Multimodal Approach to Human Detection in Unconstrained Environments using YOLOV7 for Conventional, Infrared & Thermal Cameras

by

Maymuna Rukaiya
19101142
Md. Muhtadee Faiaz Khan Soumik
19101491
Sazzad Hossan Sakib
22241131
Md. Ashikul Islam
22241137
Mohammad Farhan Ishrak
22241187

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
January 2023

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Maymuna Rukaiya

19101142

Md. Muhtadee Faiaz Khan Soumik

19101491

Sazzad Hossan Sakib

22241131

Md. Ashikul Islam

22241137

Mohammad Farhan Ishrak

22241187

Approval

The thesis titled "Multimodal Approach to Human Detection in Unconstrained Environments using YOLOV7 for Conventional, Infrared & Thermal Cameras" submitted by

1. Maymuna Rukaiya (19101142)
2. Md. Muhtadee Faiaz Khan Soumik (19101491)
3. Sazzad Hossan Sakib (22241131)
4. Md. Ashikul Islam (22241137)
5. Mohammad Farhan Ishrak (22241187)

Of Fall, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 19, 2023.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Khalilur Rhaman

Professor
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

Mr. Md. Tanzim Reza

Lecturer
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Search and rescue operations in disaster-stricken areas are often hindered by challenging environmental conditions, such as poor visibility, limited lighting, and high levels of noise and clutter. These conditions can make it difficult to locate and rescue survivors in a timely manner, which can have significant implications for their survival and recovery. Traditional methods of human detection, such as visual observation, can be ineffective in these environments, and new and innovative approaches are needed to address these challenges. This research presents a novel multimodal approach to human detection in unconstrained environments using YOLOv7 for conventional, infrared and thermal cameras. The proposed approach aims to improve human detection performance in challenging environments, such as post-disaster situations, where traditional methods may fail. A unique dataset of 7,087 images was created for this research, including both conventional and thermal images, which were collected to capture the realistic scenario of disaster environments. The dataset was used to train various CNN models for human life detection, and the results were evaluated using standard metrics. Additionally, to further enhance the search and rescue operations in post-disaster situations, a Bangla speech recognition model was integrated into the system. The results of this research demonstrate the effectiveness of the proposed approach in detecting humans in challenging environments, such as low-light and obscured conditions. The use of thermal imaging in particular, has the potential to significantly improve human detection in disaster scenarios where visibility is limited. This research provides a valuable contribution to the field of human detection in unconstrained environments and has the potential to improve search and rescue operations in the future.

Keywords: Human Detection; Machine Learning; YOLO v7; Faster R-CNN; Bangla Speech Recognition; Thermal Image; Primary Dataset.

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our supervisor Dr. Md. Khalilur Rhaman sir for his consistent support and valuable advice in our work. He helped us whenever we needed help. Thirdly, we would like to express our sincere gratitude to our co-supervisor Md. Tanzim Reza sir for his kind guidance and technical support. And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ϵ Epsilon

υ Upsilon

DNN Deep Neural Network

HMM Hidden Markov Model

IoU Intersection over Union

$LSTM$ Long Short-Term Memory

$LWIR$ Long Wave Infrared

mAP Mean Average Precision

NMS Non-Maximal Suppression

ROI Region Of Interest

RPN Region Proposal Network

Chapter 1

Introduction

The concept of 'Disaster' is easier to grasp than its definition. Disaster is a summative idea or a 'sponge word' for certain scholars. Burton et al. (1978) called it a 'collective stress condition'[2], while Quarantelli and Dynes (1977) called it a 'social crisis era'[1]. It is characterized as a crisis that depletes society's ability to manage disasters and, if not addressed, leads to massive loss of life and property. A disaster, in literary terms, is an event that causes widespread destruction, sorrow and catastrophe.

Bangladesh is a highly disaster-prone country with a population of around 160 million people and an area of around 147,570 km2. The country is exposed to many geo-hazards and hydro-meteorological hazards or disasters due to its geographical location and meteorological features. Floods, cyclones, tidal surges, tornadoes, nor 'westers, earthquakes, river erosion, fire, infrastructure collapse, arsenic contamination of groundwater, water logging, water and soil salinity, cold wave, building collapse, epidemic, and various forms of pollution are among the major disasters affecting the country.

Natural catastrophes have occurred frequently in Bangladesh. It was struck by 219 natural disasters between 1980 and 2008. Bangladesh is extremely sensitive to natural hazards due to its geographic location, land characteristics, abundance of rivers, and monsoon climate. Natural hazards have an impact on Bangladesh's coastline morphology[31]. Bangladesh has had at least minor-to-moderate magnitudes 465 earthquakes, according to the Geological Survey of Bangladesh between 1971 and 2006.Besides natural disasters, huge casualties are being faced due to accidents and man-made disasters. Building collapse is a major phenomenon. On April 24, 2013, the collapse of Rana Plaza was Bangladesh's most horrific industrial tragedy, killing 1,135 people and injuring 2,500 more. Following the event, the Bangladesh Army, in collaboration with the Bangladesh Navy, Fire Service, BGB, and Police, launched a rescue mission. Furthermore, several volunteers assisted with the rescue efforts.

However, while carrying out this evacuation process, there have been various sorts of difficulties and obstacles. It becomes very tough for evacuation teams to penetrate certain devastated areas during disaster and rescue people. Moreover, it becomes more difficult to even detect living human beings trapped under ruins. To solve this sort of problem various sorts of research has been conducted over the years. With

inclusion of Convolutional Neural Network and Microwave sensing, various models have been proposed. With each proposed model or principle comes more flaws and problems. In this paper we have gone through all of those proposals and came up with possible approaches.

1.1 Problem Statement

We cannot prevent natural catastrophes from occurring, but we can prepare for them using contemporary technologies and respond to them as best we can. Natural catastrophes are capable of causing substantial harm to human lives and their resources. A natural disaster's aftermath can result in financial, environmental, physical, and mental harm to a person. While social and medical aid can compensate for mental and financial losses, the loss of life cannot be compensated or restored. People can get buried within a mound of bricks, cement, and rocks due to building collapses and structural problems. Survival beneath structurally damaged and collapsing structures is extremely difficult since oxygen and sunshine are almost never available. Natural disasters such as earthquakes and tsunamis can trap individuals beneath collapsing structures, making it difficult for rescue crews to extricate people from these locations. This is particularly evident in metropolitan settings where a big population resides in tall, enclosed structures. Natural catastrophes create destruction, but they do not immediately endanger human life. Typically, subsequent occurrences, such as structure collapses and fire outbreaks, cause fatalities and permanent injuries. Multiple reasons, such as earthquakes and avalanches, structural defects and gas explosions, etc., can result in structure collapse and fire outbreak.

In such situations, timely and reliable information is important for rescuing trapped and injured humans. Researchers propose that a person trapped within a collapsed structure has a good chance of survival provided we can collect accurate information about their location. With oxygen and timely rescue, a human life can recover within 72 hours after being rescued. After 72 hours of being trapped beneath a fallen structure, the survival probability decreases, and without access to water, the majority of victims are unlikely to live beyond 120 hours. Typically, the first attempts at rescue are made by survivors on the scene, followed by large-scale search and rescue operations by local, national, and international rescue organizations. The effectiveness of the whole search and rescue operation ultimately hinges on the precise and prompt finding of survivors within the collapsed building. There are essentially three common search techniques for discovering survivors of a catastrophic event: Physical search including both visual and audible techniques, Canine search with scent-detecting canines and electronic search utilizing different electronic equipment and sensors.

Several human forces are utilized throughout the rescue effort, including firemen, police officers, and medical personnel. All of them are exposed to very hazardous scenarios produced by the destructed environment in which they operate, such as collapsing structures, landslides, and craters. Consequently, the rescuer may become a victim in need of rescue. Therefore, the rescue operation poses a significant danger to the rescue troops. In this light, the search for alternatives to human rescues has

been in high demand.

In quest of completely and partially autonomous alternatives to the human element, we are in a mission of creating a robotic module-based solution that can be combined with a quadcopter or a small-sized robot that can visit these tough places and discover living and imprisoned humans as soon as feasible.

1.2 Research Objectives

The research aims to develop a module that can detect alive human subjects under obstacles during natural disasters using image processing and sound recognition. Image processing using day, night and thermal image classification and CNN models will give better results about human identification and sound recognition will ensure the presence of alive victims. The objectives of this research are -

1. To deeply understand the CNN models used in image processing.
2. To explore algorithms for more efficient implementation of image processing.
3. To understand sound recognition and the algorithms associated with it.
4. Our approach is divided into 4 categories:
    - To test day cam
    - To test night cam
    - To test thermal cam
    - To use Bangla Speech Recognition.
5. To provide a better solution in improving the models.

Chapter 2

Literature Review

The necessity for human life detection has become more widespread in Bangladesh as this country is extensively recognized to be one of the nations with the greatest susceptibility to the effects of circumstances beyond individual's control in the whole globe. According to the Asia Pacific Disaster Report 2015 (UN-ESCAP), Bangladesh is one of the most vulnerable nations among the fifteen countries that have high exposure to (10th position) and risk from (5th position) natural calamities [10]. In the aftermath of natural disasters like earthquakes, hurricanes, explosions etc., there remains a good probability that victims will be able to survive in the voids that are naturally made in fallen structures. This is the case because these voids are created by the buildings themselves.

On a yearly basis, thousands of Bangladeshi people lose their lives as a result of being entrapped in collapsing buildings in Bangladesh. The magnitude of the devastation brought on by natural disasters is, without a shadow of a doubt, huge. These natural disasters will not only cause death and injury to people, but they will also hinder our capacity to export goods and force millions of people to evacuate their homes. Consequently, in the aftermath of a catastrophe such as this one, it is of the utmost importance to search for and rescue those who may be buried alive or otherwise trapped under debris. And so finding and rescuing those who are buried or trapped beneath rubble is a top priority.

2.1 Brief History of Unconstrained Disasters

Bangladesh, a country thickly packed with people, is frequently struck by a large number of natural catastrophes such as earthquakes, fire accidents, floods, tsunamis etc. As a result, the nation needs to deal with the fallout from these catastrophic catastrophes. There were more than two hundred natural catastrophes that occurred during the years between 1980 and 2008 [17]. In addition to this, the country sits at the intersection of three distinct tectonic plates, making it one of the most geologically active places on Earth. Over the course of the previous four decades, Bangladesh has been shaken by more than 250 earthquakes, some of which have registered on the Richter scale as having a magnitude that is more than 6.0 [15]. This is a direct consequence of the location of Bangladesh, which is close to the region where the Indian, European, and Burmese tectonic plates smash with one another. In addition, the collapse of the Rana Plaza building in Dhaka, Bangladesh,

on April 24, 2013, led to the deaths of at least 1,132 people and the injuries of over 2,500 more people. These victims were not rescued immediately; rather, it took many days to determine their precise location and carry out the rescue operation. Eventually, they were saved. Moreover, Five months earlier, at least 112 workers had lost their lives in another sad occurrence, when they were trapped inside the blazing Tazreen Fashions complex on the outskirts of Dhaka[19]. They had been unable to get out of the building when it was on fire, which led to this tragic outcome.

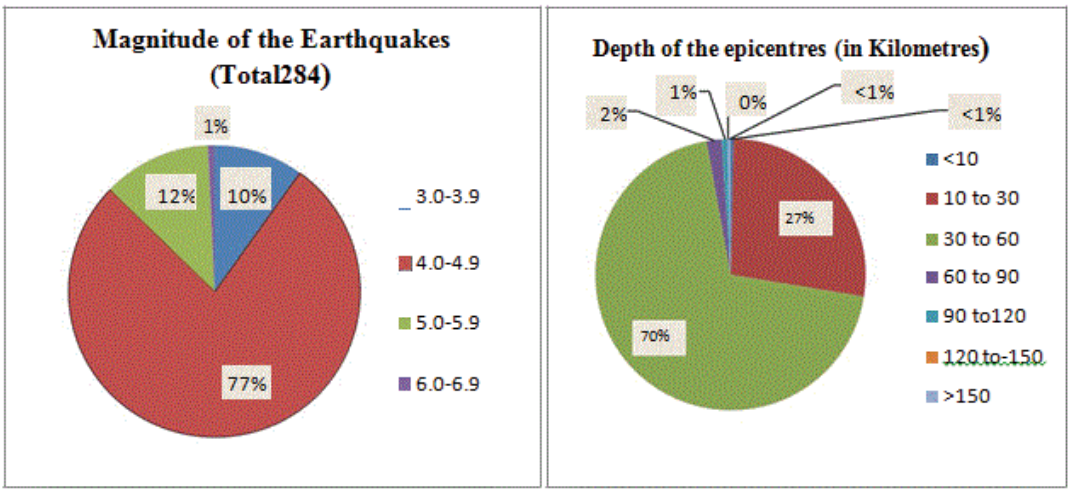


Figure 2.1: Earthquakes of Different Magnitudes & Depths of Epicentres of the Earthquakes

This depicts magnitude and depth of every earthquake that has happened in the last forty years[15].

2.2 Artificial Intelligence & Machine Learning against Disasters

Artificial intelligence (AI) can be defined as simulation of human intelligence carried out by machines. In AI, programs are developed in order to carry out discrete, well-defined tasks, which are then utilized in a variety of contexts and have been used to advance in a variety of fields. Using its built-in, highly advanced algorithms, AI technology is able to quickly and accurately measure distances, locate and track optimal paths, identify humans despite the presence of obstacles, and retrieve data from the image processing procedure. Additionally, the technology is able to identify humans even when they are hidden behind obstacles. It is being investigated as a potential answer for easing the suffering caused by natural disasters and finding a way to put an end to them altogether. In addition to this, if the data model is trained properly and the obtained data is processed in the appropriate manner, Machine Learning may be able to assist in finding a solution more quickly in the aftermath of an unconstrained catastrophe. Assuming the data we are getting is accurate, it has the potential to assist in the automation of damage mapping technologies and contribute to the development of improved responses to natural catastrophes.‘Artificial

intelligence has the potential to help all countries achieve major advances in disaster management that will leave no one behind,' said Jürg Luterbacher, the Chief Scientist and Director of Science and Innovation at the World Meteorological Organization (WMO)[28]. Drones, robotics, and sensors that are powered by AI may now deliver information that is both intelligent and accurate about regions that are prone to disaster. As a direct consequence of this, rescue workers and first responders have gained a better understanding of the seriousness of a natural disaster. As a consequence of this, drones and other sensing technology can assist rescue workers in locating victims and getting to them as swiftly as possible. In recent years, numerous researchers have achieved success in constructing rescue robots that are tailored to the structural characteristics of their respective countries. Taking into account the fact that Bangladesh is unique in a variety of ways, including human voice, movement, postures, structural characteristics of buildings, and so on, we need to take these factors into consideration when developing our model.

2.3 Related Works

To aid with search and rescue efforts, advanced technologies are being constructed. The purpose of this section is to critically evaluate previous works in the subject of Human life detection in trapped structures due to unconstrained events. Here, many different approaches will be analyzed which are used to achieve the main outcome.

M. Donelli[8] proposed a lightweight microwave system that is presented for earthquake and other disaster-related search-and-rescue operations. With the use of an independent component analysis (ICA) technique, which permits to efficiently provide clutter cleaning and removing background noise, the suggested system based on a continuous wave X-band radar can identify respiratory and heart fluctuations. Here the life signals are extracted from the modulated back-scattered wave. The suggested rescue radar is tiny enough to be installed on a lightweight unmanned aerial vehicle (UAV) for access to hazardous or inaccessible locations. The heart of the system is a bi-static continuous-wave (CW) radar with a 10.45 GHz local oscillator, a low-noise amplifier, and a transmitting antenna made up of two patches. When a human is discovered as an object of interest, the radar emits an electromagnetic wave and receives the reflected signal, which carries information about the target's breathing and heartbeat. The back-scattered signal is picked up by a two-patches antenna, amplified, and sent to an I/Q detector. Due to their use of I/Q signals as two distinct data sources, they utilize an orthogonal detector to satisfy the ICA's need for a number of observation points equal to the number of the original signals. Then, the I/Q signals are sent through a low-frequency wireless channel as input to a remote elaboration system. In addition to this, the amplitude and phase of the received back-scattered wave are varied according to the movement of breathing and heartbeat.

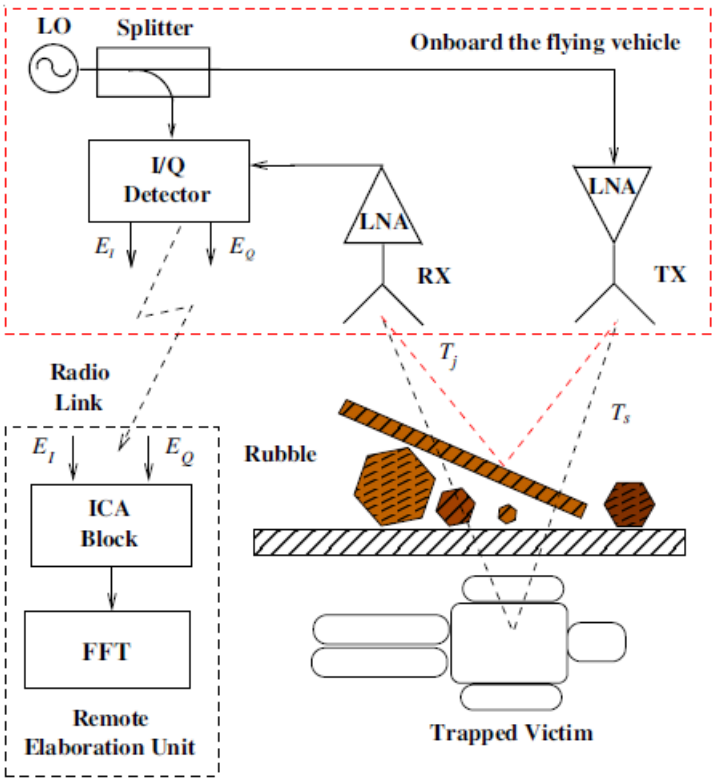


Figure 2.2: Diagram of the microwave life detection system proposed by M. Donelli

The researcher tested their work on a person who was placed in a hollow concrete pipe with a diameter of 1.5m and a thickness of 0.15 m. The acquired experimental findings confirm the viability of the suggested detection approach in locating trapped persons with a respectable degree of accuracy.

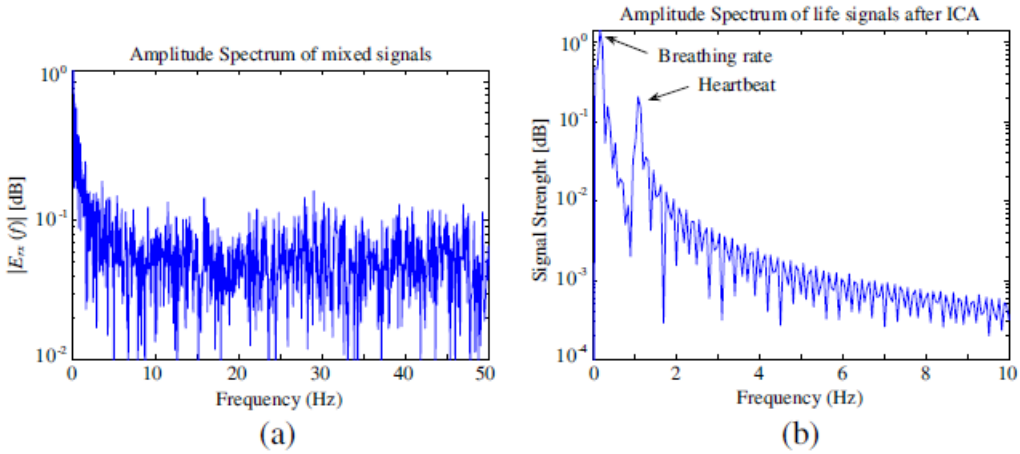


Figure 2.3: Experimental results. Frequency spectrum of the measured (a) mixed signals noise plus signals life obtained from experimental setup, (b) signals life extracted after the application of the ICA algorithm.

Sharma et al.[18] reviewed two unique algorithms for human identification in the thermal domain. Hot-Spot and Background Subtraction Algorithms are implemented here. The Hot-Spot Algorithm uses thermo-graphic cameras ability to de-

tect objects in any setting. Temperature increases item radiation, which may detect humans. The program finds the active zone to detect humans. The picture's "hot-spot" is where the pixels' intensity matches a human's body temperature. For pixel intensities, an image is grayscaled - 0 (very cold) to 255 (very hot) (extremely hot). Then the thresholding method assigns white values to pixels with intensities over the threshold to create a binarized image. And Background Subtraction Algorithm for thermal imaging human detection has three features. First, median and average filtering removes people from the backdrop. Pixel and edge difference images of input and background photos discover humans. Difference image threshold depends on background brightness. Experimentally, environment determines the threshold parameter and it gets adjusted before applying algorithm. Finally, candidate region size and height-to-width ratio separate humans from other objects.

Figure 2.4: Hot-spot algorithm - (a) the original image and (b) binary with humans detected

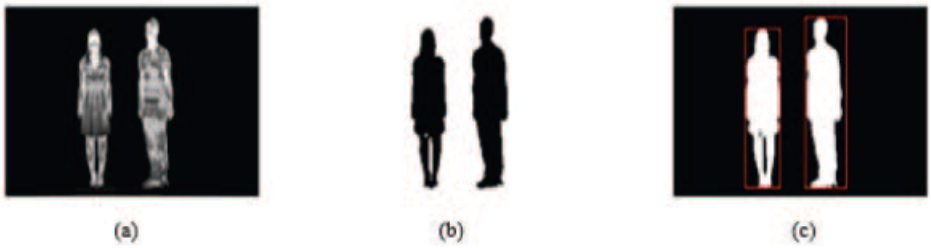


Figure 2.5: Background subtraction Algorithm - (a) the original image; (b) the difference image and (c) binary with humans detected

Two videos are included in the experimental dataset. There are thermal photographs in there of people and other living things in a variety of settings. Two of the clips had a frame rate of 24 frames per second, while the other two had a frame rate of 30. Two films' picture sizes varied between 1280p and 960p. The TPR, FPR, and Accuracy are used as benchmarks against which the effectiveness of these two methods can be assessed.

$$TPR = \frac{TP}{TP+FN}$$
$$FPR = \frac{FP}{FP+TN}$$
$$Accuracy = \frac{TP}{TP+FN+TN+FP}$$

TP is the fraction of frames in which human detection occurred when it was present, and FP is the total number of false positives, or times a human detected something that wasn't there. Number of frames where a human was mistakenly detected. Humans missed an object in FN frames. And TPR is the TP rate whereas FPR is FP rate.

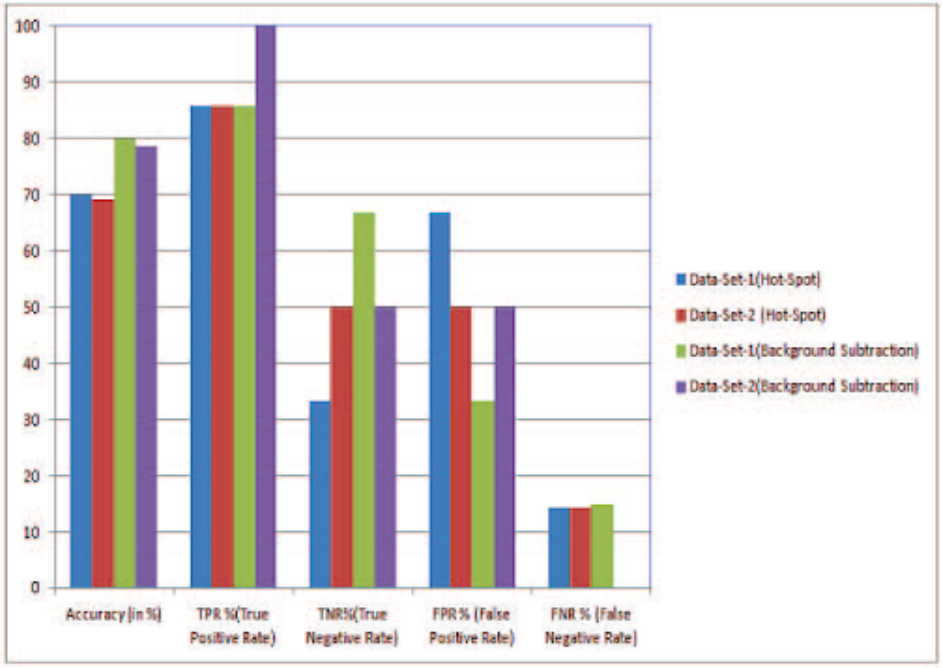


Figure 2.6: Graphical representation of the experimental output of the two algorithms proposed by Sharma et al.

Background Subtraction has 79% accuracy and Hot-Spot has 69%. Background Difference Method outperforms the Hot-Spot algorithm here. Background conditions and items in frame affect the outcome. Background Subtraction resists frequent background illumination and brightness variations.

In their paper[23], Di Zhang et al. proposed a human rescue sensor system which includes CO2 sensor, thermal camera, and microphone. This method was evaluated in a high-fidelity simulated disaster region to identify alive victims. The CO2 sensor reduces the probable affected region, while the thermal camera locates the sufferer. Using microphones with other sensors may help discover victims. The researchers also built and tested an algorithm to distinguish voices or human audio beneath debris. The system was made up of three types of sensors: 1. Gas sensors (O2 and CO2) for detecting human breath and air quality, 2. Microphones to detect voices, human-made noises, and ambient noise, 3. Thermal vision camera to see local temperature patterns in specific areas.

Table 1 given below displays the average detection time across all trials, which comes to nearly an hour. For eight out of nine tests, the researchers were able to identify the victim. These nine tests took three days to conduct. Since 48 hours is the threshold at which 80 percent of survivors can be saved alive, speed and accuracy in casualty discovery are crucial. Moreover, in a noisy environment, the accurate

Test	Execution Time	Result
Day 1 Morning	1 h 35 min	Success
Day 1 Afternoon	56 min	Success
Day 1 Evening	1 h 25 min	Success
Day 2 Morning	33 min	Success
Day 2 Afternoon	50 min	Success
Day 2 Evening	1 h 12 min	Failed
Day 3 Morning	2h 13 min	Success
Day 3 Afternoon	20 min	Success
Day 3 Evening	31 min	Success

Table 2.1: Global result of the tests of the Human rescue sensor system

voice recognition rate is 89.36%. Human-made suspicious noises, such as scratching and coughing, have a 93.85% success rate when classifying correctly. Each sensor's efficacy was measured and validated in this research. Still there remain limitations like some regions cannot be immediately accessed with a telescopic pole or directly observed due to the existence of impediments, therefore a sensor system using merely a thermal camera is not robust.

In a paper by Aparna U et al.[30], the concept of deep learning helps Spotter, the proposed model, detect humans. The USB camera implanted into the rubble will record live video of the collapsed building after a tragedy. These USB Cameras are imaging cameras that use USB 2.0 or USB 3.0 technology, to transfer the image data. USB Cameras use the same USB technology as most laptops to readily connect to dedicated computer systems. Here, USB 2.0 cameras are appropriate for imaging applications due to their 480 Mb/s transfer rate and USB 3.0 cameras can send data at 5 GB/s. After receiving, YOLO slices the movie into frames by using OpenCV, a real-time computer vision programming framework and Contrast Limited Adaptive Histogram Equalization(CLAHE) improves visual resolution. The percentage of possibility of being human is then disclosed during the YOLO object detection process. For training, the victims of entrapment from the COCO dataset are used.

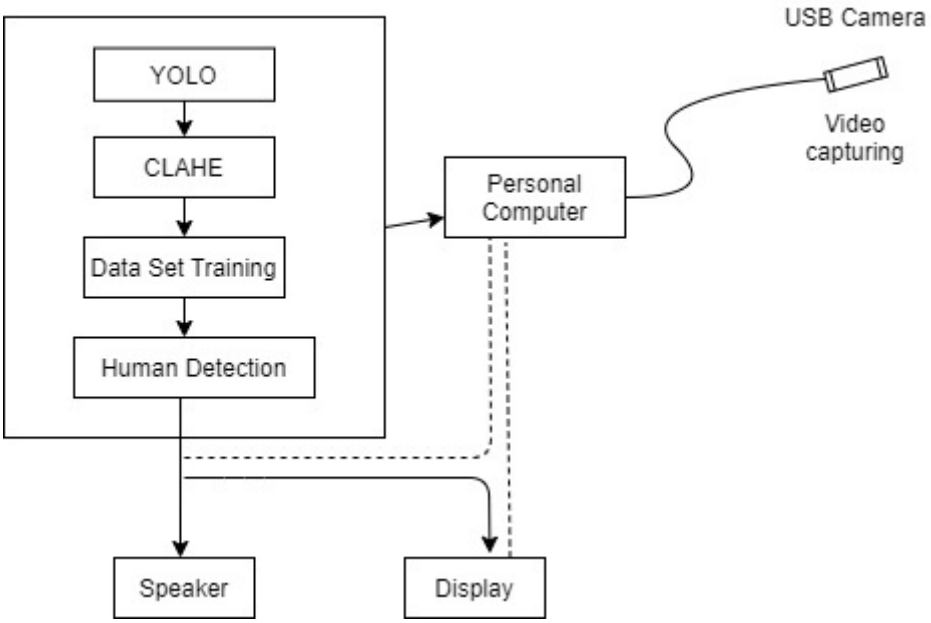
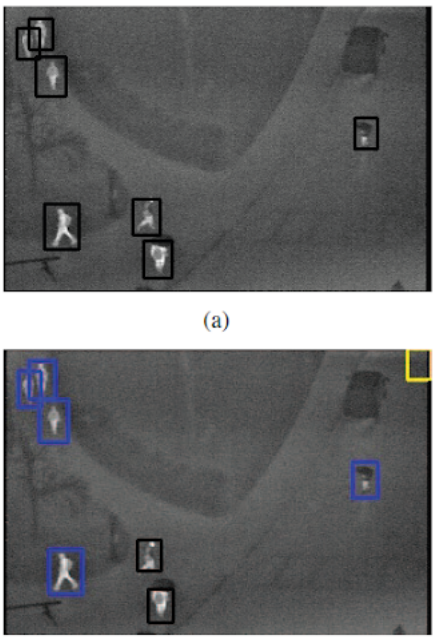


Figure 2.7: System architecture of Spotter

There is a 0.91 probability that the victim will be discovered precisely and promptly. The USB camera is small enough to fit through gaps in the rubble, and when the flash is activated, video may be captured even in the pitch-black places under the debris. USB cameras with low-power USB connections, such as those found on laptops, may require an external power source. When the system is low on power, this slightly reduces its effectiveness.

Weihong Wang et al.[7] used the Shape Context Descriptor (SCD) for local edge feature extraction and boosting algorithms with a cascade framework for human classification in thermal images which all together is called 'Shape Context Descriptor based AdaBoost Cascading Classification'. The Canny edge detector transforms training samples during the Adaboost cascade's training phase. After that, the SC descriptor for each pixel in the training samples is created. All of the SC descriptors are Adaboost classifier candidate features. The SC descriptors gather detailed and efficient human shape information. Each descriptor is then projected to one dimension using Fisher's linear discriminant (FDA). This data can be used to depict a specific aspect of the human form. A weak classifier is trained from the collection of SC descriptors at each Adaboost cycle. Each weak classifier can then be specified as h(x) = +1 if wT x> θ ; and h(x) = -1 otherwise, where h(.) is a weak classifier, w is the FDA weight vector, x is a SC descriptor, and θ is an optimal threshold for minimizing the number of misclassified examples. For training, About 500 high-quality training samples are extracted from the dataset photos and used in the learning process. There are also around 500 negative samples.

When the number of false positives is approximately 20 (which corresponds to a false positive rate of 5%), the suggested detector can achieve a detection rate more than 70%. When the detection rate for the suggested detector reaches 80%, the false positive rate is slightly above 40 (10% false positive rate).



(a)

(b)

Figure 2.8: Detection results of the proposed detector - (a) is from our Shape Context based detector. (b) is from rectangle feature based detector

Chapter 3

Methodology

The ultimate objective of our proposed system is to look for hints and evidence of human bodies or distress calls made by victims in order to identify areas where victims are trapped, deliver this information to a remote platform automatically after a period of time, and repeat the process until there are no more victims to search for. To accomplish this, we are equipping our machine with the capacity to quickly search for human bodies while ignoring other oddities using machine learning. We must ensure that the machine remains focused and provides reliable data so that rescuers may operate precisely with the information presented. We rely on sensors and machine learning algorithms to maximize our system's detection capabilities.

- **Conventional Camera**: It is essential to have good images to work with, since without these, our system could lose key information that is essential to the detection process. We are employing a conventional day camera to capture the highest quality images, which we will then analyze and compare to our data-set for the most accurate findings.

- **Infrared Camera**: We want our system to function properly under all conditions. In addition, disasters do not just occur during the day, thus for our machine to complete the detection procedure in the evening, it requires a night camera and a night data-set for training and detection.

- **Thermal Camera**: The heat is an excellent source of evidence when searching for victims. By attaching a thermal camera to the system, we can increase the system's searching efficiency.

- **Speech Recognition**: Visuals alone cannot maximize the efficiency of the rescue operation. Occasionally, clues and evidence of victims are so minute that they evade camera and machine detection. For this reason, we introduced Speech Recognition. We may make our machine human-finding-proof by receiving and identifying distress calls on location. In this research, we focused on Bangla Speech Recognition in disastrous environments.

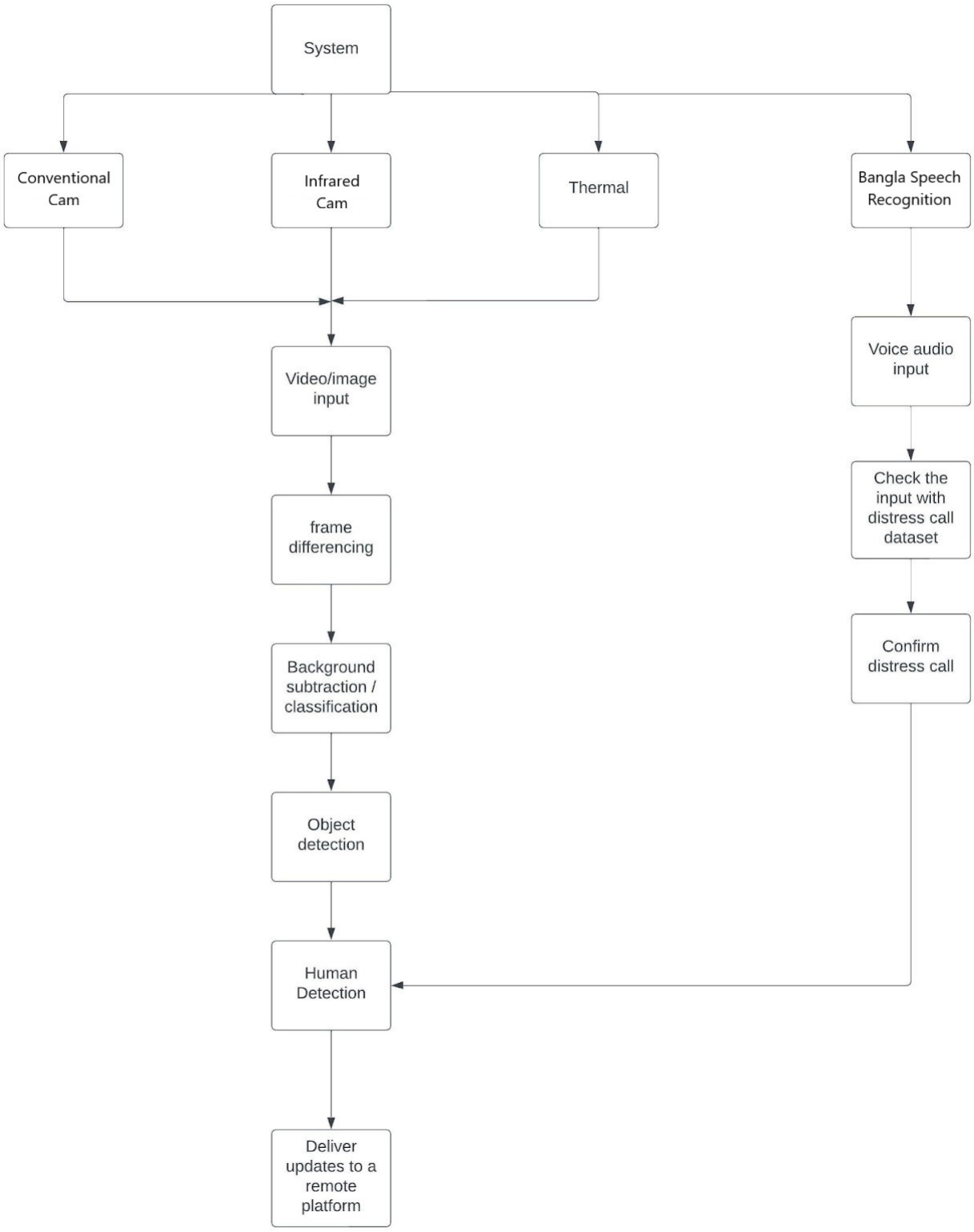


Figure 3.1: Workflow of the Multi-modal Human Life Detection System

After gathering information from different sensors and process them by training and checking with different levels of data our system gets the result containing specific are of victims being stuck that need help. Our machine delivers this data frequently with a short interval of time to a remote platform which is far from site but can be

accessed by the rescue team to work with this information instantly.

3.1 Characterization of Data

It can be difficult to find existing datasets that are fully suitable for our specific research. This is because datasets are often created for specific purposes, and may not contain all the necessary data or may contain extraneous data that is not relevant to the current research. Additionally, the data in existing datasets may not be collected in the same way or under the same conditions as the current research, which can make it difficult to compare the results. In the case of our research on human life detection in unconstrained environments using conventional, infrared, and thermal cameras, we have had difficulty finding existing datasets that were collected in a variety of unconstrained environments or that use all three types of cameras. For example, the thermal datasets that we found were not shot in an unconstrained environment. The test subjects were not blocked by any obstacles and their postures were not similar to a trapped or injured human. As a result, we had to collect our own primary data in order to conduct the research from a pragmatic point of view.

We mainly used two datasets for this research. In one dataset we kept all the conventional and infrared images and the other dataset is of thermal images. The reason we kept conventional and infrared images together is because our training model will work on both the images in the same way. Based on the lighting of the area, our cameras will switch but the final images will present a clear vision and based on that frame our model will try to detect alive humans. As, both types of images are trained with exactly the same class names hence they are placed in the same dataset. This dataset has a total of 5,428 images and they are splitted into three sets. Among them 4,428 images are in the training set (82%), 500 images are in the validation set (9%), and 500 images are in the testing set (9%). The other dataset that we used is for thermal images only which consists of 1,659 thermal images which is again our primary dataset. The splitting ratio among Training Set, Validation Set and Testing Set is respectively 70%, 20% and 10%. So our testing set has 1,159 images, the validation set has 330 images and the testing set has 170 images in total. There are many challenges that can arise when collecting data for a research, especially when the research is being conducted in unconstrained environments. Some of the challenges that we have encountered in our research include:

- **Access to the necessary equipment**: In order to collect data using conventional, infrared, and thermal cameras, we needed access to these types of cameras. For the conventional photos, we used our phone camera but we did not have access to any infrared or thermal camera. As a result our data collection was delayed till we arranged funding to buy an infrared camera and thermal camera. Moreover, while data preprocessing and training our device ran out of resources sometimes as most of the data related work was done in an old desktop from 2017. The processor of our desktop was Intel i5-7500, ram 16 GB DDR4 and a 4 GB GTX 1050ti graphics card.

- **Environmental factors**: Unconstrained environments can be unpredictable

and may present various challenges to data collection. For example, the weather, lighting conditions, and terrain can all affect the quality of the data that is collected. It was impossible to actually visit an incident zone to capture data as we cannot predict/interrupt an incident zone beforehand. As a result we had to stage the scenarios to take data that is close to the actual disastrous environment.: Unconstrained environments can be unpredictable and may present various challenges to data collection. For example, the weather, lighting conditions, and terrain can all affect the quality of the data that is collected. It was impossible to actually visit an incident zone to capture data as we cannot predict/interrupt an incident zone beforehand. As a result we had to stage the scenarios to take data that is close to the actual disastrous environment.

- **Ethical considerations**: When collecting data from human subjects, it is important to obtain their informed consent and to protect their privacy. As we needed our test subjects to play the role of trapped or injured humans, we needed as many realistic postures as possible. Many actors were not comfortable with that, so finding people who voluntarily will help us in this cause was a big challenge.

- **Data storage and management**: Collecting a large amount of data can be challenging, as it requires careful planning and organization to ensure that the data is stored and managed properly. As we had to create data from different places and different devices. All the data was needed to manage at a central repository to reduce confusion. As most of the online repositories do not offer large storage for free we initially faced some difficulties. Especially some video footage that we collected to extract data that weighed a few gigabytes.

Overall, data collection can be a challenging aspect of any research, but it is especially so when the research is being conducted in unconstrained environments.

3.1.1 Conventional Image Data

In a typical disaster scenario, there may be sufficient daylight to spot humans without an infrared or thermal camera. For this scenario, we have collected a substantial number of images captured by a few smartphone cameras with sensors capable of capturing more information. Our exclusive dataset comprises two distinct image types. A small amount of this collection comprises photographs taken from several film scenarios portraying humans stuck in an unrestricted environment, as well as a few stock images. The majority of this dataset is clicked and organized by our team. We attempted to recreate a disastrous location and used objects to limit the vision of the entire human body. Moreover, various human body positions are mimicked when the photographs are taken. The devices that were used to take these photographs are IPhone 13 pro max, IPhone 12, IPhone 11 pro max, Samsung Galaxy

A71, One Plus 8 and Samsung Galaxy A01. All the photographs were taken with their built in camera application, we did not use any third party camera applications or filters while collecting the images. We used people of different age groups and skin tones of our region.

Figure 3.2: Conventional Data (Before Annotation)

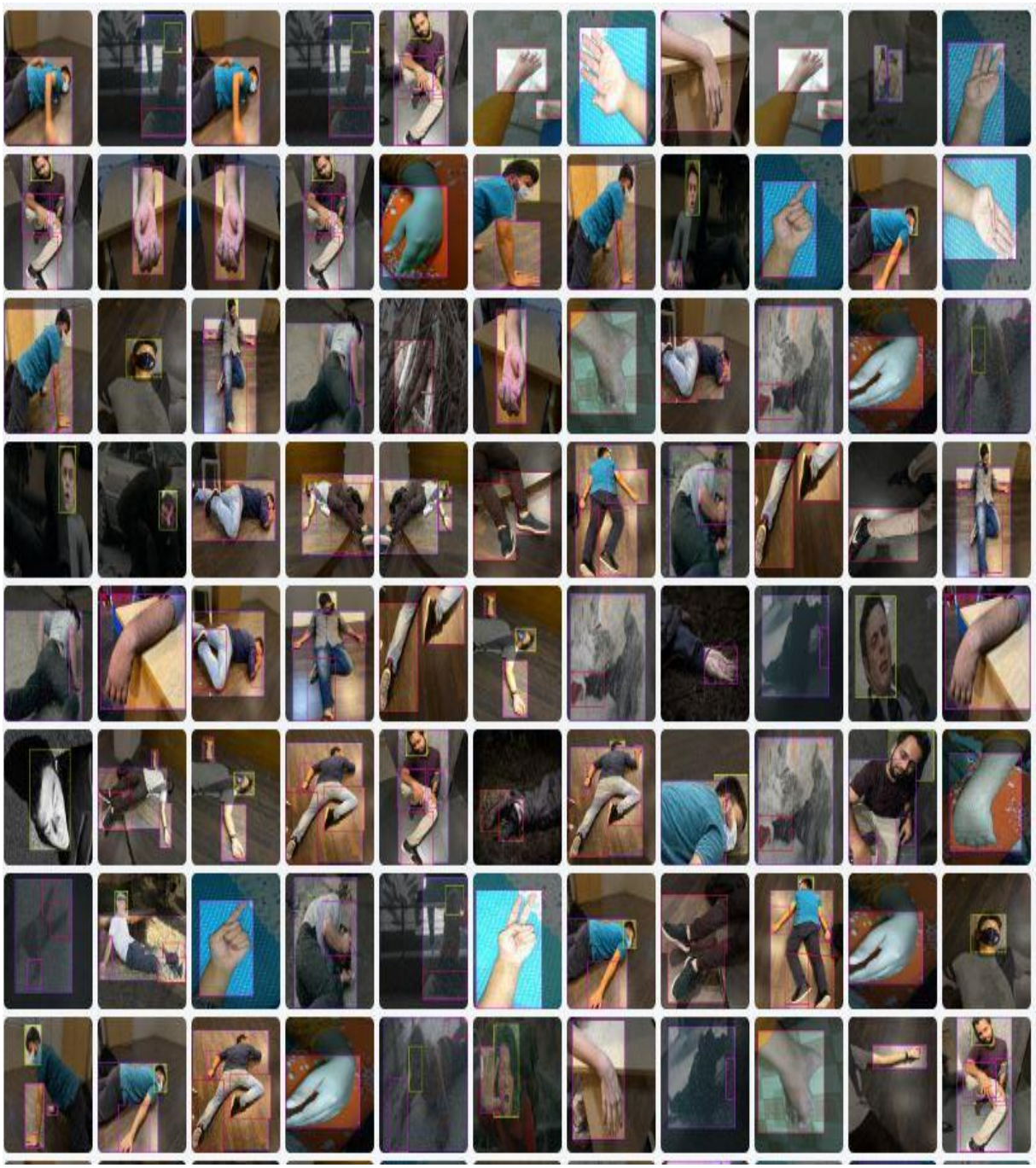


Figure 3.3: Conventional Data (Annotated)

3.1.2 Infrared Image Data

It is much more challenging for rescue workers to help victims of a disaster if the incident takes place at night or if the area where it occurred is immediately obscured from the sunshine. To this day, one of the most difficult challenges involves identifying people throughout the night or in dimly lit locations. This problem will be remedied when we integrate a night vision camera into our module. This camera will capture images during the night and run them through a processing system to identify individuals. A night vision device, also known as an infrared (IR) sensor, detects and amplifies low levels of infrared light not visible to the human eye. This allows the device to create an image of the environment in complete darkness or low-light conditions. There are two main types of night vision technology: thermal imaging and image intensification.

Thermal imaging uses a thermal camera to detect the infrared radiation emitted by objects in the scene, and then converts it into an electrical signal that is processed to create a thermal image. This image is typically displayed as different colors, with warmer objects appearing as brighter or redder colors and cooler objects appearing as darker or bluer colors.

Image intensification uses a special tube called an image intensifier tube to amplify the available light in the scene. This tube contains a photocathode which converts the available light into electrons, these electrons are then amplified and projected onto a phosphor screen which produces an image visible to the human eye. However, the dataset that we require in order to train our model is quite rare as most of the night datasets contain pedestrians or living beings roaming around. We required a dataset that was a match for our research environment and contained some obstructions in the form of objects. As a result, we developed a specialized solution just for this issue. We used a V380 2MP HD 1080P Night Vision Wireless Wi-Fi Ip Camera, which comes equipped with its own infrared sensor, so that we could generate this individualized data set. The wide-angle rotating feature of this camera, which covers 360 degrees horizontally as well as 120 degrees vertically, enables it to capture photographs in total darkness at a distance of roughly up to 16 feet, making it ideal for low-light situations. In order to make the data more applicable to our investigation, we simulated a dark environment by positioning some objects so that they partially obstructed views of the human body. The environment was the one that came the most close to actually being a disaster zone.

Figure 3.4: Infrared Image Data (Not Annotated)

Figure 3.5: Infrared Image Data (Annotated)

3.1.3 Thermal Image Data

Unlike traditional cameras, thermal cameras capture images based on the temperature of the objects in their field of view by detecting the infrared radiation that is being emitted by the objects and converting it into an image. Traditional or in another word Conventional cameras produce images based on the visible light that is absorbed by the remaining objects, whereas thermal cameras produce images based on the temperature of the objects within the scope of their vision. This allows thermal cameras to see in complete darkness as well as through some types of substances like smoke and fog that would normally obscure their view. The resulting picture, which is known as a thermogram, illustrates the distribution of the temperature which is shown area wise by making use of different colors to signify the various temperatures, was made as a consequence of the experiment.

The use of thermal imaging, a relatively recent development in image technology, is able to detect and visualize the radiation in the far infrared band of the electromagnetic spectrum which enables the cameras to capture images of the surface's temperature. The idea that all objects with temperatures above absolute zero Kelvin (-273 ° C) emit infrared energy is the foundation of the black body radiation theory. This theory states that the intensity of an object's infrared emission increases as the temperature of the object rises above absolute zero Kelvin. This can reveal information that is not easily obvious to the naked eye, such as the position of hot spots in an electrical panel or the presence of a person hidden in the dark. The information can be revealed in this way that would otherwise be hidden from plain sight. Even if the specific target doesn't have a good visual contrast, which makes it difficult to differentiate the target from its surroundings, and thermal cameras make it possible to quickly notice heat emanating from the target rather than watching light. This is the case even when viewing light. In these circumstances, the use of thermal cameras is a viable alternative that may be considered.

But there were not enough thermal datasets that could match the environment related to our research. So, to make the data more pertinent to our research, we created our primary thermal dataset. To create the primary thermal dataset, we have used A-BF RX-450 Infrared Portable Thermal Imaging Camera, a small and portable camera, which possesses the ability to measure objects temperature in between -15℃ to 150℃ as well as 8∼14µm spectrum band. This 8∼14µm spectrum band indicates Long-Wave infrared - LWIR band which can also be called "thermal IR radiation". Imagers with LWIR band can detect distinct temperature differences or radiated temperatures that indicate the operator with critical information about the target. Moreover, the image captured with this imager has resolution of 256×192. The thermal imaging portable camera was connected to a Samsung Galaxy to M31 and for capturing the thermal images we used an android application called "bufan" as the default camera application is not able to use the thermal camera.

Figure 3.6: Electromagnetic spectrum with illustrated IR segments wavelengths in μm [26]

We created different kinds of settings where some portions of the data-set were captured in a completely dark environment and some were in daylight. Additionally, various things were employed to obscure the shape of the human body in some places. Non-human living things as well as items with high temperatures were also used as objects to partially block the human body. These things contribute to an improved ability to differentiate between human components like in the trapped places, there will be a huge possibility of existing a good number of ACs, wires, machines etc. In that case, the false data which was created in our data-set will help to distinguish if those shapes match human parts or not. Since the background is unimportant in thermal photos, we have aimed to capture several anatomical features, such as the hand, face, leg, and body, from every conceivable viewpoint.

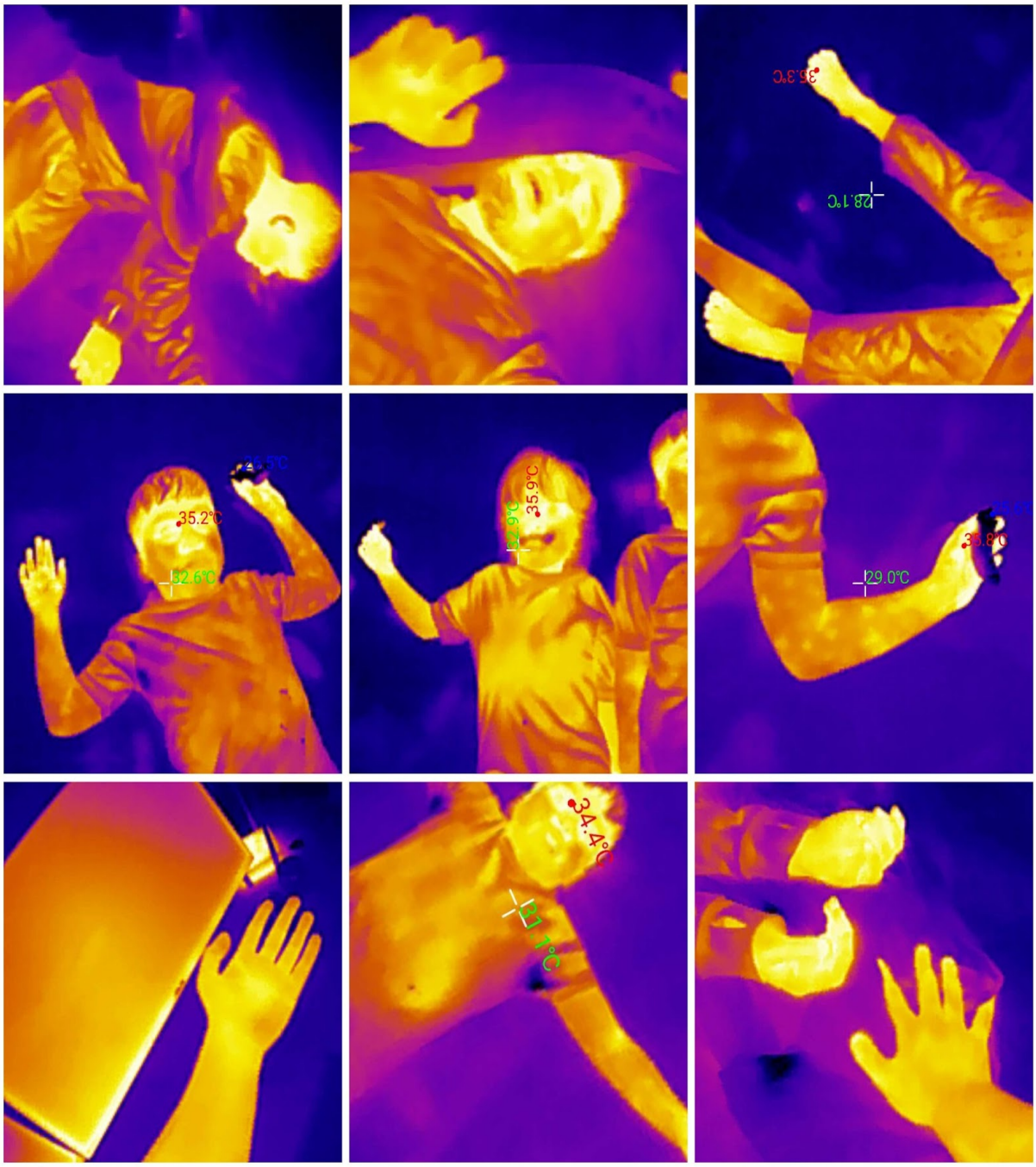


Figure 3.7: Thermal Image Data (Before Annotation)

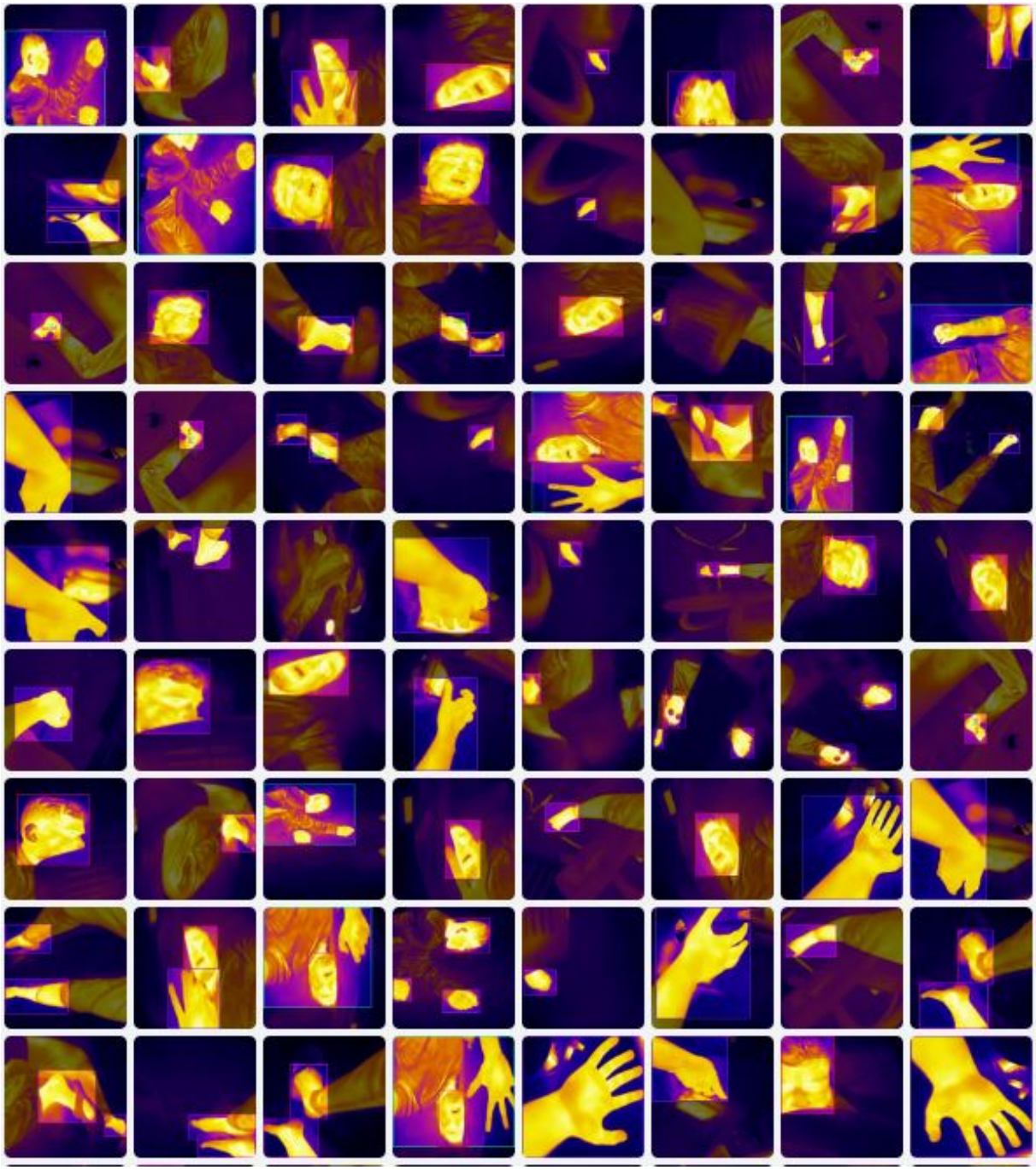


Figure 3.8: Thermal Image Data (Annotated)

3.1.4 Data Annotation

Annotation of data is the process of labeling and categorizing the data to make it more useful for machine learning and other analysis. There are many annotation tools available but we chose Roboflow for several reasons. Firstly, Roboflow is a cloud-based platform that allows for easy collaboration among our team members, making the annotation process more efficient. Secondly, Roboflow offers a variety of annotation tools, including bounding boxes and polygons, which allows accurately annotating our data, even for difficult objects like the human body. Moreover there are some additional features like data pre-processing, augmentation etc.

Here is a step-by-step breakdown of the annotation process that we followed:

- First, we uploaded all of our primary data images to the Roboflow platform.
- Next, we created a new project and defined the classes that we would be using for annotation (hand, leg, face, and human body).
- We then assigned each image to ourselves or our team members for annotation.
- For annotation, we used the bounding box tool to draw a rectangle around each object in the image. We labeled each bounding box with the appropriate class.
- We repeated the annotation process for all images in our dataset.
- After all the images were annotated, we reviewed the annotations to ensure that they were accurate. We made any necessary corrections.
- We exported the annotations in the format required for our machine learning model.
- Lastly, we used these annotations to train our machine learning model.

By using Roboflow, we were able to quickly and accurately annotate our primary data, which is essential for training a machine learning model. The platform's collaboration feature and annotation tools made the process efficient and streamlined. It helped us to work with our team members in real-time and make sure that the data is properly labeled.

3.1.5 Data Preprocessing

As part of our research, we needed to preprocess our primary data images before using them to train a machine learning model. One of the issues we encountered was that the images were of different sizes, which can cause problems during the training process. To address this issue, we used the data preprocessing capabilities of Roboflow. We selected all the images in our dataset and applied a transformation to stretch them to a fixed size of 416 x 416 pixels. This step was important as it ensured that all images in our dataset were of the same size, which is a requirement for many machine learning models. We also used the feature auto-orient to match the orientation of all the images.

In addition to resizing the images, we also performed other preprocessing steps using Roboflow such as data augmentation, normalization, and cropping of images. This helped us to improve the performance of our model by increasing the size of the dataset and reducing overfitting. The platform's user-friendly interface made it easy to apply various preprocessing steps, which saved us a lot of time and effort. As part of our research, we wanted to increase the amount of data available to train our

machine learning model. To achieve this, we used the data augmentation feature of Roboflow.

One of the data augmentation techniques we used was blurring. Using Roboflow's user-friendly interface, we selected all the images in our dataset and applied a blur with a maximum of 1.25 pixels. This helped us to create additional data by slightly altering the original images, which can improve the performance of our model by reducing overfitting. Another data augmentation technique we used was adding noise. We added random noise with a maximum of 5% to our images. This helped us to create additional data by introducing variations in the original images.

Overall, the use of Roboflow's data preprocessing capabilities allowed us to quickly and easily prepare our primary data for use in training a machine learning model. By using these data augmentation techniques, we were able to increase the amount of data available for training our model, which can lead to better performance. The use of Roboflow allowed us to easily and quickly apply these techniques to our primary data, which saved us a lot of time and effort.

3.1.6 Audio Data

Various audio formats are considered for the speech recognition process. Three popular audio file formats were used in this research namely MP3, FLAC and WAV. MP3 format is a lossy compression format. During the process of compressing the data, we can lose information. On the other hand, FLAC is a loss-less compression format. So, it also compresses the data but it allows us to perfectly reconstruct the original data. WAV is an uncompressed format. It stores the data in an uncompressed way resulting in the audio quality being the best but also the file size being the largest[13].

In our research, we have used 6 phrases which are predominantly used by the victims during post disastrous situations. These 6 phrases contain 24 words and 17 unique words. In Figure-3.9, the phrases that are used for research are given. These phrases were recorded in a real life environment which includes commute noises as a background noises.

Bangla Phrases	Number of words
হ্যালো আমাকে কি শোনা যাচ্ছে	5
আমি বিল্ডিঙের নিচে আছি	4
আমাকে সাহায্য করুন	3
আমার এখানে খাবার ও পানি দরকার	6
আমি এখানে আছি	3
আমাদের সাহায্য দরকার	3
	Total: 24

Figure 3.9: The predominant Bangla Phrases heard from victims during disastrous situations

Recording Audio

The voice is recorded using a microphone. For recording the voice, PyAudio library is used. PyAudio provides bindings for PortAudio v19 which is a cross-platform audio I/O library. After setting up the parameters named frames per buffer, frame rate, channels. The initial time for recording the voice is 60 seconds. The recording is made in WAV format as no noise gets lost in WAV format.

Plotting Audio

After recording the audio in wav format, it is then plotted to get a clear understanding of the audio properties. An audio spectrogram is created based on the recorded audio. The spectrogram contains frequency on the y-axis and time on the x-axis. The spectrogram shows color change and density whenever there is prominent sound present in the audio.

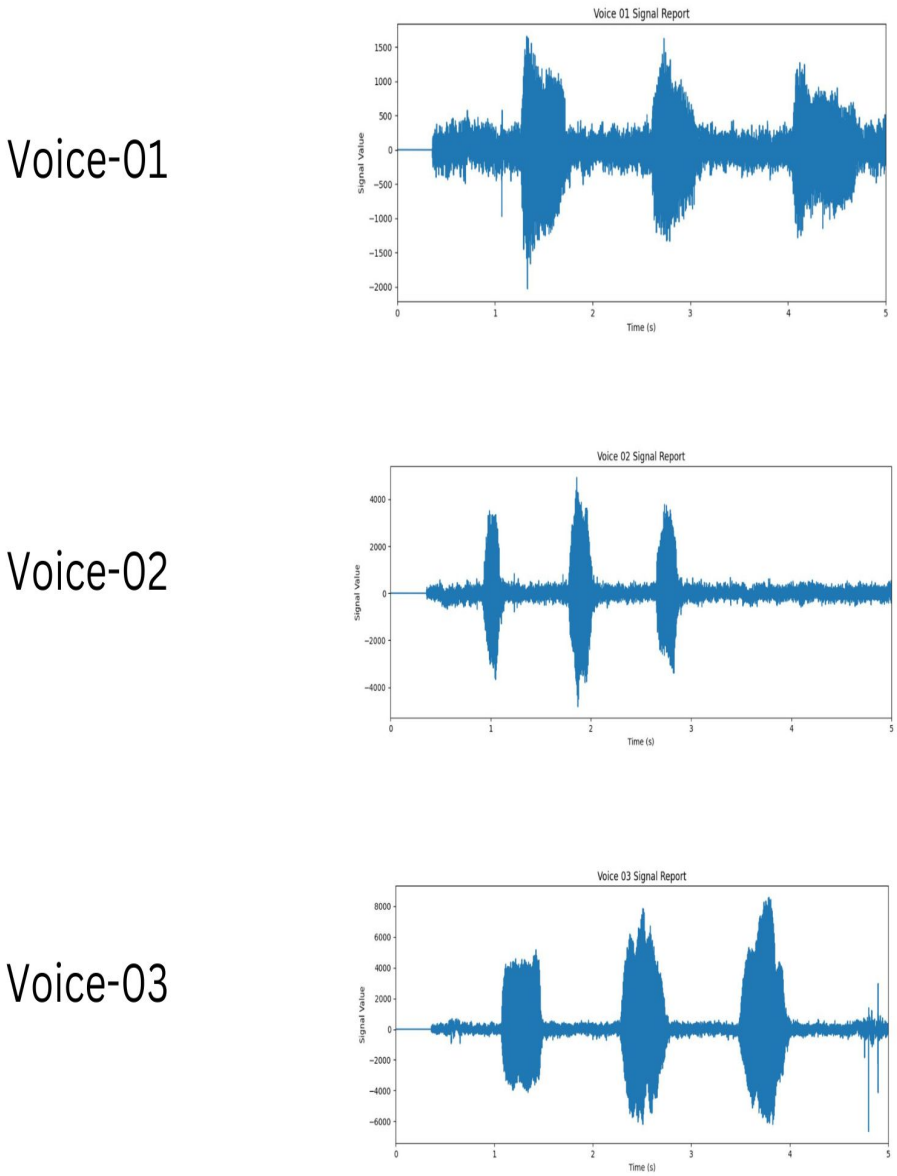


Figure 3.10: Plotted Signal of various Audio Inputs.

Speech Detection

For speech detection we utilized the Speech Recognition library in our system. The SpeechRecognition library is an open-source library for performing speech recognition in Python. It provides a set of methods for recognizing speech from audio input and for adjusting the audio input for background noise. The system employs an ambient noise reduction feature that helps to improve the accuracy of the transcription by adjusting the audio input for background noise. This is accomplished by using the adjust_for_ambient_noise() method of the Recognizer class, which analyzes the ambient noise in the audio input and applies a noise reduction algorithm to the

audio before processing it further.

The system continuously listens to the microphone using the listen() method of the Recognizer class. This method listens to the audio input from the microphone and records it until silence is detected. The duration of the recording can be controlled using the duration parameter, but in our system, it is set to none so that the system can continuously listen to the microphone. Once speech is detected, the system proceeds to the next step of text conversion. The system also includes error handling mechanisms to handle situations where speech cannot be detected. For instance, if there is no speech input or the audio input is too low, the system will raise an exception, and the user will be prompted to try again.

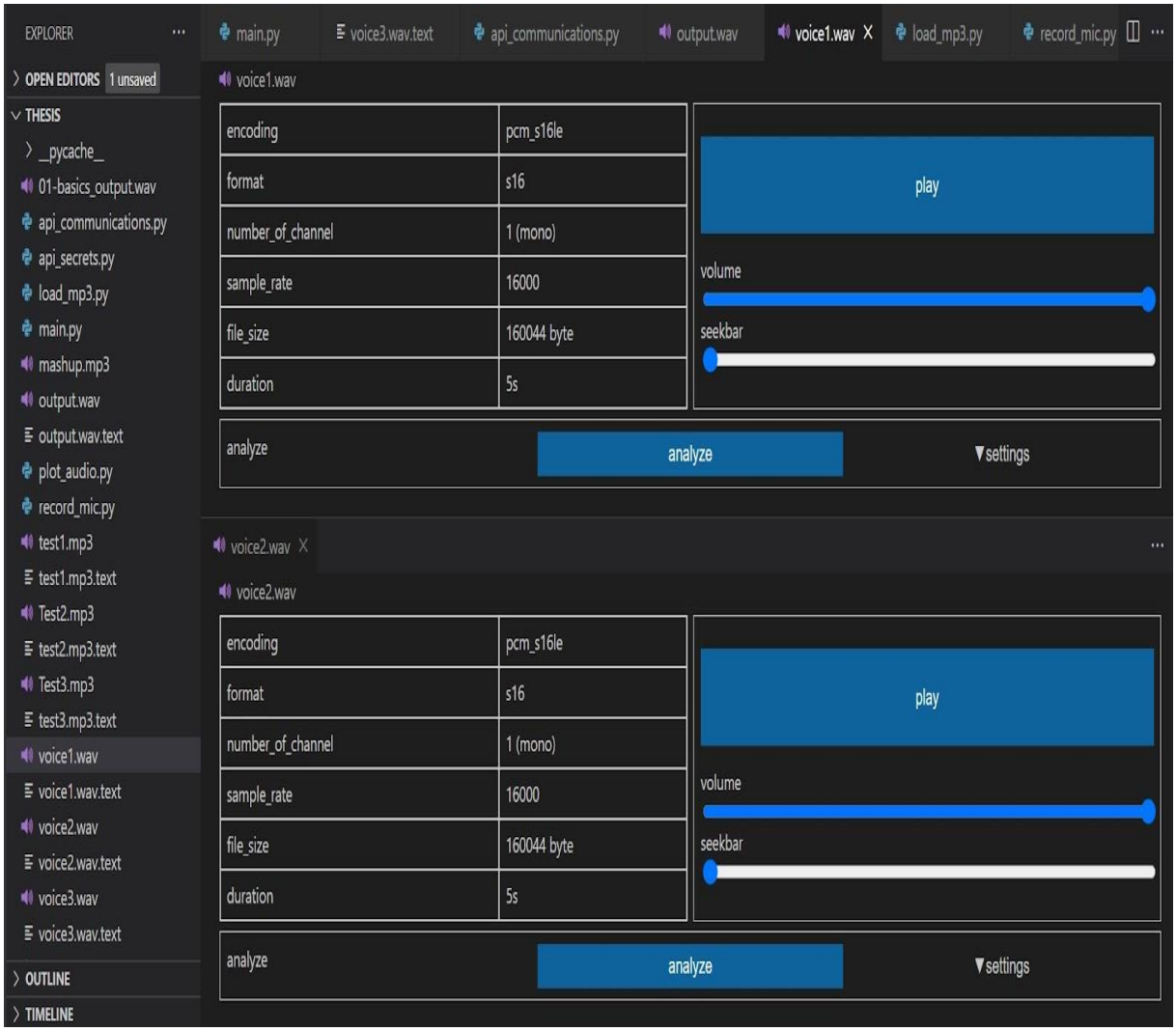


Figure 3.11: Audio Data Processing in WAV format

Text Conversion

For text conversion we used Google's Speech-to-Text API to transcribe spoken words into text. The API is capable of recognizing a wide range of languages, including Bangla. The system uses the recognize_google() method of the Recognizer class, which sends the audio input to the Google Speech-to-Text API and receives the

transcribed text in response. The language parameter of the method is set to "bn-BN" to indicate that the input speech is in the Bangla language.

The system also includes error handling mechanisms to handle situations where the API is unavailable or the speech cannot be recognized. If the Google Speech-to-Text API is unavailable, the system raises a RequestError exception. If the speech is not recognized, the system raises an UnknownValueError exception. The system has been tested on a variety of spoken inputs and has been able to accurately transcribe spoken words with a high degree of accuracy. Additionally, the system can also be used to transcribe speech from pre-recorded audio files.

Chapter 4

Description of the models

4.1 Mask R-CNN

Convolutional Neural Networks (CNNs) such as Mask R-CNN are cutting edge when it comes to segmenting images. This particular Deep Neural Network variation recognizes items in a picture and creates superior segmentation results for every one of them. Mask R-CNN is an extension of Faster R-CNN. Mask R-CNN was created on top of Faster R-CNN. It is an instance segmentation model. Faster R-CNN is a type of region-based convolutional neural network that provides bounding boxes and a confidence score for each object's class label[21]. Mask R-CNN adds a branch for predicting ROI in parallel with the remaining branch which is used for recognizing bounding boxes.

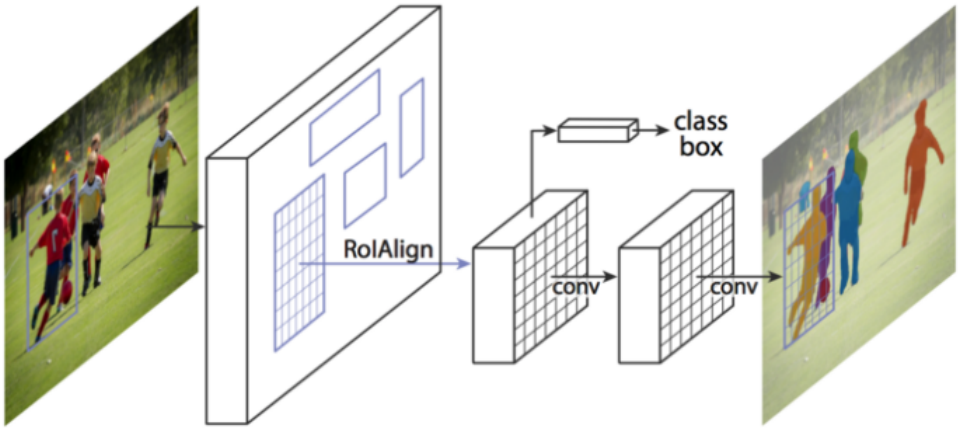


Figure 4.1: Mask R-CNN framework[16]

Using the Matterport Mask R-CNN implementation[32], we conducted the experiment. NVIDIA GeForce GTX 1050 Ti is used for a single GPU to run each training session. Training our night image dataset model which completed 20 epochs (5 steps per epoch). We ran object detection on a day-light image to make a prediction.

```
Classes: ['BG', 'hand', 'leg', 'face']
Processing 1 images
image                    shape: (416, 416, 3)         min:    0.00000  max:  255.00000  uint8
molded_images            shape: (1, 1024, 1024, 3)    min: -123.70000  max:  151.10000  float64
image_metas              shape: (1, 16)               min:    0.00000  max: 1024.00000  float64
anchors                  shape: (1, 261888, 4)        min:   -0.35390  max:    1.29134  float32
```

Predictions1

Figure 4.2: Human detection using Mask R-CNN

4.2 Faster R-CNN

Faster R-CNN has two modules. First module is RPN which is used for generating region proposals and second one is fast-rcnn for detecting objects in the proposed regions. RPN produces regional proposal ideas. The ROI Pooling layer is used to extract a fixed-length feature vector from each of the image's area suggestions. The Fast R-CNN is then used to classify the extracted feature vectors. Along with their bounding-boxes, the class scores of the discovered objects are also provided. Faster R-CNN also uses anchor boxes to handle multi-scale object detection, which allows the model to detect objects of different sizes in the same image.

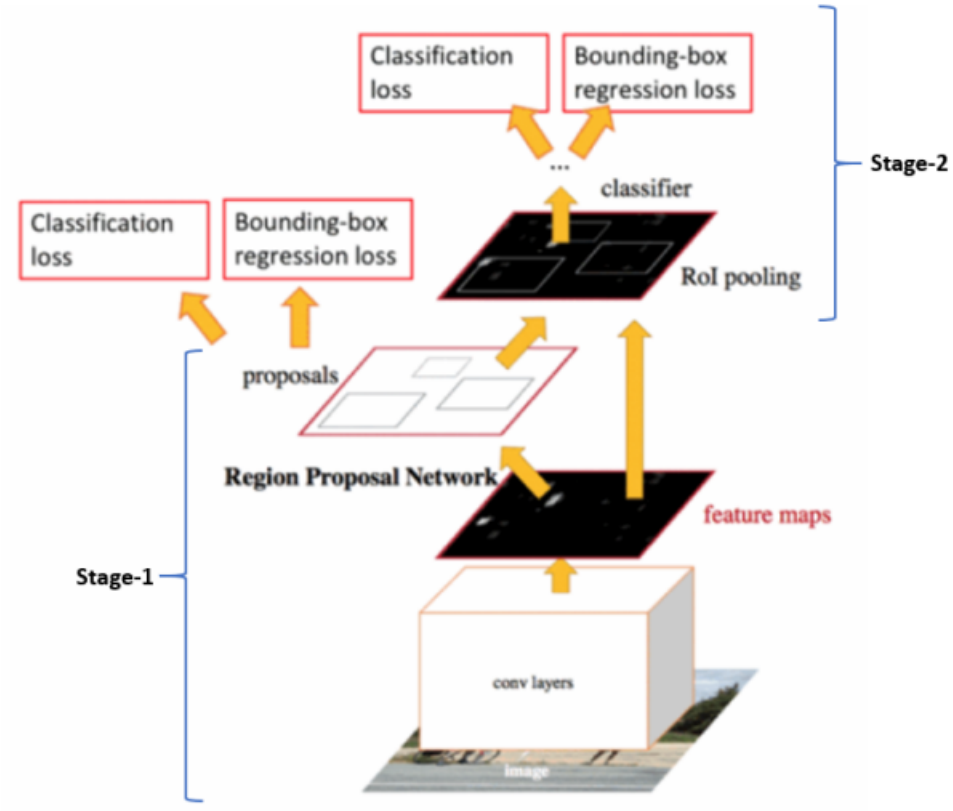


Figure 4.3: Faster R-CNN architecture[21]

During training, the model will be evaluated on the validation dataset after each training epoch and the performance metric, like mAP, is reported. This allows the developer to monitor the model's performance during training, and stop training when performance on the validation dataset stops improving.In machine learning, a batch size refers to the number of samples that are processed by the model before the model's parameters are updated.

We have used the dataset I which contains conventional and infrared images. The settings used to train the dataset are below:
Number of epochs: 100, Batch-size=16, img-size= 416 * 416.

```
"mode": "val",
"epoch": 16,
"iter": 500,
"lr": 0.02,
"bbox_mAP": 0.449,
"bbox_mAP_50": 0.804,
"bbox_mAP_75": 0.45,
"bbox_mAP_s": 0.151,
"bbox_mAP_m": 0.346,
"bbox_mAP_l": 0.496,
"bbox_mAP_copypaste": "0.449 0.804 0.450 0.151 0.346 0.496"
```


Figure 4.4: Accuracy on conventional and infrared images using Faster R-CNN

From the figure, we can see that we got 0.804 Mean Average Precision(mAP) @0.5. Here, "bbox_map_50" likely refers to the mean Average Precision (mAP) of a bounding box object detector, evaluated at an Intersection over Union (IoU) threshold of 0.5. Here, we took the 16th epoch because it has the maximum mAP value. Also, because the performance on the validation dataset had saturated, and further training would have led to overfitting.

We have used the dataset II which contains thermal images. The settings used to train the dataset are below:
Number of epochs: 100, Batch-size=16, img-size= 416 * 416.


```
"mode": "val",
"epoch": 21,
"iter": 330,
"lr": 0.002,
"bbox_mAP": 0.658,
"bbox_mAP_50": 0.919,
"bbox_mAP_75": 0.776,
"bbox_mAP_s": 0.0,
"bbox_mAP_m": 0.631,
"bbox_mAP_l": 0.658,
"bbox_mAP_copypaste": "0.658 0.919 0.776 0.000 0.631 0.658"
```


Figure 4.5: Accuracy on thermal images using Faster R-CNN

From the figure, we can see that we got 0.919 Mean Average Precision(mAP) @0.5. A mAP score of 0.919 suggests that the model has a high precision in identifying objects and drawing accurate bounding boxes around them. A mAP score of 1.0 would indicate perfect performance, but it's very rare to get such high scores. 0.919 is considered a high score and the model is performing well in object detection task. Here, We took the 21st epoch because the performance on the validation dataset had saturated, and further training would have led to overfitting.

4.3 YOLO v5

YOLO is a cutting-edge object identification technique that is so quick that it has become a de facto standard in computer vision for detecting things. Previously, sliding window object detection was used, and subsequently faster variants such as RCNN, fast RCNN, and faster RCNN were developed. In 2015, however, YOLO was developed, which surpassed all prior object identification algorithms, and it will be used for the implementation. Let's examine the architecture in detail to determine why YOLOv5 is quicker than any other algorithms in this sector. You Only Look Once is the full meaning of the acronym YOLO[12].

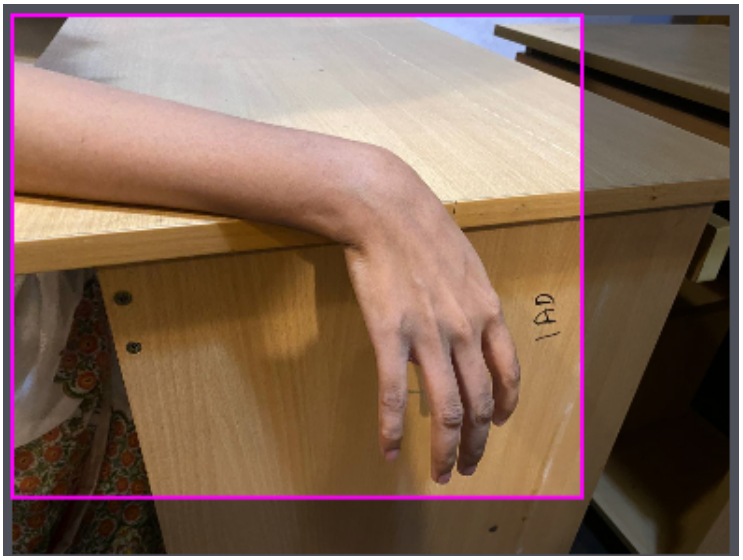


Figure 4.6: YOLO Object Detection Bounding Box

Our research's objective is to identify the human body in challenging circumstances, hence the output of the neural network for this condition is very straightforward. Suppose we are detecting the human hand, in which case the human hand is equal to one and the entire human body is equivalent to zero. However, when we employ object localization, we are not only identifying the class of the item. Therefore, the method determines the bounding box or location of an object within the Image.

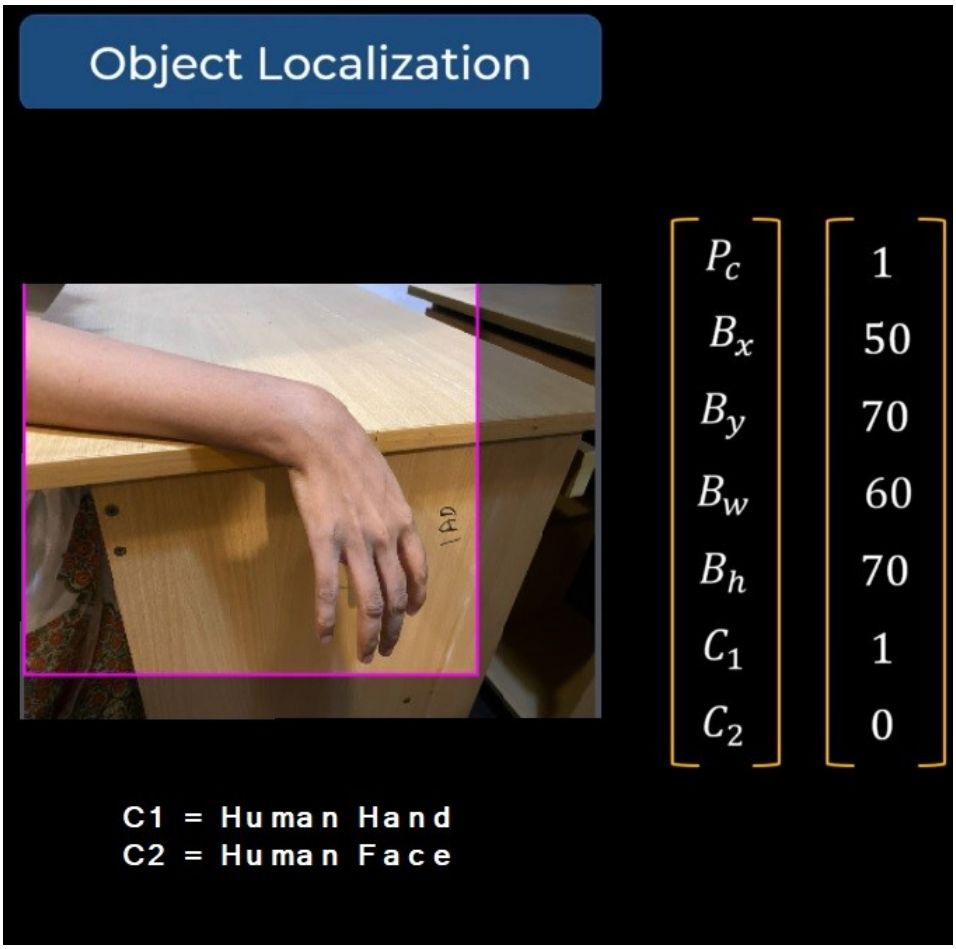


Figure 4.7: YOLO Object Localization

So in terms of neural network output, you can have a vector like this where pc is the probability of a class.

If a human hand or body is present, this number will equal one. If there are no human body parts, this number will equal 0 and the enclosing box will be empty. Bx, By is the center's coordinate, which is denoted by the Yellow circle. The width and height of this purple box are 60 and 70. C1 is the class for human hands. So here it will be one, but c2 will be 0 for the Human body (whole body).

In another image similar to this. Here is a whole human body. The pc probability of any class is 1, as there is always at least one object, and these probabilities are analogous to bounding box coordinates. C1 is 1 since it contains a human hand, and C2 is also 1 because it is a human body. If there is no object in the image, pc will be 0 and the other values are irrelevant. To train a neural network that classifies the item and the bounding box, several photos (such as these three images) must be parsed by the neural network, and since this is a supervised learning issue, bounding boxes must be given for each of these images.

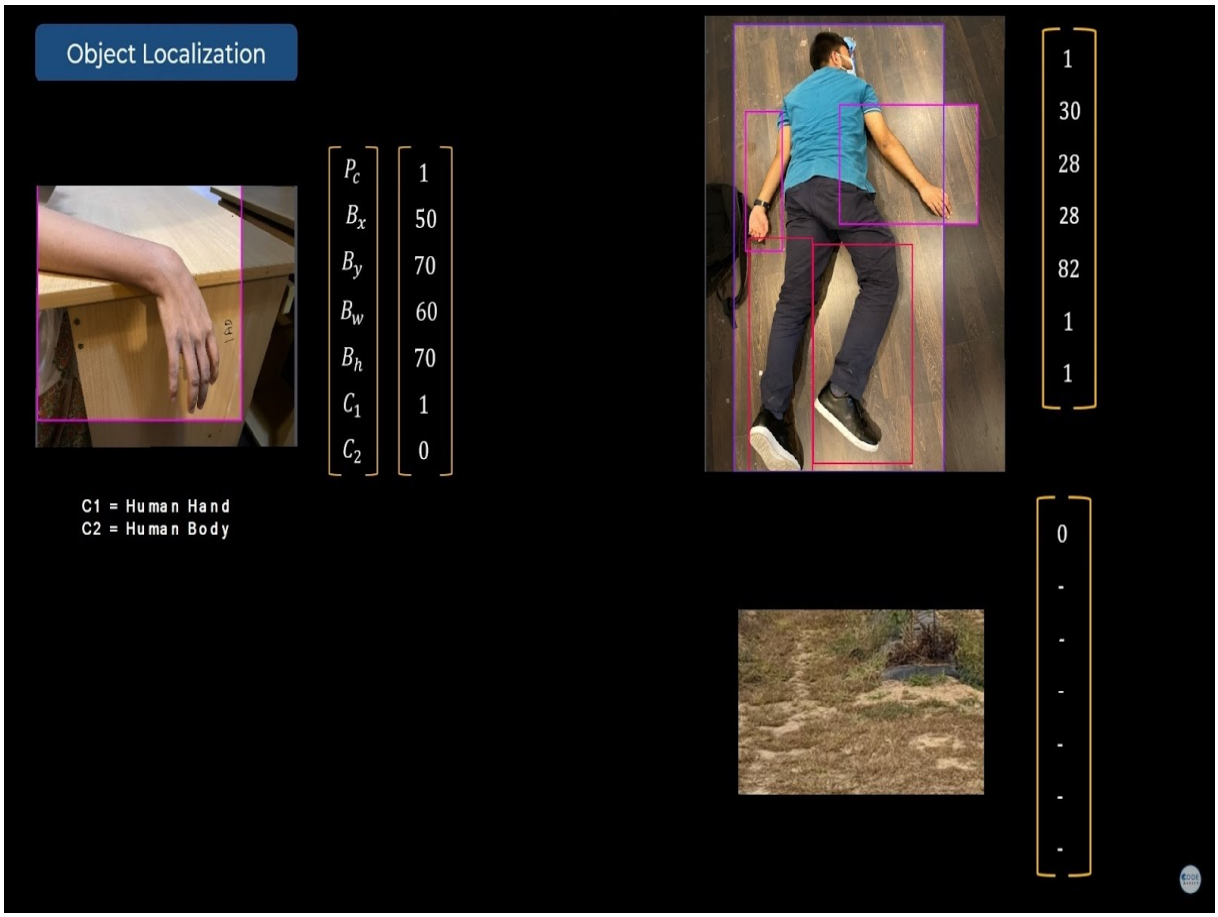


Figure 4.8: YOLO Object Localization

Providing neural networks with bounding boxes is insufficient. A neural network can only comprehend integers, thus this must be converted into this type of vector. There exists a vector of size 7 for each related picture, therefore x strain and y train will each be vectors of size 7. You can possess 10,000 similar photos. You can train a neural network such that if you input a new image, it will tell you that particular vector. This vector tells you that this is a dog since c1 is set to 1, and it also tells you the bounding box, so it provides the answer for your object recognition or localization. This graphic depicts both a human hand and the human body. A picture may contain n number of items, such as two human hands and three human bodies, or five human hands and a single human body. We cannot anticipate with certainty how many hidden objects and classes may exist on the premises.

Therefore, it is difficult to estimate the output dimension of a neural network with a single object. Determining the size of the neural network's output is difficult if you do not know the number of objects, n. The top limit is 10, thus if there are only 10 things, you can fit them into a vector of size 70, but what if there are 11 objects? Suppose that you have this image, which has two bounding boxes. What the YOLO algorithm will do is divide this image into grid cells of this type. Therefore, I'm using a 4x4 grid here. It might be either 3x3 or 19x19.

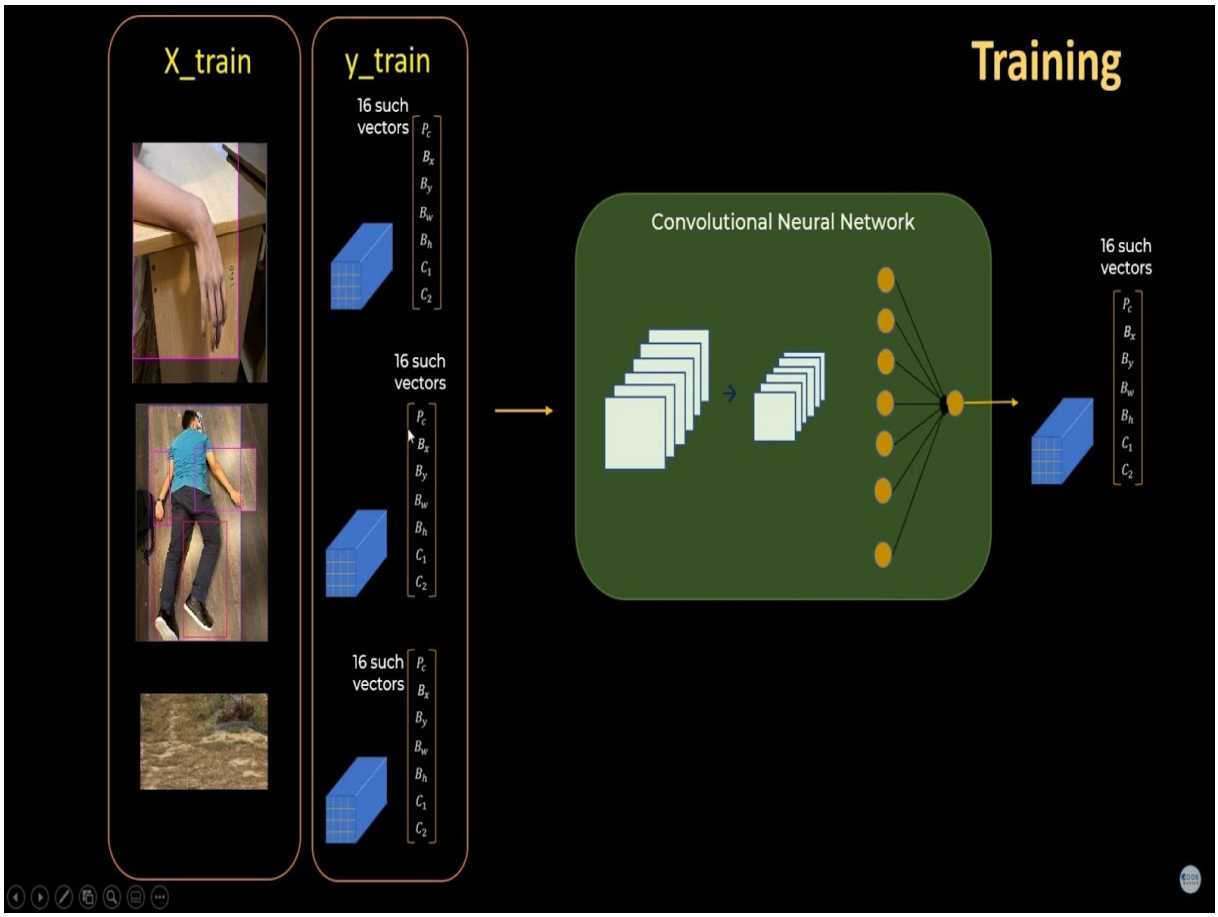


Figure 4.9: YOLO CNN Network

For each grid cell, such as this one, it is possible to encode and derive the vector representing the PC's bounding box. C1 and C2 There are no objects present, thus the probability of class is 0 and the other numbers are irrelevant. However, for each grid cell containing items, the algorithm must locate the hand's central location, as each hand corresponds to a certain grid cell. This specific cell links with the coordinates and considers each point to have zero Coordinates. Then this vector is created. The size of C is the breadth of this grid cell, which is three times larger than the previous cell. the percent will be 0 remaining Because there are 4x4r grid cells in total. Each cell contains a vector of size seven. This 7-dimensional vector is the result of expanding the top-left 4x4x7 cell in the Z direction. The picture is followed by the surrounding rectangles. The data will then serve as the training data set. A training data cell will include a certain number of photos.

Based on the bounding rectangle of each image, the algorithm attempts to extract its contents. It will initially create a 4x4 grid, a 3x3 grid, or a 19x19 grid. There will be one target vector for each cell, for a total of sixteen vectors per training sample or picture. Using this, you can train your neural network, and after training, you can do predictions.

This sort of picture may generate 16 such vectors and y 16 since it resembles a 4x4 grid, which informs the network of the bounding rectangle for each of these items. This is therefore the YOLO algorithm. It is not the same as entering it 16

times and performing 16 iterations in one forward pass. It is able to make all predictions, hence the name You Only Look Once. There may be a few problems with this strategy, as this is a basic algorithm that needs modification. We are using the fifth version of this algorithm. The model head of yolov5 is similar to its predecessors. It has also other two important components which are the model backbone and model neck.

The design enables faster detection of large amounts of data in a short amount of time. After examining a large diversity of algorithms, we determined that yolov5 is an approach that we should experiment with to achieve our prestigious aim. As we gathered images resembling the state and setting in which we want our system to operate, we were able to construct a dataset that is entirely our own. So, following the successful creation of this dataset, we intended to train the yolov5 algorithm and, using supervised learning, discover item classes in various photos. We did not need to perform any data preparation in order to train our yolov5 model because we built our dataset with roboflow and technological advancements have allowed us to forego this step. We exported the data in a selected yolov5 format, analyzed the dataset, and attempted to train the model. Initially, we felt that 10 epochs of 380-sized photos on 32 batches would serve as a solid foundation. As this was supervised learning, we utilized a dataset from roboflow that had bounding boxes for the four specified classes to train the yolov5 architecture.

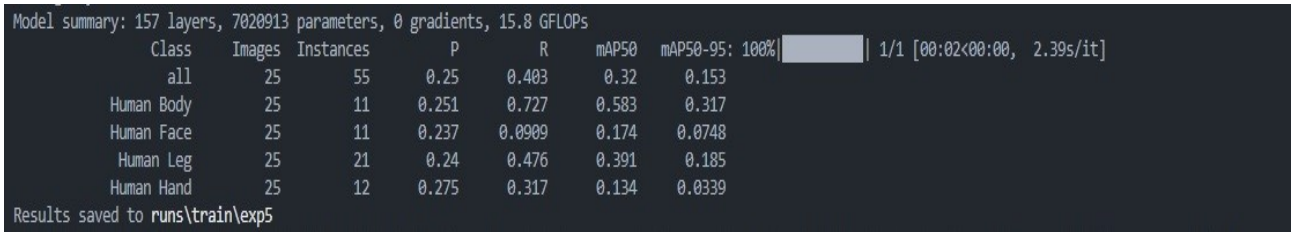


Figure 4.10: YOLOv5 training

After successful training, the model saved a concise summary of checkpoints indicating the model's training progress. We repeated the training with varying numbers of epochs and batches, and in certain cases, we also experimented with resolutions to determine whether or not the algorithm produces half-bounding boxes. For this training, however, we discovered that the bounding box must be orientated in a certain manner and that the whole bounding box must be collected.

After obtaining the training results, we used this model learning to the detection of classes on various photos. We utilized a variety of photos to determine whether or not our system can recognize human body parts under varying but equivalently challenging settings.

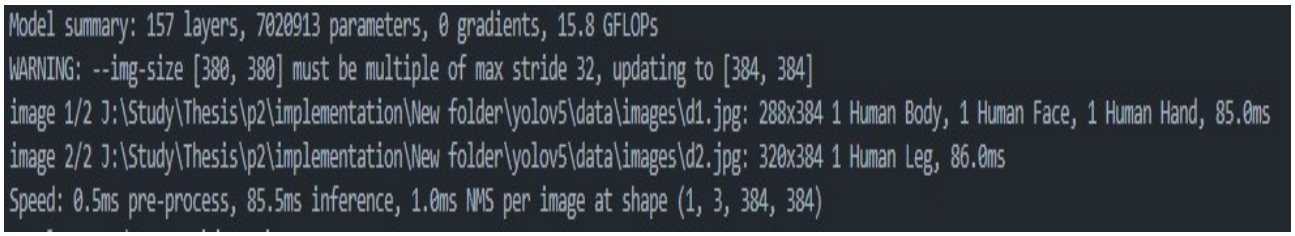


Figure 4.11: YOLOv5 class detection

We started modestly to determine whether or not our system is capable of identifying specific classes, and it was. Using the provided photos and the previous training, it correctly identified all classes. But we were not content with the accuracy score of yolov5 and our research found out tat there are better version of this backbone available for this type of training. Our primary dataset plays an important role for distinguishing the environment than we normally would see around us. So we wanted that our the proposed model that comes from our research should use this opportunity to detect humans in these unconstrained environments efficiently.

4.4 YOLO v7

After searching for a better model that understands what our proposed model needs and also use the full ability of our dataset to detect every classes we designed for human detection we came to know about YOLOv7. The Darknet-53 convolutional neural network (CNN), which has numerous convolutional layers followed by a few fully connected layers, serves as the foundation for the YOLOv7 architecture[22]. YOLOv7 can accurately and quickly recognize objects in a video or image thanks to this architecture. Small items can be found in huge photos by using the YOLOv7 object detector, which can identify objects of various sizes[12].

In contrast, YOLOv5 has a less sophisticated architecture based on the EfficientNet-B0 CNN and is less effective at real-time object recognition than YOLOv7. Furthermore, YOLOv5 is less proficient at detecting items of various sizes, which may hinder its ability to find little objects in huge photos.On the other hand, to recognize objects of various sizes and forms, YOLOv7 employs an image processing method known as anchor boxes, which are pre-defined boxes with various aspect ratios and scales. By using this method, YOLOv7 is able to recognize objects with a wider range of sizes, angles, and aspect ratios. Additionally, YOLOv7 uses non-maximum suppression (NMS) to get rid of overlapping boxes, increasing the object detection's overall accuracy[24]. Because of its more sophisticated design, YOLOv7 is better able to recognize things in real time and with high accuracy, especially small items in huge photos. Furthermore, the usage of anchor boxes and NMS in YOLOv7 enhances the object detector's robustness and general accuracy for object detection.

Even though yolov7 is successor of yolov5 but with Extended Efficient Layer Aggregation, Model Scaling Techniques, so many parameters to work with that handles image processing in so many ways in so little time we realized YOLOv7 is the model

that we should try our dataset on. YOLOv7 is a versatile object detector that can handle different types of images. We used day camera, night camera, and thermal camera so our model got to work with day images, night images, and thermal images. However, each type of image has its own unique characteristics and challenges, and YOLOv7 may need to be adapted to handle these different conditions.

One of the advantages of YOLOv7 is its high detection speed, which makes it suitable for real-time applications such as video surveillance and self-driving cars[14]. Another advantage is its ability to detect objects of different scales, which is helpful for detecting small objects in large images. in order to detect human bodies in an unconstrained environment using YOLOv7, we trained the model on a dataset of images that contain human bodies. This dataset included a variety of different poses, angles, and lighting conditions to ensure that the model can detect human bodies in different environments. The model gains the ability to recognize patterns and features in the photos that are unique to human bodies during the training phase. Once trained, the model may be used to identify human beings' pictures taken by the thermal, day, and night cameras. The model may not generalize well to diverse locations, stances, and lighting situations depending on the quality of our dataset and the quantity of training data. As a result, it's crucial to validate the model on data that is reflective of the intended environment. But in order for the model to understand that the data we are feeding it is genuinely effective at detecting humans in extremely limited environments, we made every effort to reproduce as many different environments as possible and prepare our dataset accordingly.

In the cases of conventional-day images, due to the ideal lighting, YOLOv7 can recognize humans in daytime photographs with great accuracy. The model is capable of quickly recognizing aspects of human bodies, such as limbs and facial features, as well as items of various sizes. It's crucial to remember that the model should be trained on a dataset that accurately depicts the surroundings of the target. For infrared/night images, the images have poor lighting. So YOLOv7 may have trouble identifying human bodies in night-time photos. In these situations, image processing methods like edge detection kicks in to increase the detection accuracy of the human body. Edge detection improved the image's contrast, making it simpler for the model to recognize aspects of the human body. We have used dataset-I which contains conventional and infrared images. The settings used to train the dataset are below: Number of epochs: 150, Batch-size=16, img-size= 416 * 416.

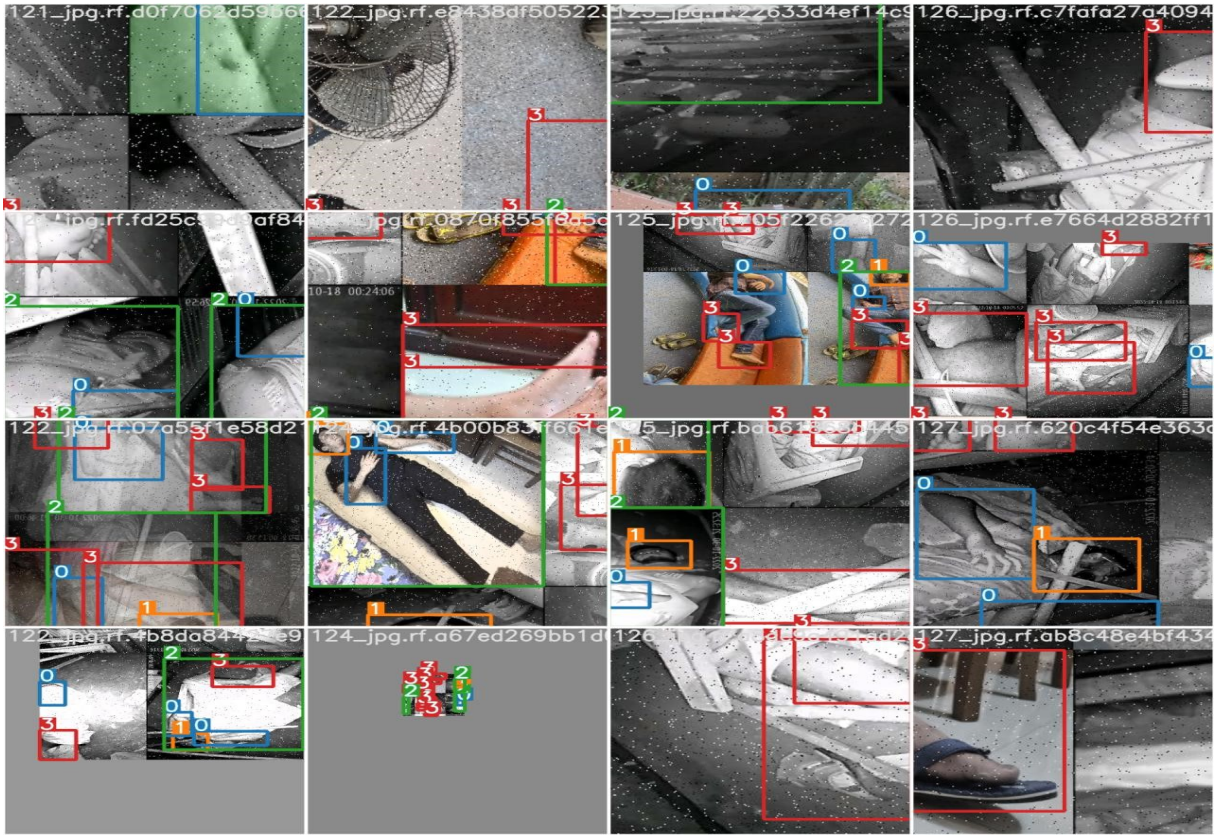


Figure 4.12: Train Batch of conventional and infrared images

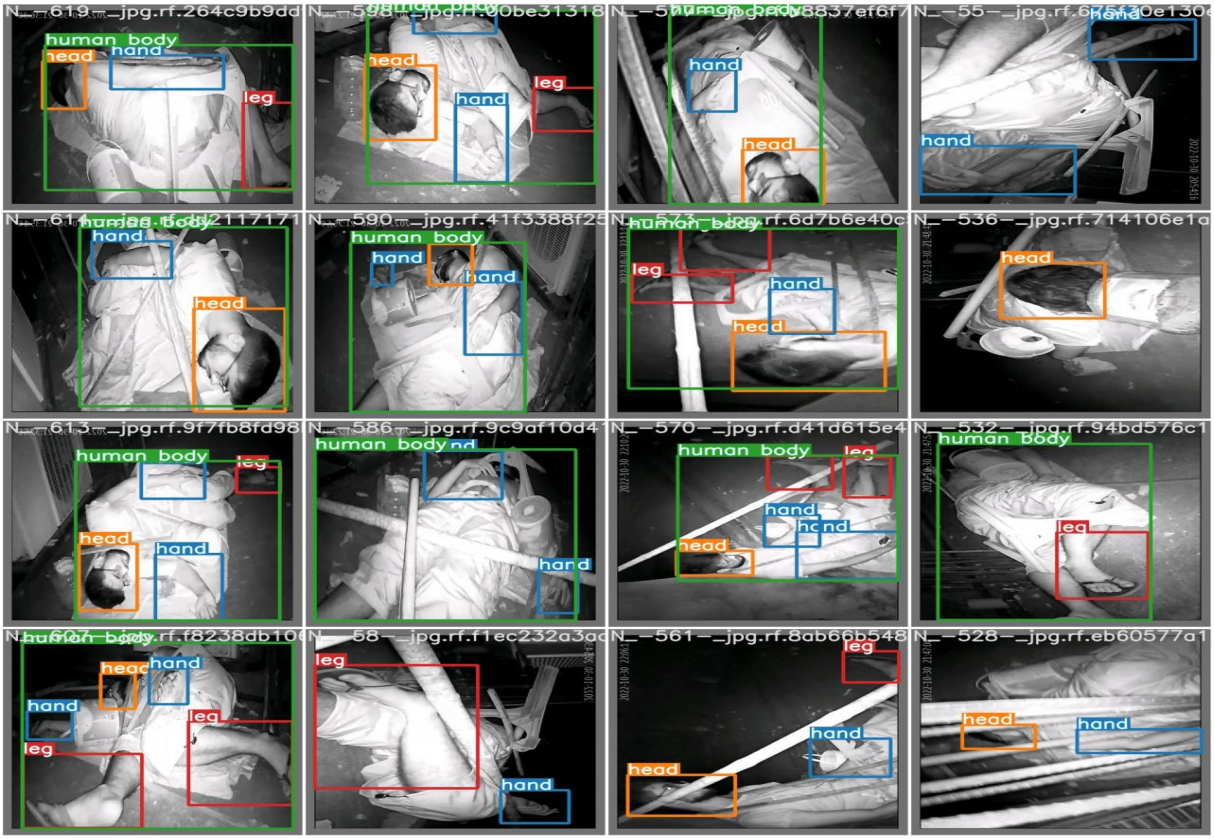


Figure 4.13: Test Batch of conventional and infrared images

The following graph illustrates the precision-recall trade-off for a particular classifier is called a precision-recall curve. The proportion of accurate positive forecasts among all positive predictions is known as precision, whereas the proportion of accurate positive examples is known as recall. The classifier's threshold is changed, and the precision and recall are plotted at each threshold to create the curve[25]. On our primary image dataset, we are measuring the performance of YOLOv7 using the precision-recall curve. The curve demonstrates the trade-off between maximizing the number of items detected (high recall) and reducing the number of false positives (high precision)[6].

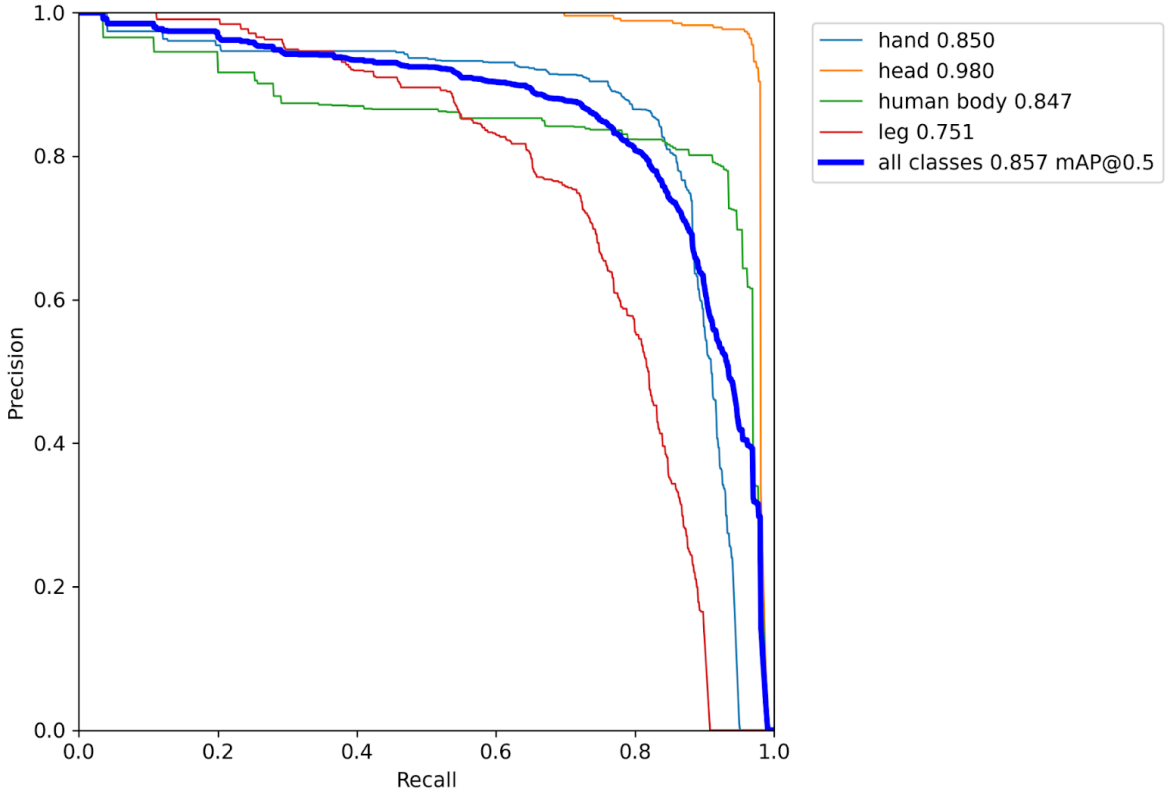


Figure 4.14: Precision-Recall curve of conventional and infrared images

The next figure represents the confusion matrix. In order to express how well a classification model performs on a set of test data for which the true values are known, it is common to use a matrix, which is a table. The matrix, which may also be used for multi-class classification, is frequently used to describe the results of a binary classification model. The count of true positives, false positives, true negatives and false negatives are represented by four separate cells in the matrix[6]. We can see in the matrix how many times the model successfully or erroneously identified an object. A clear image of the model's performance in terms of precision, recall, accuracy, and other metrics are also provided by the matrix.

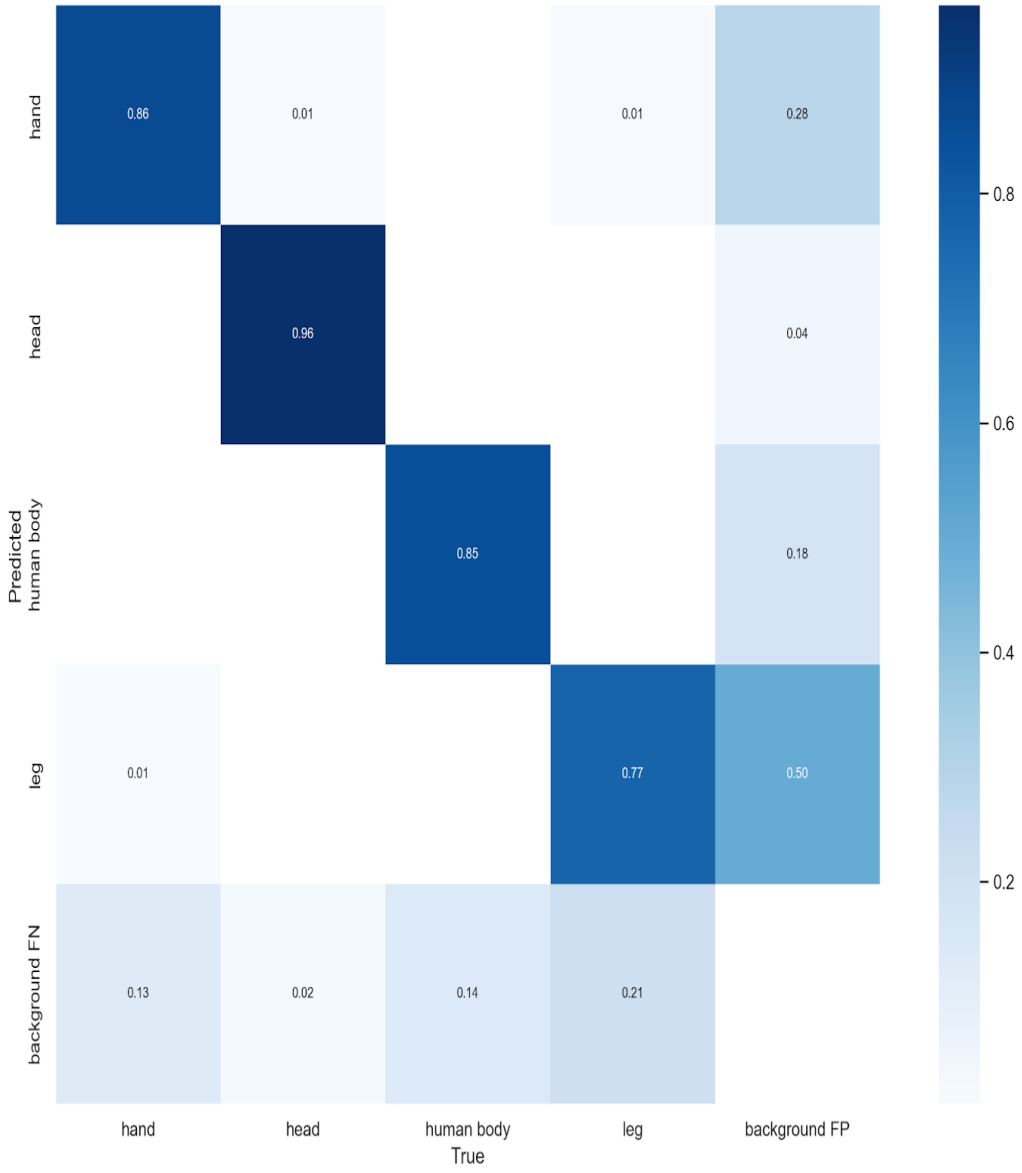


Figure 4.15: Confusion Matrix of conventional and infrared images

Mean Average Precision (MAP) is used to assess how well the model performs in terms of the average precision across all object classes. By averaging the precision at various recall levels, the average precision is determined for each class[6]. To get the overall MAP, the mean of all class-specific average precisions is calculated. In contrast to only examining the performance of the model on one class at a time, the MAP provides us with an overall picture of how well the model is performing by displaying the average performance of the model over all classes of objects[27].

So, from the Precision-Recall curve figure, we can see that we have got 0.857 Mean Average Precision (mAP) from training the dataset.

The human body will have a different temperature than other objects, hence YOLOv7 can detect people in thermal photos with excellent accuracy. The model may increase the traits of human bodies while filtering out non-human things via color thresholding. The model should be trained on a dataset that accurately reflects the conditions of the target environment, as thermal images have their own distinct properties, such as the absence of color information. We have used dataset-II which contains thermal images. The settings used to train the dataset are below:
Number of epochs: 150, Batch-size=16, img-size= 416 * 416.

Figure 4.16: Train Batch of Thermal Images

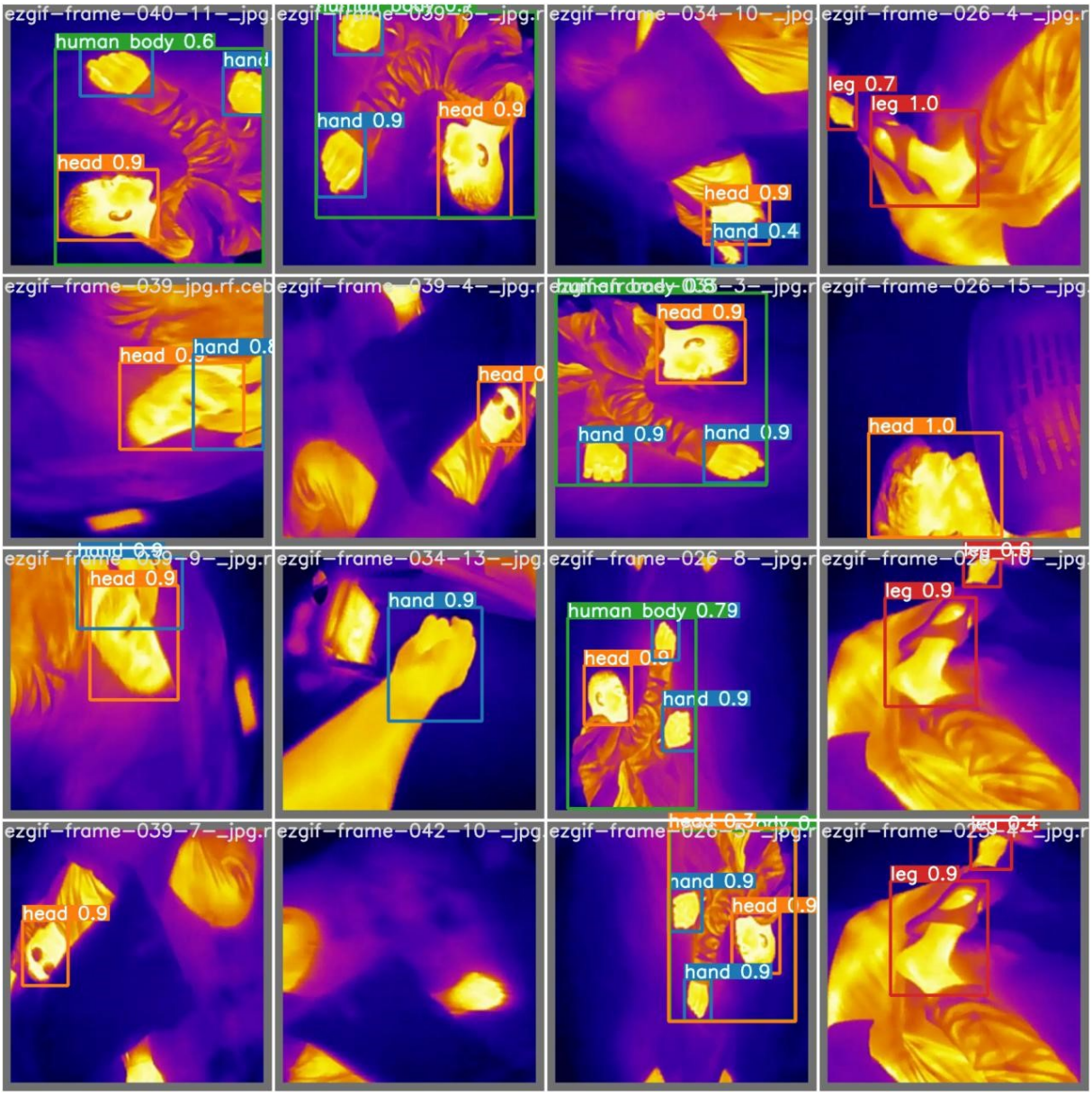


Figure 4.17: Test batch predictions of thermal images

So, we can see that we have got 0.924 Mean Average Precision (mAP) from training the dataset.

4.5 Speech Recognition

During a post disastrous situation, voice detection is an effective way for search and rescue operations. Using speech recognition, alive human beings can be identified. Moreover, analyzing a speech, the position, state of a human being can also be understood. For this reason, speech recognition is considered as an integral part of this research. In our research we have focused on Bangla Speech Recognition System.

There are various packages that are used for speech recognition. The fundamental work of speech recognition is to convert spoken languages in the form of texts[4]. Firstly, the speech recognition internally converts the physical sound into a signal of electricity and then the electric signal, with the help of analog to digital converter is converted into a digital signal. Then using a digitalized model this signal is transcribed into a text format.

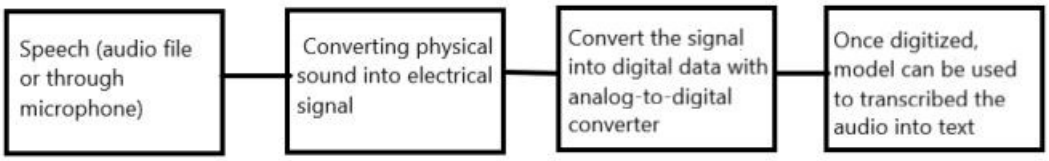


Figure 4.18: Speech Recognition Flow[9]

There are multiple audio tools and APIs for converting audio data to text format. Among such APIs, some are good at speech recognition but are slower in comparison of performance with other APIs. Among APIs named Google Speech Recognition API, IBM Watson API, AssemblyAI API, etc. comparative analysis has been made[5]. After reviewing the comparative analysis, Google Speech-to-Text API was chosen for Bangla Speech Recognition.

Audio	IBM(Waston)	Amazon	Google	Microsoft Azure	Speechmatics
1	14.72	8.97	10.34	8.21	9.38
2	32.57	24.3	35.29	31.47	24.25
3	14.83	12.34	14.76	19.48	8.39
4	9.37	8.19	10.92	9.55	6.79
5	31.75	19.34	23.31	20.93	17.48
6	10.22	7.36	7.01	7.49	6.13

Table 4.1: Comparative Analysis of different APIs

Google Speech-to-Text API is a cloud-based machine learning service that enables users to convert audio to text in over 120 languages and variants. The model behind Google Speech-to-Text API is based on recurrent neural networks (RNNs), specifically long short-term memory (LSTM) networks. These networks are well-suited for processing sequential data, such as speech, and have been shown to be effective in speech recognition tasks.

The API uses a combination of deep neural networks (DNNs) and traditional hidden Markov models (HMMs) to perform speech recognition[20]. The DNNs are trained on a large dataset of audio and corresponding transcriptions, which allows the model to learn the mapping between audio and text[11]. The HMMs are used to model the temporal dynamics of speech, such as the transitions between phonemes. Google Speech-to-Text API also uses a technique called transfer learning, which allows the model to take advantage of pre-trained models and fine-tune them for specific tasks. This allows the API to achieve high accuracy in speech recognition, even for rare or new languages and dialects.

Chapter 5

Result Analysis

In our thesis, we proposed using YOLOv7 for conventional, infrared, and thermal cameras as part of a multimodal strategy for people detection in unrestricted contexts. In order to find human remains in uncontrolled conditions, such as a spot where an earthquake has just occurred or where a building has burned down, we used day, night, and thermal cameras. We utilized Mask R-CNN initially, followed by Faster R-CNN, YOLOv5, and subsequently YOLOv7.

5.1 Comparative Analysis of CNN Models

As we depend on machine learning models to detect human bodies we had to make sure that the models know what to look for and what to avoid. So we needed data that served our purpose. Creating, curating, and maintaining our own dataset was not easy but it sure was fruitful and a great addition to our thesis and in this aspect of ML. Our dataset focuses on human body parts in diverse and difficult environments. Trying out our dataset and then fixing and recommending a model is a huge task and we wanted to make sure that we suggest the right model that not just fits our dataset but also serves our purpose. So we switched between models in order to compare their effectiveness on our dataset and determine which one performed the best. For our Dataset I (day and night images or conventional camera images) we got 80.4% accuracy using Faster R-CNN on dataset I and for dataset II we got 91.9% accuracy using Faster-R-CNN to detect all the classes of human bodies. For YOLOv5, it detected all the classes but not all images so the accuracy was not up to the mark. But when we used YOLOv7, it's the sophisticated architecture of object detection really worked on our dataset for our Dataset I we got 85.7% accuracy and for our Dataset II(Thermal images) YOLOv7 got us 92.4% accuracy which is not just greater than other models but also greater score than our conventional camera images score. The thermal image dataset was a tough job to arrange. We could not find any proper dataset that feeds our needs so we had to make our own to complete our research. We needed a rich dataset that checked all the difficult requirements. Because YOLOv7 had the highest accuracy of any of the models we tested—a sign that it did the best job on our dataset—we improved it. So when we analyzed our training results we found some insightful information regarding these models.

In our research, Faster R-CNN, Mask R-CNN, and YOLOv5 were less accurate in human body detection than YOLOv7 for a variety of reasons. To begin, for

Faster R-CNN, we found many papers that claimed that Faster R-CNN definitely cuts the most accurate spot when it comes to object detection or in our case human body detection. But in our research, we found something different than usual and we also figured out why it happened. As we mentioned earlier for both conventional and thermal images YOLOv7 outperformed Faster-R-CNN because Faster-R-CNN works well detecting generic objects or objects which are easily distinguishable from the surrounding objects but when it comes to a difficult dataset like ours where the environment is so unconstrained and the desired object is covered by something, in that case, YOLOv7 is already faster than Faster-R-CNN but in our case more accurate too[14]. Faster R-CNN and Mask R-CNN are two-stage object identification models with very sophisticated architecture and a greater number of parameters to be trained. This intricacy may have increased the risk of overfitting, which could have harmed the performance of these models on our dataset[3].

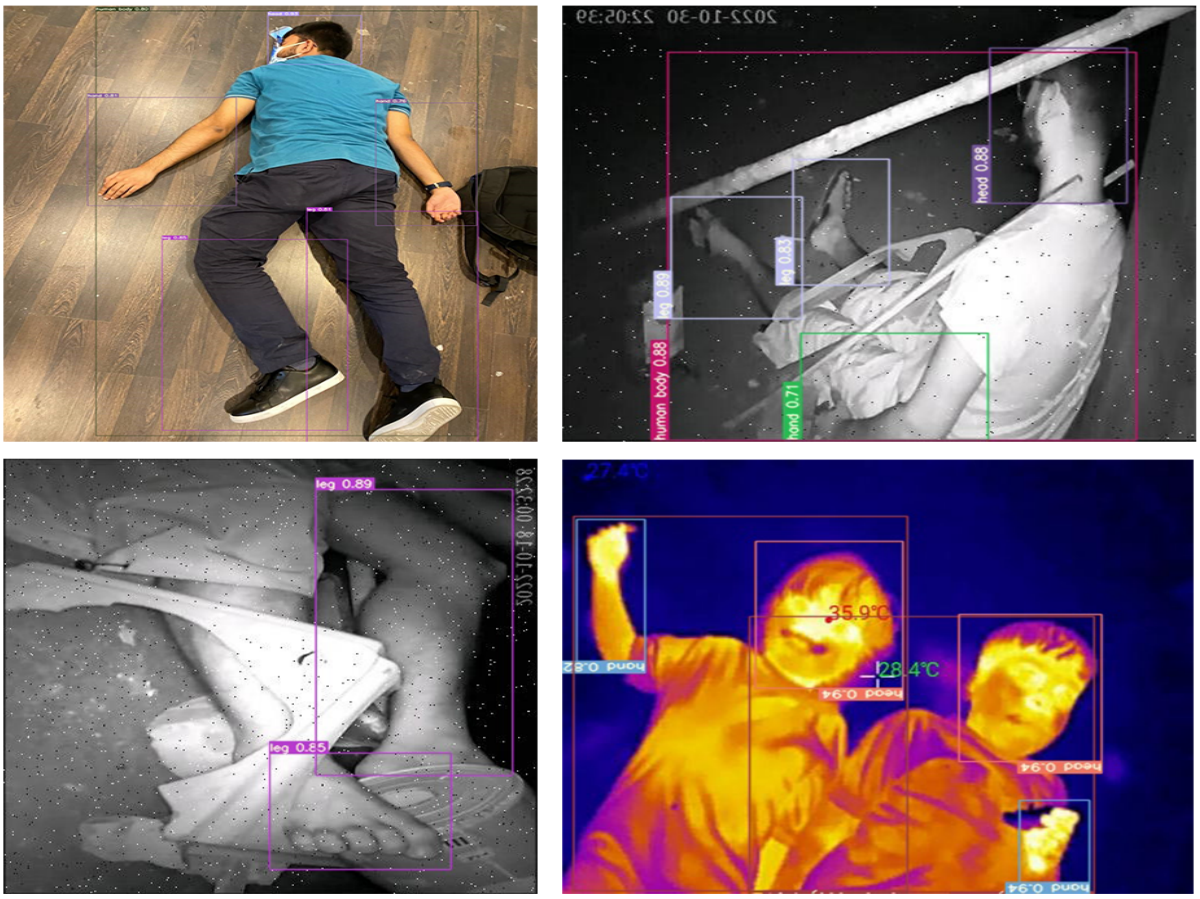


Figure 5.1: Human Detection results using YOLO v7

Second, YOLOv5 is a one-stage object detection model, which is less complicated and faster than two-stage models such as Faster R-CNN and Mask R-CNN. However, it may not be able to learn as many characteristics as the two-stage models, resulting in lesser accuracy on our dataset.

Finally, YOLOv7 features a more powerful and accurate architecture and a greater number of convolutional layers than YOLOv5. Furthermore, YOLOv7 includes an attention mechanism that can aid in the detection of items in a complicated and congested environment. YOLOv7 was already trained on a huge dataset of diverse photographs of individuals and other objects in various backdrops and environments, which helped the model learn better features and generalize well on our new images[29].

Therefore, the lower accuracy of Faster R-CNN, Mask R-CNN, and YOLOv5 in human body detection compared to YOLOv7 might be attributed to their comparatively complex design, lower ability to learn features, absence of attention mechanism, and lack of exposure to various images. On the other hand, YOLOv7's architecture, attention mechanism, and exposure to a wide range of images helped it perform better on our dataset.

5.2 Speech Recognition Analysis

Speech recognition was conducted in the Bangla Language over a particular set of phrases. These phrases are specially taken into consideration in disaster situations. These phrases are widely heard from the victims during search and rescue operations in post disastrous situations. Thus, by analyzing these phrases, we can get a better understanding of the proposed system during such unconstrained disastrous environments.

After completing the Speech Recognition of the Bangla Language, we obtain a spectrogram analysis and a transcription report of the audio inputs.

5.2.1 Spectrogram Analysis

- The spectrogram analysis shows different frequency levels based on the confidence of the audio. Comparing the spectrogram frequency level, a stress call or a life in rubble can be identified.

- The confidence parameters give floating values. Audio showing distress signs shows very less confidence value. Moreover, confidence values can also be used to determine the situation like position or stress of the audio.

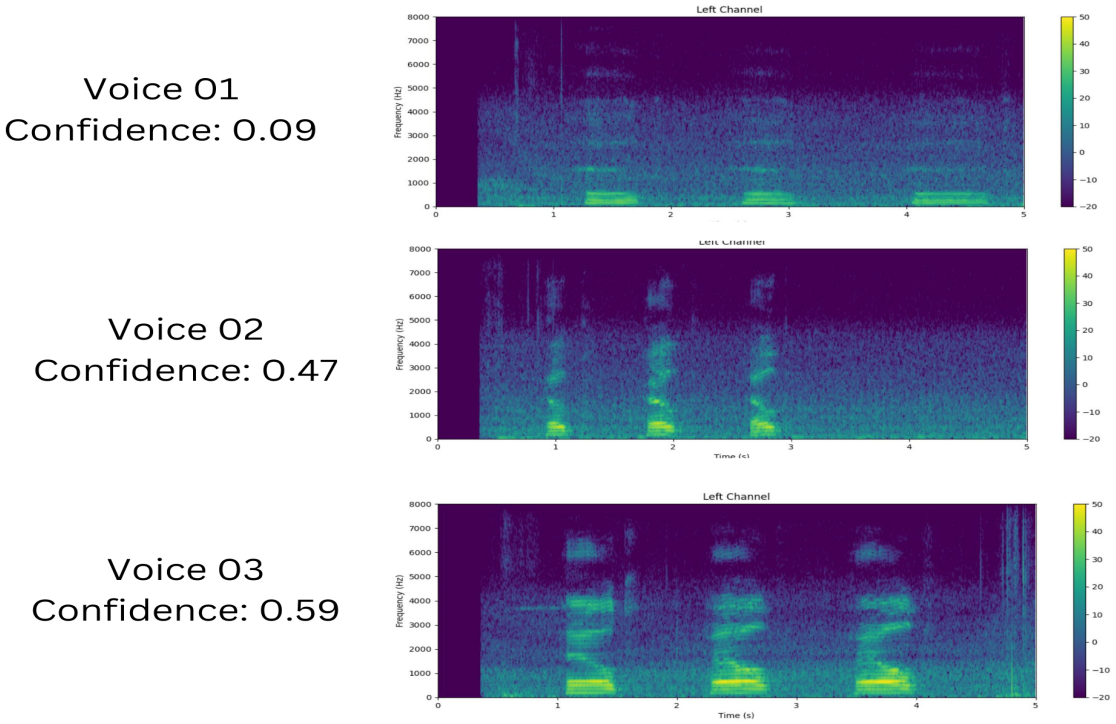


Figure 5.2: Spectrogram comparing Frequency levels with Confidence values of different voice records

5.2.2 Transcription Analysis

As per Figure 3.9, we have used speech recognition on 6 Bangla phrases spoken by different people in a real life commute environment. After analyzing the speech, it was found that the Transcription showed an average accuracy of 84.72%. The transcription showed disturbances on joint words that are often pronounced consequently. However, the transcription accuracy is 100% on phrases with words that can be pronounced independently.

Bangla Phrases	Accuracy Percentage(%)	Average Accuracy(%)
হ্যালো আমাকে কি শোনা যাচ্ছে	100	84.72
আমি বিল্ডিঙের নিচে আছি	75	
আমাকে সাহায্য করুন	66.67	
আমার এখানে খাবার ও পানি দরকার	66.67	
আমি এখানে আছি	100	
আমাদের সাহায্য দরকার	100	

Figure 5.3: Accuracy comparison among transcription of the audio speeches.

Chapter 6

Conclusion

In this research, we proposed a pioneering strategy for enhancing the effectiveness of rescue operations in unconstrained environments by utilizing cutting-edge machine learning techniques. Specifically, we employed the YOLOv7 algorithm for swift and accurate recognition of human bodies, while disregarding extraneous obstructions. Our system utilizes a combination of primary data of 7,087 images of various categories such as a day camera for high-quality images, an infrared camera for nighttime detection, and a thermal camera to improve search efficiency. Furthermore, we integrated a Bangla Speech Recognition System to detect distress calls from any type of environment. This information is then transmitted to a remote platform that can be accessed by the rescue team in real-time, allowing for quick and precise responses. Through our studies, we achieved an accuracy of Precision of 85.7% and 92.4% using conventional-infrared images and thermal images respectively, demonstrating the accuracy and reliability of our proposed model. Additionally, we achieved an average accuracy of 84.72% in detecting and transcribing Bangla Speech.

Our proposed system is a significant step forward in the field of rescue operations, as it allows for efficient and accurate detection of human bodies in unconstrained environments. The use of multiple cameras and speech recognition technology improves the robustness of the system, making it more reliable in various scenarios. Our results indicate that this approach has the potential to significantly improve the efficiency and effectiveness of rescue operations. We believe that this system can be integrated into existing rescue operations and aid in saving lives in emergency situations.

Bibliography

[1] E. L. Quarantelli and R. R. Dynes, "Response to social crisis and disaster," *Annual review of sociology*, pp. 23–49, 1977.

[2] I. Burton, R. Kates, and G. White, "The environment as hazard.,(oxford university press: New york.)," 1978.

[3] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sensing of Environment*, vol. 62, no. 1, pp. 77–89, 1997, ISSN: 0034-4257. DOI: https://doi.org/10.1016/S0034-4257(97)00083-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425797000837.

[4] H. Erdogan, R. Sarikaya, S. F. Chen, Y. Gao, and M. Picheny, "Using semantic analysis to improve speech recognition performance," *Computer Speech & Language*, vol. 19, no. 3, pp. 321–343, 2005.

[5] S. Furui, "50 years of progress in speech and speaker recognition research," *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 1, no. 2, pp. 64–74, 2005.

[6] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

[7] W. Wang, J. Zhang, and C. Shen, "Improved human detection and classification in thermal images," in *2010 IEEE International Conference on Image Processing*, IEEE, 2010, pp. 2313–2316.

[8] M. Donelli, "A rescue radar system for the detection of victims trapped under rubble based on the independent component analysis algorithm," *Progress In Electromagnetics Research M*, vol. 19, pp. 173–181, 2011.

[9] B. Singh, N. Kapur, and P. Kaur, "Speech recognition with hidden markov model: A review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 3, pp. 400–403, 2012.

[10] Z. Zhongming, L. Linong, Y. Xiaona, Z. Wangqiang, L. Wei, *et al.*, "Asia-pacific disaster report 2015: Disasters without borders-regional resilience for sustainable development," 2015.

[11] V. Beat and J. Novet, "Google says its speech recognition technology now has only an 8% word error rate," *Venture beat*, 2016.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[13] I. Siegert, A. F. Lotz, L. L. Duong, and A. Wendemuth, "Measuring the impact of audio compression on the spectral quality of speech data," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2016*, pp. 229–236, 2016.

[14] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2110–2118.

[15] M. Al Zaman and N. Monira, "A study of earthquakes in bangladesh and the data analysis of the earthquakes that were generated in bangladesh and its' very close regions for the last forty years (1976-2016)," *J Geol Geophys*, vol. 6, no. 300, p. 2, 2017.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[17] M. H. Rahman, M. S. Rahman, and M. M. Rahman, "Disasters in bangladesh: Mitigation and management," *Barisal University Journal Part*, vol. 1, no. 4, p. 1, 2017.

[18] S. K. Sharma, R. Agrawal, S. Srivastava, and D. K. Singh, "Review of human detection techniques in night vision," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, IEEE, 2017, pp. 2216–2220.

[19] *The rana plaza accident and its aftermath*, Dec. 2017. [Online]. Available: https://www.ilo.org/global/topics/geip/WCMS_614394/lang--en/index.htm.

[20] N. Boyko, O. Basystiuk, and N. Shakhovska, "Performance evaluation and comparison of software for face recognition, based on dlib and opencv library," in *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, IEEE, 2018, pp. 478–482.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, *Mask r-cnn*, Jan. 2018. [Online]. Available: https://arxiv.org/abs/1703.06870.

[22] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[23] D. Zhang, S. Sessa, R. Kasai, *et al.*, "Evaluation of a sensor system for detecting humans trapped under rubble: A pilot study," *Sensors*, vol. 18, no. 3, p. 852, 2018.

[24] A. Bochkovskiy, H. Wang, A. Shvets, and V. Lempitsky, *YOLOv7: Optimal Speed and Accuracy of Object Detection*. 2020.

[25] D. M. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[26] N. C, *Friday night lights: Night vision oob (out of band) - fact or fiction?* -, Jun. 2021. [Online]. Available: https://www.thefirearmblog.com/blog/2021/06/11/oob-out-of-band/.

[27] C. François, *Deep learning with python*. Manning, 2021.

[28] P. B. .-. K. Taylor, *Ai in natural calamities for disaster response and recovery*, Jun. 2022. [Online]. Available: https://www.hitechnectar.com/blogs/ai-natural-calamities-disaster-response-recovery/.

[29] J. Solawetz, *Yolov7 - a breakdown of how it works*, Jan. 2023. [Online]. Available: https://blog.roboflow.com/yolov7-breakdown/.

[30] U. Aparna, T. Kodakara, B. Athira, A. MV, A. Ramakrishnan, and R. Divya, "Spotter: Detection of human beings under collapsed environment,"

[31] *Information on disaster risk reduction of the member countries*. [Online]. Available: https://www.adrc.asia/nationinformation.php?NationCode=50Lan+g.

[32] Matterport, *Matterport/mask_rcnn: Mask r-cnn for object detection and instance segmentation on keras and tensorflow*. [Online]. Available: https://github.com/matterport/Mask_RCNN.