shimming et al.**	2017	LLE-CNN	MAFA	--------
6	**Er-Yang et al.**	2016	CNN+class. Algorithm	Gathered manually from hospitals	65.29% accuracy
7	**Ross Girshick et al.**	2014	CNN to bottom-up region proposals	PASCAL VOC	53.3% mAP
8	**Karen S. et al.**	2014	ConvNet+multitasking learning	UCF-101, HMDb-51	37% miss rate
9	**Andreas et al.**	2015	RGB-D	RGB-D object dataset	84.7% accuracy
10	**Jorg Wagner et al**	2016	CNN	KAIST	43.80% miss rate

11	**Jingjing Liu t al.**	2016	Faster-RCNN	KAIST	37% miss rate
12	**Daneil Konig et al.**	2017	Fusion RPN + BDT	KAIST	29.83% log Avg. miss rates
13	**Dan Xu. Et al.**	2017	CMT-CNN	KAIST	10.69% miss rate

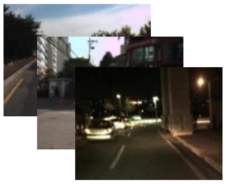
Chapter 3 : DATASET

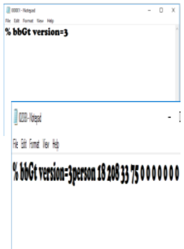
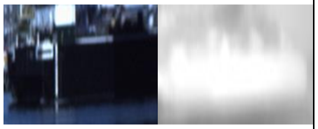
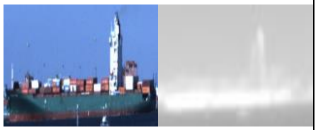
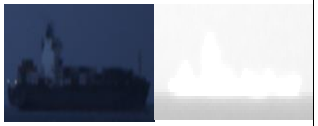
Data is the main focus and major requirement in the attainment of considerable results in the field of convolutional neural network. The basic idea of the conventional Neural network was proposed in the late 1990s by Le Cun et al [48]. And its architecture was almost similar to today's Alex Net, But the reason why it did not gain much popularity in solving vision problems was lack of data availability for training of neural network. Gradually, with time a huge amount of data exchange started to take place via the internet, which helped computer scientists to gather data and organize it in form of manageable databases which were then used for the purpose of training neural networks, in turn their performance increased and error rate decreased by a noticeable amount. We are still far away from making an intelligent system of vision which can be as responsive as a human being in a scene or an image.

Nevertheless, scientists and engineers are constantly trying to build better systems than before, either by providing multi-spectral databases or by improving the architecture of the neural network. V. V. Kniaz et al. Addressed a problem of the minimal amount of multi-spectral datasets, and the datasets which are already present have issues of geometry and time frame. To solve this big problem they also proposed its solution, which is to generate data of different spectrum by using the very own convolutional neural network. SqueezeNet architecture is used for transformation of images from visual to infrared images. They have used some layers of ThermalNet to devise a new network named fire module. Semantic image segmentation is used in conventional neural network for image transformation. Training of CNN was done using NIVIDIA digits. The resultant image is regarded as infrared image which is then matched with the original infrared image. The results were pretty impressive and thermal emission of images was matched with actual geometrically aligned infrared images which is a major step towards multi-spectral datasets [49].

In above mentioned approach, infrared images were produced by ThermalNet architecture, which turned out to be quite effective. But by the comparison of infrared and thermal rays characteristics, it is concluded that thermal rays are more effective in detection of humans than infrared rays as humans are more visible in thermal images corresponding to long infrared images. Long infrared images are prone to occlusion due to interference of traffic lights ad headlights of vehicles. The solution to aforementioned problem was given by Soonmin Hwang et al. at KAIST. They developed a hardware solution consisting of visual camera, thermal camera, three axis camera jig and a beam splitter, which provided the solution for alignment of visual and thermal images, fused together multi-band images for better results and done experimentations on multiple possible combinations of ACF (Agregatted Channel Features) to check which gives the best outcome. There are several properties related to the KAIST dataset like Scale, Position, Occlucion and change in lighting conditions. KAIST provided solution for scaling as it generates bounding boxes of different sizes in accordance with the distance between human and image capturing apparatus be it a moving car or anything else. It also provides position density, which shows at which side most of the crowd occurs. It divides occlusion problem in three parts i.e.: no occlusion where no occlusion occurs, partial occlusion where an object is partially occluded by another object or light source and fully occluded where an object is totally covered by another object or light/heat source. It also provided a solution for different lighting conditions as in daylight visual cameras provides better detection while in low light or night time thermal cameras take hold. By combining ACF with T, T+TM+O and T+THOG it is concluded that ACF+T+THOG gave best results so it is chosen as standard solution and is named as multi-spectral ACF.

Table 3.1: Comparison of Datasets

	Data Set	Size of Dataset	Resolution	Sample images	Annotation	Aim of Dataset
1	**KAIST multispectral data**	95k	640x512	 Visible images Thermal images	Annotations are given in the form of a text file as shown below: Text file indicating no person in a frame	Pedestrian detection and localization

					Text file indicating the presence of a person  % bbGt version=3 % bbGt version=3person 18 208 33 75 0 0 0 0 0 0 0	
2	**OSU color and thermal database**	9k	360x240	 Thermal and Visible images	No annotations are given	Moving target tracking and pedestrian classification
3	Maritime imagery in visible and infrared spectrum	IR: 1242 Visible: 1623 Total Images: 2865 Total Pairs: 1088 Number	Visible: 149x30 – 5056x5056 Infrared: 44x16 – 1024x768	  	Each line contains seven space delimited pieces of information, in this Format: Visible_Image, IR_Image, Fine_Grained_Label, Basic_Label Unique_ID, Is_Training, Is_Night,	Data Set for Recognizing Maritime Imagery in the Visible and Infrared Spectrums

		of unique ships: 264 Number of Night IR Images: 154 Number of Fine-Grained Categories: 15 Number of Basic Categories: 6		Pairs of Visible and Infrared images	Visible_Image is the relative path to the visible (EO) image and IR_Image is The relative path to the corresponded IR_Image. Fine_Grained_Label is a string that contains the fine grained label for the Visible and IR images. Basic_Label is similarly defined for The basic level categories.	

Chapter 4 : PROPOSED METHODOLOGY

This chapter includes the proposed architecture and all the baseline methods which are used in this thesis. Our proposed methodology consists of three main modules. The first module is a fusion module which consists of two encoders for visible and thermal image encoding respectively. Both modules take the input images and after encoding the feature maps carry out the process of fusion. Once the features are fused the fused feature vector is then transferred to the decoder block which decodes it back into the image. From this module a final fused image is obtained which is then transferred to the next module called ResNet block. This block takes the fused image and classifies it as a person class or a no person class. Images classified as no person class are discarded while the ones having person class are transferred to the third and final module of image localization. These three models are explained in detail below. **Figure 4.1** shows the flow diagram of our proposed technique which clearly shows all the three blocks of the model and how images are encoded, fused, decoded, classified and localized. The sections below discuss the three modules of the proposed technique in detail.

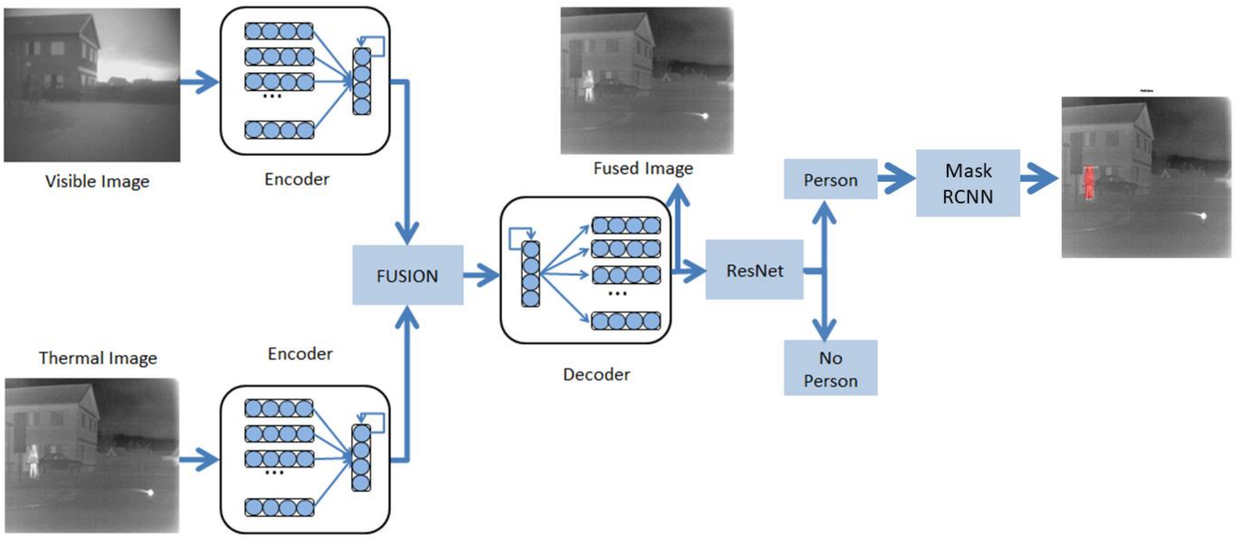


Figure 4.1: System level diagram of proposed fusion technique

4.1 Fusion

In this section we will discuss the proposed technique for fusion of thermal and visible images. For fusion of multi spectral images, we took the Dense Fuse architecture of image fusion as a baseline [50]. This procedure involves three major blocks: first one is encoder which takes image inputs in separate sub-networks and then after encoding them pass the results to the next block, the second block is known as fusion block which carries out the fusion of information acquired by encoder, finally the last block which is called decoder will get the fusion result and decode the feature information back to image content. Figure 4.1 shows the top level view of the dense fuse network. *I1* and *I2* are referred as Thermal and Visible images. *C1*, *C2*, *C3*, *C4* and *C5* are convolution layers. Encoder contains a *dense block* which will be discussed in detail in Encoder explanation. *If* is the final fused image.

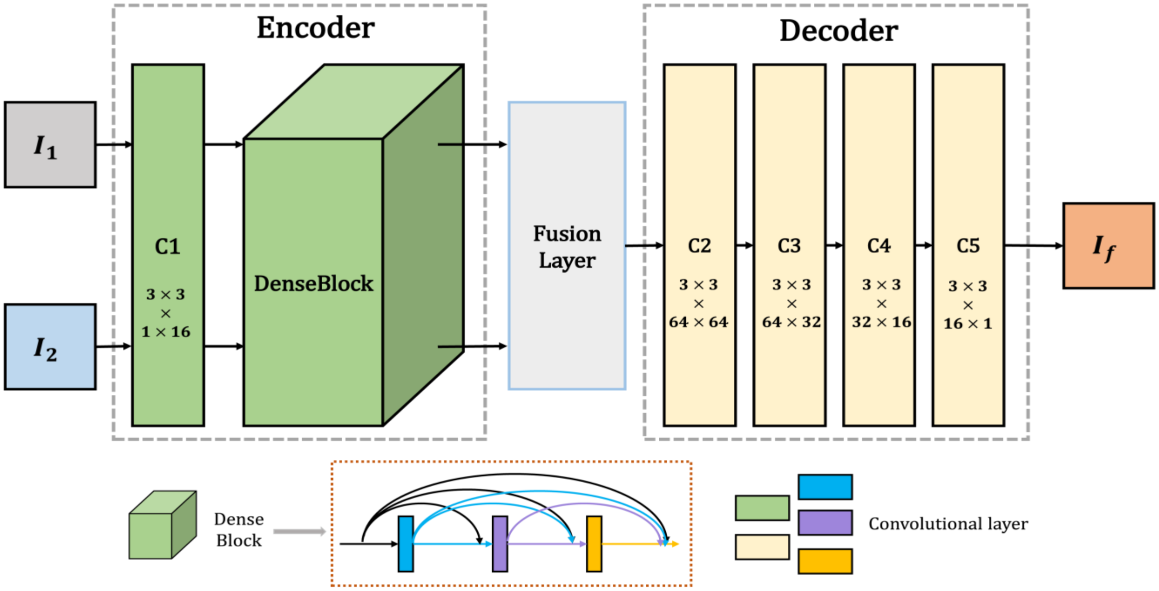


Figure 4.1: Densefuse architecture [51]

4.1.1 Encoder – Decoder Architecture

In Machine learning algorithms the method of feature representation is very important, the way a feature vector is represented says a lot about the efficiency of the technique used. Figure 4.2

shows the Cartesian coordinate and polar coordinate representation of feature maps. These features are then fed into neural networks for classification and decisions. Handcrafted features were used in machine learning applications before CNN. With the advancement in neural networks algorithms are formed which helped in feature learning by the network itself. The initial layers of CNN extracts the useful features from the input which are then fed to fully connected layers for classification and decision purposes.

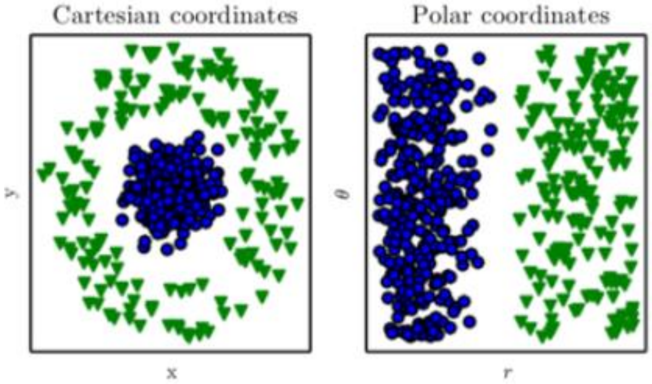


Figure 4.2: different feature maps [52]

Encoder-Decoder network architecture is a very common term in neural network representations. These kind of architectures is built upon two components: encoder and decoder. Encoder takes the inputs in forms which can be perceivable by humans and form a coded representation of those inputs which are much less in terms of dimensions. While decoder takes the encoded code as input and process it to produce an output which is again in human perceivable form. Encoder-Decoder architectures can be built using different techniques: CNN, RNN and MLPs. The choice of base technique depends on the requirement of the task. For image processing and image related tasks CNN based Encoder-Decoders are preferred while for text related tasks RNNs and LSTMs are used. Figure 4.3 shows the basic layout of Encoder-Decoder network architecture.

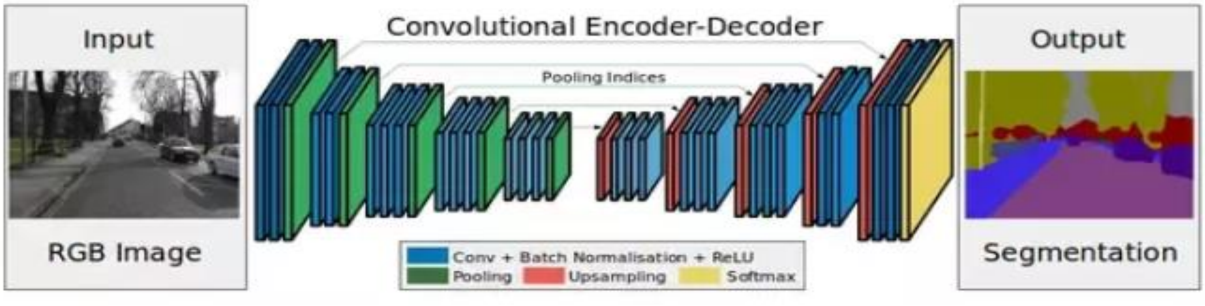


Figure 4.3: CNN Encoder-Decoder architecture [53]

4.1.2 Siamese Architecture

The network architecture used in encoder part of Densefuse is a Siamese architecture. Siamese architecture is a neural network, which is different from usual neural networks. It does not classify the input fed into it, rather it is designed to take decisions by comparing the similarities and dissimilarities of multiple inputs. Siamese architecture consists of two sub networks whose feature maps are compared and final encoded output is decided on the basis of that comparison.

Figure 4.4 shows the basic illustration of Siamese network architecture. Two sister networks with same properties take same or different inputs and these inputs pass through convolution layers for feature extraction. Once the features are extracted from both inputs, then they are passed through a contrastive loss function which actually calculates the similarities and differences between the inputs.

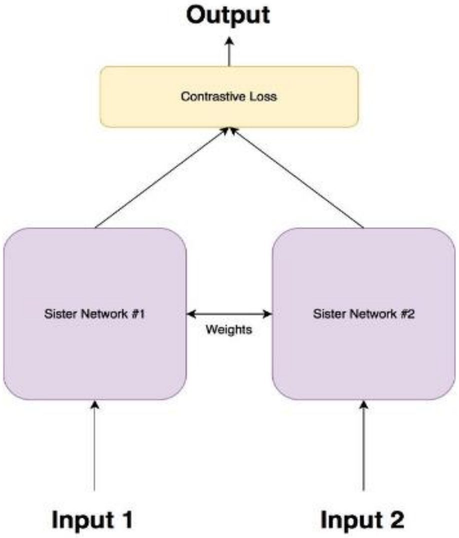


Figure 4.4: Basic illustration of Siamese architecture [54]

The reason behind the choice of contrastive loss function is that the purpose of Siamese architecture is not to classify the inputs but to differentiate between them. Loss functions like cross entropy and SGD will not be helpful in that case. The performance metric of Siamese architecture is that how accurately, it differentiates between the inputs.

4.1.3 Fusion Architecture

Dense fuse Encoder-Decoder network is discussed in detail in this section. We are aware of the functionality of Encoder and Decoder along with Siamese architecture which is used for Encoding purpose. Let us take a deeper look into the dense fuse architecture. Figure 16 shows the top level view in which I1 and I2 are referred as Visible and Thermal images. It is assumed that both the images are geometrically aligned before feeding them to the network. Dense fuse is built upon three basic blocks: Encoder block, Fusion block and Decoder block. The Siamese Encoder network consists of two channels C11 and C12. The output of C11 is fed into Denseblock11 while Output of C12 is fed into Denseblock12. Dense block itself consist of three internal convolution layers. The output of C11 is fed into each input of dense block convolution layers. The advantage of this cascaded technique is that the maximum number of features is preserved for better contrast. All the convolution layers of Encoder consist of 3x3 filters and 16 feature maps each. Encoder takes the image of any size. Once the encoding is done fusion part comes. Fusion is done by using two techniques. After fusion the fused output is sent into the decoder block which decodes it back into an image. Figure 4.5 shows the architecture for the training depicting one channel.

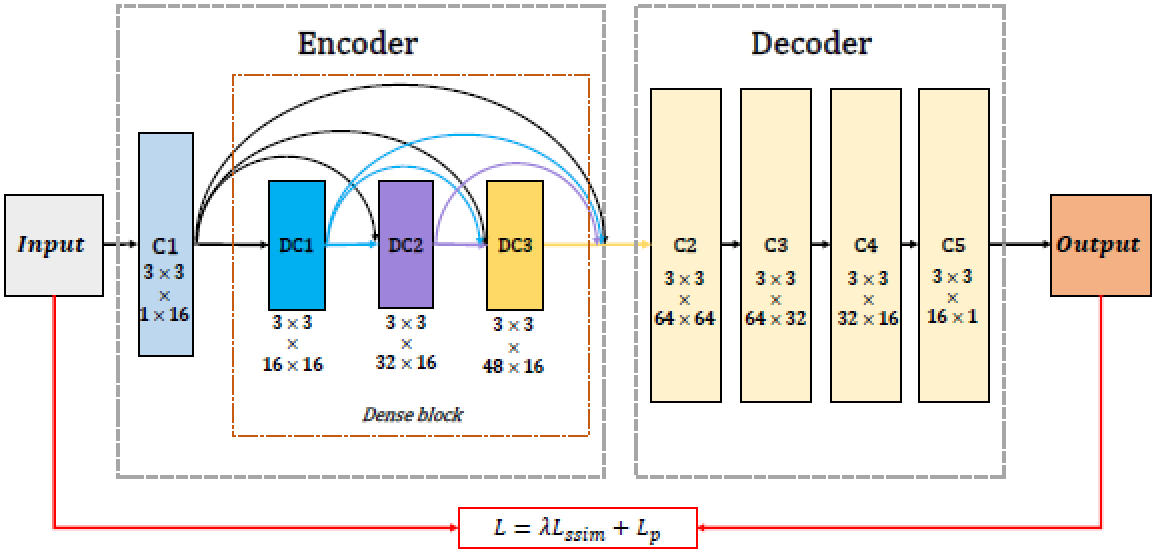


Figure 4.5: Training architecture [51]

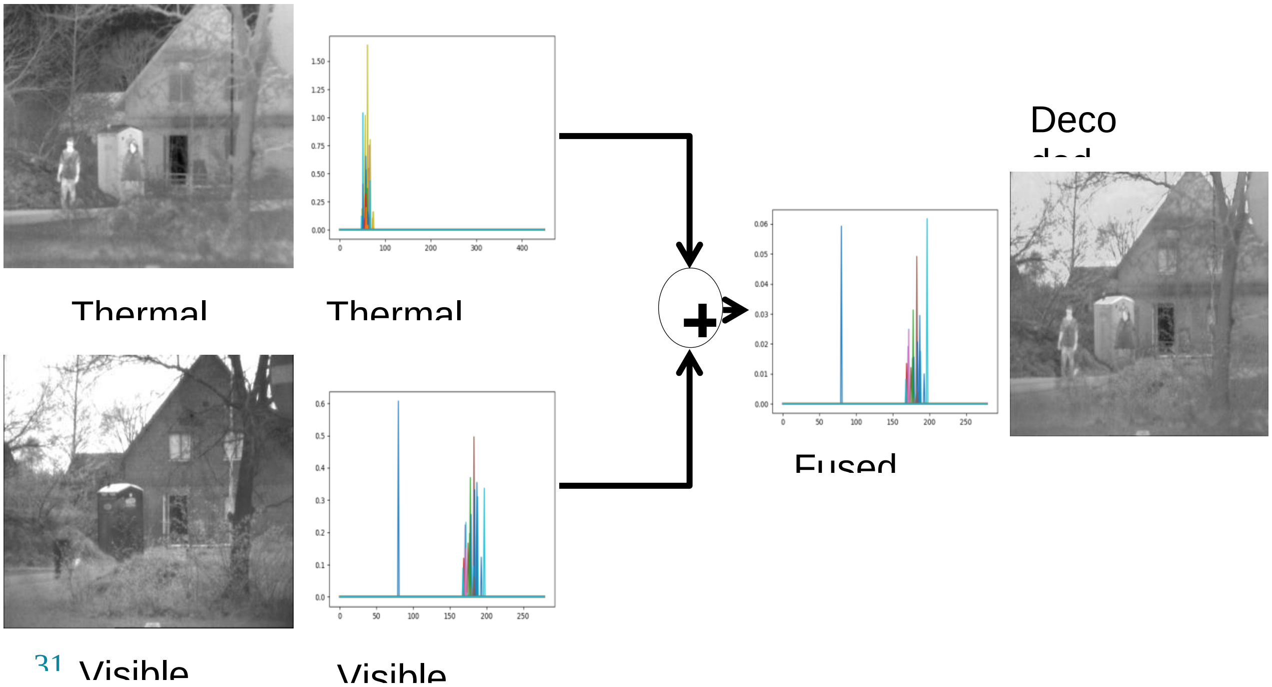


Figure 4.6: Graphical representation of feature maps

4.2 Classification

Generally Fusion is categorized in three branches: Pixel level fusion, Feature level fusion and decision level fusion. Table 4 provides the comparison of all three levels of fusion based on various features [55].

Table 4.1: Fusion levels and their performance comparison

Sr. No	FEATURES	PIXEL LEVEL	FEATURE LEVEL	DECISION LEVEL
01	Information content	Maximum	Medium	Minimum
02	Information loss	Minimum	Medium	Maximum
03	Fault tolerance	Worst	Medium	Best
04	Immunity	Worst	Medium	Best
05	Dependence of Sensor	Maximum	Medium	Minimum
06	Method difficulty	Hardest	Medium	Easiest
07	Pre-treatment	Minimum	Medium	Maximum
08	Classification performance	Best	Medium	Worst
09	Open system	Worst	Medium	Best

Pixel level image fusion is the most basic level of fusion. Image are made of pixels and when it comes to the fusion of images first thing that comes in one's mind is fusion of pixels. It is also referred as data level fusion and provides the highest level of accuracy, but involves huge amount of data processing. Second level of fusion in known as fusion of features. In feature level fusion, feature information for each image in multi-sensor system is extracted, which is then fused using different algorithms based on different applications. The main advantage of feature level fusion over pixel level fusion is that it reduces processing overhead by only taking selected feature for fusion. Decision level fusion is the highest level of fusion. In the decision level fusion feature information is extracted from images and it is used for making decisions based on different algorithms. In this section we will implement feature-level fusion strategy.

4.2.1 ResNet-152

For feature level fusion the technique we opt to use is Convolution Neural Network. The reason behind choosing convolution neural network is that this technique outperforms all other handcrafted feature acquisition techniques as per our review in chapter 2. Going deeper in convolution neural network architectures, we found that the best architecture with respect to accurate classification and minimal miss rate scores is Resnet. ResNet is further divided in three versions ResNet-50, ResNet-101, ResNet-152 and among these three architectures ResNet-152 provides least miss rate score. That is why ResNet-152 is chosen for feature level fusion architecture.

ResNet was the winner of ILSVR 2015 classification challenge. Its ensemble model resulted in 3.57% top-5 error rate. Along with this ResNet also won imagenet detection, classification and COCO detection, classification challenges of ILSVRC and COCO in 2015. It is observed that the accuracy of Faster-RCNN [56] is increased by 28% when its VGG16 layers were replaced by ResNet-101 layers. When we go deeper in network layers upon convergence accuracy starts to decrease instantly which is also known as the accuracy saturation problem. ResNet provides the solution to the accuracy saturation problem.

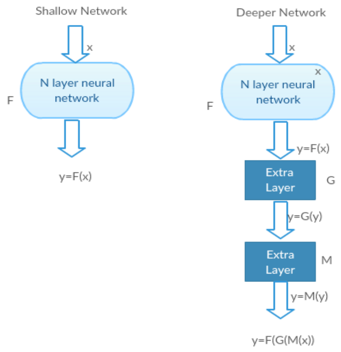


Figure 4.7: (a) Shallow layered network (b) Deep layered network [57]

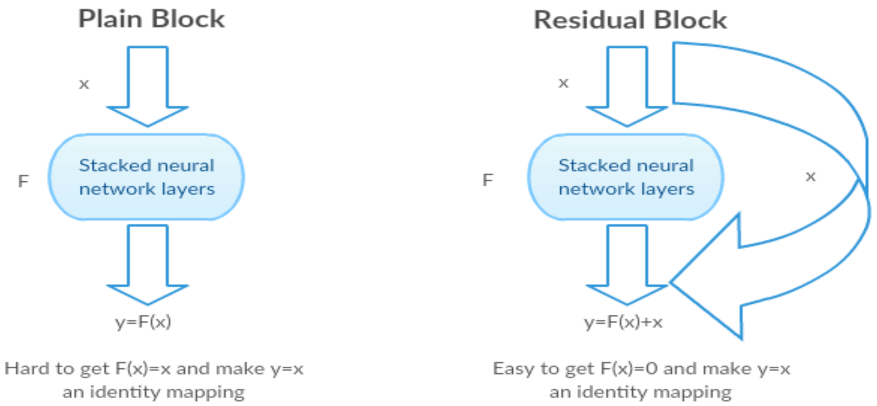

Plain Block
x
F
Stacked neural network layers
y=F(x)
Hard to get F(x)=x and make y=x an identity mapping
Residual Block
x
F
Stacked neural network layers
x
y=F(x)+x
Easy to get F(x)=0 and make y=x an identity mapping

Figure 4.8: Plain block and Residual block [58]

Figure 4.6 shows (a) network architecture of shallow networks and (b) shows the network architecture of deeper network. Both these networks are going to give the same output, in deeper network initial layers are replaced with shallow networks and the rest of the network just acts as an identity function which will propagate the output of initial shallow layers. Deeper networks are supposed to perform better, but it is observed through experiments that in reality their results were less accurate than shallower networks. The solution to this problem was residual networks [59]. An alternative way of mapping is introduced in residual networks in which residual function is calculated by subtracting input x from hidden function H(x). This helps us in finding hidden function H(x) by adding residual function F(x) with input x.

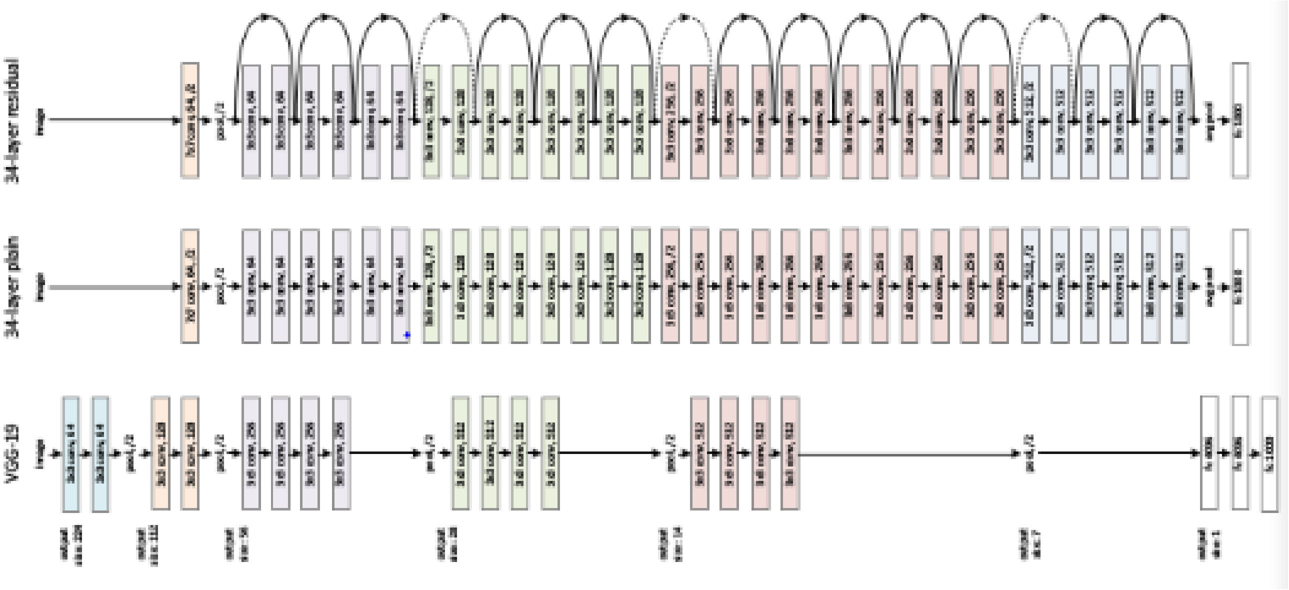
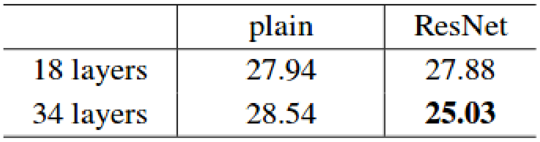


Figure 4.9: Plain block VGG and Residual block VGG [19]

To test the hypothesis made by authors in [17], the author presented multiple experiments. They took two VGG like networks with 18 and 32 layers, and another 32 layered network with residual connections. These networks were mostly consisted of 3x3 filters with stride 2 down sampling, average pooling layers, a fully connected layer of 1000 features and a softmax at the end.

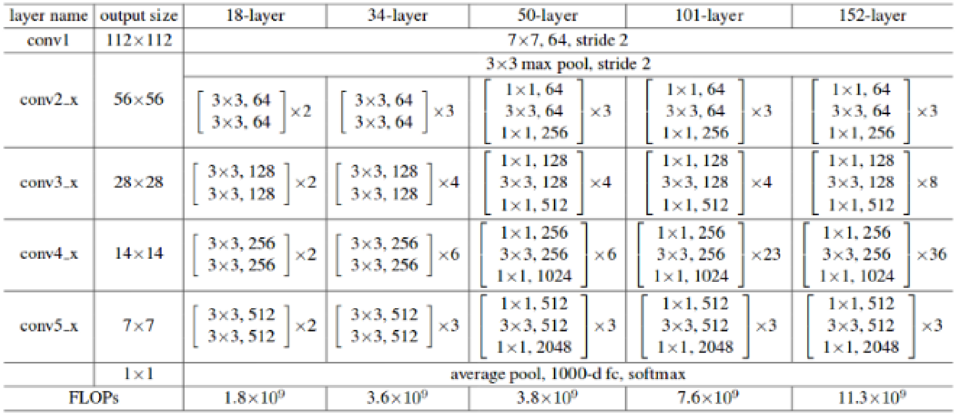
From the outcomes of above networks, it was deducted that from both plain networks VGG-18 performed better than VGG-34, while residual networks outperformed both plain networks with a clear difference.



	plain	ResNet
18 layers	27.94	27.88
34 layers	28.54	**25.03**

Figure 4.10: results of plain block and ResNet [58]

In the less deeper ResNets e.g. ResNet-18 and ResNet-34 each block is two layers deep. While in more deeper networks, e.g. ResNet-50, ResNet-101 and ResNet-152 ResNet blocks are 3 layers deep. The ResNet architecture, we implemented in our feature level image fusion is ResNet-152. Figure 4.7 elaborates the structure of deeper ResNet architectures. In ResNet-152 input size of feature vector of each block is half of the output size of feature vector of the previous block. There are total five convolution blocks and each block is multiple times as shown in figure. After convolution blocks there is an average pooling layer, a 1000-d fully connected layer and a softmax layer for providing prediction classes.



layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 4.11: Resnet Architectures [60]

4.3 Localization

Classification is done on the basis of features in section 4.1. In feature level classification objects are classified, but not detected and localized. Convolution neural networks are capable of object detection and excelled in this field over the past few years. Object detection caught the eyes of researchers in 2014 after the introduction of the basic object detection technique introduced by Ross Girshick [60]. Figure 4.11 shows the basic illustration of object detection and classification performed by RCNN. After the comparison of results of RCNN and HOG based classifiers, it was clear that CNN outperformed other techniques with a clear margin.

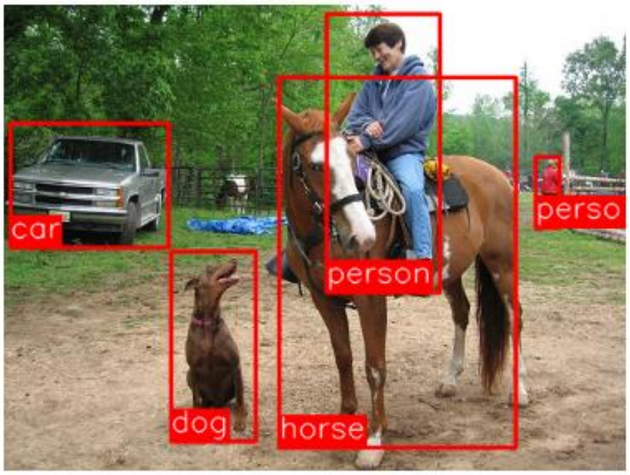


Figure 4.12: object detection and classification illustration [62]

RCNN takes image as an input, identifies the region of interest and provides bounding box output and also a classification score. Figure 4.12 shows the architecture of RCNN. But someone new to object detection must be in extreme ambiguity that how to get accurate classification, or even how to get bounding boxes. The mechanism behind RCNN is that it creates a lot of bounding boxes or square boxes of different sizes of the image by the selective search method. The Selective search method propagates square and rectangular windows of different sizes over the image and select the boxes in which adjacent pixels show a pattern of potential object.

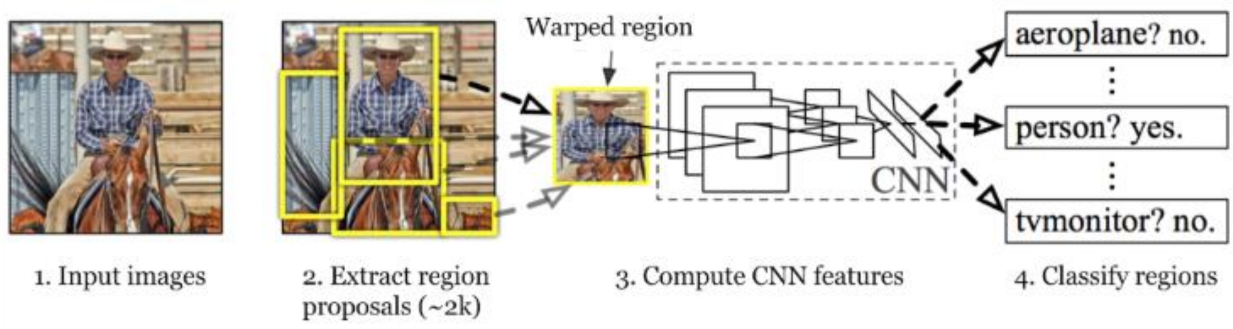


Figure 4.13: RCNN architecture [63]

Convolution neural networks like VGG and ResNet needs to be pre-trained over imagenet before RCNN starts classification. Selective search will propose approximately 2k proposals for each image. Some of those proposals can possibly contain the desired object. Region proposals are then down sampled or wrapped to get a standard size for CNN input. Than CNN will be fine-tuned with wrapped proposals for the given number of classes. A feature vector is produced by every propagation of CNN, which are then fed to SVM classifiers for classification. If the IOU overlap > 0.3 than the proposal contains the required objects otherwise it does not. The error rate in detections is reduced by the regression model. But the bottleneck of RCNN is speed overhead, which also makes it an expensive algorithm. There are three to four separate models which do not share any computation power. Generation of 2000 proposals per image causes a huge processing overhead when dealing with large data sets.

4.3.1 Fast RCNN

To address the issues of RCNN, Girshick proposed an improved model in 2015 which is known a Fast-RCNN [61]. The first problem which was solved in faster_RCNN was that all the separate models were combined as one model. Fast-RCNN calculated separate feature vectors for each proposal, but Fasster-RCNN combined all the feature vectors of an image in one vector. That feature vector is then used by CNN for classification and bounding box regression. These two solutions turned out to be effective in terms of speed. Figure 4.13 shows the architecture of Faster-RCNN.

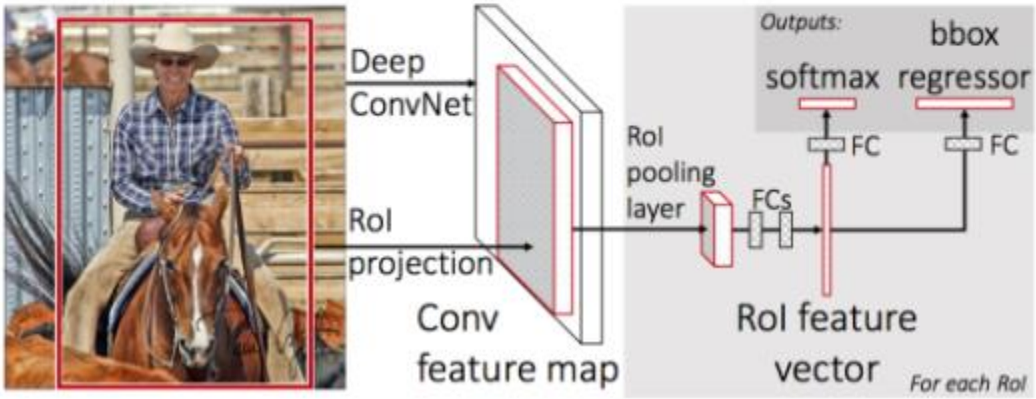


Figure 4.14: Fast-RCNN architecture [65]

ROI pooling is the process in which an image is converted from h x w matrix to H x W matrix by applying max pooling, but the image is dealt in the form of small windows and each window is max-pooled to get the output image window and the whole image likewise. Figure 4.14 shows the mechanism of ROI-pooling layer.

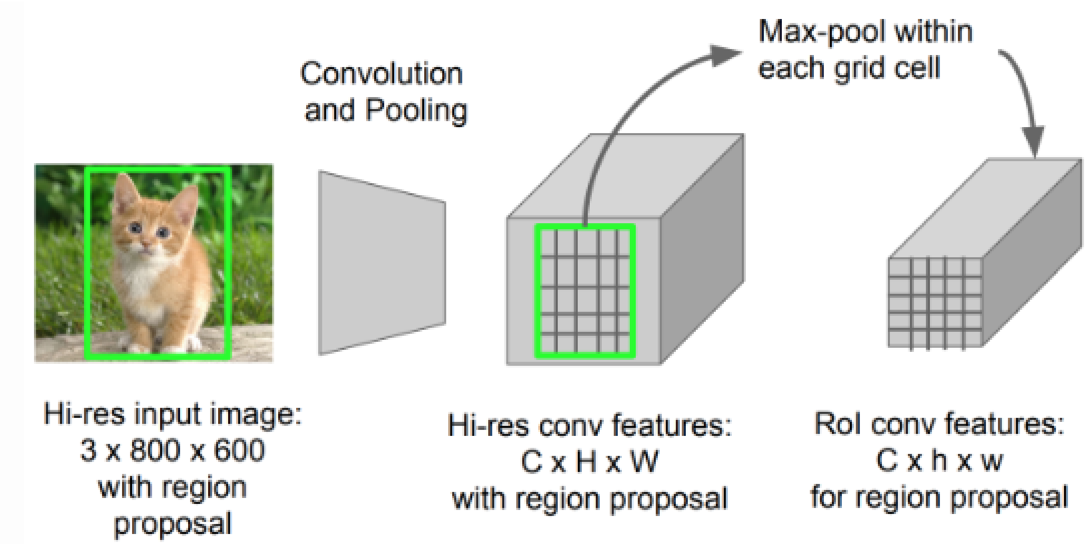


Figure 4.15: ROI pooling [66]

Similar to RCNN, Fast-RCNN is also pre-trained over imagenet. 2k proposals are generated for each image. The difference from RCNN starts when the max-pooling layer of pre-trained RCNN is replaced with ROI-pooling layer. This helps to reduce the almost redundant proposals as ROI-pooling layer output a fixed shape feature vector which cuts out the redundant information. The output of the network shows the bounding boxes for each object as well as the softmax classification for K+1 classes. Speed overhead is almost overpowered by Fast-RCNN but the cost overhead still remains as there is a separate model for proposal generation.

4.3.2 Faster RCNN

The solution to above two problems was proposed in Faster_RCNN [56]. Faster-RCNN is a combined model which provides a classification score and bounding boxes with the help of RPN and Fast-RCNN.

CNN is pre-trained on imagenet for classification. For the generation of Region proposals, Region proposal Network RPN is a fine-tuned end-to-end. Proposals having IOU overlap >0.7 are positive while the ones having IOU overlap <0.3 are negative. Fast-RCNN is trained using RPNS from Region Proposal Network. RPN is then trained using Fast-RCNN. They have some shared convolution layers. After this Fast-RCNN is fine-tuned, but only the unique layers of it are fine-tuned. Figure 4.15 shows the layout of Faster-RCNN.

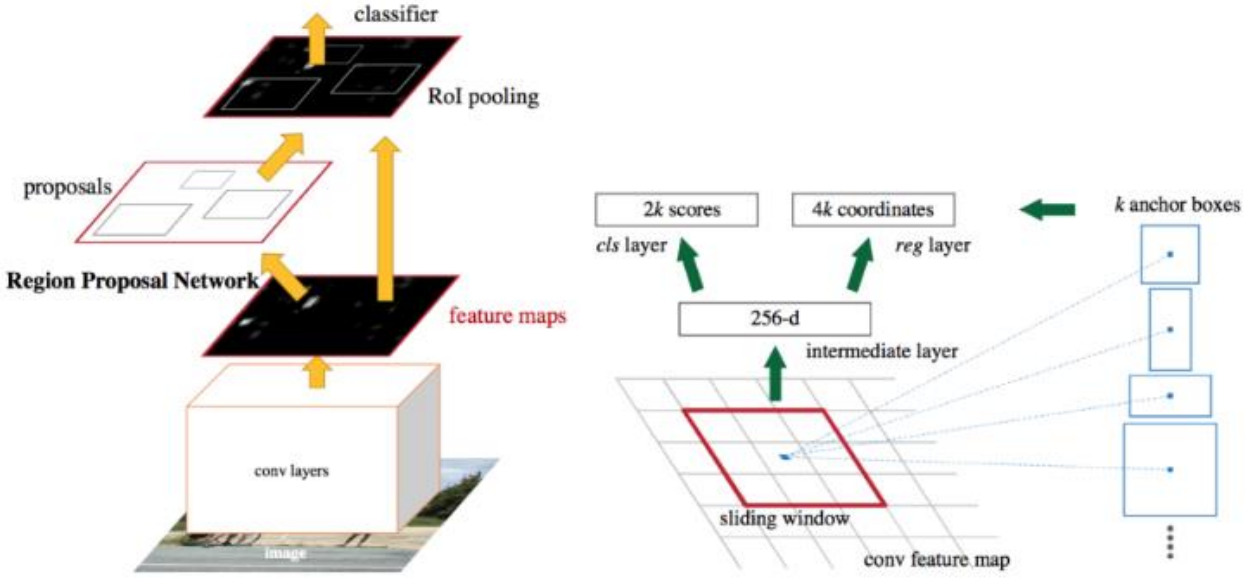


Figure 4.16: Faster-RCNN [67]

4.3.3 Proposed architecture

In this section we will discuss the model we are using in this thesis for object detection from fused images. Mask-RCNN [62] is an extension of Fasater-RCNN. Apart from object detection and classification Mask-RCNN also produces instance segmentation masks. Along side the Classification and RPN branch Mask-RCNN introduces a new branch for the instance segmentation which outputs the masks of detected objects. The instance segmentation branch is a Fully connected network, which provides pixel to pixel segmentation on the basis of the Region of Interests. A fully connected network is a network in which every node is connected to every other node. As detecting and overlaying a mask over the object is much more complex than just drawing the bounding boxes around them, Mask-RCNN introduces a new layer called ROI-Align layer in place if ROI-pooling layer.

Both ResNet 50 and ResNet 101 can be used in Mask_RNN. Initial part of ResNet act as a low level feature extractor while the later part act as High level feature extractor. Low level features include corners and edges while high level features are complete objects like a person, car, street lights, etc. Figure 4.16 shows the illustration of Mask_RCNN architecture.

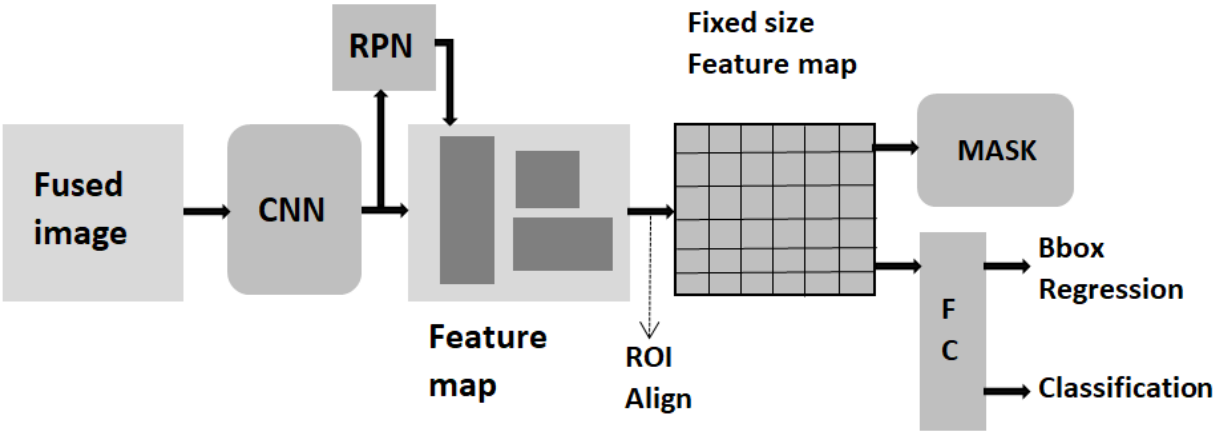


Figure 4.17: Mask-RCNN architecture

After backbone architecture of convolution layers comes Feature Pyramid network FPN. FPN takes in the feature vector from backbone architecture, and pass the high level features down from another FPN so that high level features can be accessed by initial layers. In that way initial layers have both the low level features and well as high level features. Figure 4.17 shows a high level view of FPN. Mask-RCNN implemented in this thesis uses ResNet 101 and FPN backbone.

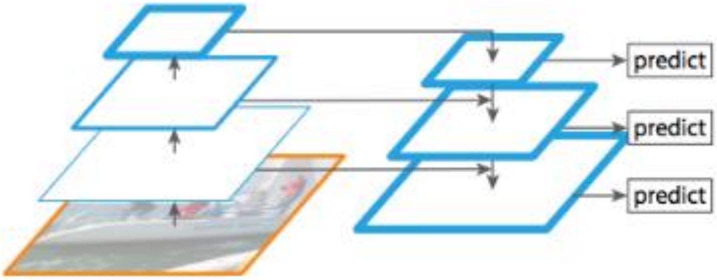


Figure 4.18: FPN [69]

RPN introduced in Faster-RCNN takes a sliding window and propagates it throughout the image. ~2k anchors are formed in the result. RPN usually takes about 10ms to skim an image, but in case of Mask-RCNN it might take longer than that as the input image to Mask-RCNN is relatively of bigger size hence it needs more anchors. Anchors are divided into two classes based on the IOU overlap. The foreground class is highly likely to contain an object while background class does not. Foreground anchors are then refined to get an exact bounding box of the object present inside the anchor. The problem of bounding box is solved, but the problem of classification still remains. To solve this problem ROI is used, the ROI takes the foreground object and classifies it into its actual class. Than ROI pooling is done to resize the image so that it can be sent into the classifier network. The segmentation branch takes in the ROI results and gives the mask for objects in the output.

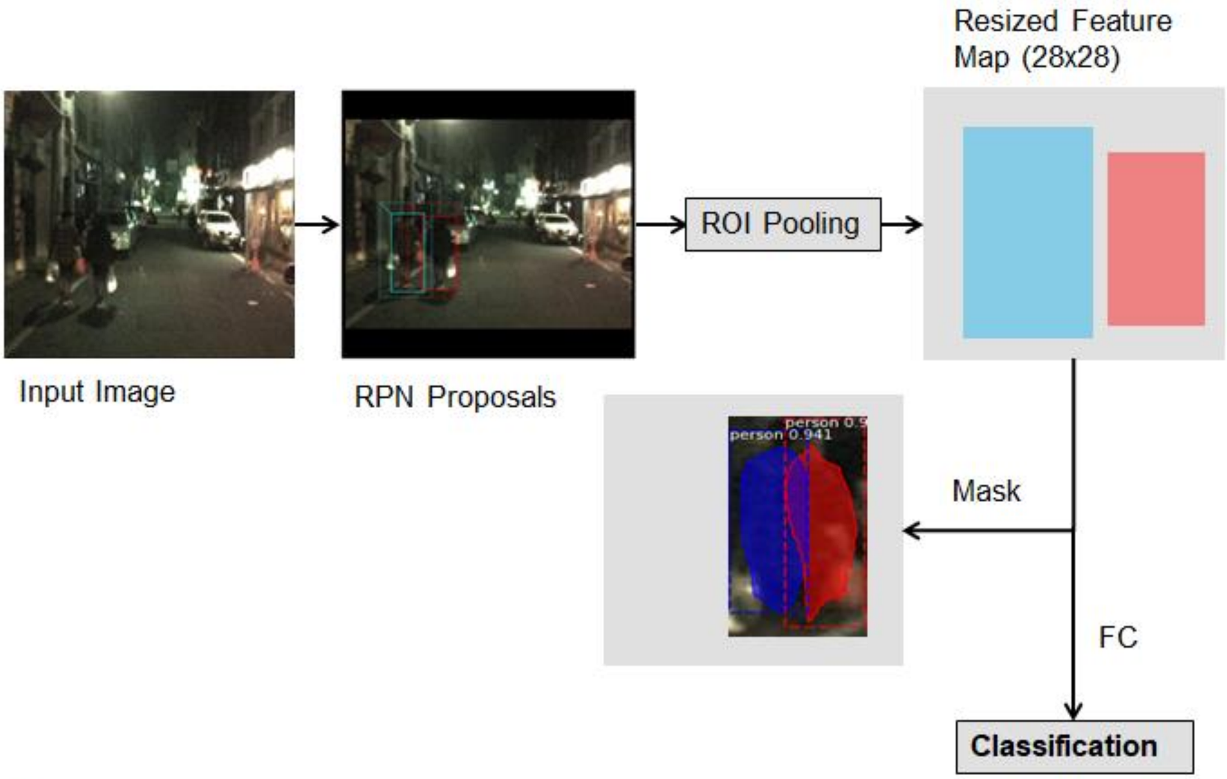


Figure 4.19: Mask-RCNN step by step results

Chapter 5 : EXPERIMENTS AND RESULTS

In this chapter, we evaluate the experiments done and their results on KAIST and local data set. A brief overview of both the datasets is given and then evaluation parameters are discussed. Results of feature level fusion and classification are shown in the form of accuracy curves. The results of dense fuse and Mask-RCNN are depicted in the visual image form.

5.1 Dataset

5.1.1 KAIST multi-spectral dataset

KAIST multi-spectral data is acquired by mounting the visible and thermal cameras on a car and an additional beam splitter is used to align the LWIR and RGB image data. Different images are taken in different lighting condition to observe the effect of light on the scene and on the object detection. Figure 5.1 shows randomly chosen images from KAIST multi-spectral data set and it ground truth bounding box representation. KAIST consists of 95000 Visible Thermal image pairs. The size of every image is 640x512 and they are aligned geometrically. The dataset is divided into 60 to 40 ratio for training and testing.

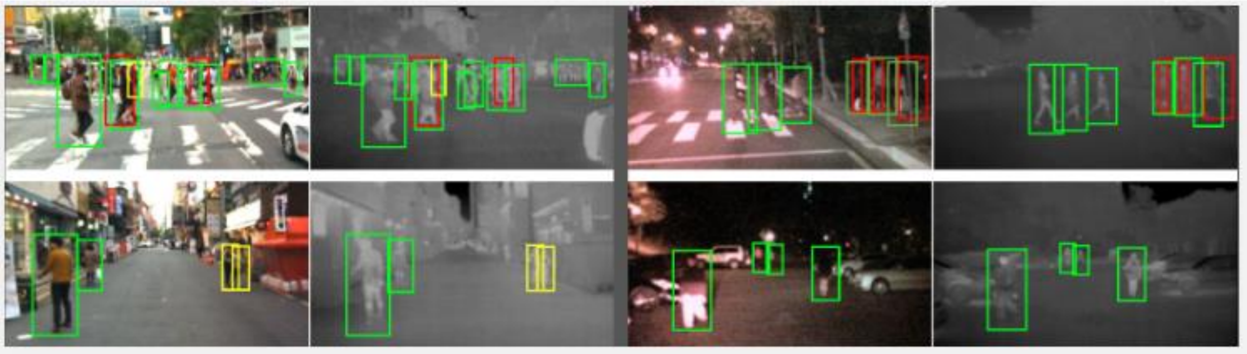


Figure 5.1: KAIST multi-spectral data [70]

5.1.2 Local dataset

Apart from KAIST multi-spectral data, we used another local dataset which consisted of visible thermal pairs of 20 images. This dataset contained some images which showed a clear difference of a person’s presence in the scene. In the visible image a person might be hidden behind the bush or might be wearing a camouflage which is invisible to the naked eye. In such cases Thermal images come in handy. Figure 5.2 shows some random images from the local dataset.

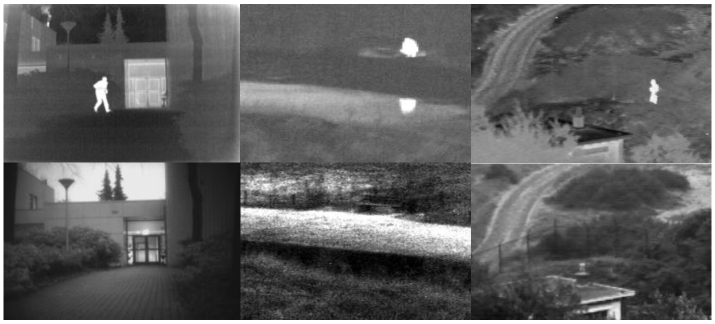


Figure 5.2: Thermal-Visible image pair of local dataset

5.2 Fusion Results

The basic purpose of fusion is to extract all the necessary information from input images and combine them in the output image. To measure the accuracy of fusion method we are using eight parameters. These parameters are used for the comparison of efficiency between addition and L1 Norm methods. The evaluation parameters used in this paper include: Entropy (EN), Fused Mutual Information of Discrete Fourier Transform (FMI-dct), FMI (pixel), FMI (wavelength), Mutual Information (MI), Mutual Information of Structural Similiarity (MI-SSIM), Structural Similarity (SSIM) and Quality of the fused image where a and b are referred as thermal and visible images respectively (Qabf). L1 norm gives best results in terms of EN, MI, Q(abf). While the addition method gives better results in terms of FMI (dct, pixel, w), MI-SSIM and SSIM. Due to high values of Qabf and entropy L1 norm fused images are less noisy and looks similar to thermal images. Images fused by the addition method contain more information about the structure and features present in individual images. Table 5.1 shows the comparison of both methods by assigning different values of Lamda.

These two methods can be chosen based on their specific characteristics according to the need of application. To illustrate the results of our proposed fusion model using an encoder-decoder technique thermal, visible and fused images are displayed in Figure 5.3. It can clearly be seen that thermal images are adding value to visible images and more persons are clear in fused images. Here the network is trained over KAIST multi spectral dataset while is tested on random images to visualize the results of images having occlusion and less visibility cases. It can clearly see that certain objects (such as persons) which cannot be seen in visible images, can be viewed when the visible image is fused with a thermal image.

Fusion techniques proposed are Addition strategy and L1-Norm strategy. Upon experimentation, it is observed that L1-Norm performs better than Addition strategy. For training purpose the dataset used is MS-COCO and NVIDIA GPU GTX 1080-Ti graphics card is used with 64 GB RAM. The learning rate of 1x10-4 is adopted. When tested over KAIST and a local dataset Densefuse provided phenomenal results. Figure 5.3 shows the fusion results over the local dataset. Image (a) is visible image, image (b) is thermal image and image (c) is fused image.

Table 5.1: Comparison of Addition and L1 Norm methods

Methods	**Addition**				**L1 Norm**			
	Densefuse_1e0	**Densefuse_1e1**	**Densefuse_1e2**	**Densefuse_1e3**	**Densefuse_1e0**	**Densefuse_1e1**	**Densefuse_1e2**	**Densefuse_1e3**
EN	6.6347	6.6383	6.6363	6.6354	6.7175	6.7596	6.7739	6.7378
FMI (dct)	0.4340	0.4339	0.4338	0.4340	0.4096	0.4092	0.4124	0.4122
FMI (pixel)	0.9098	0.9094	0.9097	0.9095	0.8989	0.8976	0.8982	0.9016
FMI (w)	0.4306	0.4306	0.4306	0.4307	0.4186	0.4198	0.4186	0.4196
MI	13.269	13.276	13.272	13.270	13.434	13.519	13.547	13.475
MI_SSIM	0.9089	0.9089	0.9089	0.9089	0.8662	0.8673	0.8713	0.8678
Q (abf)	0.4126	0.4128	0.4131	0.4125	0.4535	0.4572	0.4528	0.4543
SSIM	0.7827	0.7827	0.7827	0.7827	0.7691	0.7664	0.7686	0.7693

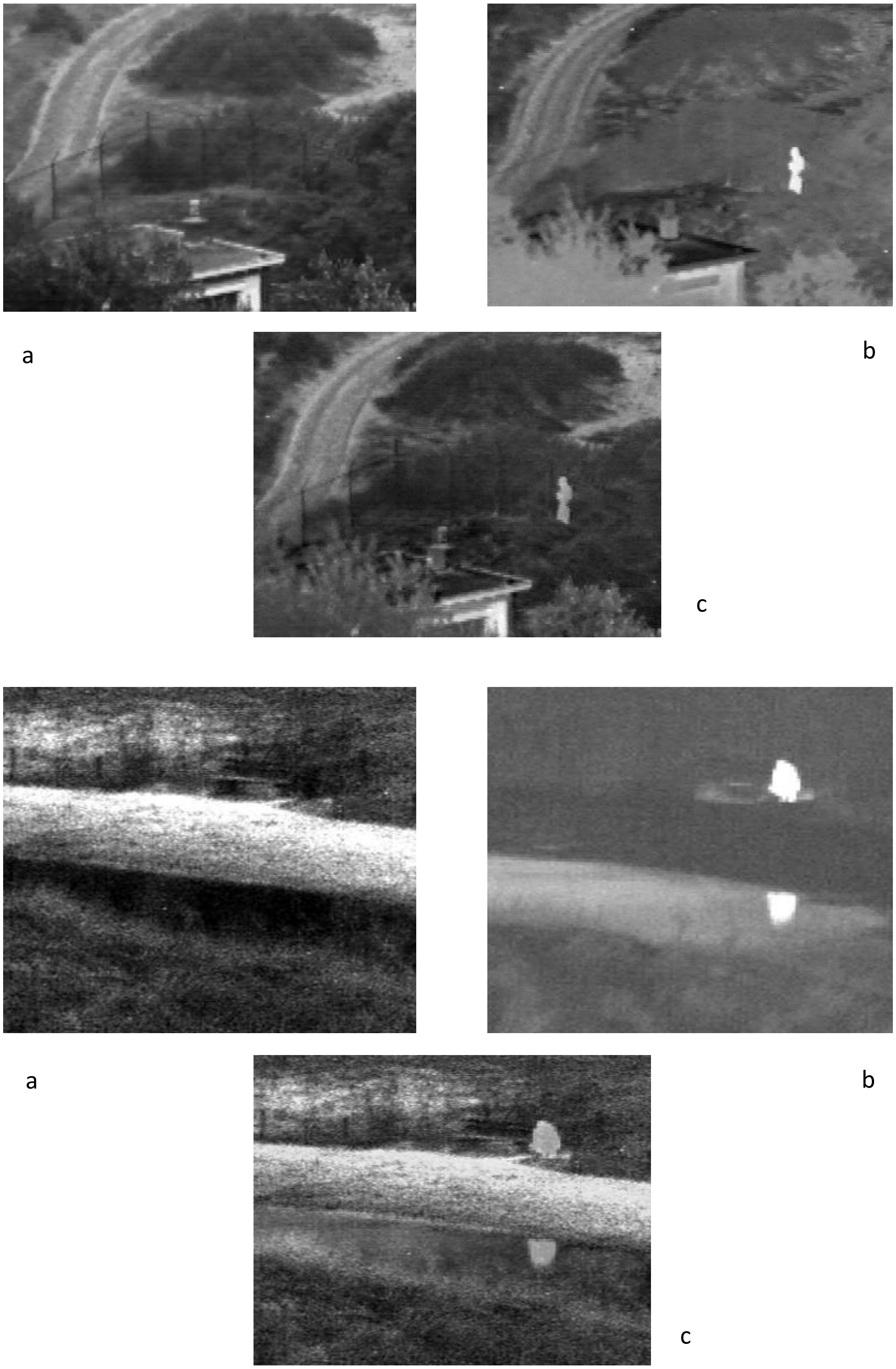


Figure 5.3: (a) Visible image (b) Thermal image (c) Fused image

5.3 Classification Results

Resnet-152 is used in this paper for fusion of images and classification of KAIST dataset. The reason behind choosing Resnet over other CNN architectures was that it is most accurate CNN architecture according to ILSVR challenge with a top 5 error rate of 3.57 percent . There are two types of blocks in deeper models, one is plain block and the other is residual block. Plain blocks maps input, X directly to the H (x) by passing through some stacked layers. Residual blocks use residual mapping of F(x) = H(x) – x, instead of direct mapping to H(x) we try and learn some function F(x) that needs to be added in x to get H(x). In this way we can find H (x) indirectly. A skip connection is built from x to be added to next block having the residual content including weight layers. If no weight layer is used in the network, then skip connection acts as an identity, but when weights are included, then learned weights are added in the next block. Resnet is composed of residual blocks stacked on top of each other.

We are using three different classification models. Evaluation of resnet is done on three models, i.e.Visible, thermal and fused models. For classification fused model is designed using Resnet architecture.

5.3.1 Evaluation metrics

There are two types of evaluation metrics we chose for result evaluation.First is training and testing accuracy plots of RGB, Thermal and Fused models for comparison of results. Second is confusion matrices for each model. Confusion matrices provide four types of results: true positives (TP), true negative (TN), false positive (FP), false negative (FN). In our case we are referring negatives as person class and positives background or no person class. True positive TP is the prediction in which the region inside the bounding box is a no-person class and our CNN model detect it as a no-person class. True negative TN is a case in which the region in the box is a person and detector also detects it as a person. False positive FP is a detection in which the region inside the bounding box is detected as a person when there is no person in the box or there is actually a person inside the box but ground truth annotation does not recognize it. False negative FN are those regions which are detected as background, but actually there is a person in it. Figure 5.5 shows the illustration of the confusion matrix.

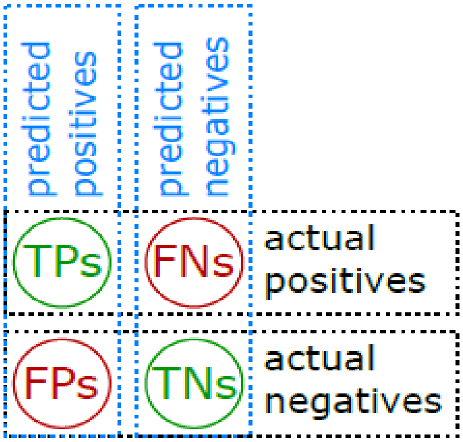


Figure 5.4: Confusion matrix representation

With the help of confusion matrix we calculated multiple other evaluation metrics for detailed analysis of results. For calculation of these parameters we applied following formulas.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5.1}$$

$$Miss\ Rate = \frac{FP+FN}{TP+TN+FP+FN} \tag{5.2}$$

$$TPR = \frac{TP}{TP+FN} \tag{5.3}$$

$$FPR = \frac{FP}{FP+TN} \tag{5.4}$$

$$SPE = \frac{TN}{TN+FP} \tag{5.5}$$

Proposed fusion architecture consists of two sub networks, which are fused by Late fusion technique. One sub network gets Visible image input while the other gets LWIR/Thermal image input. Both networks extract features from their respective input images. These feature arrays are then extracted and saved, which are later loaded in a separate network with just one dense/Fully connected layer. Before applying dense layer, feature arrays extracted from Visible and Thermal models are concatenated. Concatenated array is taken as input to dense layer.

Finally the model is tested on the test feature array. Figure 5.4 shows the CNN architecture of the proposed technique. This architecture consists of a detailed overview of ResNet-152 used in our technique. Resnet-152 consists of five convolution blocks which are iterated according to the proposed ResNet method, and a fully connected layer as last layer after average pooling layer.

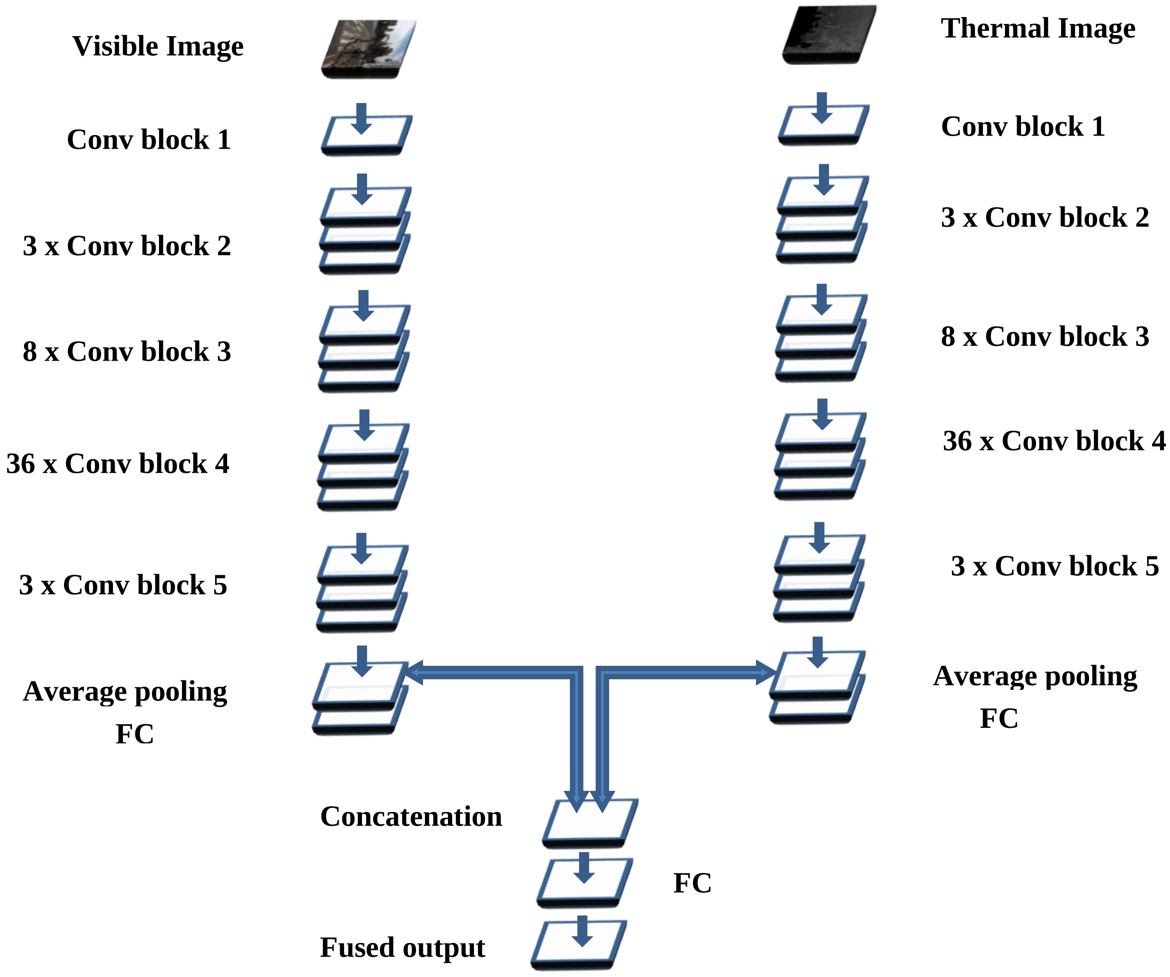


Figure 5.5: Fusion architecture using ResNet-152

5.3.2 Pre-training and fine-tuning

Before training our sub models on KAIST dataset, we need to pre-train Resnet for efficient results. For this purpose, we used a Resnet model which was pre-trained on imagenet [19]. Classification model of Resnet takes image labels as 0 or 1 to detect if there is a person in the given image (1) or there is no person in the image (0). To label our dataset according to binary labels we had to re-annotate our data. KAIST dataset annotations were in the form of bounding box location point of the objects present in the image. Matlab was used to convert these annotations to simple binary labels. "bbGt version=3s" string showed that there is no object in the image while strings like "bbGt version=3 person 446 221 27 87 0 0 0 0 0 0 0" represents the presence of objects like a person or cyclist. A simple code was written to read the annotation string of each image and generate the output label 1 if the string exceeds the limit of 16. After a generation of labels for all the images in KAIST, the data set was divided into training, validation and testing sets by the ratio of 60 percent training, 20 percent validation and 20 percent testing data. From KAIST dataset 50.2k images were used for training while for testing 40.1k images were used.

KAIST dataset consists of 95000 images. To handle such large dataset HDF5 data format was used. H5py arrays of training, testing and validation sets were created with shapes as: number of data, image height, image width and image depth according to tensorflow format. 3 channel input data with two numbers of classes, batch size 15 and 10 epochs were used for the training process. Training, testing and validation arrays were generated for labels with shape: number of data. After creating the arrays, images were read from a database and resized from 640x512 to 224x224 pixels. The purpose of resizing is that neural network only takes image input of resolution 224x224. Now the saved images in h5py arrays can be read in batches. These h5py arrays were used in the fine-tuning process. Loaded batches were used as input to the Resnet-152 for fine-tuning. Visible model was fine-tuned on visible images from KAIST and its feature maps from convolution layers were saved as numpy array. Similarly LWIR model is fine- tuned over thermal images from KAIST and its feature maps were saved.

Once the fine tuning is done. A separate model is created with just one dense layer which is also known as a fully connected layer. Softmax activation is applied with dense layer for classification.

Table 5.2 shows a comparison of different evaluation metrics from thermal, visible and fused architectures. It is clearly seen that the accuracy of fused model is greater than other two. Figure 5.6 shows the confusion matrices for each model and their normalized forms as well. The miss rate calculated from confusion matrix is 0.7%, which is 9.9% better than the last best model evaluated over KAIST multi-spectral data. Figure 5.7 shows the training and testing accuracies of visible, thermal and Fusion models.

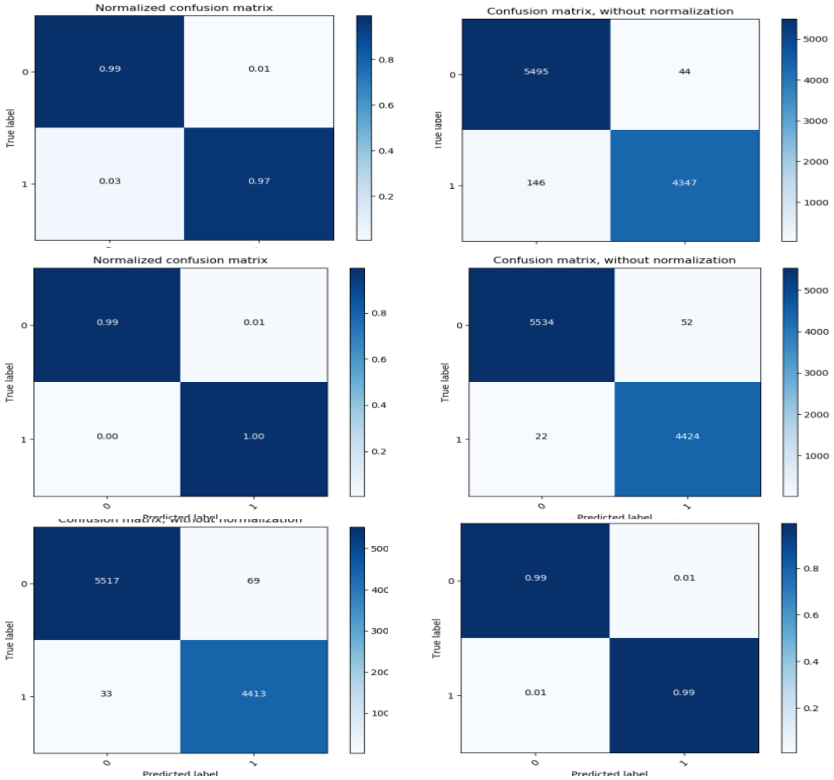


Figure 5.6: Confusion matrix for visible, thermal and fused models

Table 5.2: evaluation matrices

Models	Accuracy (%)	Miss Rate (%)	TPR (%)	FPR (%)	Specificity (%)
Thermal	98.98	1.0	99.40	0.73	98.4
Visible	98.16	1.8	97.41	3.32	98.99
Fused	99.26	0.7	99.71	0.73	98.70

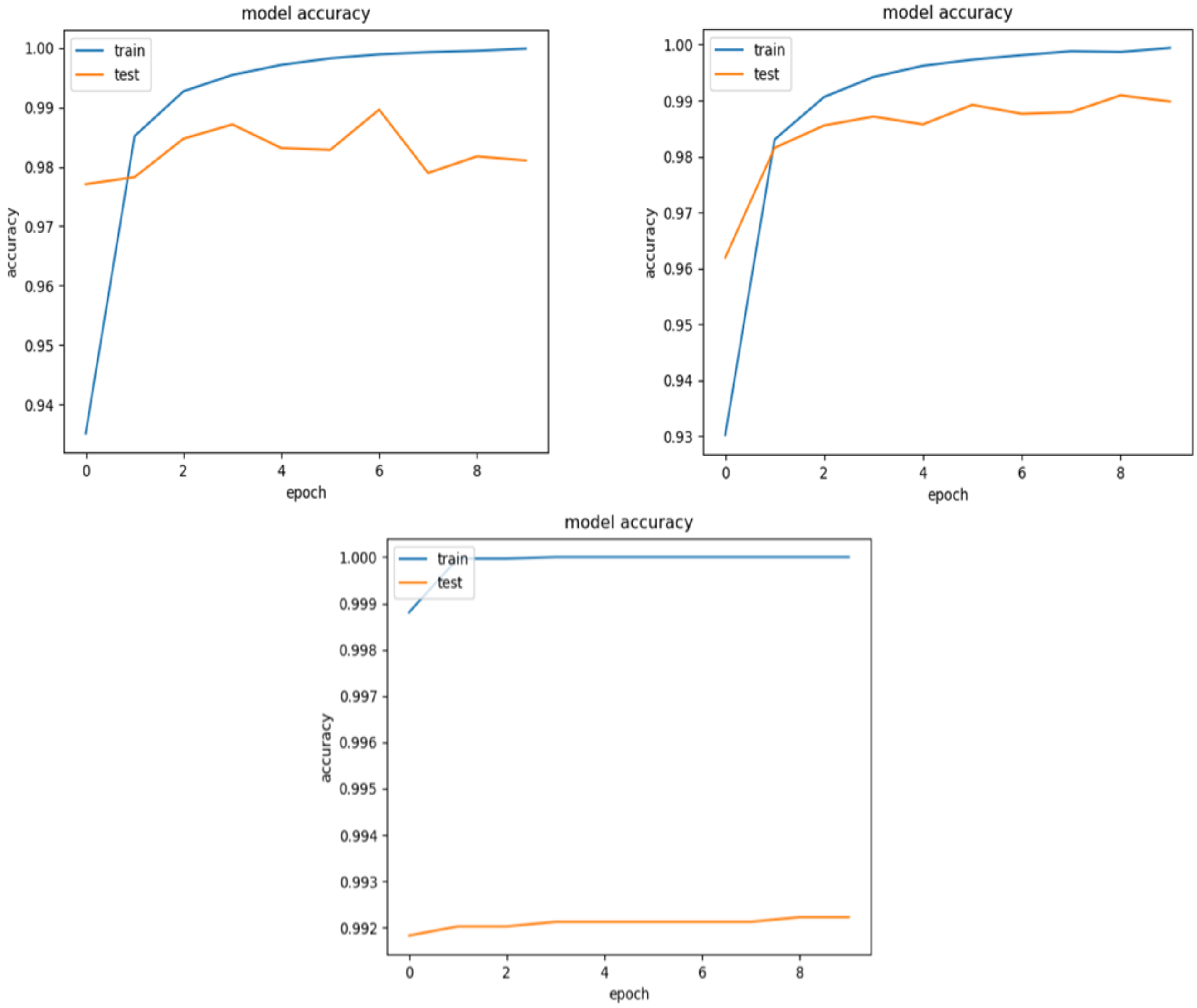


Figure 5.7: Training and Testing accuracy of visible, thermal and fused models

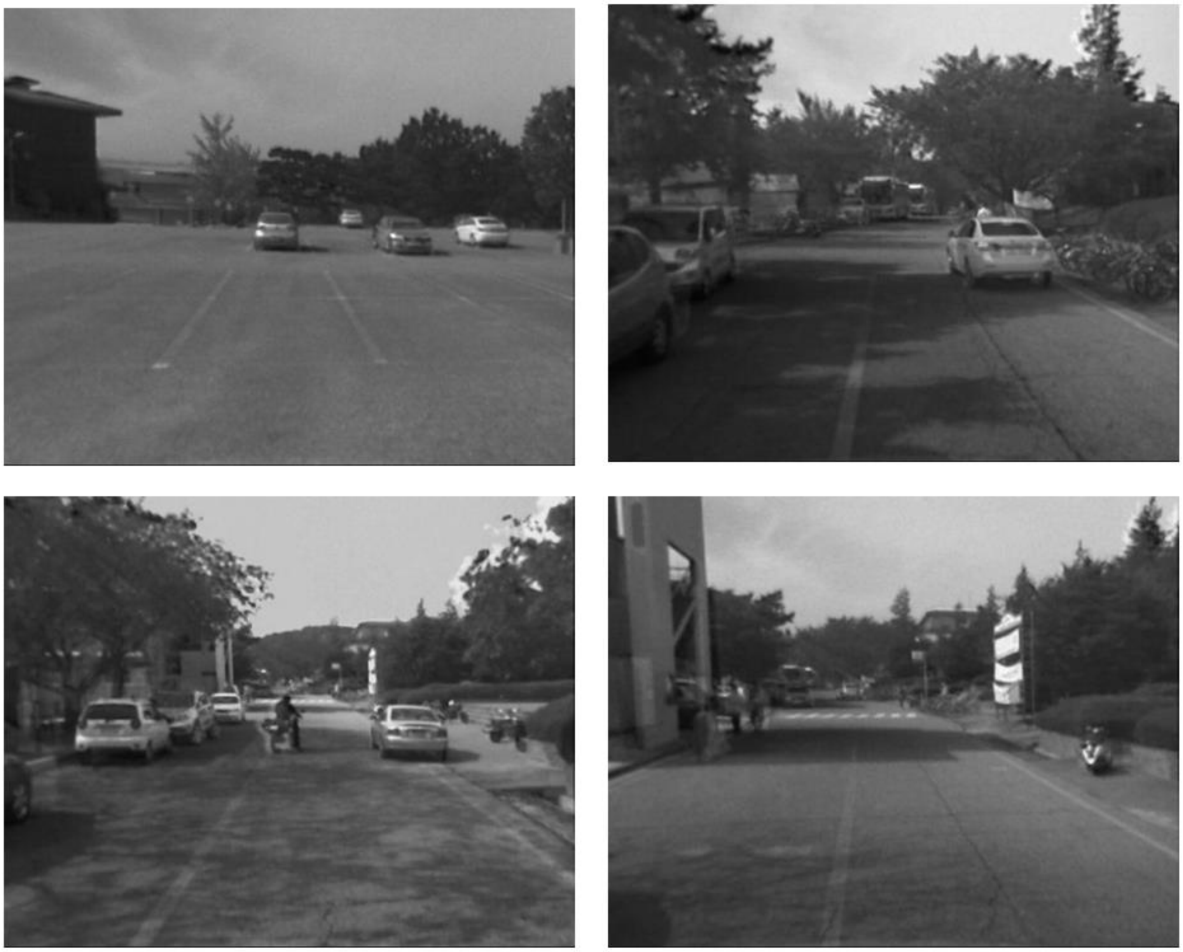


Figure 5.8: Images without persons, Row 2: Images with persons

The classification results are divided into two parts: Resultant images having persons present in them are stored in one folder while images having no persons present in them are stored separately. This step is done to separate instances for localization module. Figure 5.8 shows two pairs obtained from classification module. First row contains images which were classified as class 0 with no person present in them. Second row contains images which are classified as class 1 with a person or number of persons present in them

5.4 Localization Results

The images having persons present in them is then transferred into the Mask-RCNN model for localization.For training KAIST dataset provides bounding box annotations, but in case of Mask-RCNN we need binary masks for training. For this purpose, we used the VIA VGG annotator tool. Using this interface we created image masks for our data set. VIA tool saves json model for training and validation data sets separately. There is no need for extensive training by a huge bulk of data because Mask-RCNN model can utilize pre-trained weights of imagenet and MS-COCO. ResNet architectures need high computation power in the training process. NVIDIA GPU GTX-1080 Ti with 64 GB RAM is used in the training process.. Mask-RCNN provides promising results for localization of objects.

The Figure 5.9 shows the results of Mask-RCNN over KAIST multi-spectral dataset and local dataset. Images in the first two columns are from KAIST dataset while in images in the third and fourth column are from local dataset. Row a consists of Visible images and their respective detections are shown in row b. It can be seen that there are a lot of False detections in visible image present in the first column of the row. While its respective fused image and its detection scan be seen in c and d. First column of row d shows that there is only one false detection in the fused image. From fourth column it is noted that the person present inside the box is hidden while in fused image it can be seen partially. Detection results show that in visible image the person inside the box remains undetected when passed through Mask-RCNN model while both the persons are detected in fused image. There are some false detections in the fused image as the model is trained over KAIST and these test images are randomly chosen.

Table 5.3: Comparison of Localization techniques

S. No.	Author	Year	Technique	Data set	Reported results
1	Jorg Wagner et al	2016	CNN	KAIST	43.80% Miss Rate
2	Jingjing Liu t al.	2016	Faster-RCNN	KAIST	37% Miss Rate
3	Daneil Konig et al.	2017	Fusion RPN + BDT	KAIST	29.83% log Avg. Miss rate
4	Dan Xu. Et al.	2017	CMT-CNN	KAIST	10.69% Miss Rate
5	Ours	2019	Encoder-Decoder Mask-RCNN	KAIST	5.25% Miss Rate

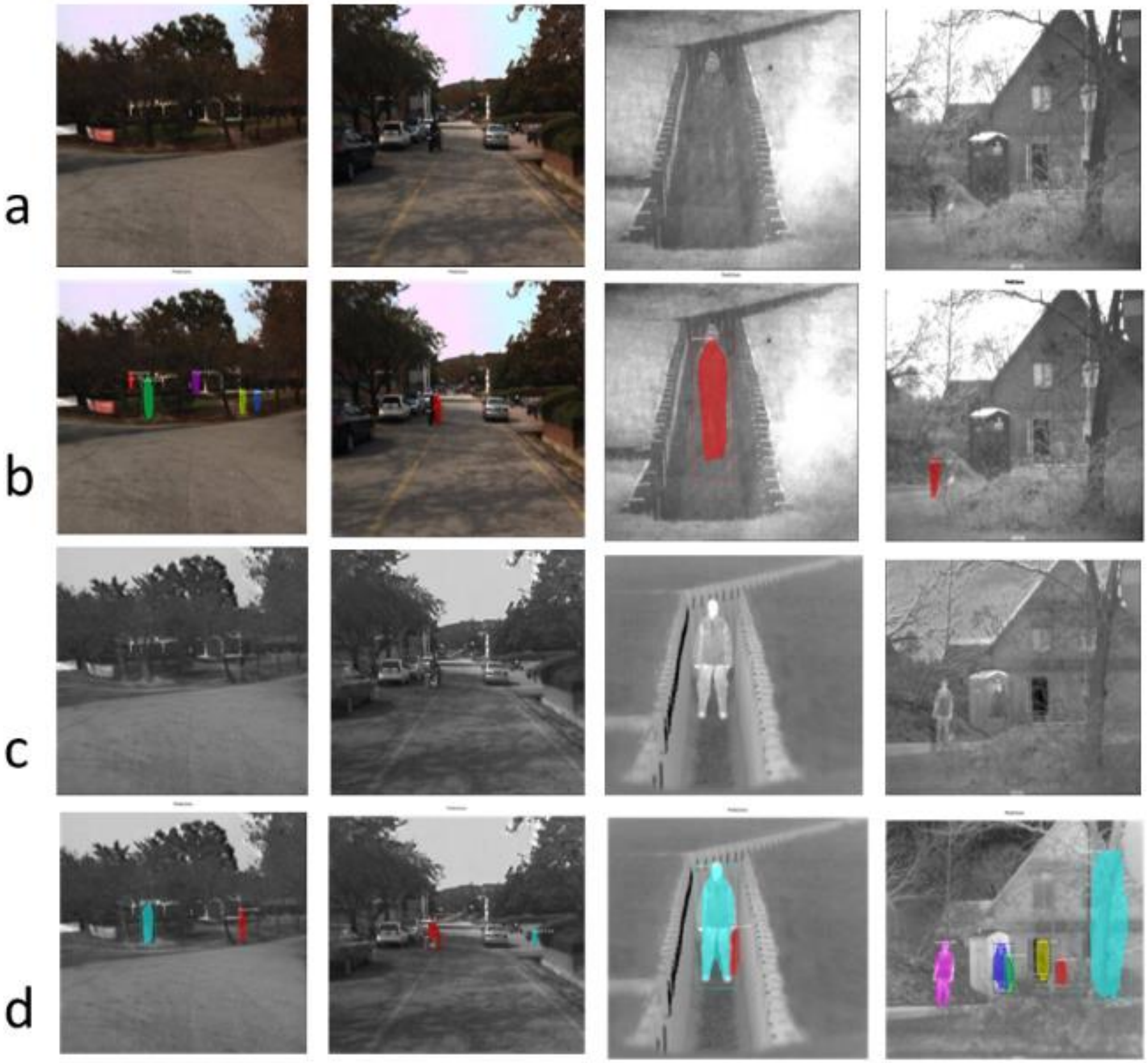


Figure 5.9: (a) Visible images (b) Localization of a (c) Fused images (d) Localization of b

Chapter 6 : CONCLUSION & FUTURE WORK

6.1 Conclusion

The purpose of this thesis is to utilize modern technology and computer vision models for efficient surveillance and the provision of foolproof security in organizations, schools, hospitals and military zones. For this purpose, we are utilizing visible and thermal cameras to obtain images of the premises as well as heat maps of suspicious intruders. These images are then fused together to get a combined more informative output for detection of a doubtful presence. The motivation behind fusing both kinds of images is to have clear information for the purpose of person detection during day and night time. During day time visible images come in handy while during night time, foggy weather, storms and occlusion cases, thermal (LWIR) images provide better object information. The approach chosen for fusion is a convolution neural network because of its efficiency and gives accurate results. We are using Encoder-decoder CNN architecture for fusion of visible and thermal images and ResNet architecture for object detection and localization. Localization of object or persons is done using Mask-RCNN model which not only localizes the object, but also provides a mask for localized object. KAIST multi-spectral data is used for the training of CNNs and local dataset is also used in the testing process. When the results of visible detections are compared with results of Fused detections, it is clearly observed Fused model outperforms the detection and localization process by giving accurate masks for KAIST and comparatively better masks for objects than Visible model.

6.2 Contribution

- Review & comparison of recent developments in object detection and localization systems using a convolutional neural network.
- Fully automated system for fusion of visible and thermal images using a convolutional neural network.
- Fully automated system for classification and localization of objects from fused images for the purpose of safety and security.

6.3 Future Work

The system proposed by us is quite efficient for fusion and classification and provides optimal masks for each instance present in the frame. This method avoids the occlusion problem efficiently. However the localization module can be improved by training the network on a diverse dataset. This system can be trained and modified a little to be used for detection of any kind of object. There is a little overhead of training time in localization module due to 101 layers of ResNet which can be minimized by using a less deeper ResNet model.

References

[1] J. Ratcliffe, "Video surveillance of public places," 2006.

[2] Wagner, Jorg and Fischer, Volker and Herman, Michael and Behnke, Sven, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2016.

[3] [Online]. Available: https://www.fbi.gov/news/stories/latest-crime-stats-released/latest-crime-stats-released.

[4] Mishra, Dhirendra and Palkar, Bhakti, "Image Fusion Techniques: A Review," *International Journal of Computer Applications ,* vol. 130, pp. (0975--8887), 2015.

[5] Wagner, Jorg and Fischer, Volker and Herman, Michael and Behnke, Sven, "Multispectral pedestrian detection using deep fusion convolutional neural networks," *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN),* pp. 509-514, 2016.

[6] Liu, Jingjing and Zhang, Shaoting and Wang, Shu and Metaxas, Dimitris N, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644,* 2016.

[7] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 770-778, 2016.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell and Jitendra Malik, UC Berkeley, Ross Girshick-Jeff Donahue-Trevor Darrell-Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation, Tech report (v5)," IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner., "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998.

[10] [Online]. Available: http://deeplearning.net/tutorial/_images/mylenet.png.

[11] Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton, "ImageNet classification with deep convolutional neural networks," in *Communications of the ACM*, 2017.

[12] [Online]. Available: https://cdn-images-1.medium.com/max/1536/1*qyc21qM0oxWEuRaj-

XJKcw.png.

[13] Karen Simonyan & Andrew Zisserman, "Very deep Convolutional networks for large scale image recognition," *arXiv:1409.1556,* 2015.

[14] [Online]. Available: data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAASsAAACoCAMAAACPKThEAAABPlBMVEX////u7u7x8fG+vr709PSoqKgAAADv7e6UlJT//v+enp65ubmrq6vCwr+0tLTt7+yYmJjV1dWqrOnAv82rrMn08viqqrh2dsKLjb12dbjx8f/v7f+EhrWgoadxcaTb29yMjIzHx8fOzs7h4eFwcHCFhcCEhIRtbb+amsDe39p6enpkZGRZW.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] [Online]. Available: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcTxBjzMtMmnS5Av1_0Z-SvRONLOqd7XV40xH9bveoXaKZcgb3sh.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] [Online]. Available: https://cdn-images-1.medium.com/max/1600/1*D0F3UitQ2l5Q0Ak-tjEdJg.png.

[19] [Online]. Available: https://github.com/KleinYuan/tf-object-detection.

[20] [Online]. Available: https://www.google.com/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwizqaK_s6HgAhUsy4UKHUDEB6YQjRx6BAgBEAU&url=https%3A%2F%2Fwww.slideshare.net%2Fanirudhkoul%2Fsqueezing-deep-learning-into-mobile-phones&psig=AOvVaw3aJKRT6z-wy1_Ly6Y57jIO&ust=1549346.

[21] Matthias Limmer,Hendrik Lensch, "Infrared Colorization Using Deep Convolutional Neural Networks," 2016.

[22] TeresaArauÂ jo, Guilherme Aresta, EduardoCastro, JoseÂ Rouco,Paulo Aguiar, Catarina Eloy, AntoÂ nio PoloÂ nia, AureÂ lio Campilho, "Classification of breast cancerhistology images using ConvolutionalNeuralNetworks," *DOI: 10.1371/journal.pone.0177544,* 2017.

[23] Gustav Larsson, Michael Maire, Gregory Shakhnarovich, *Learning Representations for Automatic Colorization,* Computer Vision – ECCV 2016 Lecture Notes in Computer Science, 2016.

[24] ”, R. Niessner, H. Schilling, B. Jutzi, "Investigations On The Potential Of Convolutional Neural Networks For Vehicle Classification Based On Rgb And Lidar Data," in *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2017.

[25] [Online]. Available: http://www.adobe.com/en/products/aftereffects.html, Refine Edge tool in Adobe After Effects CC..

[26] Ashnil Kumar, Jinman Kim, David Lyndon, Michael J. Fulham, David Dagan Feng, "An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification," *IEEE J. Biomedical and Health Informatics 21 (1),* pp. 31-40, 2017.

[27] Ronald Kemker, Carl Salvaggio, Christopher Kanan, "Algorithms for Semantic Segmentation of Multispectral Remote Sensing Imagery using Deep Learning," *arXiv: 1703.06452v2 [cs.CV], ,* 2017.

[28] Jiwen Lu, Gang Wang, Weihong Deng, Pierre Moulin, Jie Zhou, "Multi-manifold deep metric learning for image set classification," in *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2015.

[29] Simon Philipp Hohberg, "Wildfire Smoke Detection using Convolutional Neural Networks," 20 09 2015. [Online].

[30] Jin Kyu Kang, Hyung Gil Hong and Kang Ryoung Park, "Pedestrian Detection Based on Adaptive Selection of Visible Light or Far-Infrared Light Camera Image by Fuzzy Inference System and Convolutional Neural Network-Based Verification," 2017.

[31] Natalia Neverova, "Deep learning for human motion analysis," 2016. [Online].

[32] Samer Hijazi, Rishi Kumar, and Chris Rowen, "Using Convolutional Neural Networks for Image Recognition," in *IP Group, Cadence*.

[33] C. a. J. L. V. G. Pohl, "Review article multisensor image fusion in remote sensing: concepts, methods and applications," 1998.

[34] Er-Yang Huan, Gui-Hua Wen, Shi-Jun Zhang, Dan-Yang Li, Yang Hu,Tian-Yuan Chang, Qing Wang, and Bing-Lin Huang, "Deep Convolutional Neural Networks for Classifying Body Constitution based on face image," 2016.

[35] Dhirendra Mishra, Bhakti Palkar, "Image Fusion Techniques: A Review".

[36] Mohammad Hanif, Usman Ali, "Optimized Visual and Thermal Image Fusion for Efficient Face Recognition," in *9th International Conference on Information Fusion*, 2006.

[37] Shuo-Li Hsu, Peng-Wei Gau, I.-Lin Wu, Jyh-Horng Jeng, "Region-Based Image Fusion with Artificial Neural Network," *Proceedings of World Academy of Science: Engineering & Technolog,* vol. 53, p. 156, 2009.

[38] Liuhao G, Hui Liang, Junsong Yuan, Daniel Thalmann, "3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single, Depth Images," 2017.

[39] Habibollah Agh Atabay, "Binary Shape Classification Using Convolutional Neural Networks," *IIOAB J. 7 (5),* pp. 332-336, 2016.

[40] Sachin Sudhakar Farfade, Mohammad Saberian, Li-Jia Li, "Multi-view Face Detection Using Deep Convolutional Neural Networks," vol. 3, 2015.

[41] Shiming Ge, Jia Li, Qiting Ye, Zhao Luo, "Detecting Masked Faces in the Wild with LLE-CNNs," 2017.

[42] Shutao Li, James Kwok, Yaonan Wang, "Multifocus image fusion using artificial neural networks," 2002.

[43] Karen Simonyan, Andrew Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *NIPS 2014,* pp. 568-576, 2014.

[44] Andreas Eitel, Jost Springenberg, Luciano Spinello, Martin Riedmiller, Wolfram Burgard,, "Multimodal deep learning for robust RGB-D object recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[45] F. V. H. M. B. S. Wagner J., "Multispectral pedestrian detection using deep fusion Convolutional neural networks," in *Proceedings of the European Symposium on Artificial Neural Networks;*, Bruges, Belgium, 2016.

[46] Liu J., Zhang S., Wang S., Metaxas N., "Multispectral Deep Neral Network for Pedestrian Detection," in *Multispectral DNNs for pedestrian detection*, 2016.

[47] Xu D., Ouyang W., Ricci E., Wang X., Sebe N, "Learning Cross-Modal Deep Representations for Robust Pedestrian Detection," Hong Kong.

[48] Y. a. Y. B. LeCun, "Convolutional networks for images, speech, and time series," 1995.

[49] [Online]. Available: https://soonminhwang.github.io/rgbt-ped-detection/.

[50] ”,Hui Li, Xiao-Jun Wu, "DenseFuse: A Fusion Approach to Infrared and Visible Images," vol. 5, 2018.

[51] [Online]. Available: https://www.researchgate.net/publication/324717482_DenseFuse_A_Fusion_Approach_to_Infrared_and_Visible_Images.

[52] [Online]. Available: https://www.google.com/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwimy-SjuaHgAhVHxxoKHfcXDNUQjRx6BAgBEAU&url=%2Furl%3Fsa%3Di%26source%3Dimages%26cd%3D%26ved%3D%26url%3Dhttps%253A%252F%252Ftowardsdatascience.com%252Fdeep-learning-d5fe55326e57%26ps.

[53] [Online]. Available: https://qph.fs.quoracdn.net/main-qimg-78a617ec1de942814c3d23dab7de0b24.

[54] [Online]. Available: https://www.google.com/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwjitoDNuaHgAhXmzIUKHZfxA_0QjRx6BAgBEAU&url=https%3A%2F%2Fhackernoon.com%2Fone-shot-learning-with-siamese-networks-in-pytorch-8ddaab10340e&psig=AOvVaw0efctyUvzfBANvdN1CbHyl&ust=1549.

[55] Wu D., Yang A., Zhu L., Zhang C. (2014) In: Ma S., Jia L., Li X., Wang L., Zhou H., Sun X, "Survey of Multi-sensor Image Fusion," in *Life System Modeling and Simulation. ICSEE 2014, LSMS 2014. Communications in Computer and Information Science, vol 461. Springer,* , Berlin, Heidelberg, 2014.

[56] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun,, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," vol. 1, 2015.

[57] [Online]. Available: https://cdn-images-1.medium.com/max/1600/1*1y9hueMSZAeo1Hbp9KYKiw.png.

[58] [Online]. Available: https://cdn-images-1.medium.com/max/1600/1*WVs9ywVLLKjSUBZ_mnfFrw.png.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition," 2015.

[60] [Online]. Available: https://www.google.com/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwib3_Wlu6HgAhWrxYUKHeo2COAQjRx6BAgBEAU&url=%2Furl%3Fsa%3Di%26source%3Dimages%26cd%3D%26ved%3D%26url%3Dhttps%253A%252F%252Fdatascience.stackexchange.com%252Fquestions%252F33022%252F.

[61] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, , "Rich feature hierarchies for accurate object detection and semantic segmentation," vol. 5, 2014.

[62] [Online]. Available: https://www.google.com/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwjs_qi3u6HgAhUDLBoKHYPYBiAQjRx6BAgBEAU&url=%2Furl%3Fsa%3Di%26source%3Dimages%26cd%3D%26ved%3D%26url%3Dhttps%253A%252F%252Fhackernoon.com%252Fevolution-of-image-recognition-and-obje.

[63] [Online]. Available: https://medium.com/@umerfarooq_26378/from-r-cnn-to-mask-r-cnn-d6367b196cfd.

[64] R. Girshick, "Fast R-CNN," vol. 2, 2015.

[65] [Online]. Available: https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwj3pcXdvKHgAhXryIUKHTh-CzkQjRx6BAgBEAU&url=https%3A%2F%2Fwww.slideshare.net%2FJinwonLee9%2Fpr12-faster-rcnn170528&psig=AOvVaw24qA-hChI3p8UeXaqKYx5T&ust=1549348942404518.

[66] [Online]. Available: https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwiQ3r_6vKHgAhVM3RoKHdT6D9gQjRx6BAgBEAU&url=https%3A%2F%2Flilianweng.github.io%2Flil-log%2F2017%2F12%2F31%2Fobject-recognition-for-dummies-part-3.html&psig=AOvVaw1OKszHaMkD3sAs82yTS4mV&ust=1549348.

[67] [Online]. Available: https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwiC_-aKvaHgAhWXgM4BHRf2AkoQjRx6BAgBEAU&url=https%3A%2F%2Flilianweng.github.io%2Flil-log%2F2017%2F12%2F31%2Fobject-recognition-for-dummies-part-3.html&psig=AOvVaw2c9vKYmryzDEtNhkMpZjeJ&ust=1549349.

[68] ”, Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, "Mask R-CNN," *springer,* vol. 3, 2018.

[69] [Online]. Available: https://www.google.com/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=&url=https%3A%2F%2Fmedium.com%2F%40jonathan_hui%2Funderstanding-feature-pyramid-networks-for-object-detection-fpn-45b227b9106c&psig=AOvVaw0HMez_B7D_lUErgvb3fOL7&ust=1549349070578142.

[70] [Online]. Available: https://www.google.com/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwjlnpGpv6HgAhWKo48KHaB7B6sQjRx6BAgBEAU&url=%2Furl%3Fsa%3Di%26source%3Dimages%26cd%3D%26ved%3D%26url%3Dhttps%253A%252F%252Fsites.google.com%252Fsite%252Fpedestrianbe

nchmark%252F%26p.

[71] [Online].