Multi-Spectral Image Classification using Deep Learning Techniques

A thesis submitted in the partial fulfilment of the requirement for the

Degree of Master of Computer Science and Engineering

of

Jadavpur University

By

PRIYAM SARKAR

Registration Number: 140741 of 2017-2018

Examination Roll Number: M4CSE19011

Under the guidance of

Dr. Nibaran Das

Associate Professor

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

May, 2019

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

JADAVPUR UNIVERSITY

TO WHOM IT MAY CONCERN

This is to certify that the thesis entitled "MULTI-SPECTRAL IMAGE CLASSIFICATION USING DEEP LEARNING TECHNIQUES" has been satisfactorily completed by Priyam Sarkar (University Registration No.: 140741 of 2017-18, Examination Roll No.: M4CSE19011). It is a bonafide piece of work carried out under my guidance and supervision and be accepted in partial fulfilment of the requirement for the Degree of Master of Computer Science and Engineering, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, Kolkata.

____

Dr. Nibaran Das (Thesis Supervisor)

Associate Professor
Department of Computer Science and Engineering
Jadavpur University, Kolkata-700032

Countersigned:

____

Prof. Mahantapas Kundu

Head, Department of computer Science and Engineering
Jadavpur University, Kol-700032

____

Prof. Chiranjib Bhattacharjee

Dean, Faculty of Engineering and Technology
Jadavpur University, Kol-700032

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

DECLARATION OF ORIGINALITYAND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains literature survey and original research work done by the undersigned candidate, as part of my ME studies.

All information in this document have been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material results that are not original to this work.

Name	: PRIYAM SARKAR
University Registration No.	: 140741 of 2017-18
Examination Roll No.	: M4CSE19011
Thesis Title	: MULTI-SPECTRAL IMAGE CLASSIFICATION USING DEEP LEARNING TECHNIQUES

Signature with date

JADAVPUR UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE OF APPROVAL

The foregoing thesis is hereby accepted as credible study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the thesis only for the purpose for which it is submitted.

____

Signature of Examiner

Date:

____

Signature of Supervisor

Date:

ACKNOWLEGEMENT

First and foremost, I would like to start by thanking God Almighty for showering me with the strength, knowledge and potential to embark on this wonderful journey and to persevere and complete the embodied research work satisfactorily.

I am pleased to express my deepest gratitude to my thesis guide, **Dr. Nibaran Das**, Department of Computer Science and Engineering, Jadavpur University, Kolkata for his invaluable guidance, constant encouragement and inspiration during the period of my dissertation.

I am highly indebted to **Jadavpur University** for providing me the opportunity and the required infrastructure to carry on my thesis. I am also grateful to the **Center for Microprocessor Applications for Training Education and Research** for giving me the proper laboratory facilities as and when required. I am thankful to all the teaching and non-teaching staff whose helping hands have smoothed my journey through the period of my research.

Last but not the least; I would like to thank my family members, classmates, seniors and friends for giving me constant encouragement and mental support throughout my work.

____

Priyam Sarkar

University Registration No.: 140741 of 2017-18
Examination Roll No.: M4CSE19011
Master of Computer Science and Engineering
Department of Computer Science and Engineering
Jadavpur University

CONTENTS

CHAPTER ONE

INTRODUCTION

The field of computer vision has developed considerably in the last couple of decades. Especially since the introduction of deep learning approaches like convolutional neural networks [1], many complicated tasks such as object detection [2], localization [3], segmentation [4], natural scene [5] understanding have received a significant boost in terms of performance, speed, and scalability. However, deep learning approaches are heavily dependent on the availability of an abundant amount of good quality data. In this work, we introduce a Multi-spectral image dataset with three spectrums, namely, visual spectrum (RGB), near infrared spectrum and thermal spectrum. Multi-spectral images are used in various domains such as surveillance, medical imaging, aerial imagery, remote sensing, and so on. Our eyes are capable of processing only a limited spectrum of light. With multi-spectral information, many other properties of objects can be utilized to enhance computer vision applications. Spectrums beyond the visible spectrum can reveal many properties such as surface quality and thermal characteristics. The proposed dataset not only brings in 3 different spectrums into the learning process but also other associated factors that come with working with different sensors such as different focal lengths, perspectives, sensitivity to noise, image resolution, sharpness and so on. The purpose of the dataset is to provide researchers with access to multiple spectrums to access various properties that might not be available in the visible spectrum such as surface quality, texture, material properties, thermal characteristics, and so on.

1.1 DEFINITION OF MULTI-SPECTRAL IMAGE

The electromagnetic spectrum of the image data captured within a specific wavelength ranges is multi-spectral image. The captured data or wavelength are disunited by filters or by the utilization of sensors that are sensitive to particular data or wavelengths, including light from frequencies beyond the visible light range , i.e. infrared and ultra-violet. When the human ocular perceiver fails to capture with its receptors for red, green, and blue the multi-spectral imaging can grab information from frequencies beyond the visible light range.

Accumulation of multiple monochrome images of the same scene is multi-spectral imagery. Each individual image is referred to as a band. Discretely this single-band image could be processed in a multi-spectral image. An Image with m number of bands, we can say that a vector of length m can represent the effulgence of each pixel in m-dimensional space.

In this modern era, we can overcome the disadvantages in multi-spectral images of having additional information to process, which leads to increase in the required computation time and memory significantly. However, currently with the availability of machine power, this not an issue in the field of computer vision.

1.2 IMPORTANCE OF MULTI-SPECTRAL IMAGE

Multi-Spectral imagery is a kind of photography that captures images of things that we cannot see. An RGB sensor captures the light that falls onto the sensor. The sensor captures images in the same way that our eye perceives color. To do this, our camera uses wideband filters to divide the light into three channels: red, green, and blue (RGB). On the other hand, a multi-spectral image grabs data that is neither visible to the human eye nor to a typical RGB sensor. Their utility in different fields is as follows.

- Space-Based Imaging: Infrared Images of the earth on space programs. Satellites that are used for observation of earth uses multi-spectral imaging cumulates into a single optical system.

- Military Target Tracking: A lot of unconventional threats are faced by the military. In this scenario, multi-spectral imaging will give advantage to monitoring the local terrain from a distant space. Its useful in locating improvised explosives, any movement of the enemy during the night time and depth of hidden bunkers could also be measured.
- Land Mine Detection: Multi-spectral images captured from drones on battlefields utilize the information to detect sensitive information like underground missiles and land mines.
- Ballistic Missile Detection: Multi-spectral images are utilized to intercept and detect the ballistic missiles.
- Document and Painting Analysis: Advantage of using multi-spectral imaging in document and painting Analysis is it interpret ancient papyri and other documents from antiquity by imaging the documents in the infrared range.
- Farming: Multi-spectral images captured from drones on agricultural fields utilize the information to detect sensitive information about soil, crops, fertilizing and irrigation to minimize the use of sprays, fertilizers, and irrigation while increasing the yield from the fields.
- Healthcare: Diagnostic medicine is a key area in healthcare where multi-spectral imaging is being utilized. It provides sharp information about the presence of diseases that is very challenging to detect.
- Forensics: Multi-spectral images are captured on a crime scene or in a forensic laboratory provide sensitive information for forensic evidence.
- Environmental Monitoring: Multi-spectral data from perilous and inaccessible areas utilized to monitoring glacial, deforestation, and depth sounding of the ocean and coastal depths.

1. 3 PROBLEMS THAT MAY BE SOLVED USING MULTI-SPECTRAL IMAGE

In the field of computer vision, material detection is a challenging problem. With the use of a multi-spectral image, we have a lot of information to work with, for material detection problems. In the real world scenario problems like the classification of materials in natural scenes [5] and scene category recognition [6] is quite challenging. In the era of machine learning, State-of-the-art technology like Deep Learning will come handy to solve this type of challenges with the availability of machine power.

CHAPTER TWO

PREVIOUS WORK

There are several recent datasets in the multi-spectral domain (see Table 1). Multi-spectral datasets on street images have been utilized before for many different fields such as is object detection, semantic segmentation. Multi-spectral datasets on natural scenes have been utilized before for many different fields such as color restoration, semantic segmentation, object localization, object tracking etc. Most of these multi-spectral datasets are composed of 2 spectrum among visible, near infrared or thermal.

Table 1. SOME OTHER DATASETS CORRESPONDING TO MULTI-SPECTRAL IMAGES

Authors	**Year**	**Spectrums**	**Type of Images**	**Objective**
Takumi et al. [7]	2017	RGB, NIR, MIR, FIR	Street Images	Object Detection
Ha et al. [8]	2017	RGB, Thermal	Street Images	Semantic Segmentation
Aguilera et al. [9]	2018	RGB, NIR	Natural Scenes	Color Restoration
Alldieck et al. [10]	2016	RGB, Thermal	Street Images	Object Detection
Brown et al. [6]	2011	RGB, NIR	Natural Scenes	Multi-spectral SIFT

Choe et al. [11]	2018	RGB, NIR	Natural Scenes	Semantic Segmentation
Davis et al. [12]	2007	RGB, Thermal	Natural Scenes	Background Subtraction
Hwang et al. [13]	2015	RGB, Thermal	Natural Scenes	Object Localization
Li et al. [14]	2018	RGB, Thermal	Natural Scenes	Object Tracking

CHAPTER THREE

THEORY AND EXPERIMENTAL SETUP

3. 1 DATASET PREPARATION

The dataset consists of objects or combinations of object categories captured in an indoor environment. The objects selected are mostly static objects that often occur on a desk. The objects are made up of single or multiple types of materials such as wood, plastic, glass, rubber, synthetic, and metal. In the subsections below the procedures related to the data acquisition is described in details.

3. 1. 1 CAPTURING DEVICES

The images were captured using three separate devices (Refer Figure 1) as described below.

- Visible Spectrum: Nikon D3200 DSLR Camera with Nikkor AF-S 3.5-5.6G standard lens.
- Near Infrared Spectrum: Watec WAT-902H2 Camera, 24mm lens (SV-EGG-BOXH1X), Schnieder 093 IR Pass Filter (830nm).
- Thermal Spectrum: FLIR A655SC Thermal Camera.

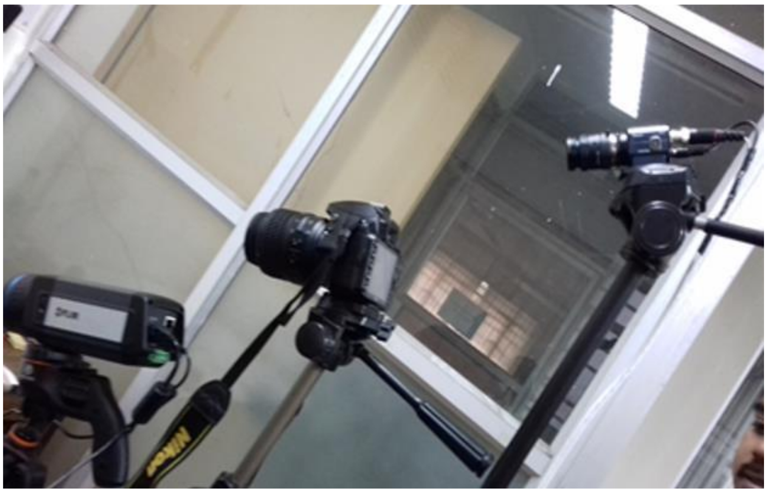


Figure 1: Camera Setup showing the cameras corresponding to the thermal, visible and near-infrared spectrum (left to right)

3. 1. 2 CAMERA SETUP

The three cameras (Refer Figure 1) are setup in a way such that the images are vertically aligned. The horizontal alignment was not possible as the cameras were mounted on tripods of different heights so that they do not block each other. The objects that were supposed to be captured were kept on a table covered with white paper. To minimize the shadow which cast while capturing photos in indoor lighting condition, we have mounted an infrared light source on the ceiling. The infrared source was kept when capturing with the other two cameras.

3. 1. 3 IMAGE CAPTURING PROTOCOL

The entire image capturing process was divided into two phases. The first phase was for one object at a time, and the second phase involved taking two objects at a time. Each different object or combination of the object was initially kept in the center of the table. Three vertical metallic markers were kept on around the object that shall be used later during the image registration process to provide reference points. One image is captured this way with each of the three cameras. Then the markers are removed, and 8 sets of images with each of the three cameras are captured. The 8 sets of images correspond to the 8 different orientation angles at which the object or the combination of objects are kept at the table. Among these 8 images, the first images correspond to the orientation of the marker image. From there the objects or the combination of objects are rotated clockwise at an angle of 45° before taking the next 3 captures from the 3 cameras. It is to be noted when the combination of objects are rotated their relative position is kept constant. For the single object phase we have a total of 25 object categories, 8 different angles, an additional image per object with the markers, and 3 different spectrums. That result in a total of 600 images (25 X 8 X 3) and 75 marker images (25 X 3). For the combination of 25 objects taking 2 objects at a time, we have a total of 300 different combinations ($^{25}C_2$). For each of them, we have 8 different angles, an additional image per combination with the markers and 3 different spectrums. That results in a total of 7200 images (300 X 8 X 3) and 900 marker images (300 X 3).

3. 1. 4 IMAGE REGISTRATION

Since the images of the different spectrum are captured with different cameras of different focal lengths and kept at different positions, the objects in the images of the different spectrum are not registered by default. In other words, there is no pixel to pixel correspondence across the RGB, NIR or Thermal images. During the capturing process of the dataset for every object or combination of objects, 8 different angles were captured along with 1 extra image with 3 long metallic markers kept vertically on the table surface surrounding the object. This marker image is used in the process of image registration. The tips of the 3 vertical markers are taken as a reference point to calculate an affine transformation matrix for that set of objects or combination of objects. Among the three spectrums, the RGB image is used as a fixed image, and the NIR and thermal images are treated as moving images that must be registered with the RGB image.

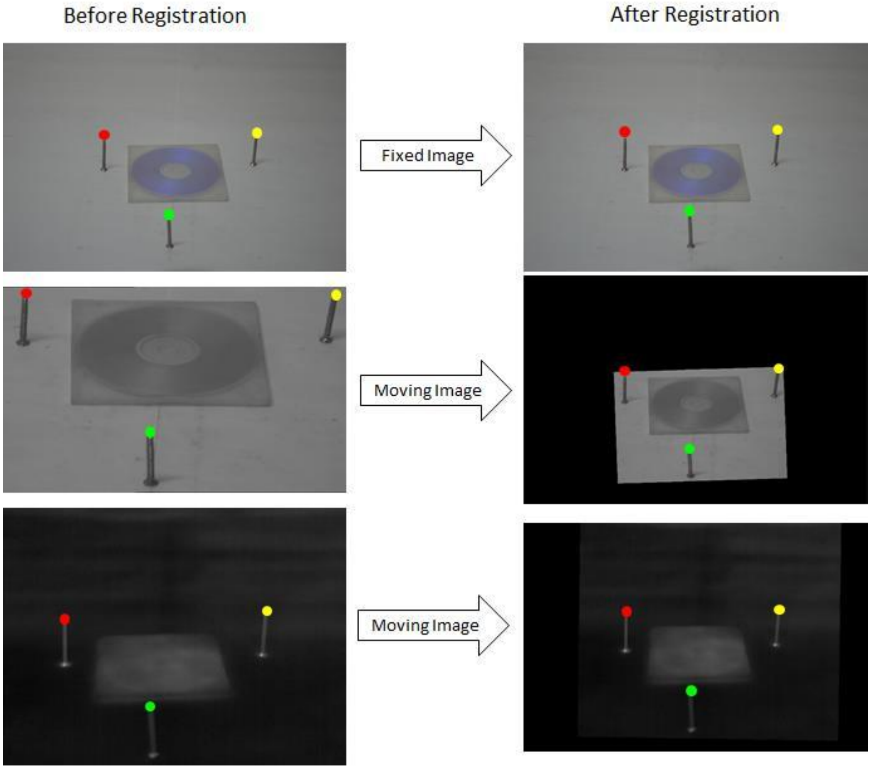


Figure 2: Image Registration by Affine2d transformation using 3 markers (colour coded in red, yellow and green)

Let P = (x_1, y_1), Q = (x_2, y_2), R = (x_3, y_3) be the vertices of a triangle in the space $\mathbb{R}^2$ and T be a transformation matrix which takes the to another triangle P' Q' R' whose vertices are P' = (x'_1, y'_1), Q' = (x'_2, y'_2), R' = (x'_3, y'_3).

Hence

$$T(x_i, y_i, 1) = (x'_i, y'_i; 1) \qquad \forall i = 1, 2, 3 \text{ ………............Equation 1}$$

Where (x_i, y_i, 1) is the extended representation of point (x_i, y_i). Writing the above equation in a matrix form we have

$$T\begin{bmatrix} x_1 & x_2 & 1 \\ y_1 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix} = \begin{bmatrix} x'_1 & x'_2 & 1 \\ y'_1 & y'_2 & 1 \\ x'_3 & y'_3 & 1 \end{bmatrix} \text{ …………………………………....Equation 2}$$

Since the points *P, Q, R* are non-collinear, the row vectors (x_i, y_i, 1) are linearly independent to each other. Hence the matrix $\begin{bmatrix} x_1 & x_2 & 1 \\ y_1 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix}$ is invertible (by the theorem at page 309 in [15]). Multiplying both side to above Equation 2 by the inverse of the matrix $\begin{bmatrix} x_1 & x_2 & 1 \\ y_1 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix}$ we have transformation matrix *T* as

$$\mathrm{T} = \begin{bmatrix} x'_1 & x'_2 & 1 \\ y'_1 & y'_2 & 1 \\ x'_3 & y'_3 & 1 \end{bmatrix} \begin{bmatrix} x_1 & x_2 & 1 \\ y_1 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix}^{-1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{Equation 3}$$

Using this transformation matrix, the other 8 images of the corresponding set are registered using the ’affine2d’ function in Matlab, that performs an affine transformation of an image based on the provided affine transformation matrix T. The registered image is further finely refined using manual supervision to further increase the quality of registration. The manual supervision is necessary as the affine transformation matrix can perform rotation, scaling or skewing but not translation. Though this procedure does not guarantee a pixel-perfect registration, it provides us with a decent enough set of images to work with. It should also be noted that since the objects are being captured by three different cameras looking at the image from a slightly different angle, it is technically impossible to perform accurate pixel registration due the different viewpoint perspectives.

3. 1. 5 IMAGE CROPPING

Once the images are approximately registered, they are cropped to remove unnecessary regions from the image that was introduced during the registration phase. Due to the different resolution, it was noticed that the NIR image shrunk more in size to fit with the RGB image as compared to the thermal image. Hence an optimal cropping window calculated on the NIR image would also be applicable for the other spectrums. During the registration process as the image is shrunk, the void around the registered region is filled with zeros. We binarize the registered NIR image using a very low threshold to separate the image region from the surrounding void. Then the horizontal and vertical histograms were computed. The registered region resembles a slightly rotated rectangle in most cases. Hence the horizontal and vertical histograms had a trapezoidal shape. By calculating the coordinates corresponding the vertices of the shorter edge of the trapezium, a cropping window can be drawn that can fit well within the region of the image (Refer Figure 3). This same cropping window is also used for the RGB, and thermal images and the three images are cropped accordingly.

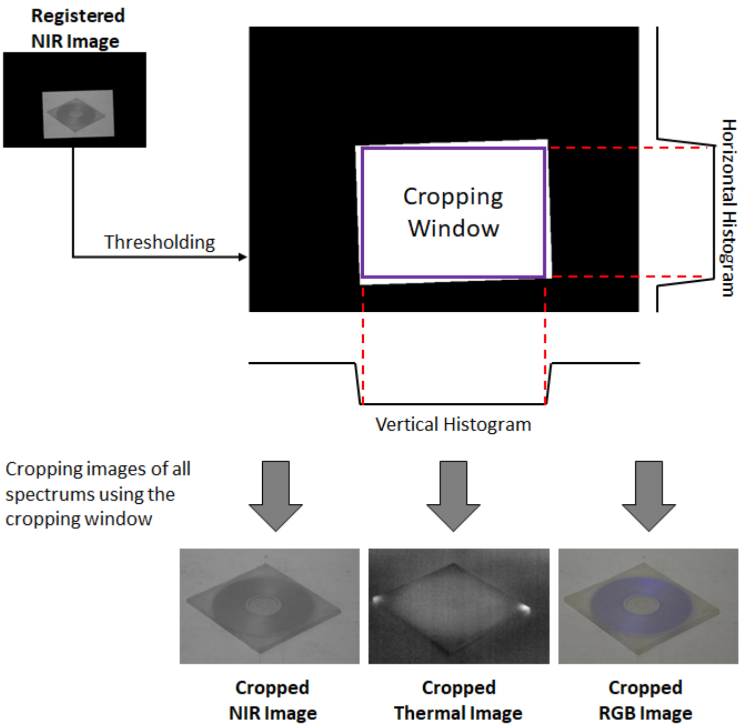


Figure 3 : The process of cropping registered images to remove unnecessary regions

3. 2 DATASET ANALYSIS

In this work, we introduce two datasets with images from 3 different spectrums, namely, visible, near infrared and thermal spectrum. The first dataset, or the single object dataset, consists of a total of 600 images spread across 25 classes and 3 spectrums and 8 different angles. The multiple object dataset comprises of 7200 images spread across $^{25}C_2$ object combinations and 3 spectrums and 8 different angles.

Table 2. DESCRIPTION OF OBJECT CLASSES

Serial No.	Object ID	Object Name	Materials
1	brsh	Brush	Wood, Metal, Synthetic
2	btp	Black Tape	Paper, Plastic
3	cd	CD inside case	Plastic
4	cup	Cup	Ceramic
5	esr	Eraser	Rubber
6	kr	Key Ring	Paper, Metal
7	mb	Marble Block	Marble
8	mem	Memory Card	Plastic, Metal
9	mt1	Scissor	Plastic, Metal
10	mt2	Stapler	Plastic, Metal
11	mt3	Key	Metal
12	mt4	Lock	Metal
13	ob1	Paper Weight	Glass
14	ob2	Paper Weight	Glass
15	ob5	Marker Pen	Plastic
16	pb	Power Bank	Plastic, Metal
17	pl3	Pen	Plastic
18	pncl	Pencil	Wood
19	pnd1	Pen Drive	Plastic
20	pnd2	Pen Drive	Plastic
21	shpr	Sharpener	Plastic, Metal
22	sng	Sunglass	Plastic, Metal

23	spec	Spectacles	Plastic, Metal, Glass
24	vas	Box	Plastic
25	whnt	Whitener Pen	Plastic

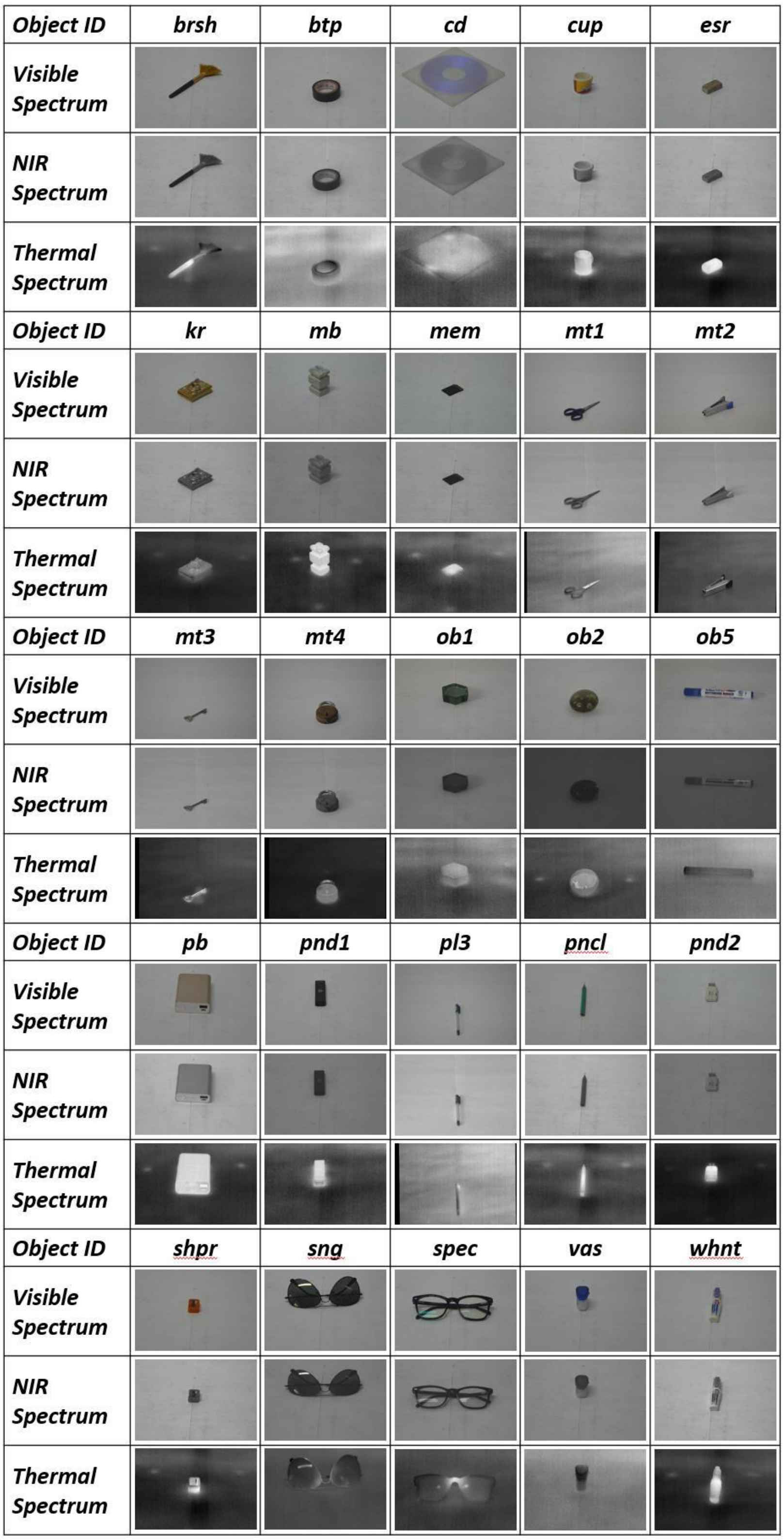


Figure 4 : Samples from 25 object classes and 3 spectrums.

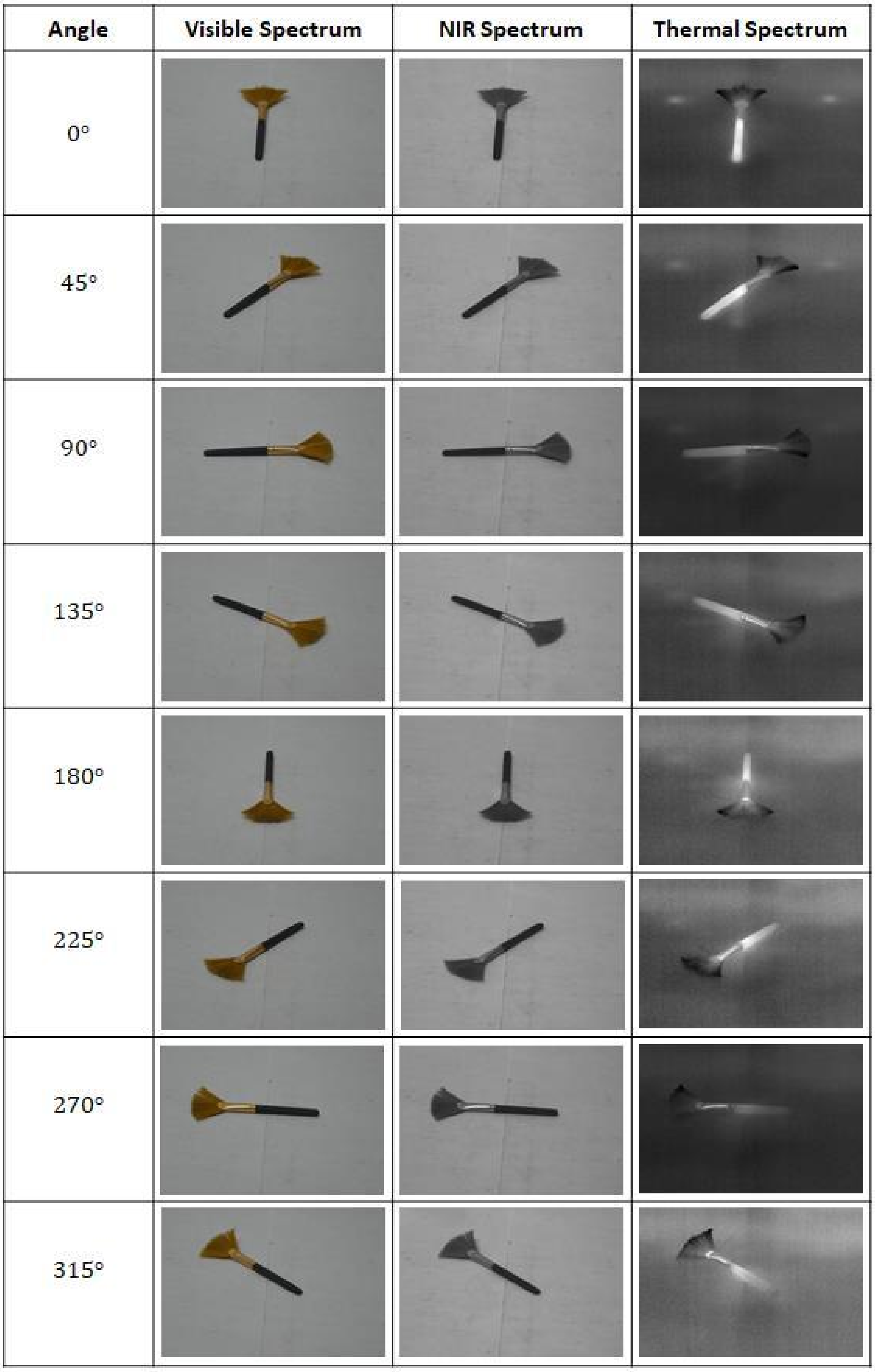


Figure 5 : Images from each object(single objects) or combination of objects(multiple objects) are captured from 8 different angles at rotated at an interval of 45°.

3. 2. 1 OBJECT DESCRIPTIONS

We have used 25 different objects often found a desktop environment composed of different kinds of materials such as plastic, wood, glass, ceramic, metal and so on. A summary of the various objects is provided in (Refer Figure 5). The selected objects vary greatly in their sizes and the amount of region occupy in the frame (Refer Figure 6) shows the average area covered by the objects in the frame.

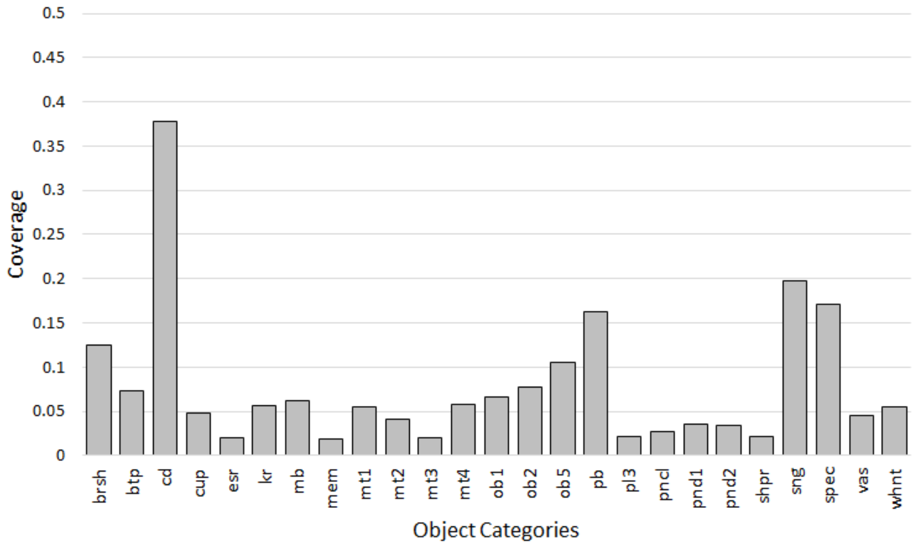


Figure 6 : Ratio of average area covered by the various objects with respect to the frame.

3. 2. 2 CONTRIBUTION OF VARIOUS SPECTRUMS

The various spectrums covered in this dataset bring forth several unique properties.

a) Visible Spectrum: The visible spectrum brings forth those components of an image that are also recognizable to the naked human eye. The primary distinguishing factors that define an object is its colour, texture, and shape. In visible spectrum different colored parts of an object can create false bounderies which could be challenging to work with. Another issue is the role of refraction or reflection in the visible spectrum. These issues are tackled to a great extent in other spectrums.

b) Near Infrared Spectrum: The images corresponding to the near-infrared spectrum are captured with a special camera that consists of a sensor which is sensitive to larger band of spectrum as compared to normal DSLR cameras. To compensate for the lack of natural light in indoor conditions, an infrared light source is used to illuminate the objects. An IR Filter that only allows wavelength over 830nm to pass through is attached in front of the NIR camera to block out the entire visible spectrum. The near infrared spectrum behaves differently as compared to the visible spectrum. For example, it is much more robust to refractions through glasses.

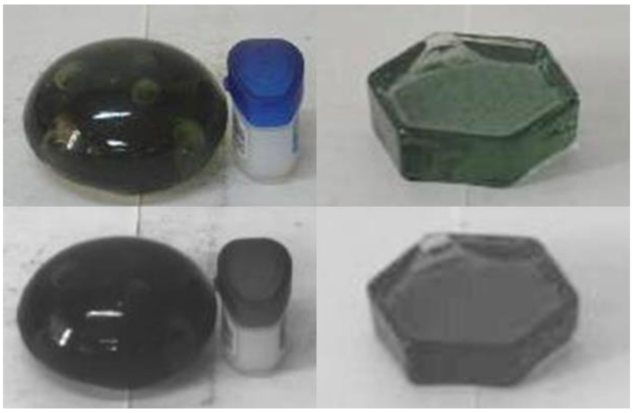


Figure 7 : A comparison of visible spectrum(top) and NIR spectrum(bottom). NIR images show a higher degree of robustness against refractions.

c) Thermal Spectrum: The thermal camera captures produces grayscale image with the intensities proportional to the temperature of a region in the frame. The thermal image demonstrates various thermodynamic properties of the objects in the image. Because different materials have different heat capacities, even in the same room temperature, the temperature of different objects vary slightly. This result intensity gradients that are much robust to factors like prints in the surface, or transparency of materials.

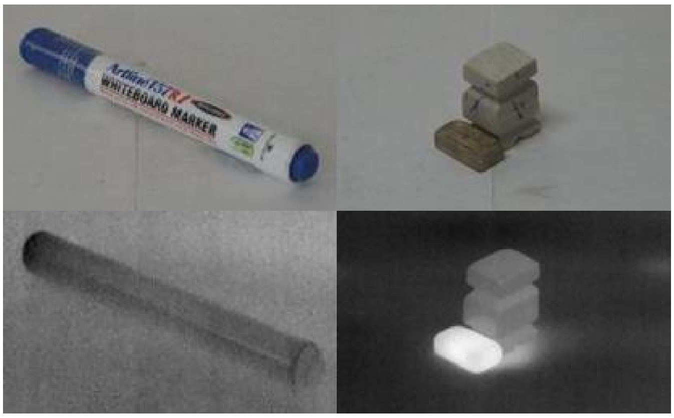


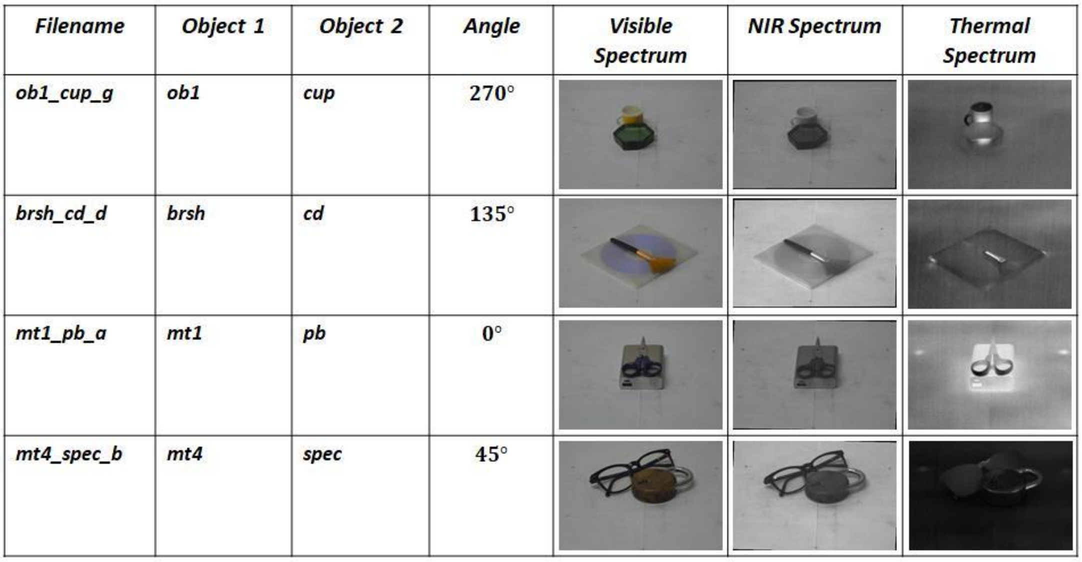
Figure 8 : A comparison of the visible spectrum(top) and thermal spectrum(bottom). Thermal images are invariant to visible properties of objects. Variations due to prints do not show up in the thermal images, and different objects with similar color composition have different intensities in the thermal image due to the different properties.

3. 3 ANNOTATIONS

To promote various supervised tasks, object classification and localization ground truths have been provided.

3. 3. 1 OBJECT CLASSIFICATION

The dataset has been designed to allow both single and multiple object classification. For single object classification objects of 25 different classes have been captured in three different spectrums, and 8 different angles. The images are organized into folders corresponding to class names or object id as shown in Figure 5. For multiple object classification, all possible combinations of objects, taking two at a time, were considered. For each combination, there were images from 3 different spectrums and 8 different angles. All the images were stored in directories corresponding to the object ids of the two objects in the form of <object_id_1>_<object_id_2>, where object_id_1 and object_id_2 refer to the ids corresponding to the object classes as per Figure 5.



Filename	Object 1	Object 2	Angle	Visible Spectrum	NIR Spectrum	Thermal Spectrum
ob1_cup_g	*ob1*	*cup*	270°			
brsh_cd_d	*brsh*	*cd*	135°			
mt1_pb_a	*mt1*	*pb*	0°			
mt4_spec_b	*mt4*	*spec*	45°			

Figure 9 : Samples images from combined object datasets.

3. 3. 2 OBJECT LOCALIZATION

Annotations for object localization has also been provided in the form of bounding boxes. For both single and multiple object datasets, the bounding box information has been provided in the xml format. The xml files for each single object are defined as follows:

```
<?xml version="1.0" encoding="utf-8"?>
<tagset>
<image>
<imageName>brsh_a_RGB</imageName>
<taggedRectangles>
<taggedRectangle height="158" width="120" x="102" y="26"/>
</taggedRectangles>
</image>
<image>
<imageName>brsh_a_corrected_NIR</imageName>
<taggedRectangles>
<taggedRectangle height="158" width="120" x="102" y="26"/>
```

```
</taggedRectangles>
</image>
<image>
<imageName>brsh_a_corrected_THERMAL</imageName>
<taggedRectangles>
<taggedRectangle height="158" width="120" x="102" y="26"/>
</taggedRectangles>
</image>
.... // continued for other angles for that object
</tagset>
```

For the multiple object images, two taggerRectangle tags are used to denote the objects in the order they appear in the filename as shown below.

```
<?xml version="1.0" encoding="utf-8"?>
<tagset>
<image>
<imageName>brsh_cd_a_RGB</imageName>
<taggedRectangles>
<taggedRectangle height="151" width="104" x="170" y="46"/>
<taggedRectangle height="151" width="257" x="85" y="53"/>
</taggedRectangles>
</image>
<image>
<imageName>brsh_cd_a_corrected_NIR</imageName>
<taggedRectangles>
<taggedRectangle height="151" width="104" x="170" y="46"/>
<taggedRectangle height="151" width="257" x="85" y="53"/>
</taggedRectangles>
</image>
<image>
<imageName>brsh_cd_a_corrected_THERMAL</imageName>
<taggedRectangles>
<taggedRectangle height="151" width="104" x="170" y="46"/>
```

```
<taggedRectangle height="151" width="257" x="85" y="53"/>
</taggedRectangles>
</image>
.... // continued for other angles for that object
</tagset>
```

CHAPTER FOUR

RESULT AND CLASSIFICATION BENCHMARKS

To analyze the challenges involved in the learning process in the dataset, some commonly used deep learning based classifiers such as ResNet [16], InceptionNet [17], and DenseNet [18] are implemented.

4. 1 DETAILS OF USED DEEP CLASSIFIERS

We have given a brief description about these deep classifierss in the following section.

4. 1. 1 RESNET18

Residual networks (ResNets) have recently achieved state-of-the-art on challenging computer vision tasks. Recently proposed residual networks (ResNets) get state-of-the-art performance on the ILSVRC 2015 classification task [19] and allow training of extremely deep networks up to more than 1000 layers. In theory, the increasing number of layers may improve the accuracy of the network, but in practice, few challenges may arise.

- Vanishing gradient problem somewhat solved with regularization like batch normalization etc.
- By adding more layers, accuracy did not improve as observed by the domain experts. Training error is also increasing because it is not over-fitting.

The basic idea about the network is that, at each convolution layer the neural network learns some features about the data F(x) and feed forwords the remaining errors further into the network . So the output error of the convolution layer is H(x) = F(x) - x.

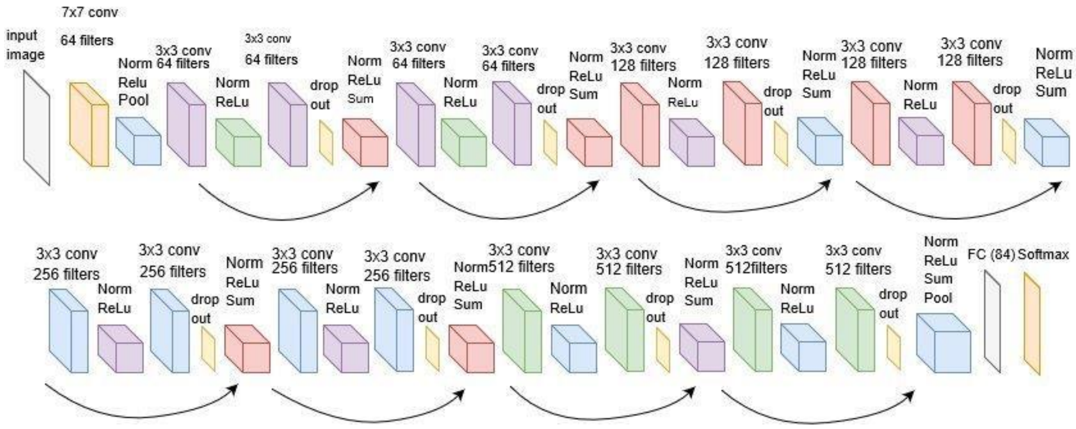


Figure 10 : Architecture of Resnet 18 [20]

4. 1. 2 INCEPTION – V3

Inception is a great architecture, and it is the result of multiple cycles of trial and error. A machine learning technique like transfer learning allows us to retrain a pre-trained model which significant leads to decrease the training time, and also the dataset size is required is reduced. Inception V3 is one of the famous models that can be used in transfer learning [21].

The knowledge can be maintained in the final layer of the pre-trained model that had learned during its original training and apply it to a smaller dataset, which results in a high accuracy without any need of extensive training and computational potency.

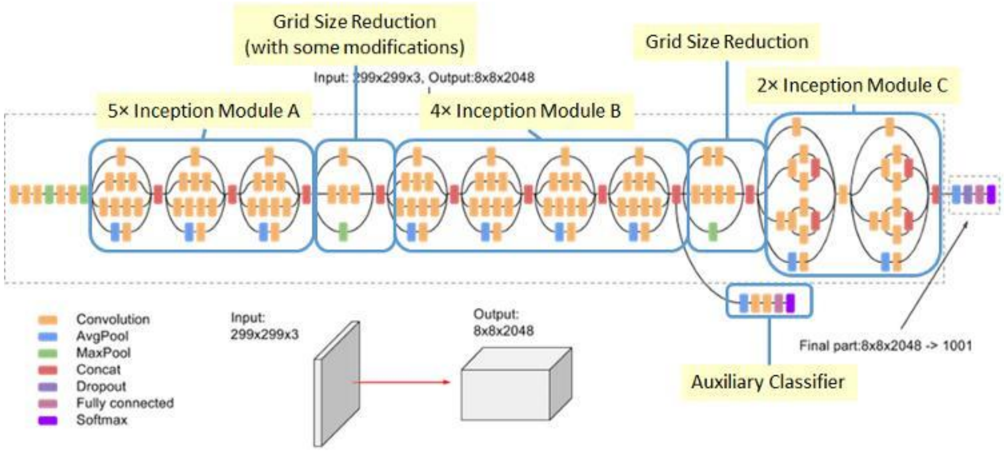


Figure 11 : Architecture of Inception – V3 [1]

4. 1. 3 DENSENET 121

Densely Connected Convolutional Networks [18], DenseNets, are the next step on the way to keep increasing the depth of deep convolutional networks. DenseNets simplifies the problem os CNNs when they dig deep. In CNNs the information travels from the input layer to the output layer (the gradient in the antithesis direction), so the information gets vanishes on either side of the network.

DenseNets reuse the extracted features to increase its network potential. DenseNets comparatively requires fewer parameter as compare to CNNs, which eliminate redundancy in feature maps.

Another problem with very deep networks was the problems to train, because of the mentioned flow of information and gradients. DenseNets solve this issue since each layer has direct access to the gradients from the loss function and the original input image.

[1] https://medium.com/@sh.tsang/review-inception-v3-1st-runner-up-image-classification-in-ilsvrc-2015-17915421f77c

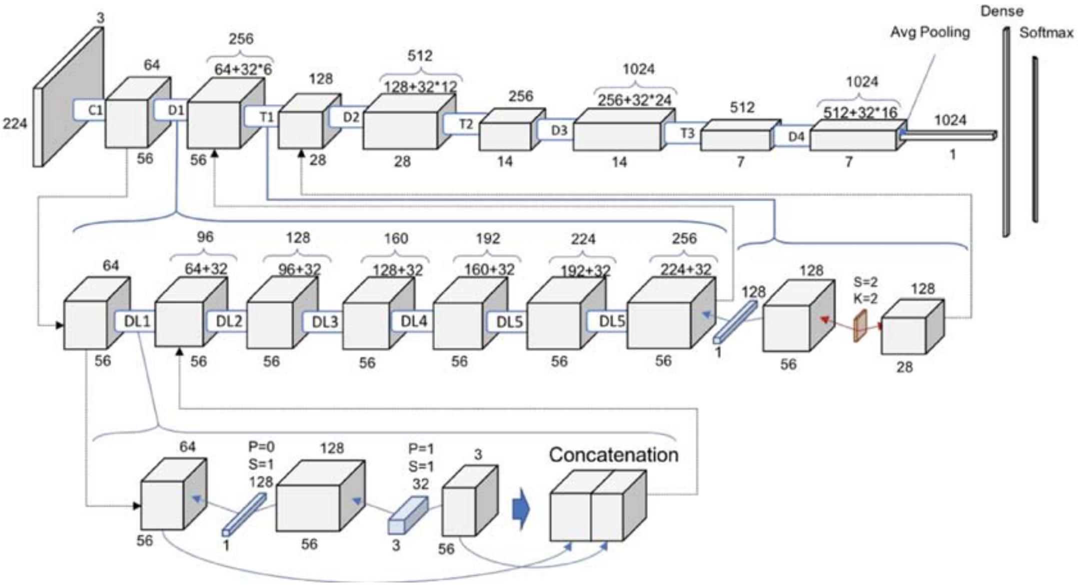


Figure 12 : Architecture of DenseNet 121 [2]

4. 2 RESULTS

The datasets were randomly divided into 3:1 ratio for the training and the test set. The best model from training was tested on the test set images. We also analyzed the performance of the datasets after applying data augmentation, where we used random flipping to capture a greater variety of angles. The data augmentation was more effective for the multiple object classification scenario. The multiple object dataset is more challenging than the single object recognition in terms of the accuracies obtained. For single object recognition, Inception-V3 network was the most effective while Densenet121 was the best for multiple object recognition. Table 4 shows all the results in details and visualization of the result is given in Figure 13 and Figure 14.

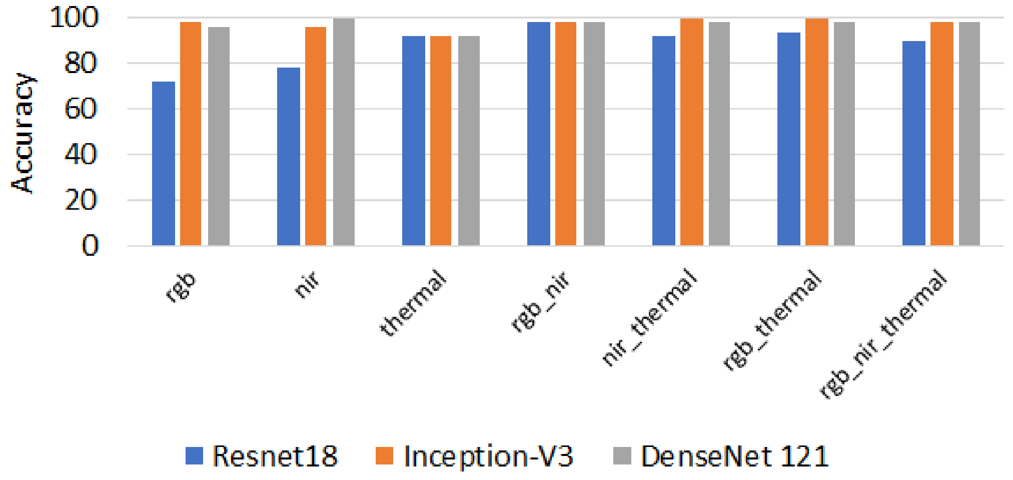
[2] https://towardsdatascience.com/understanding-and-visualizing-densenets-7f688092391a

Table 3. PERFORMANCE BENCHMARKS FOR SINGLE OBJECT CLASSIFICATION ON THE 2 DATASETS USING RESNET18, INCEPTION-V3 AND DENSENET-121 NETWORKS

Spectrum	**Single object dataset**							
	Without Data Augmentation				**With Data Augmentation**			
	Resnet-18	**Inception-v3**	**DenseNet 121**	**Mean**	**Resnet-18**	**Inception-V3**	**DenseNet 121**	**Mean**
Rgb	72.00	98.00	96.00	88.67	74.00	100.00	92.00	88.67
Nir	78.00	96.00	100.00	91.33	82.00	100.00	94.00	92.00
thermal	92.00	92.00	92.00	92.00	94.00	92.00	92.00	97.33
rgb_nir	98.00	98.00	98.00	98.00	92.00	100.00	100.00	97.33
nir_thermal	92.00	100.00	98.00	96.67	82.00	98.00	96.00	92.00
rgb_thermal	94.00	100.00	98.00	97.33	90.00	96.00	92.00	92.67
rgb_nir_thermal	90.00	98.00	98.00	95.33	94.00	100.00	98.00	97.33
Mean	88.00	97.43	97.14	94.19	86.86	98.00	94.86	93.24

Table 4. PERFORMANCE BENCHMARKS FOR MULTIPLE OBJECT CLASSIFICATION ON THE 2 DATASETS USING RESNET18, INCEPTION-V3 AND DENSENET-121 NETWORKS

Spectrum	**Multiple object dataset**							
	Without Data Augmentation				**With Data Augmentation**			
	Resnet-18	**Inception-V3**	**DenseNet 121**	**Mean**	**Resnet-18**	**Inception-V3**	**DenseNet 121**	**Mean**
rgb	92.00	92.67	95.67	92.97	93.42	94.67	98.17	95.42
nir	83.08	77.75	90.33	83.72	82.58	75.91	92.25	83.58
thermal	59.92	51.50	62.25	57.89	64.33	52.92	71.33	62.86
rgb_nir	89.58	91.50	97.25	92.78	96.17	96.83	98.08	97.03
nir_thermal	81.75	81.92	86.92	83.53	81.92	79.92	91.08	84.31
rgb_thermal	87.83	92.75	96.17	92.25	93.42	89.08	97.92	93.47
rgb_nir_thermal	92.75	87.50	95.83	92.03	97.08	91.17	97.92	95.39
Mean	83.64	82.23	89.20	85.02	86.99	82.93	92.39	87.44



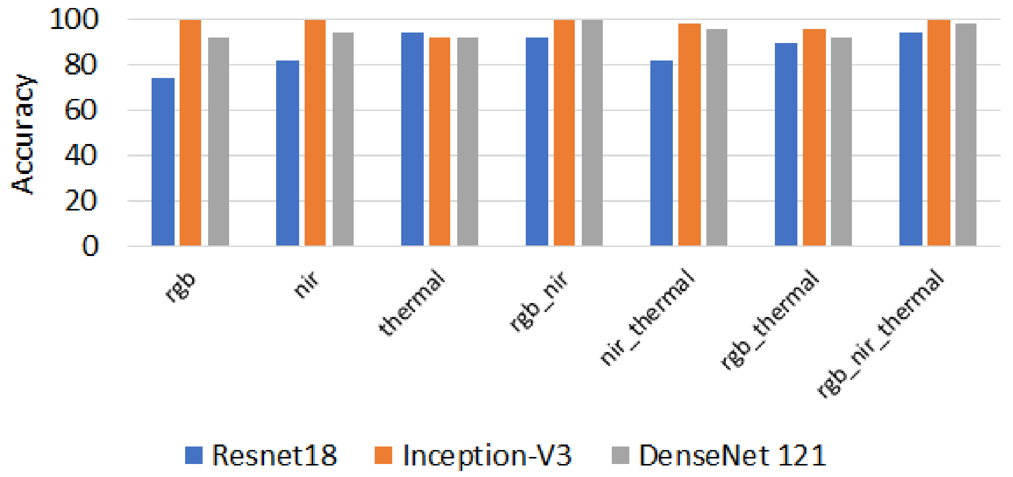
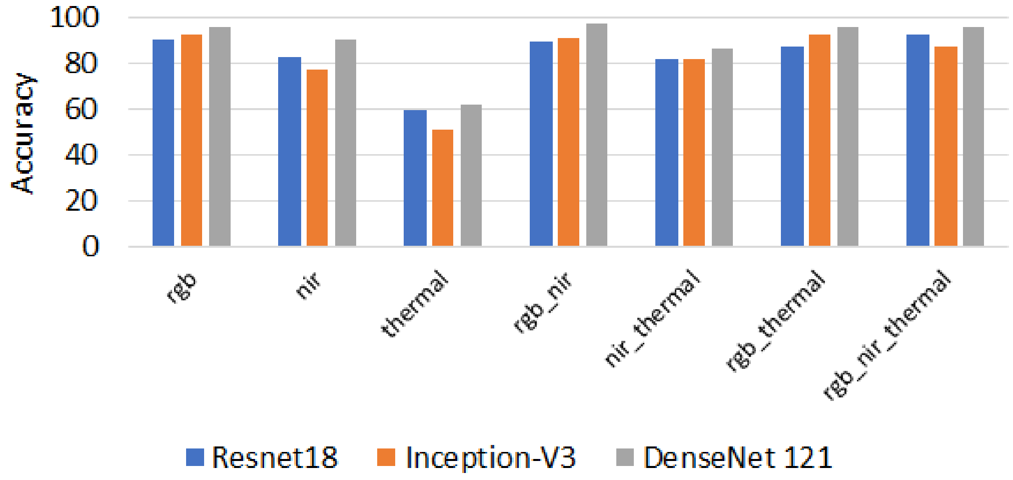


Figure 13 : Performance benchmarks for the dataset with Resnet18, Inception-V3, and DenseNet121 Architectures. Top: Single Object without data augmentation, Bottom: Single Object with data augmentation



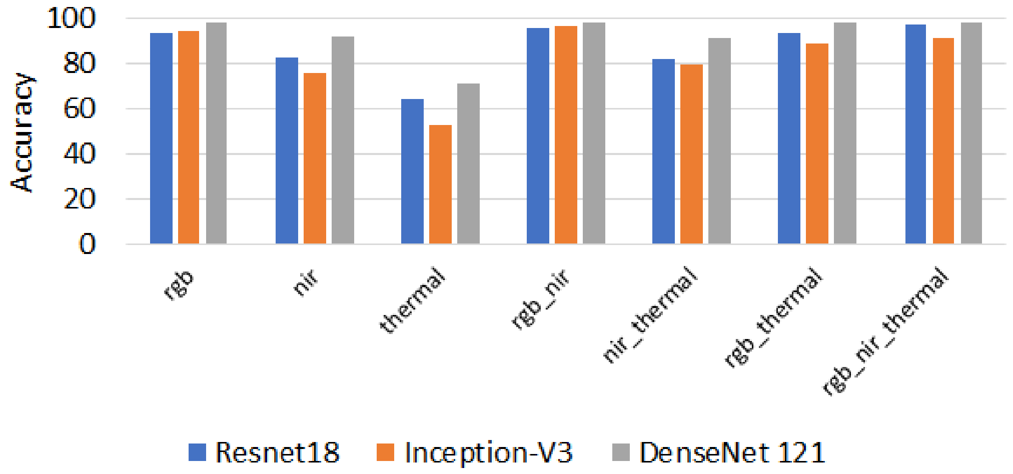


Figure 14: Performance benchmarks for the dataset with Resnet18, Inception-V3, and DenseNet121 Architectures. Top: Multiple objects without data augmentation and Bottom: Multiple objects with data augmentation.

CHAPTER FIVE

CONCLUSION

The proposed dataset presents several challenges that can be addressed. Some of the possible avenues of research that can be explored using the proposed dataset are single class and multi-class supervised object classification, single and multiple object localization, image to image translation across spectrums, generating images to simulate object rotation and so on. While several datasets exist that focus on RGB and NIR or RGB and Thermal images, this dataset provides 3 different spectrums that are RGB (visible), NIR and Thermal images together. We have provided approximately registered images across three different spectrums, for 25 object categories captured from 8 different angles. A total of 7800 images are provided, thus making it one of the largest multispectral dataset in the current era.

REFERENCES

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Ha, “Gradient-Based Learning Applied to Document Recognition,” no. November, pp. 1–46, 1998.

[2] P. F. Felzenszwalb, I. C. Society, R. B. Girshick, S. Member, D. Mcallester, and D. Ramanan, “Object Detection with Discriminatively Trained Part-Based Models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[3] S. L. Meeks, F. J. Bova, T. H. Wagner, J. M. Buatti, W. A. Friedman, and K. D. Foote, “Image localization for frameless stereotactic radiotherapy,” *Int. J. Radiat. Oncol.*, vol. 46, no. 5, pp. 1291–1299, 2000.

[4] T. Liu and T. Stathaki, “Enhanced Pedestrian Detection using Deep Learning based Semantic Image Segmentation.”

[5] S. T. Namin and L. Petersson, “Classification of materials in natural scenes using multi-spectral images,” *IEEE Int. Conf. Intell. Robot. Syst.*, pp. 1393–1398, 2012.

[6] M. Brown, S. Sabine, and D. L. Epfl, “Multi-spectral SIFT for Scene Category Recognition,” *CVPR 2011*, pp. 177–184.

[7] K. Takumi, K. Watanabe, Q. Ha, A. Tejero-De-Pablos, Y. Ushiku, and T. Harada, “Multispectral Object Detection for Autonomous Vehicles,” pp. 35–43, 2017.

[8] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, “MFNet : Towards Real-Time Semantic Segmentation for Autonomous Vehicles with Multi-Spectral Scenes,” pp. 5108–5115, 2017.

[9] C. Aguilera, X. Soria, A. D. Sappa, and R. Toledo, “RGBN Multispectral Images: A Novel Color Restoration Approach,” in *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection - 15th International Conference,*

PAAMS 2017, 2018, pp. 155–163.

[10] T. Alldieck, C. H. Bahnsen, and T. B. Moeslund, "Context-Aware Fusion of RGB and Thermal Imagery for Traffic Monitoring," 2016.

[11] S. Kim, S. Im, and J. Lee, "RANUS : RGB and NIR Urban Scene Dataset for," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 1808–1815, 2018.

[12] J. W. Davis and M. A. Keck, "A Two-Stage Template Approach to Person Detection in Thermal Imagery Stage-2 : AdaBoost Classification."

[13] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So, "Multispectral Pedestrian Detection : Benchmark Dataset and Baseline."

[14] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T Object Tracking : Benchmark and Baseline," pp. 1–13.

[15] G. B. Gustafson and C. H. Wilcox, *Analytical and computational methods of advanced engineering mathematics*, vol. 28. Springer Science & Business Media, 2012.

[16] K. He and J. Sun, "Deep Residual Learning for Image Recognition," pp. 1–9.

[17] C. Szegedy *et al.*, "Going Deeper with Convolutions," 2014.

[18] G. Huang, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks."

[19] K. Lyman, "Resnet in Resnet: Generalizing Residual Architectures," pp. 1–7, 2016.

[20] M. Al, R. Alif, and S. Ahmed, "Isolated Bangla Handwritten Character Recognition with Convolutional Neural Network," no. March 2018, 2017.

[21] C. Szegedy, V. Vanhoucke, and J. Shlens, "Rethinking the Inception Architecture for Computer Vision," 2014.