AUTOMATION OF DRONE SURVEILLANCE FOR SEARCH AND RESCUE

Thesis submitted in fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

By

BALMUKUND MISHRA

Enrollment No.: E17SOE801

Department of Computer Science Engineering

BENNETT UNIVERSITY

(Established under UP Act No 24, 2016)

Plot Nos 8-11, Tech Zone II,

Greater Noida-201310, Uttar Pradesh, India

Month: September, Year: 2021

AUTOMATION OF DRONE SURVEILLANCE FOR SEARCH AND RESCUE

Thesis submitted in fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

By

BALMUKUND MISHRA

Enrollment No.: E17SOE801

Supervisors:

Prof. Deepak Garg

Dept. of CSE, Bennett University, Greater Noida, India

Dr. Pratik Narang

Dept. of CSIS, BITS Pilani, Pilani, India

Department of Computer Science Engineering

BENNETT UNIVERSITY

(Established under UP Act No 24, 2016)

Plot Nos 8-11, Tech Zone II,

Greater Noida-201310, Uttar Pradesh, India

Month: September, Year: 2021

Dedicated to my beloved parents.

Declaration by the Scholar

I certify that the research work reported in the Ph.D. thesis entitled **"Automation of Drone Surveillance for Search and Rescue"** submitted at **Bennett University, Greater Noida, India** is a authentic record of my work carried out under the supervision of **Prof. Deepak Garg** and **Dr. Pratik Narang**. I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my Ph.D. Thesis.

Date: 20/06/2021

(Signature of the Scholar)

Balmukud Mishra

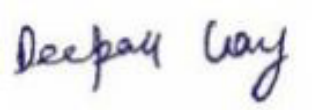
Department of CSE

Bennett University, Greater Noida, India

Roll No. E17SOE801

Supervisor's Certificate

It is recommended that the thesis title **Automation of Drone Surveillance for Search and Rescue** submitted by **Balmukund Mishra** at **Bennett University, Greater Noida, India** under our supervision is a bonafide record of his original work. This work has not been submitted for any previous degree to anybody.



Date: 20/06/2021

(Signature of Supervisor)

Dr. Deepak Garg

Professor and Head CSE,

Bennett University, Greater Noida, India

(Signature of Co-Supervisor)

Dr. Pratik Narang

Assistant Professor, Dept. of CSIS

BITs Pilani, Pilani, India

Abstract

Search and rescue is a critical application of disaster management and executed after the disaster. The primary objective of this operation is to save human lives by shifting them to safer places. Disaster response is always a race against a moment to identify the catastrophe victims as quickly as possible. Recent natural calamities such as the earthquake in Nepal's, Tohoku, and Haiyan, or floods in Europe and India have shown that local civil officials and emergency services are having difficulty managing the crisis properly. To handle these situations and find the stuck humans quickly, it is required to search the wide affected area quickly and respond to the needy people in terms of medicine, food packets, or sending their location to the rescue team to shift them to a safer place.

Drone capability is improving day by day, and they can scan the wide affected remote area quickly. Using the existing computer vision algorithms with the drone to help people is a promising use case of technology, and it can save millions of lives in the disaster. Using a high-quality drone for disaster response is the best possible solution for disaster management. Manual drone based monitoring of disaster area necessitates a lot of human effort, so, the automation is ideal way to scan large affected area for saving more lives. The recent development of AI and ML algorithms has opened the door of automation. However, drone surveillance has its own challenge and need to focus on various level such as increasing the capability of drone, development of accurate dataset, and efficient aerial surveillance algorithms.

This thesis investigated the current methodology of disaster management techniques and suggested various realistic solutions using computer vision and artificial intelligence techniques for automating drone surveillance. However, a primary prerequisite for such automation is the availability of suitable datasets, which is lacking in the literature for drone surveillance. therefore, we developed a few datasets for drone surveillance automation in this study, and suggested different strategies for locating disaster victims, including identifying humans and their immediate actions, such as waving hands. We've also developed a method for determining

whether a given scenario is a help or a non-help scenario. We tested various action recognition models in this study and discovered that models with spatio-temporal features perform better with the type of action needed to recognise help situations in disasters. Disaster relief in the form of food and medication is critical for survival in disaster-affected areas. Based on current capabilities, it is possible to provide such immediate assistance via drone. However, sometimes disaster victims cannot appear physically in open spaces due to weather conditions and disaster types. Here in this thesis, we have developed a text-classification-based approach to recognize the emergency. The proposed method requires a dataset, and we have created an emergency text classification dataset in which images have text such as "Help", "SOS", and "Emergency" written on the ground, wall, roof, and sand.

An efficient action recognition module in drone surveillance may aid disaster relief by identifying people in disaster areas and determining whether they need assistance. For this, we have investigated the prior object detection and action recognition model on our developed dataset and some existing datasets and found that the current approach of directly applying these modules in drone surveillance is ineffective. To handle such challenges in drone surveillance, we have developed a novel framework for using the action recognition models in drone surveillance. Individual human features in crowd surveillance videos are modelled using the proposed system. Hence, the human features are explored enough to classify human action at the time of surveillance. Since actual disaster data is not available for model training, the proposed model introduces another feature by reducing the impact of background, and making it more practical.

Contents

List of Tables

List of Figures

List of Abbreviations

AAR	Accurate action recognition
AI	Artificial intelligence
AP	Average precision
ANN	Artificial neural network
CNN	Convolution neural network
DL	Deep learning
FMFM	Faster motion feature modeling
FN	False negatives
FP	False positives
GPU	Graphical processing unit
INSARAG	International search and rescue advisory group
IOU	Intersection over union
JPG	Joint Photographic Group, a form of image
LSTM	Long short time memory
mAP	Mean average precision
ML	Machine learning
MP4	A digital multimedia container former, used to store video
MSE	Mean square error
OD	Object detection
PSO	Particle swarm optimization
R-CNN	Region based convolution neural networks

RMSprop	Root mean square propagation, an activation function in DL
RNN	Recurrent neural network
RPN	Region proposal network
SAR	Search and rescue
SGD	Stochastic gradient descent, an activation function in DL
SSD	Single shot detection
SVM	Support vector machine
TN	True negatives
TP	True positives
UBSAR	Urban search and rescue
UNDAC	United nations disaster assistance and coordination
UNHCR	United nations high commissioner for refugees
YOLO	You look only once

List of Symbols

$AP(i)$	Average precision for class i
Pr	Area under precision recall curve for evenly distributed between 11 points
TP	True positive cases
TN	True negative cases
FP	False positive cases
FN	False negative cases
L	Log loss for the classification problem
z_i	Outcome of text-classification classifier in terms of image labels
$f(x,y)$	Representation of image for the input to the classifier
I	Output image with bounding box of object detection module
$f(h,w)$	Input to the object detection module
X_i	Represents the output to the OD and extraction of bounding boxes
$I^{c*l*h*w}$	Video with width l, h , and w
Y_i	Output label from the object detection model such as car, bus, human
$O(I)$	Applying object detection model O in Image I
$L_{loc}(y,y')$	Localization loss
$L_{class}(y,y')$	Classification loss
$L(y,y')$	Object detection loss
S	Action classification with spatial features
T	Using temporal features for action classification
$\sum_{r=1}^{w} A\ (\ I_r\ )$	Concatenating the output of classification module for multiple image

Chapter 1

Introduction

1.1 The perspective

Search and rescue is an essential operation that is generally executed after the disaster occurs in the perspective of natural calamities. It is a combination of two different words, "Search" and "Rescue." Each one is individually performed by the local volunteers or with the help of army teams traditionally. SAR operations aim to rescue the utmost number of people in the shortest time while minimizing risk to rescuers. A natural disaster such as flood, earthquake, heavy snowfall, and landslide typically damage infrastructure, causes injury, and sometimes massive loss of human life. An immediate life-saving response is needed in these cases to rescue those who are trapped and stabilize them by shifting to a safer place. In the context of disaster and its management, Dr. Robin Murphy once said, "If you can reduce the initial response by one day, you can reduce the overall recovery by 1000 days"[1]. Reducing response time of the disaster relief and recovery is a critical phase through which millions of lives could be saved. Figure 1.1 shows a list of potential solutions for the disaster response techniques. Out of all the techniques shown here, the drone-based strategy for disaster relief and recovery has the

advantage of covering a wide range in a short period and hence most suitable for an outdoor disaster such as flood and earthquake. It is possible to deliver the item and perform civilian surveillance due to the availability of commercial drone with a large payload. As a result, it encouraged us to use it for disaster relief and search and rescue. The automated drone can save human lives, particularly in disaster, wild-life, and border-area search operations. However, to apply it on a bigger scale, this operation should be fully automatic and work without human intervention. Advances automation technology, such as deep-learning-based image and video processing, may be able to solve this issue. In order to provide better search results in disaster, a specific image and video processing component is required to embedded in drone. Developing drone based solution for surveillance leads to various potential attacks also, as the nature of communication in drone is wireless and specially in remote areas makes it easier to attack for attackers. To protect the drone from such attacks, [181] proposed a private block-chain based mechanism for secure communication in IoT based UAV devices.

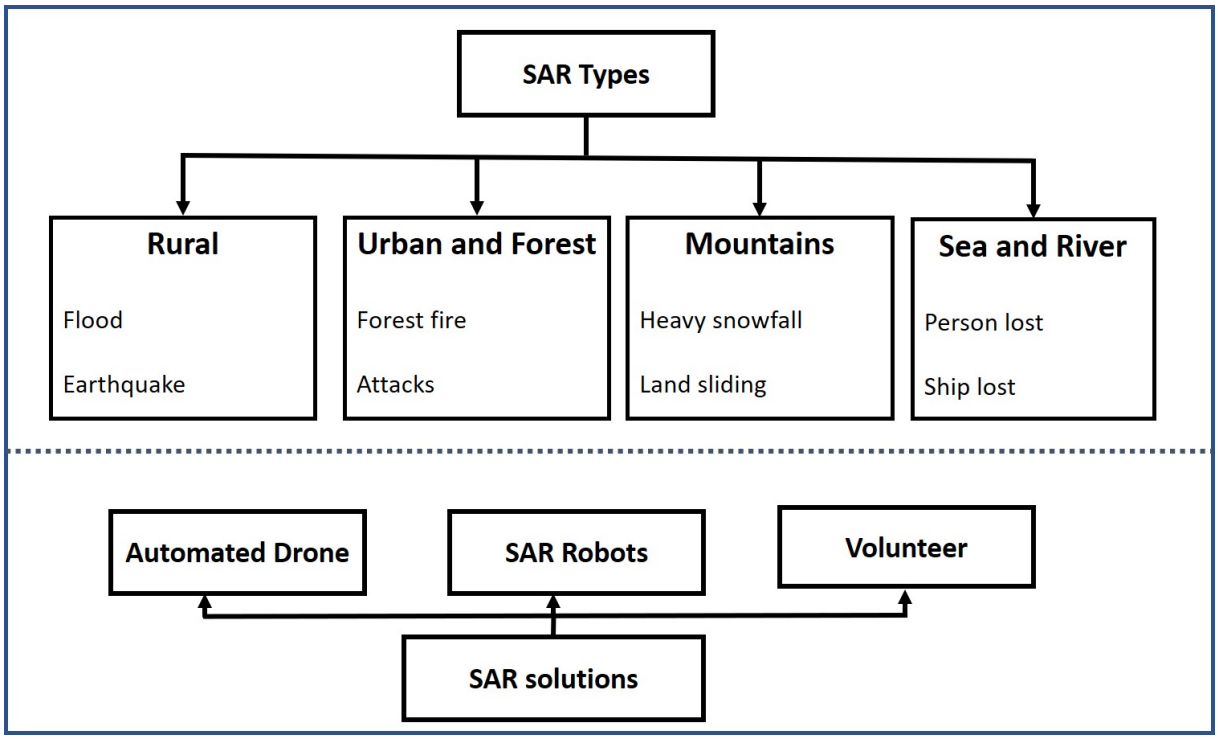


Figure 1.1: Types of SAR methods and its probable solutions

Autonomous drone surveillance of the affected area is a potential solution for providing an effective search for humans or places where rescue is needed and image and video processing of various types may support in the automation of this automation. There are different aspect of

image and video analysis for the automation of search operations in disaster are as follows:

1.1.1 Classification based approach for help situation identification

Classification based methods are now-a-days very useful to solve various real-life problems. For the disaster management applications such as search and rescue, it could be used to classify the humans and estimate the emergency situation.

1.1.2 Object detection based approach for help situation identification

Classification-based methods, on the other hand, are not always the best option for drone surveillance, where images are typically taken from a large distances and several objects appear in a single frame. Another deep learning technique called object detection is ideal for recognising humans in disaster relief and rescue when multiple objects appear in a single frame.

1.1.3 Action recognition based approach for help situation identification

The field of computer vision and deep learning is vast, and there are several techniques available to manage such automation. However, action detection, which is focused on the natural behaviour of humans in disasters, is another area that needs to be discovered for decision making at the time of drone surveillance.

1.1.4 Scene classification based approach for help situation identification

Action recognition is a difficult task in computer vision that has been attempted in various ways in the past. Particularly for aerial and drone images where multiple people are performing the action and sometimes these actions are disjoint, making it difficult to detect a specific action accurately. In this case, action recognition modules typically do not perform well; thus, a

scene classification technique that can accurately identify a specific scene as a help situation is required.

1.1.5 Text classification based approach for help situation identification

Another aspect of search and rescue is determining the need for assistance when humans are not completely visible to the drone. In this case, we can consider identifying those situations by using emergency text such as "Help," "Emergency," and "SOS." These texts could assist the surveillance vehicle in recognizing the scene as an emergency and a person who is not visible due to weather conditions.

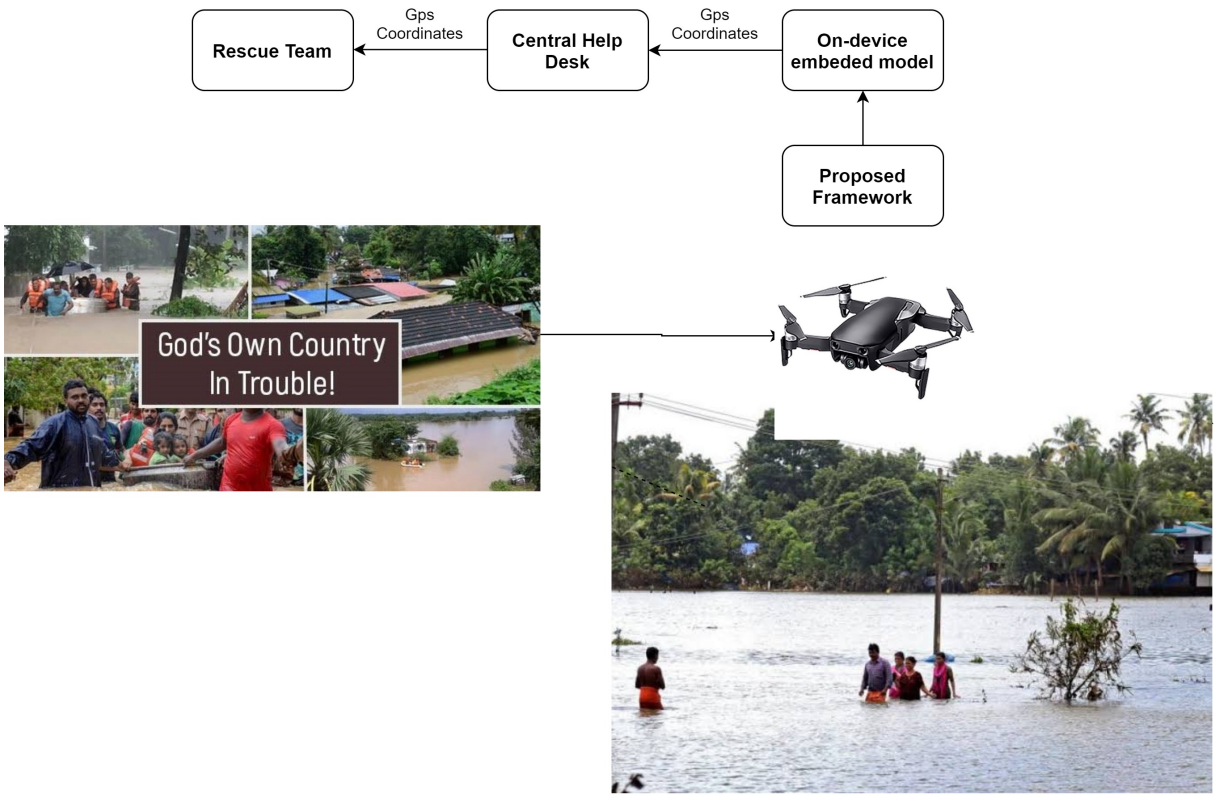


Figure 1.2: Architecture of automated SAR

1.2 Overview of natural disaster and disaster response

Recent natural calamities such as the earthquake in Nepal's, Tohoku, and Haiyan, or floods in Europe and India have shown that local civil officials and emergency services are having difficulty managing the crisis properly [2]. The search for human survivors on the site is paramount for rescue services. Previous search and rescue attempts are mainly volunteer-based and and often time-consuming. Also, SAR through the human-operated mission is a complicated and harmful job which may even cause the human crisis executives themselves to lose their life. Disaster response is always a race against a moment to identify as quickly as possible the victims of the catastrophe. Rapid scanning of impacted areas through drones equipped with high definition camera and navigation system is capable to save human lives in disaster without taking the risk of volunteer's life. Several fundamental factors must be taken into consideration to provide an automation of drone-based search operations, and a couple of them is the quality of data and environmental hazards. Different methods of SAR are represented in Figure1.1, starting from volunteer-based human-intensive SAR to fully automated drone-based SAR. SAR robots are a solution for indoor disasters that is currently gaining popularity because they are intelligent, never get tired, and can penetrate enclosed areas. The type of SAR technique required is determined by the geographical location of the disaster; for example, SAR robots are less effective in hills. In the category of surveillance and rescue through aerial vehicle, the first rescue operation was performed on 29 November 1945 by Sikorsky's chief pilot Dmitry "Jimmy" Viner, in the cockpit [3]. In this operation, before the ship sank, all 5 crew members of the oil boat were rescued through helicopter. The issue with this approach is that it is extremely expensive as compared to the drone or robot based SAR. Since its capabilities are improving all the time, the drone could be a very useful tool for SAR.

The drones are now commercially available and can be used by civilians. More investments are coming in the drone technology these days, and research is going on for the development of drone-based application, which can help humans to ease their life [4,5,6]. Looking at the demand and capability of drones to be used for different real-time applications including SAR, the automation is required to apply it at a bigger scale. For the automation of these surveillance application, deep-learning is the key. However, basic requirement of deep learning algorithms is the availability of appropriate data in ample amount for training the network.

This thesis investigates the existing literature in the field of search and rescue and the automation of drone surveillance. With the help of various computer vision techniques, this thesis provides different solution of this in terms of human detection in disaster affected area, their action recognition to estimate the actual needy person. This thesis also investigates different scene classification approaches and provide an approach to identify the actual needy person in disaster based on video-level scene classification and text based emergency situation identification.

1.3 Theoretical background of automation techniques

Deep learning is a popular area of machine learning, where hierarchical architectures learns the high-level abstraction of data. It learns the insights of data in both ways (Supervised and unsupervised). Recent success of deep learning algorithms is due to the availability of data and the processing power. Deep learning models are now a days very popular in several areas such as handwritten digital classification, audible or visual signal analysis, facial recognition, recognition of disasters, voice recognition, computer vision and automatic language processing. Some researchers divided the deep learning algorithms into four categories: convolutional neural network, restricted boltzmann machines, auto encoders and sparse coding [7]. CNN based solutions are best suited for image based problem such as object classification, object detection, and action recognition.

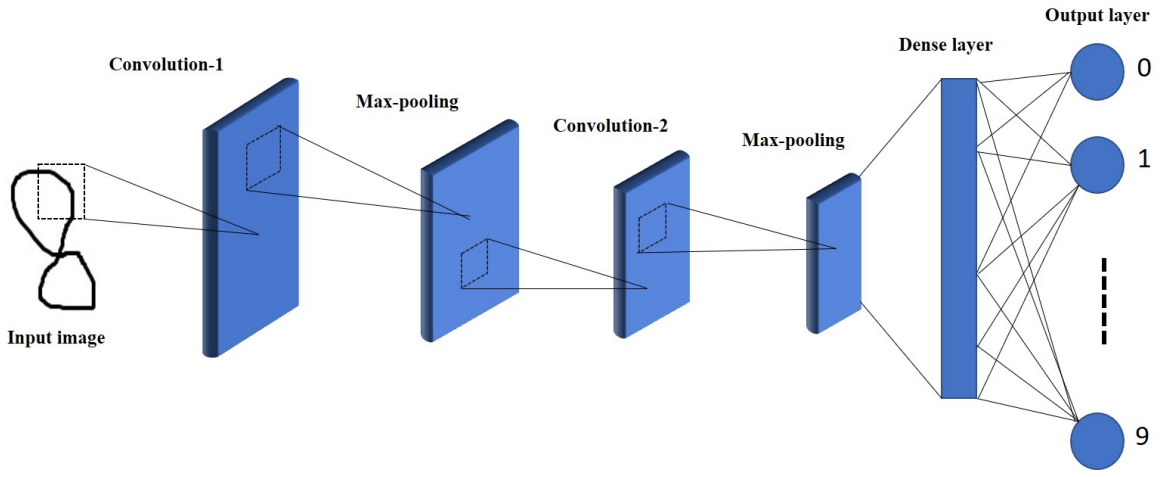


Figure 1.3: An example of digit recognition using CNN

Convolution neural network is a combination of mainly three different operation such as convolution, pooling, and padding at different layer of network. Using these operations at different layers, network is trained with the two different strategies ie: forward and backward stage. Out of all the convolution layer in the network, initial layers extract the low-level feature of the image such as edges, line, and corners. Other layers contribute in extracting mid-level and high-level features such as structure, objects, and shape. An example of digit classification using convolution neural network is represented through Figure1.3. In this, input image of digit is passed through the convolution layers and the last dense layer is recognizing the digit based the extracted feature throughout the convolution layers.

1.4 Problem specification

Drone surveillance for SAR could be an ideal solution for saving human lives in natural disaster. However, existing approaches of SAR are either human intensive or manual screen monitoring. Automation of drone surveillance can improve the performance of SAR missions and able to save more lives. The challenge in automation of drone surveillance for SAR is the unavailability of dataset which plays a key role. In addition, existing algorithm of automation are mostly developed for the bigger object and for the ground level images where substantial features are available for the recognition. Human covers smallest ground level area compared to the other objects such as car, aeroplane, animal etc. and hence hard to get the necessary feature in drone surveillance for object classification and detection.

1.4.1 The statement of the problem

The present research work can be described by the following statement of the problem:

To investigate the current scenario of search operation in natural disaster and make it advance using computer vision and deep-learning algorithms for the automation of drone surveillance.

1.4.2 Research objectives

The research objectives are divided into following sub steps to achieve the automation of search operation in natural disaster using drone surveillance:

- As dataset is an essential part of the automation using deep learning, our first objective is to develop the dataset required for the automated drone surveillance in SAR applications.
- Our second objective is to develop the novel technique of automation for search operation using text extraction, human detection, action recognition, and scene classification.
- To device the technique useful for identifying the help situation in different terrain of disaster using action recognition.

1.5 Assumptions and dataset

Following assumption and dataset has been made to successfully achieve our research objectives:

- It has been assumed that if human is not visible or didn't want to explore his body in case of disaster due to the real time circumstance such as weather condition, they just write emergency text such as Emergency, SOS, or Help on roof, wall, or ground.
- For on one of the proposed technique in this thesis, presence of human in disaster is assumed that they are stuck and required help in terms of medicine, food, and rescue.
- Human action such as waving hand in the direction of moving drone is considered as the person asking for help in disaster situation.
- A dataset is developed for the automation of emergency text recognition such as SOS, HELP, EMERGENCY for the automation of search operation.
- Images of human present individual and in group are collected and annotated for the development of human detection dataset for drone surveillance.

- Dataset of human action is developed for the identification of help situation. In this, two different datasets is developed and annotated as two action dataset and six action dataset.
- A video dataset is also developed for the identification of situation as help-situation. Developed dataset contains video clip in help class with the frames containing single or multiple human waving hand in the direction of drone.

1.6 Thesis outline

The subject matter of the thesis is presented in the following seven chapters,

✓ Chapter-1 gives an overview of the advantage of SAR system for disaster situations. It also describes the motivation of this research. Further, the existing problem in the field of search and rescue system is outlined and specifies the research objective for this thesis.

✓ Chapter-2 elucidates the principle and important parameters of a SAR system and its characteristics. This chapter thoroughly investigates the existing approaches for SAR and useful techniques in the field of AI, ML, DL for the automation of drone surveillance.

✓ Chapter 3 describes the development of dataset for human detection and object detection based action classification for search and rescue in natural disaster. In this chapter, a enhanced object detection model is used for recognizing action in drone surveillance videos.

The algorithms and result of this chapter is published in an international journal "Computer communication" of elsevier Volume 156, p,1-10 with title "Drone-surveillance for search and rescue in natural disaster", April 2020, SCIE, IF-2.6.

✓ Chapter-4 describes the algorithms and dataset developed for video classification based action recognition in drone surveillance. In this, a video classification model using enhanced 3DCNN is used, and hyper-parameters were optimized using PSO.

The result and algorithms of this chapter is published in international journal "Neural computing and application" of springer, S. I : Hybridization of Neural Computing with Nature Inspired Algorithms, June, 2020, SCIE, IF-4.7.

✓ Chapter-5 highlights the methodologies developed for finding the places where instant help is required in case of natural disaster using aerial emergency text classification. In this, a dataset of emergency text classification is developed, where each images has text written such as "Help", "Emergency", "SOS". In addition, a convolution neural network is proposed for classification of such text in drone images.

The dataset, algorithm, and result of this chapter is submitted in 6th International conference on computer vision and image processing, CVIP, IIT Ropar.

✓ Chapter-6 describes the algorithm developed for background invariant fast motion modeling for finding the places where rescue is required. In this chapter, a unified end-to-end trainable approach is developed for fast motion modeling of temporal action recognition required for identifying action in drone surveillance.

The proposed algorithm and result of this chapter is submitted in a Elsevier Journal, Computer Vision and Image Understanding.

✓ Chapter-7 gives a summary of overall research done in this thesis. This chapter also summarizes all the techniques and datasets developed for search and rescue using drone surveillance. In addition, this chapter point out the future scope in this field for the development of an accurate drone surveillance system.

Chapter 2

Review of Literature

This chapter discusses the life cycle of disaster management and gives a list of some recent natural calamities which has caused the severe damage to the loss of human lives and capital. Search and rescue is one of the important phase of disaster management life cycle, and this chapter gives a brief description about the current scenario of search and rescue, some important search and rescue event in the past, and tools and technology helpful to develop a powerful search and rescue system.

2.1 Disaster management

Disasters are becoming more common on the Indian subcontinent, resulting in severe casualties. The following is the response of the Indian Air Forces to a major disaster that occurred recently:

- **2010 Ladakh floods:** A flash flood occurred in the Leh Ladakh area of North India on August 6, 2010, at 12 a.m., due to a cloud burst. The water level rose 14 inches in two hours, causing massive damage to human lives and capital [53]. That was one of the toughest situations of the era. In addition, the hospital in Leh was in poor shape and

unable to assist the victims. The Indian Air Force, on the other hand, launched a rescue operation almost immediately after the tragedy, saving many lives.

- **2013 Uttrakhand floods:** Uttarakhand is a state of India and as a part of the Himalayan region, is extremely vulnerable to geological disasters. On June 16, 2013, the state of Uttarakhand was struck by one of the worst natural disasters in Indian history, resulting in major casualties [31]. The tragedy occurred between the 14th and the 18th of June as a result of heavy rain. A huge landslide and cloud burst were caused by heavy rain. The Indian Armed Forces have assisted in the crisis and with the help of army, the were able to save some lives

- **2015 Nepal earthquake:** A powerful earthquake hits Nepal on April 25, 2015. It caused extensive damage, killed approximately 9000 people and injured 22,000 others, according to [184]. The Indian Air Force has participated during rescue and mobilized 1xIL 76, 2xC-130J Hercules, 4xC-17 Globemaster transporters, 2 x Advanced Light Helicopters 8xMi-17 helicopters starting 25 April.

- **2018 Kerala flash flood:** In August 2018, Kerala, an Indian state, experienced heavy rains, resulting in a flood [61]. This flood killed approximately 400 people and displaced millions of people. SAR Operations by IAF have helped about 1,000 ladies, girls, older people. Also, the IAF helicopters have delivered food and water packages to stranded residents.

- **2018 USA hurricane:** In 2018, Central America and Florida state have seen a wild face of Hurricane and hundreds of people died with approximately damage of 25.5 billion dollor [37]. Earlier, in 2005, These states have suffered the biggest damage by a huricane when thousands of people have lost their lives with maximum capital damage of 125 billion dollor.

- **2011 Tohoku earthquake:** In 2011, Japan has seen the biggest earthquake of this era in which approximately 15,881 people have lost their lives[185].

All of these accidents, along with their destruction, indicate that a strong rescue system is needed that can reduce the loss of human life. Major organizations came into focus for global disaster management providing disaster relief are as follows:

- International Search and Rescue Advisory Group (INSARAG): This is a United Nations Organization for the Administration of Humanitarian Affairs (OCHA). It is a network of disaster-prone and disaster-responding countries and organizations dedicated to urban SAR and operational field coordination. Its goal is to provide standards and classification for international USAR teams as well as international response planning methods in the aftermath of earthquakes and structural disasters.

- United Nations High Commissioner for Refugees (UNHCR): The United Nations High Commissioner for Refugees (UNHCR) was established in 1951 to assist Europeans who had fled or lost their homes. They later assisted millions of refugee families in Asia, Africa, and the Middle East in the twenty-first century.

- United Nations Disaster Assistance and Coordination (UNDAC): It was established in 1993 to assist the United Nations and governments of disaster-affected countries. At the national level, UNDAC also assists in the management of incoming foreign relief.

Natural disaster is uncontrollable, however, an effective disaster management technique may reduce the losses. The methods for dealing with a disaster are divided into three parts [24]: before the disaster, impact analysis, and post-disaster operation. The following are the main components of this life-cycle:

- Mitigation: It involves reducing or eliminating the likelihood or consequences of hazards.

- Preparedness: It involves equipping the people who may be impacted by a disaster or who may help the impacted people.

- Response: Disaster response involves the quick action is to be taken to control and reduce the loss after the disaster happens.

- Recovery: It involves returning normal lives to the disaster victims by following the impact of disastrous consequences.

In past, different methods have been applied to gather the information about the disaster. As it is an important aspect of disaster management to collect the information about when and in which area disaster has happened, for this recently, [68] proposes a technique which uses the social media analysis such as tweets for quickly information gathering. Likewise, there are multiple

aspects which need to be strengthen for a quick and effective disaster management strategy. Here, in this study, We focused on emergency management strategies, with search and rescue being one of the most important components.

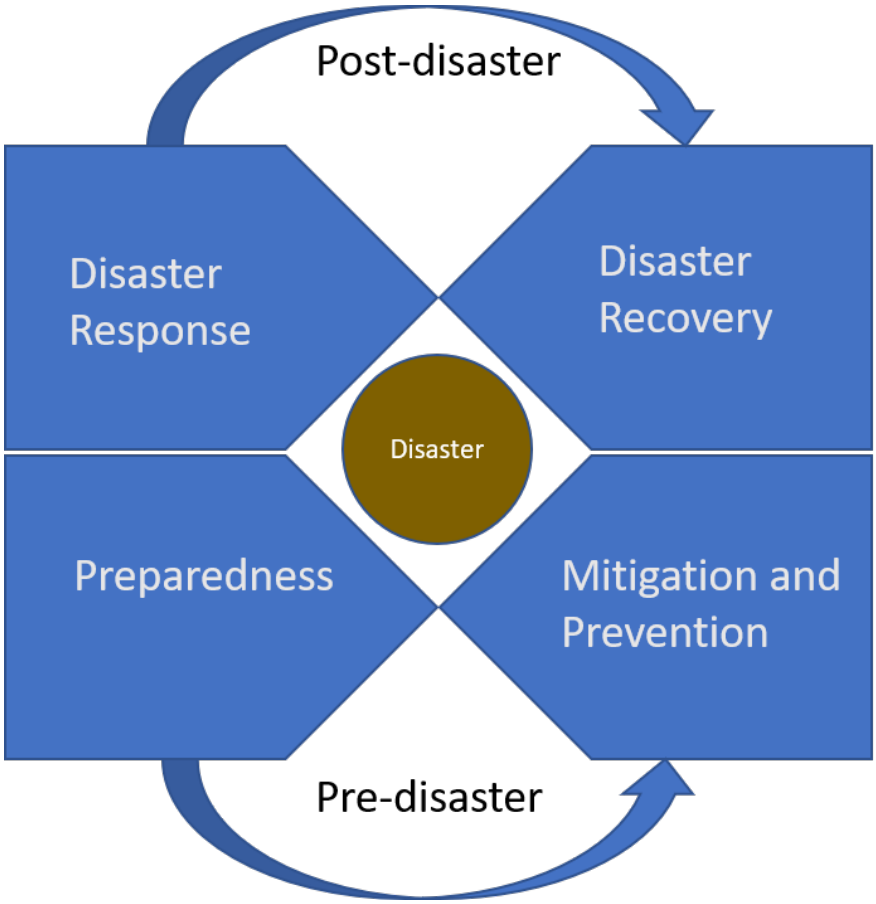


Figure 2.1: Life cycle of disaster management methods

2.1.1 Search and rescue

The fundamental goal of Search and rescue operations in disaster is searching for people in trouble or imminent danger and assisting. Canadian force defined the search and rescue as "Search and Rescue comprises the search for, and provision of aid to, persons, ships or other craft which are or are feared to be, in distress or imminent danger "[38]. Following the 1656 wreck of the Dutch merchant ship Vergulde Draeck off the west coast of Australia, one of the earliest well-documented SAR attempts in the world ensued [102]. Survivors requested assistance, and, without success, three separate SAR missions were performed in response.

However, based on the geographical location and environmental conditions, the search and rescue process can be subdivided into five categories:

- **Ground(Lowland) search and rescue:** People may go missing for a variety of reasons, some, due to problems such as domestic violence, while some die as a result of involuntary causes such as mental illness, infection, an accident, or suicide. In these situations also, some search and rescue events are used to be initialized and falls into the lowland search and rescue category.

- **Mountain rescue:** SAR operations in the rough and rocky areas mainly relate to mountain rescue. This search method is carried out by skilled and well-trained trekkers or mountainers supported by ground navigation and air support. This is one of the most difficult geographical position for SAR operations.

- **Cave rescue:** It is an advanced SAR method for injured, streaked and lost cave explorers.

- **Urban search and rescue (UBSAR):** The mission of UBSAR is to locate and rescue people from falling buildings or other cities and industries. Due to its specialised nature, most teams include police, fire and emergency care personnel and are multidisciplinary.

- **Combat search and rescue:**It is a search and rescue operation conducted during a war within or near a battle zone. This is one of the most challenging operation as it has to be performed under the nose of army persons in war zone.

- **Maritime search and rescue:** It is carried out at sea to rescue stranded sailors and passengers, as well as survivors of crashed aircraft.

Looking at the past and current trends for search and rescue worldwide, Canada has the largest SAR component, with duties shared between the Canadian Coast Guard and the Defense Forces [39]. They have allocated five aircraft squadrons plus three combat support squadrons with SAR functions. The United States was the first to develop SAR technologies. Countries such as Israel, Germany, Russia, the United Kingdom, Turkey, and others currently have a well-developed SAR network. Using cargo aircraft, the Indian Air Force (IAF) have also performed various relief operations to distribute food and medical supplies.

2.2 Technological solution of disaster response

As a result of the hasty evacuation and rescue process, normal life is stressful and painful. Multiple methods for post-disaster rescue and SAR based on the robot are accessible and are one aspect of this. A range of ground and aviation robots are proposed by [121, 48] for SAR. In addition, [169] offers a fully independent indoor and outdoor rescue UAV robot. A broad range of work exists on rescue and navigation in complex GPS[122] or an unknown environment[5]. In order to reduce the effects of the natural disaster on humans, significant research efforts have been made in addition to robotics. To strengthen disaster management techniques, address extreme environmental circumstances by performing quick rescue work, [56] proposed a technique where UAVs were first time implemented for disaster management. Besides, [50] suggests a wireless network on UAV to extend the coverage of wireless devices. In this, a wireless sensor network (WSN) node deployed in the drone is recommended to extend the the coverage of network in remote areas. In addition to this, [110] proposed a technique to design the trajectory of aerial vehicle, which can help to design customize drone with camera fitted on UAV drone. The author in [176] analyzes the impacts of variables influencing UAV-based SAR systems efficiency and studied the optimization requirements of various search algorithms. In [153], a multipurpose UAV was proposed for rescue operations in the mountains. The suggested UAV is equipped with a high-performance camera capable of capturing both visual and thermal images. This multi-rotor drone is intended to satisfy environmental demands such as low temperature, high altitude, and powerful winds for mountaineer's terrain. This drone of about 5 kg capacity is capable of fully automated landing and take-off operation. These research have shown a path to the development of a strong and capable SAR system, and motivated us to think about developing an automated aerial and drone surveillance in disaster. In addition, a communication protocol proposed in [12] for supporting the communication and identifying the needy people though the network communication happening in the affected area. Here a drone-assisted Internet of Things (IoT) environment, called ACPBS-IoT is proposed.

2.3 Background study

Here, we studied the background of existing techniques in AI and DL to develop a robust and accurate automated system for search and rescue. This section outlines the methodologies for neural network and machine learning-based systems, beginning with the fundamentals of artificial intelligence to the fine-grained descriptions of current deep learning-based models for image classification and object detection.

2.3.1 Artificial intelligence

The origin of AI is from the first mathematical model of neural network proposed by Warren McCullough and Walter Pitts. However, in 1950, Allen Turing raises the question "Can a machine think ?" [101]; by this question, Allen Turing defines the fundamental goal and vision of artificial intelligence. At its core, AI is the branch of intelligence that aims to answer Turin's question. It is the ability to reason, discover meaning, generalize, or learn from experience. A modern definition of AI is "the study of agents that receive precepts from the environment and perform actions" [144]. Certainly, after the emergence of the first computer, researchers were interested in developing the system which can decide on their own. Previously, humans performed some of the functions now performed by computers. In machine learning and intelligence, some techniques, such as machine learning, have made tremendous progress up to the last decades. Numerous prototypes and application have been proposed in the past using machine learning in the field of fraud detection [175, 183, 162], weather forecasting[46, 100], surveillance[117, 103, 2], and health care[41, 16, 11].

The current decade has been highly significant for creativity in AI. Nowadays, AI has been an essential part of our everyday life. We use smartphones with voice assistants and machines with intellect features that most of us take for granted are use case of AI. It is no longer a pipe dream, and has not been around for some time. However, the neural network concept is old and has been used for research in various fields. The current trend of the neural network is evolved due to the evolution of extensive data. It started from the Imagenet challenge for image classification in 2010. Dataset for large-scale image classification is published in [29]

for Imagenet challenge. Later, this dataset becomes the benchmark for training and testing the Neural network for various image classification problem.

2.3.2 Machine learning

Machine learning (ML) is a branch of AI that gives programs the ability to learn and develop from experience automatically without being programmed directly. ML algorithms build a model based on the sample data, in other words, training data to predict unseen data[81]. Today, the application of machine learning is widespread, for example: playing music based on voice command (Alexa), web recommendation system, spam filtering, and robots vacuum cleaner. These are just examples; however, actually, we are surrounded by machine learning and AI applications. In the field of medical diagnosis, machine learning algorithms have got tremendous success[78, 79, 2]. We can expect more application with high precision in every field of life as the data is evolving day by day, and computing power become cheaper.

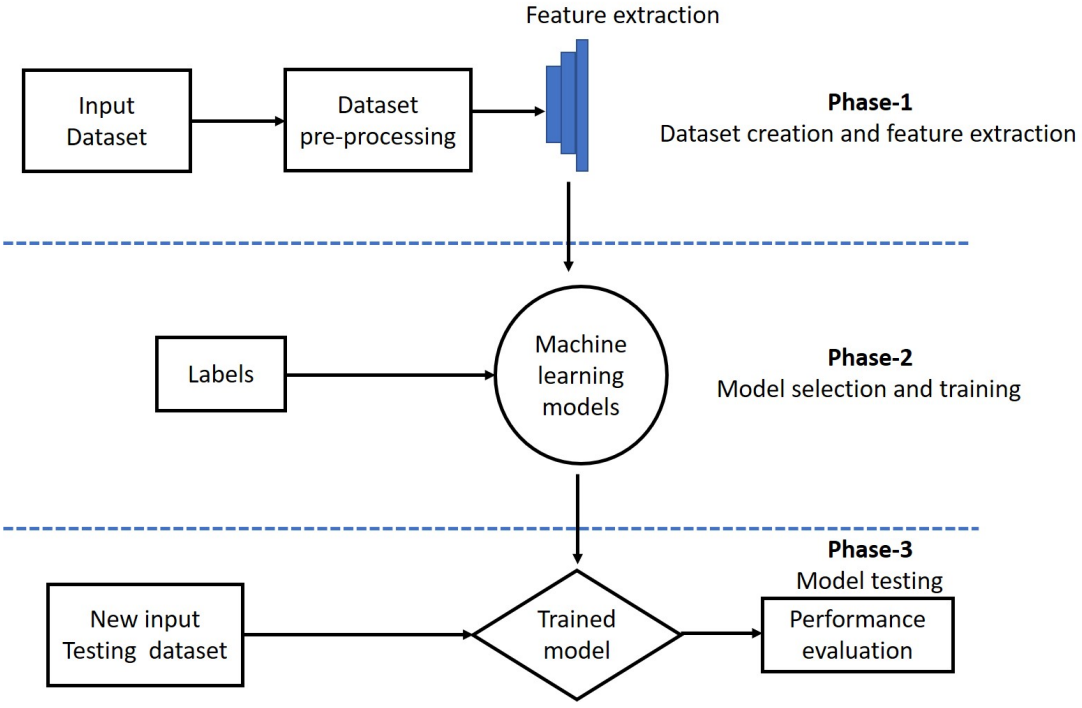


Figure 2.2: Basic workflow of machine learning algorithms

Machine learning algorithm works in three basic steps represented in figure 2.2 is as follows:

- Select and prepare an appropriate training dataset: dataset selection plays a crucial role in machine learning algorithms' success.

- Select the appropriate algorithm: The second step of developing a good application using ML is selecting the best-suited ML algorithm. ML algorithm is classified into two categories: Supervised and Unsupervised algorithm.

- Training the algorithm for creating a prediction model: Selected algorithms are then trained with the dataset. Various parameters need to be optimized at the time of training. Selecting and tuning the hyper-parameter for training is the most important part of this phase.

- Predicting and improving the model: Predicting for unseen data is the final goal of any application.

Machine learning algorithms can be divided into three main categories: Supervised learning, Unsupervised learning, and Reinforcement learning.

2.3.2.1 Supervised learning

Supervised learning model require labelled data for training. It is a kind of learning which is done through a supervisor(teacher). Both the input and output is specified in this type of learning. A complete survey of the supervised ML algorithm is presented in [120]. Binary classification, multi-class classification, regression, and ensemble learning methods are some important type of supervised learning algorithms used in various real-time applications in the past.

2.3.2.2 Unsupervised learning

Unsupervised algorithm of machine learning uses the unlabelled data and learns from the inherent features of input data. The algorithms scan the input data and find the correlation or patterns in the input data, for example, clustering. A detailed survey of unsupervised algorithms are explained in [43]. Clustering, anomaly detection, association mining, and dimensionality reduction are some important area where unsupervised learning algorithms are very useful.

2.3.2.3 Reinforcement learning

In this learning method, model interacts with the environment and learns based on the reward or acknowledgment get from the environment. Detailed examples of reinforcement learning techniques are explained in [75]. Reinforcement learning is often used for the applications Robotics, Video game, and Resource management.

These techniques have a wide variety of applications in real-life. Some vital application area of machine learning is digital assistance, recommendation system, online contextual advertisement, chatbot, fraud detection, cybersecurity, medical imaging, and self-driving cars. Few of these applications are still in the testing phase. The recent advancement of machine learning applications for remote sensing in terms of surveillance and crop disease detection increases and has an excellent power to deal with real-life problems[108]. The advancement of medical applications using Ml algorithms for diagnosis and cure have a great future ahead and articles proposed in [19, 82, 138] shows the capability and its future. Drug discovery is another application in medicine and biology which have made tremendous progress using ML algorithms [173, 189, 134]. The availability of big data for training the machine learning algorithm is one of the biggest reasons behind the success [10]. These algorithms are mostly used to automate different system and can be utilized in different ways for the drone surveillance automation as well.

2.3.3 Deep learning

This is a branch of machine learning and got recently tremendous popularity in the research community. The basic workflow of DL models are same as the ANN models, however, the number of layers in DL networks are usually higher.

2.3.3.1 The history and evolution

The origin of deep learning can be seen from 1943 when Walter Pitts and Warren McCulloch developed a model based on the human brain's neural networks. They used the mathematical algorithm "threshold logic" to understand the thought process. From 1950 to 1980, the math-

ematical model has been built in its early days, and the model is created for neural nets. The researchers have applied in various applications, including MIT's initial effort for developing the robot that could play "Ping pong" [147]. The idea of back-propagation was suggested in this phase by multiple researchers. However, there was no idea how to train multi-layer back-propagation neural networks for learning at that time. In 1982, the concept of MLP (multi-layer perceptron) was published, and later in 1987, popularized in "Learning representations by back-propagating errors" [143]. Later, in 1989 the published article "Multilayer feedforward networks are universal approximators "[58] reveals the key fact that multiple layers allow neural networks to implement any function, and certainly XOR theoretically. It comes to the reality with the application proposed in [84] for the handwritten written zip-code recognition with the back-propagation neural network for the very first time in literature. The addition of two types of layers - convolutional and pooling layers - are the primary distinctions of convolutional neural networks from original artificial neural networks. At that time, the convolution idea was called 'weight sharing.' In the early 90s, the principle of unsupervised learning was proposed with auto-encoders, which extract patterns from unlabeled data [57]. It uncovers a natural method for unsupervised learning that employs a model that describes a probability distribution over observable vectors. To learn such a model, it employs a neural network. Also, the researchers have determined that neural networks can approximate any probability distribution. The neural networks learn to make decisions, and this is where reinforcement learning begins. If supervised learning tells the learning algorithm exactly what to do, reinforcement learning 'rewards' successful choices over time rather than explicitly telling the algorithm to make the correct decisions. A PhD thesis was published with teaching robots using reinforcement learning in 1993 [91]. The thesis showed that robots could wall-pass and doors in a reasonable time.

Deep learning evolved in the late 90s from slumping neural networks and achieved significant performance in different application domain. Deep learning is a class of ML algorithms in which computers learn from experience and understand the real-world problem using a hierarchy of concepts [47]. The hierarchy of concepts allows the computers to understand small-small dots related to a complex problem. Traditional machine learning algorithms were limited to the preprocessing of data and manual crafting of object features.It took careful engineering to develop a feature extractor that transforms raw data into an understandable feature vector for pattern recognition and machine learning algorithms [83]. However, one of the most important aspects of the deep learning algorithm is that the subsequent layers automatically extract

features from data and find patterns based on those features for classification, identification, and segmentation. Deep learning applications have increased enormously in recent years and are expected to grow as the more accurate dataset is coming. The availability of benchmark datasets for object classification, object detection, and action recognition allows deep learning models to perform the task. Deep learning has made significant strides in the area of artificial intelligence. The discovery of complex structures in large amounts of data has proven to be extremely effective, and it has spread to many field of research, business application area.

2.3.3.2 Working of deep neural networks

At a fundamental level, deep neural networks for learning are how the human brain filters information and makes decisions. They use many layers of nonlinear processing units called nodes for feature extraction and transformation. So, it is a layered architecture in which each successive layers take input as the output of previous layers. In this, each level learns to transform its input into a more abstract and composite feature representation. For an image input to recognize a human, the network's first layer might encode the edges and compose pixels. The next layer might contain an arrangement of borders, and the subsequent layer may represent the encoding of eyes and noses. So, the networks learn the feature of the given label and recognize accordingly after adequate training. For training, the network learns on the dataset provided. It evaluates the performance based on a cost function which is the difference between the predicted label and the actual input label. This difference needs to be minimized, and for this, the information goes back, and the deep neural network begins to mitigate the cost by tweaking the network parameters. This process is called back-propagation. So, most of the deep learning architecture uses back-propagation strategy for the learning and error minimization. Some of the most important deep learning models are networks such as artificial neural networks, convolution neural networks, and recurrent neural networks, which can be used alone or in combination with one another.

2.3.3.3 Artificial neural network

Artificial neural networks(ANN) is inspired by a biological neural network that consists brain. ANN is built on a network of interconnected units known as nodes. Each connection between nodes can bypass the signal to other nodes. The output at each neuron is calculated by some

nonlinear function of the sum of input units. The connection are called edges, and each edge has some weights associated with it[191, 182]. ANN has multiple layers, and each layer performs a different transform on its inputs. The main component of ANN are the Input layer, None (perceptron), Hidden layer, an Output layer, Edges, Weight, and Bias. A sample of ANN is represented in Figure 2.3.

Neural networks have gained enormous attention in the recent past. However, plenty of research have been done with ANN in various fields such as financial management, trading [33], medical [1], and forensics [128]. A neural network analyzes pricing data and uncover prospects based on data processing for making trade decisions. Specific approaches of technological study do not differentiate subtle nonlinear inter-dependencies and trends from networks. The accuracy of neural networks in making price estimates for stocks varies, according to studies. Some projections estimate 50 to 60 percent of the time for the right stock values, while others are correct in 70 percent of all scenarios. Some have suggested that a 10 percent performance gain is all that an investor should wish for from a neural network [129].

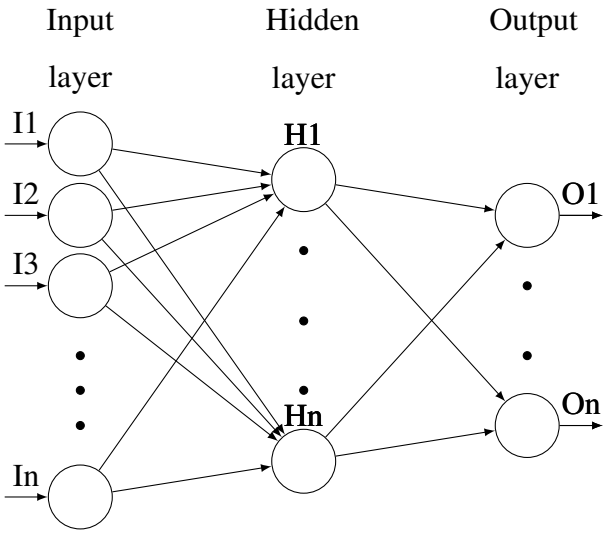


Figure 2.3: Basic architecture of ANN.

2.3.3.4 Convolution neural network

The convolution neural network architectures became prominent for complicated image-related tasks [192]. A particular type of deep neural network consists of layers with a convolution

feature, which works mainly as a feature extractor. Convolution is a point-wise multiplication operation of two superimposed pixel values. A feature map called kernel is convolved with the input image in a sliding window using the dot product. In a deep convolution neural network, multiple functions operate one by one. In combination with convolution layers, features are summarized with the help of the pooling function. The main components of a convolution neural network is Convolution, Pooling, Padding, Striding, Flatten.

The origin of convolution neural networks can be considered from LeNet[85] architecture. The original LeNet architecture consists of two convolutional layers with the two dense layers at the end. In this, each convolutional layer consists of an activation function followed by a pooling layer. It is one of the first architecture uses convolution operation for image classification tasks and is trained on the MNIST dataset for digit recognition. In this, 5*5 convolution kernel was used with 2*2 pooling function. The convolution block produces an output of the given size (batch size, channel, height, width). Each example in the mini-batch must be flattened before passing the convoluted block output to the fully connected block.

In ImageNet ILSVRC challenge 2012, the Alexnet model of convolution neural network is proposed, and it proved the ability of deep convolution networks for higher accuracy. However, it could not provide a general template for developing a new network, and then VGGNet is proposed in the 2014 ILSVRC challenge. This paper's main contribution was to believe that the depth of network plays a vital role in the network's performance, and it proposes a network with 13 convolutions and three dense layers (16 layers) and calls VGG16[155]. It uses a homogeneous filter size for all the convolution of 3*3, and for pooling, it uses a 2*2 kernel. One drawback of VGG16 is that it uses maximum memory to train the model, as it has more training parameters.

ResNet is another variant of a convolution neural network. A shortcut connection is created[55] which minimizes the exploding gradient problem in large convolution neural networks. This shortcut connection is also called residual connection.

2.3.3.5 Recurrent neural networks

The convolutional model takes a given number of inputs and produces a fixed-sized vector with a predetermined number of steps as an output. Recurrent networks, on the other hand, allow

us to work with sequences of vectors in both input and output. Directed cycles are created as a result of the connections between nodes. Unlike standard neural networks, the input and output of a recurrent neural network are linked rather than independent. Furthermore, every layer of the recurrent neural network uses the same standard parameters. You may use the back-propagation method [109] to trains the RNN in a fashion similar to the standard neural network. In this case the gradient computation is not only determined by the current step, but also by the prior step. The Figure 2.4 shows an example of the recurrent neural network.

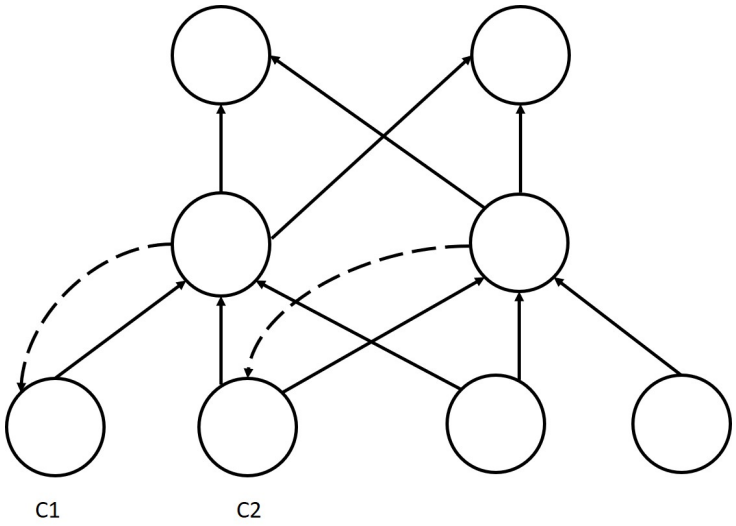


Figure 2.4: An example of simple recurrent neural network[109]

2.4 Derived application areas of deep neural networks

In modern eras, deep neural networks are utilized in almost every field of automation. The fundamental concept behind all the image and video level automation are derived from the original architecture of deep neural network architecture. some important derived architecture for various computer vision task are: object classification, object detection, object segmentation, action recognition, and video classification is explained in this section.

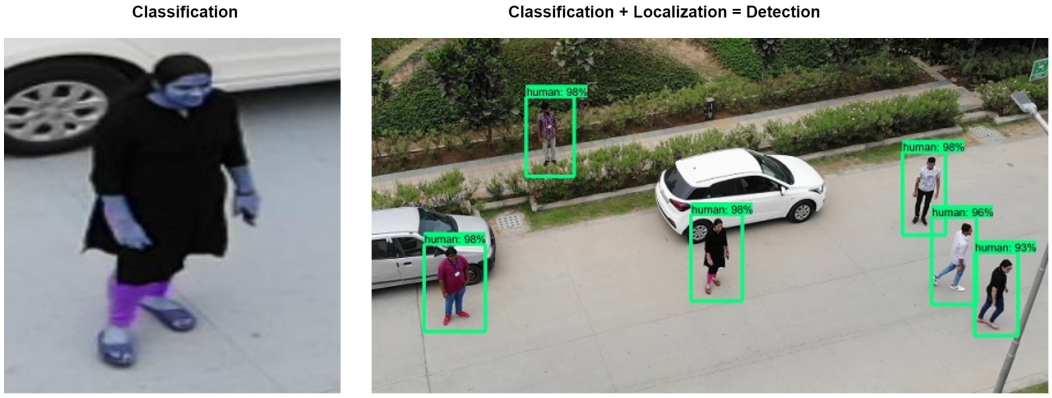


Figure 2.5: Visual example of classification and detection

2.4.1 Object classification

Object categorization is one of the most important challenges for decision making and automation in the natural image. With a powerful object classification function, computer vision provides a human thinking feature into machines which can be simulated artificially to make complicated judgments accurately. Figure 2.5 shows an example of classification problem, where human is recognized. Most image classification networks employ a convolution layer for extraction of features and fully connected layers for classification. In other areas such as object detection, action recognition also, the classification network is used inherently for object classification. For the technique used to classify the object, initially the traditional machine learning algorithms were used, like K- nearest neighbor algorithm has been used in [59]. However, traditional machine learning algorithms requires explicit feature engineering. The introduction of deep learning algorithms such as ANN, CNN, and RNN deep architectures or their hybrid designs reduces the extra work of feature engineering. Recently, the object classification architecture using deep learning achieves outstanding result in the area text classification, surveillance, and social network analysis.

The research article [51] proposes a deep learning architecture for the classification of age-group in social network analysis. Also, [186] proposed a classification network for sentiment analysis in social media. Likewise, there are plenty of work research work available in literature for social media analysis, uses different type of deep learning architectures. Multi-class image and video classification is another area where classification networks has been heavily

used in the past, such as [170] and [54] uses classification networks for video surveillance. Classification networks is able to solve many real life problems.

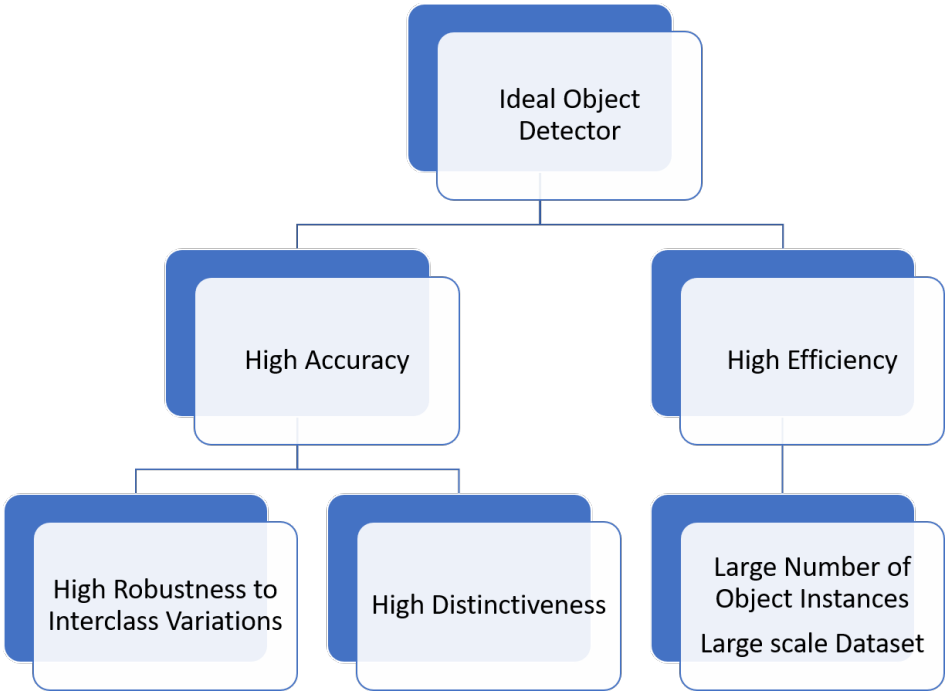


Figure 2.6: OD categorization for ideal object detection algorithms

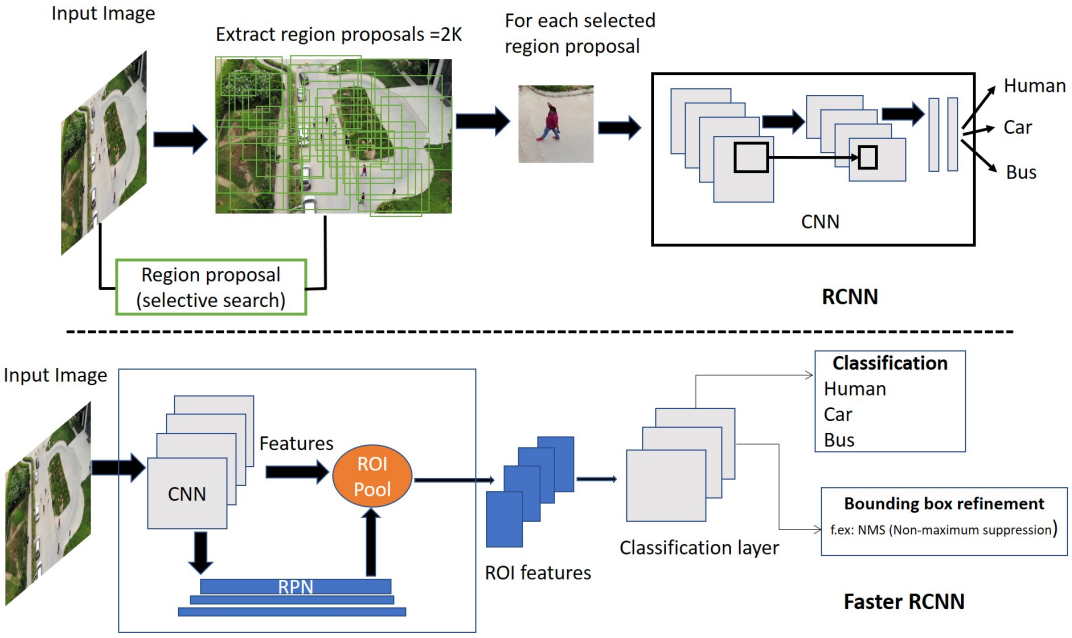


Figure 2.7: High level architectures of two-stage object detection techniques

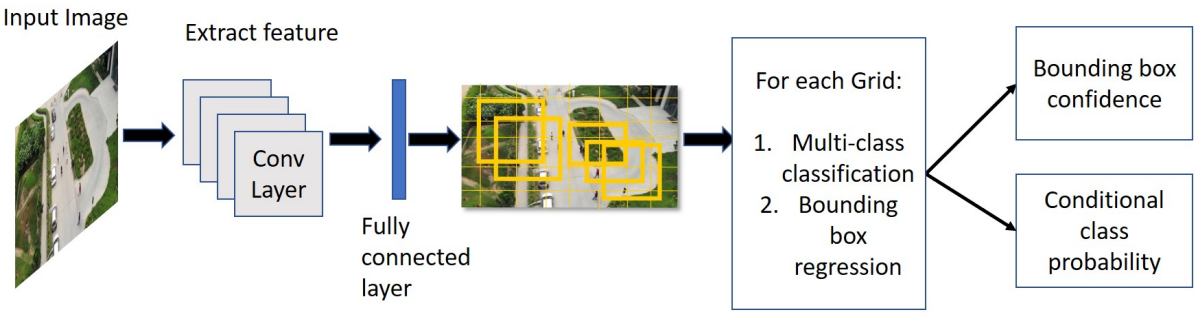


Figure 2.8: High level architectures of one stage object detection technique YOLO

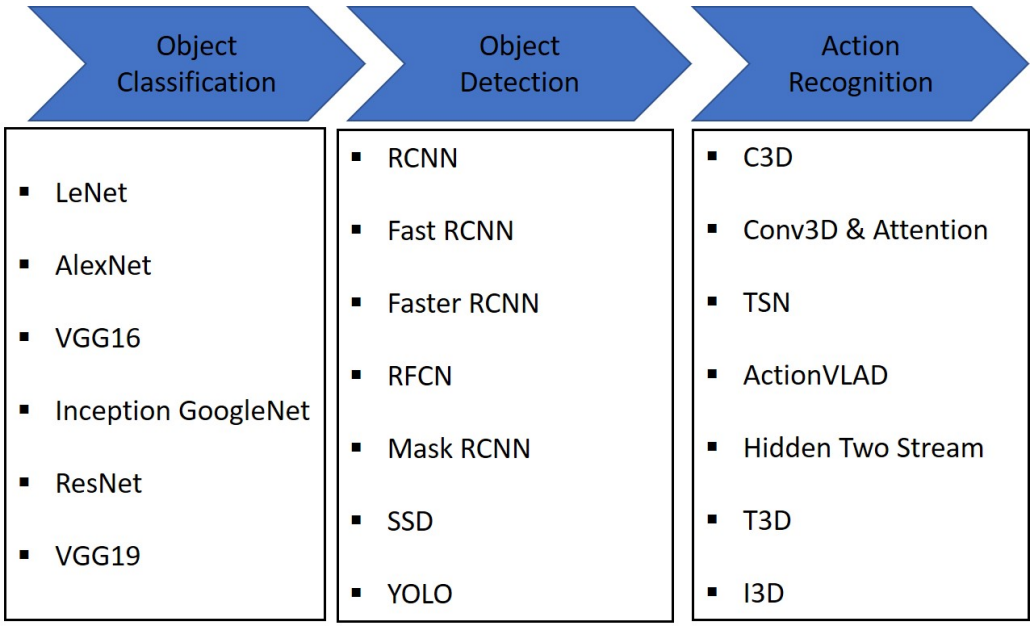


Figure 2.9: Key models using different CNN architectures for object classification, object detection and action recognition.

2.4.2 Object detection

Object detection (OD) is a fundamental concept in computer vision that seeks to recognise object instances in real pictures from a wide range of specified categories. Object detection evaluates if examples of objects from input classes (such as persons, vehicles, bikes, dogs, and cats) exist in a given image and, if so, outputs the spatial position and frequency of each occurrence (e.g., via a bounding box). Robot vision, consumer electronics, defence, autonomous driving, human-computer interaction, content-based image retrieval, intelligent video monitoring, and virtual reality are just a few of the applications for object recognition in artificial

intelligence and information technology. The goal of generic object detection is to find and categorize a wide range of natural things. The ideal objective of generic object detection is to build general-purpose object detection algorithms that meet two competing goals: high quality and high accuracy. Figure 2.5 gives a basic understanding of object detection models output.

Based on their working and internal structure, object detection algorithms are categorized into two main categories/; One stage object detection and two stage object detection.

2.4.2.1 Two stage OD architecture

It contains a pre-processing phase for the region proposal, resulting in a two-stage pipeline. This function generates object suggestions for category-independent areas in a given input picture. CNN characteristics are produced from these object proposal regions, and category-specific classifiers are then utilised to determine the category of producing object proposals. The high level architectures of some important two-stage object detection techniques are represented in Figure 2.7.The detailed working of these techniques are as follows:

- RCNN : A method to extract the limited 2000 regions for object proposal has been proposed first in [45]. This paper is a breakthrough in the direction of huge number of huge number of object region generation in object detection. In this, the extracted 2000 region proposals are warped into a square and fed into a convolutional neural network that generates a 4096-dimensional feature vector as output. Here, CNN acts as feature extractor, the output dense layer consists of image-extracted features, and features extracted are fed into SVMs in order to identify the presence of object into the object proposal candidate area. However, it still takes too much time to train the network because the network has to classify 2000 object proposal per image at the time of training.

- Fast RCNN: [44] tries to solve the issue of previous network in the section of two stage object detection. Instead of feeding the 2000 region proposal to the CNN, this method feeds the input picture to the CNN for feature map development. The generated convolutional feature map is utilized to identify the region of proposals and warp them into squares, which are then reshaped into a fixed size feature vector for the fully connected layer by the RoI pooling layer. The ROI feature map is used by the fully connected layer to estimate the class of the proposed region as well as the offset values for the bounding

box. Unlike RCNN, you do not have to feed the convolution neural network with 2000 regional suggestions every time. "Fast R-CNN" is quicker than "R-CNN."

- Faster RCNN: The article [140] proposed a novel architecture with less inference time, and it is also a kind of region based object detection approach. This creates fewer regional proposals than other algorithms of region based OD network. The VGG-16 network was initially used as a feature extractor for Faster RCNN regional proposals. However, some advanced features extractor networks have been used over the time.

- RFCN: [26] proposed an approach for object detection for fast processing and training of network. It still uses RPN network for region proposal generation. But unlike the RCNN series of network, the fully connected layer after ROI pooling is removed. In this, all major component of the network is kept before ROI pooling to generate the score maps. In this approach, no learnable layer is available after ROI pooling and hence it is a cost effective approach.

In these techniques, ROI extraction and selection is important, and since object appears in drone based surveillance usually appears small, and hence, we can relate this with the medical images ROI. An accurate region selection method is proposed in [156]. This papers presents the importance of location, shape, and size of ROI for liver tissue characterization. In addition, [80] proposes an automatic ROI extraction technique for ultrasound images, which could be useful for small object ROI extraction and selection.

2.4.2.2 One stage OD architecture

It is a single suggested approach that does not segregate detection proposals, resulting in an one step pipeline. Regardless of these advancements, region-based techniques for mobile or wearable devices with limited storage and computing capacity may be computationally costly. Therefore, researchers began to build unified detector strategies instead of refining each portion of a complex region-based object detection technique. An example of some prominent one stage architectures are shown in Figure 2.8

- YOLO: It was initially used in [139] to identify objects quicker and in real time. The object recognition issue is represented as a regression issue in this case, and a single

neural network predicts both bounding boxes and class probabilities straight from the input image. In the original paper, they have claimed it as the fastest network with 45 frame per second prediction rate with full YOLO algorithm and a fast YOLO algorithm having 155 frame per second prediction rate.

- SSD: In [95], SSD was originally proposed, and uses VGG16 feature extractor in multi-layer detection of the objects. This method combines classification and localization into a single network.

2.4.3 Video classification

Video classification model could be an utility tool for the videos available on the internet in the field of sports, music, news, and animated films etc. In some cases, it could also be utilized for action recognition. In [8], an efficient video classification technique is proposed that uses a two-level classification strategy, i.e., one for feature extraction from the local snippet and the other for global video representation. Experiments were performed on UCF-50 and KTH datasets. In another work [15], a model is designed for crowd video analysis that uses the mid-level descriptor feature of the video to classify and analyze it. Experimental result of this article validates that the same level of accuracy can be obtained by assuming the whole crowd as a single entity for feature extraction and classification. In another work [194] related to video and scene classification, the saliency feature-based approach is followed. In this, the video frame is divided into different areas according to the saliency rate, and model the static and motion feature vectors simultaneously. In this approach, adaptive weights are learned for each class, and expected that video classification would be improved as it outperformed in their experiments with UCF-101 and CVV dataset. In another work[77] for emotion recognition and video classification through short videos, a novel CNN+RNN based approach is used in which CNN is used to extract low, mid, and high-level features. Also, at the next level, RNN uses the extracted features as input for multi-task framework exploitation. The output is an intermediate level prediction, and final estimates are calculated as the mean or median values of all these predictions. This approach has outperformed and crosses the state-of-art emotion recognition and video classification performance with a one-minute Gradual-Emotion (OMG-Emotion) dataset. Convolution Neural Networks (CNNs) have been established as dominant class models for im-

age recognition problems. Various experiments has been performed in literature and evaluates the performance of CNNs on large-scale video dataset of 1 million YouTube videos belonging to 487 classes. The study of the foregoing techniques for classification approaches using CNN and suggest the various ways to use spatio-temporal information for accurate object and action classification and speed-up the training. Article [71] proposed various network in this category, and the best patio-temporal network of this article display significant performance improvements compared to the single-frame model (59.3 % to 60.9%). Further studying the best model's of this article, achieves a generalized performance by retraining the top layers on the UCF-101 action recognition dataset and observing significant performance improvements compared to the UCF-101 baseline model (63.3% up from 43.9%).

Other than this, there are many other application areas of deep learning and AI such as sentiment analysis, Image generation, Image captioning, Video summarization etc. However, these applications areas are not in the scope of thesis, so we have discussed here only the derived application related to this thesis work.

2.5 Aerial object detection and action recognition

Object detection and action recognition is one of the most challenging task in computer vision, and it becomes more difficult from the aerial and top view of surveillance. In literature plenty of efforts has been made for recognizing vehicles for aerial images, still, recognizing human and their action is challenging. Here, we give the summary of the research available in literature for the aerial and drone surveillance.

The deep object detection models have been applied for various aerial surveillance research and [36] gives a brief survey of all the existing object detection methods applied on aerial images. Out of all the object detection application, vehicle detection is an area of traffic and transportation management and has been tried mostly in the past. To support the vehucle detection, the reserachers have proposed some benchmark dataset and achieved significant result with VEDAI512 [159] and DLR 3K [136] for aerial view of vehicle detection. In [158], object detection approaches such as R-CNN, Faster R-CNN, and SSD were tried, and it was discovered that R-CNN with their proposed area proposal network outperforms for vehicle detection

Table 2.1: Comparison table for object detection and action recognition methods in drone images

S.No	Title	Type	Methodology	dataset	Results and Remarks
1	Comprehensive Analysis of Deep Learning based Vehicle Detection in Aerial Images [158]	Aerial Image Vehicle detection	SSD, Fast RCNN Faster RCNN	VEDAI512 DLR 3K	88.7 % claimed precision Computationally expensive Comparatively bigger object Easy to detect
2	Performance comparison of deep learning techniques for recognizing birds in aerial images [96]	Aerial Image Small Object Object Detection	YOLO, SSH Tiny face	LBAI	Low resolution images were taken for testing deep networks for small object (10 px to 40 px) detection in Aerial imagery
3	Towards Fast and Accurate Vehicle Detection in Aerial Images using Coupled Region-Based Convolution neural network [30]	Aerial Images Vehicle Detection	AVF, SS+RCNN, Faster R-CNN	Munich	Their VAHL technique fails for identifying the vehicle like regions for hard cases or where features are less.
4	Using deep learning and low-cost RGB and thermal cameras to detect pedestrians in aerial images captured by multirotor UAV [28]	Aerial Images Human detection Pattern Recognition	HOG+ SVM CNN	GMVRT-v1 GMVRT-v2 UCF-ARG	90 % accuracy, dataset is inappropriate as multi-person detection required for real-time UAV video analysis
5	Object Recognition in Aerial Using Convolution Neural Network [137]	Aerial Images Aeroplane Detection	YOLO	Unpublished	84% Precision for aeroplane, dataset is neither standard nor shared.
6	Car Detection in Aerial Images of Dense Urban Areas [34]	Aerial Images Vehicle detection Traditional approach	Sliding window approach, with CNN	Vaihingen with Geo-information	0.75 precision Urban environments (Complex background) will increase the false positive due to requirement of knowledge of roadmap it is computationally expensive.
7	Region Proposal Approach for Human Detection on Aerial Imagery [105]	Aerial Images Pattern Recognition Human Detection	CNN model for ROI	Unpublished	Region proposal approach works well for small object Still performance of model is not up to the mark require improvement and analysis for real time human detection.
8	Okutama-Action: An Aerial View Video dataset for Concurrent Human Action Detection [9]	Aerial Images Human Detection Action recognition,	SSD	Okutama	0.18 mAP@0.50IOU
9	Convolution Neural Networks for Aerial Multi-Label Pedestrian Detection [157]	Aerial Images, Action Detection	SSD 500	Okutama	0.28 mAP@0.50IOU
10	UAV-Gesture: A dataset for control and gesture recognition [132]	Aerial Images, Action Detection	P-CNN descriptor	New dataset 13 action for hand signals	85 % accuracy is achieved for 13 hand signals

in aerial view. Munich [93] is another relevent dataset proposed for vehicle detection in aerial imagery, however, this dataset has only 20 images for training [30]. This paper gives a separate region proposal network called AVPN (accurate vehicle proposal network) to hypothesize vehicle locations.

Other than vehicles, few other objects were also attempted to detect from aerial images and [28] proposed a YOLO-based model for airplane detection. This model uses a five-stage framework that is window evaluation, extraction, and feature encoding, classification, post-processing, and ROI extraction. Various techniques were used to detect objects such as insects, drones, and trees in previous studies. However, human detection and their action recognition from the aerial image is challenging because ground covering area of human is very low. Human identification and action recognition on aerial images have received least attention in the literature. In [9], a dataset is proposed with 12 actions, and it followed the SSD-based approach for action classification. However, more efforts have been made with the same SSD model in [157], and got a performance of 28% mAP. Still, the area of action detection in aerial drone images is contemporary and open. It needs more effort in terms of data and models that can detect humans more accurately and classify their actions precisely. One of the important application areas of object detection is obstacle detection for visually impaired people, and [67] suggests a computationally fast and straightforward obstacle detection technique. In this, a ground plane removal method was also presented to filter out the ground area from the image. In another work for obstacle detection, a kinetic sensor and 3D image processing-based system is proposed for the indoor environment [133]. Detail of the former object detection and action recognition techniques are summarized in Table 2.1.

2.6 Aerial action recognition datasets

Action recognition is an important field of computer vision and surveillance and useful for various real life applications like crowd monitoring. However, due to the limited exposure of human feature in drone surveillance, action recognition is a difficult task and require accurate dataset. A detailed comparison of various aerial and drone dataset features is represented in Table 5.5. The proposed action recognition dataset in [9] showed the way for drone surveillance. Origi-

nally, it contains 43 minutes long fully-annotated sequence. Each and every actions performed in this dataset is natural action of humans and performed by multiple persons that give it variety. Various approaches have been explored in the literature with this dataset; however, the performance of prior object identification and action recognition models was not up to par, and delivers a maximum accuracy of up to 28 percent mAP values with 12 actions presented in this.

These dataset have emerged as open direction of research for aerial and drone surveillance, however, it could be seen as opening path of drone surveillance. More specifically, the recently developed dataset by university of Central Florida, has guided many researchers throughout the world to think about these drone based surveillance. In addition, [69] have proposed a novel dataset for face recognition through drone surveillance. They have proposed the dataset with approximately 200 video clips with 786K face annotations. Recently proposed dataset in [132] have approximately 37.2k number of frames for 13 different action of hand gesture. The dataset is proposed for identifying the hand gesture for flight control at the airbase, however, it is not appropriate for recognizing action at the surveillance. In this, ucf-rooftop proposed is a useful dataset in surveillance and could be useful for drone action recognition. However, the factors such as moving vehicle, varying angle, and height is missing in drone surveillance. Recently, another addition in the drone based action recognition through [164], where they have applied a different approach with the limited training images. While, [132] proposed a mini drone dataset for UAV gesture recognition, where mostly the gesture are recorded in outdoor spaces with 13 different action. In this, the author have achieved a maximum accuracy value of 91.9 % through PCNN descriptor Other action recognition dataset available in literature were mostly captured for ground and front view of the object such as UCF101[161]and HMDB[65].

All these dataset proposed in literature are mostly captured in a fixed background space with limited actors. A benchmark dataset is proposed in [32] on complex scenario with new challenges. In this, they have captured 10 hours of video sequence with approx 80,000 frames fully annotated with bounding boxes. Here, they took 14 different kind of attributes for three fundamental computer vision task object detection, single object tracking, and multiple object tracking in aerial surveillance. In the direction of aerial action and even detection, [119] proposed a new dataset ERA (Event Recognition in Aerial videos), consisting of 2,864 videos each with a label from 25 different classes corresponding to an event unfolding 5 seconds. They claimed that the dataset have sufficient intra-class variation and inter-class similarity for the

Table 2.2: Comparison of aerial image and video datasets

dataset	Scenario	Purpose	Environment	Frames	Classes	Resolution	Year
UT Interactions [145]	Surveillance	Action-recognition	Outdoor	36K	6	320*240	2010
NATPOS [160]	Aircraft-signaling	Gesture-recognition	Indoor	N/A	24	320*240	2011
VIRAT [124]	Drone-Surveillance	Event-recognition	Outdoor	Many	23	Varying	2011
UCF101 [161]	YouTube	Action-recognition	Varying	558K	24	320*240	2012
J-HMDB [65]	Movies, YouTube	Action-recognition	Varying	32K	21	320*240	2013
Mini-Drone [14]	Drone	Privacy-protection	outdoor	23.3K	3	1920*1080	2015
Campus [141]	Surveillance	Object-tracking	outdoor	11.2K	1	1414*2019	2016
Okutama-action [9]	Drone	Action-recognition	outdoor	70K	13	3840*2160	2017
UAV-gesture [132]	Drone	Gesture-recognition	outdoor	37.2K	13	1920*1080	2018
Ucf roof-top-dataset[123]	Drone	Action-recognition	outdoor	486 video	10	1920*1080	——

better understanding of deep learning models. In addition to this, recently [90] proposed a dataset for human behaviour understanding through UAV. This dataset contains 67,428 multi-modal video sequences and 119 subjects for action recognition. It also contains 22,476 frames for pose estimation, 41,290 frames and 1,144 identities for person re-identification, and 22,263 frames for attribute recognition. They have captured the dataset through a flying UAV in multiple urban and rural district in daytime and night over three months.

2.7 CNN for text classification

Text recognition aims to take the image and identify the single word depicted inside it. In the initial days of text extraction, usually, rule based system is followed. [20] presents the comparison of rule based system and a supervised ML model trained on large dataset for document classification and postal address classification. There are few techniques available in literature of handwriting or historical document recognition[167, 49, 74, 7]. Usually, these models are not generalized and cannot give the best result for problems like generic scene text extraction [148, 40, 116, 18]. The main challenge of text extraction from a generic scene is the variable foreground and background texture. For scene text recognition, these methods can be categorized into mainly two groups which is character-based recognition and word-based recognition. Character-based models utilize the classification method per character to produce complete word recognition throughout the word picture. In another research article [190], authors propose clustering of sub-patches of characters which can learn a collection of mid-level characteristics, and characters are identified by random forest classifiers, strokeletal and HOG characteristics. Plenty of other research for text recognition [3, 6, 52, 70, 126] uses CNN as character classifier. [3] uses the word image into character regions either through supervised or unsupervised binarization technique. They use a supervised classifier. This CNN Text Classification can also be used for disaster assistance and for a new way to identify help situations.

2.8 Conclusion

This chapter has discussed the life cycle of disaster management and point outs some leading organization and their impact in previous disaster situations. We have presented a brief introduction of search and rescue methods in disaster situations and discussed the recent search and rescue operations performed by the Indian Air Force. It shows the value of a strong search and rescue operation for quick response and saving human lives.

Recent advancements in technology can be utilized to develop a strong search and rescue system. This chapter gives a brief detail of technological perspective for establishing a robust system for this. The automation technique and the advancement of recently developed AI and DL models are also discussed here. These techniques drew a path to develop a highly accurate system for prediction and analysis. The role of aerial and drone surveillance to monitor the disaster-affected area is high, requiring a highly automated surveillance system. This chapter emphasized the deep learning models for action recognition and object detection for drone surveillance automation and discussed the existing models of this area.

Automation of aerial and drone surveillance requires an adequate dataset of surveillance from different angles and heights. A comparison of existing dataset for the automation of aerial and drone surveillance is also presented in this chapter. In addition, this chapter also gives a tabular comparison of all the existing approaches applied in the literature for aerial and drone surveillance for vehicle detection, human detection, action recognition etc.

Chapter 3

Drone Surveillance for Human Detection and Action Recognition

In this chapter, we have developed a development drone surveillance dataset for human detection and action recognition. The proposed dataset is tested with various deep learning based object detection algorithms for the annotated humans and their action. Through our experimental analysis, the dataset is validated and result is compared with various earlier models for human detection and action recognition in aerial and drone surveillance.

3.1 Introduction

Drone surveillance has capability to support one of the important aspect of disaster management that us search and rescue. Recently, the police and fire departments have adopted drones and collaborated with local SAR teams. Using drone for large scale disaster events require multiple automatic drone for scanning the affected area. As a result, drone surveillance automation is necessary, which can be accomplished with the help of a camera and advanced DL model de-

ployed on the drone itself to identify the precise locations where assistance is necessary. Figure 1.2 shows an example of automated surveillance and search operations. In this diagram, after identifying the human's location, the GPS location of the human can be sent to the rescue team for the fast and productive rescue. The recent success of deep-learning techniques for object detection and activity recognition has prompted us to investigate their use in drone surveillance. A major need of a deep-learning strategy is that it must be trained on a large amount of data. Since most of the surveillance dataset available in the literature are for ground-level surveillance, such as UCF dataset [161] and unable to be utilized for aerial surveillance hence, it is our primary objective to develop a dataset of aerial action recognition for SAR. Besides, deep-learning models use these datasets to extract the feature and classify them into the labels automatically. Out of all other neural networks used for classification or localization, convolution neural networks (CNN) suit image-based feature extraction more. An example of object classification and object detection is understood through Figure 2.5. In this, detection is a combination of classification and localization. The classification problem of images is mainly to classify the image into a different category (labels), while the objective of detection is to identify the label of the object and determine the exact position of classified labels in that image. Object detection models such as Faster R-CNN [140], RFCN [177], SSD [95], and YOLO [139] are convolution neural network-based techniques who have outperformed with the ground-level images. [36] provides a survey of past neural network techniques for aerial image and video analysis. However, detecting the human in aerial and drone surveillance and recognizing their aerial and drone surveillance actions is still challenging due to the angle of surveillance and feature variation due to height. Prior attempts to recognise the behaviour of humans in aerial and drone surveillance have outline the requirement of precise data sets and the development of efficient algorithms for human detection and action recognition [130, 9].

For drone surveillance automation, the dataset plays a crucial role in deep learning and artificial intelligence algorithms. Human detection and action recognition datasets in literature such as Ucf[161], [9] is not adequate for drone surveillance due to the height and angle. This chapter's main contribution for the automation of drone surveillance is: a. develop a novel dataset for human detection and action recognition in aerial and drone surveillance. b. Testing the ground-level algorithms for human detection and action recognition and device the appropriate human detection and action recognition technique.

Figure 3.1: Humans performing different action in the proposed dataset

3.2 Dataset development

This section describes how the dataset was collected and pre-processed. In addition we have also compared our developed dataset with the most recent drone surveillance dataset and explain how it benefits real-world applications.

3.2.1 Dataset collection

The dataset is captured in a variety of places (in and out of our campus). For this, we have used a drone equipped with a high definition camera from the height between 10 meters to 40 meters. We have used a GoPro Hero 4 Black camera with an HD lens (5.4mm, 15MP, IR CUT) and a 3-axis solo gimbal for video recording. We recorded the videos with HD (1920*1080 pixel) formats at 60 fps. Participants were instructed to act independently so that multiple actions or instances of activities could be recorded in a single image frame. Six activities have been recorded while the drone is moving in the horizontal and vertical direction. Our dataset consists of all the human acting scenarios, i.e., the human is alone and an acting group of the human

performing the same action. Humans perform different activities individually. While recording the video, sometimes wind affected the drone, and the drone's varying height for capturing the multiple humans performing other actions makes it closer to the practical scenario. Figure 3.1 represents the sample of images used in our dataset.

3.2.2 Action selection

The actions were selected from a general crowd behavior for UAV capturing the signals. The selected six actions were shown in Figure 3.1 Our primary concern is to collect the drone dataset that is helpful for SAR applications. Actions were selected based on the existing action classification dataset by adding one more action as a waving hand. Waving hand action itself has variations such as single hand waving, both hand waving, person waving hand while standing at a fixed place, and person waving both hands while he is standing. The same four cases with the person were walking and waving, sitting and waving, and running with waving. Our dataset has a rich amount of sample and variation for our prime class waving a hand. For the action selection, some more factors have been taken care of as follows:

- They should be easily identifiable from a moving drone camera.
- Actions need to be crisp enough to differentiate from each other.
- Actions should represent the normal crowd behavior.
- Actions should be diverse enough to meet the requirement for different real-life applications.
- The action should be discriminative to allow spatial and temporal methods for action recognition.

3.2.3 Variation in dataset

Actors who participated in the dataset are not professionals and hence performed naturally. Since we have captured the data with different actors performing other actions in all the videos,

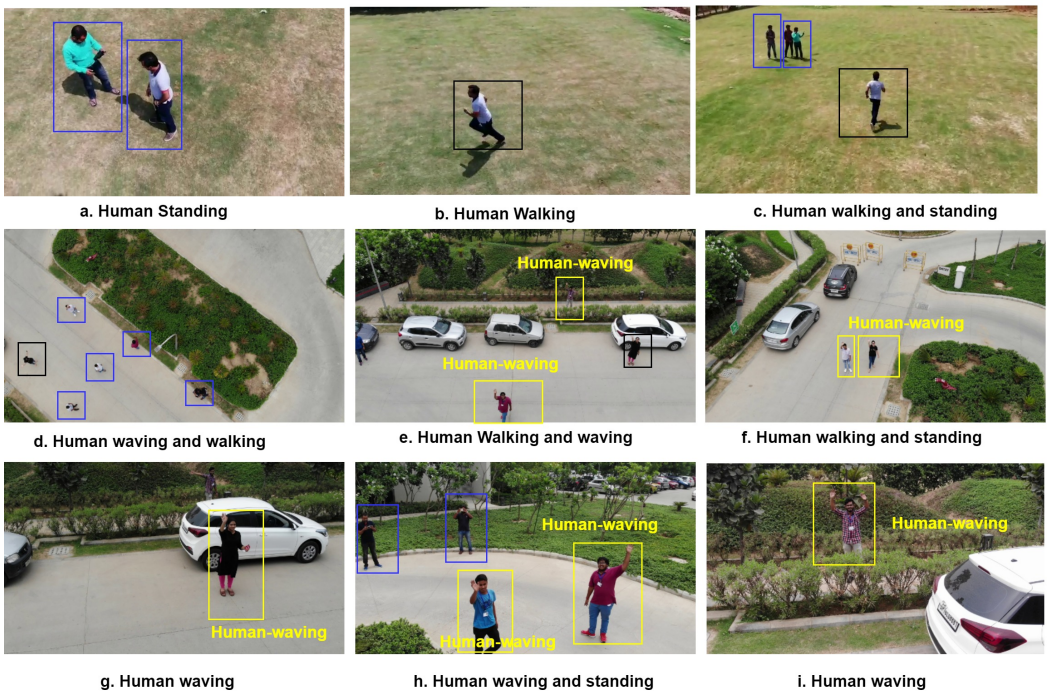


Figure 3.2: Annotations of humans performing multiple actions

there is a variation in how different people are conducting the same activity. Also, the dataset is captured from the different orientations, angles, camera movements, and height make it more generalize for real-time applications. Videos are taken at different times of day, and different lighting provides variety to the dataset in terms of weather conditions and illumination. Our actions are chosen from a variety of real-world applications that necessitate certain difficult-to-classify tasks, such as running, walking, waving, lying, etc.. The dataset is diverse because it was captured on different days of the month, with actors wearing a variety of clothing and their acting styles. These variations create a challenging dataset for action recognition; at the same time, our dataset is diverse enough for practical scenarios.

3.2.4 Pre-processing of dataset

Dataset for drone-based action recognition is collected in the form of video. Since we are targeting only spatial feature-based multi-class action recognition, all the frames are not necessary. To avoid repetition, we skipped every ten frames for one frame in our dataset. Other than our proposed dataset, the existing dataset is also re-framed according to our requirements. Okutama dataset is published initially for 13 actions. We have prepossessed the Okutama videos for our

Table 3.1: Summary of proposed six-class action dataset

Feature	Values
Number of Actions	6
Number of Images	7000
Average instance per class	1000
Frame rate	60 fps
Resolution	1920*1080
Camera motion	Yes, slow and steady
Annotation and its format	Yes, Bounding box, .xml format

experimental study by extracting the frames and annotating them in six different actions. The activities we have annotated is same as it was initially in Okutama, one action that is waving hand was added extra in the annotation.

The Preprocessing steps for dataset development after capturing video is as follows:

- **Frame extraction**: dataset was captured in the form of video and the frames were extracted using OpenCV package of python.

- **Frame selection**: Originally, the extracted frames have repetitions in the form of action features. We tested this by deleting 10, 15, 20 frames in order to retain one frame in our dataset.

- **Action annotation**: For deep learning models to work accurately, each human must be localized in the image. Figure 3.2 shows the sample of annotations. In this, there are a varying number of humans acting differently in different frames of video. However, for other UAV applications, where a simple human detector is required, this dataset can be utilized with a minor change in CSV file of the dataset. For the annotation of every image, the LabelImg application is used. Finally, we have two sets of annotations for our dataset, with two and six actions, respectively. Figure 3.2 depicts instances of humans annotated as performing various acts.

Table 3.2: Comparison of proposed dataset with the existing datasets

dataset	Scenario	Purpose	Environment	Frames	Classes	Resolution	Year
UT Interactions [145]	Surveillance	Action-recognition	Outdoor	36K	6	320*240	2010
NATPOS [160]	Aircraft-signaling	Gesture-recognition	Indoor	N/A	24	320*240	2011
VIRAT [124]	Drone-Surveillance	Event-recognition	Outdoor	Many	23	Varying	2011
UCF101 [161]	YouTube	Action-recognition	Varying	558K	24	320*240	2012
J-HMDB [65]	Movies, YouTube	Action-recognition	Varying	32K	21	320*240	2013
Mini-Drone [14]	Drone	Privacy-protection	outdoor	23.3K	3	1920*1080	2015
Campus [141]	Surveillance	Object-tracking	outdoor	11.2K	1	1414*2019	2016
Okutama-action [9]	Drone	Action-recognition	outdoor	70K	13	3840*2160	2017
UAV-gesture [132]	Drone	Gesture-recognition	outdoor	37.2K	13	1920*1080	2018
Proposed-dataset	Drone	Action-recognition	outdoor	10K sorted from 100K	6	1920*1080	——

3.2.5 Dataset summary

There are few dataset available for aerial object detection and action recognition in the literature. Table 5.3 reflects a comparison of the most recent aerial image and video datasets. Best-suited data for our drone-based action recognition is Okutama action dataset [9]. We have modified the Okutama dataset images by extracting their frames and annotating them as per our requirement for our analysis.

Also, we have created our dataset for human detection and action recognition. For the human action detection dataset, we have used a drone to capture images and videos. We have captured two different kinds of images with two actions and six actions. Existing dataset for human and their action detection is very complicated and take from more than 65-meter height. The performance of DL models with existing datasets for action recognition is unsatisfactory, which motivates us to develop our dataset for drone surveillance. Six actions were selected for our second part of experiments and those actions are:

- Person Standing
- Person Sitting
- Person Laying
- Person Handshaking
- Person Walking
- Person Waving

Table 5.2 provides an overview of this six-class action dataset. We have combined few actions to convert it into two-class action dataset other than waving a hand. So, our two-class action dataset contains the following classes: First class as "Person" waving and second class is "Other".

Human features are best visualized from a height of 10 to 40 meters. From this height, human features are visible and can contribute more to the human detection and action classification.

3.3 Performance evaluation metrics

The detail of principal evaluation metrics used to estimate the performance of developed model and dataset proposed in this chapter is as follow:

3.3.1 Intersection over union(IOU)

It is an assessment method used to measure the accuracy of an object detector. Often we view this metric for assessing performance of PASCAL VOC challenge. IOU is able to support the evaluation of any algorithms which uses bounding boxes as output. More formally, IOU requires following metric to asses object detector's performance.

- The ground-truth bounding boxes.

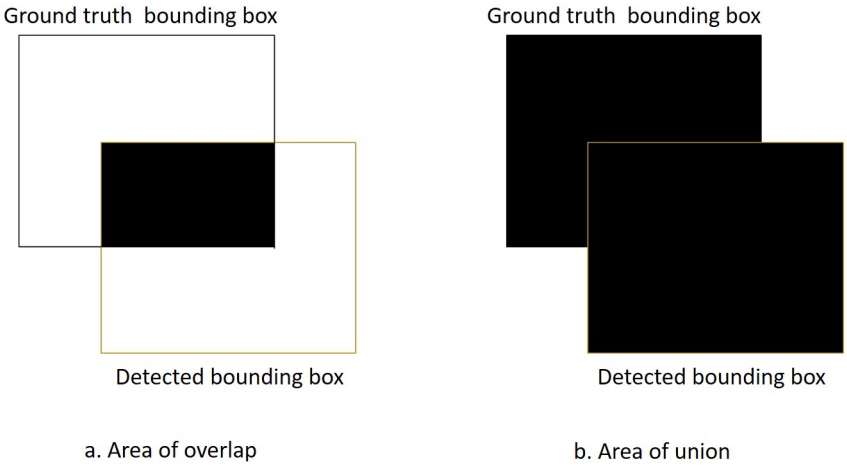


Figure 3.3: Visual example of intersection over union

- The bounding boxes predicted through model.

Figure 3.3 represents the IOU value calculation according to the ground truth and detected bounding boxes.

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{3.1}$$

3.3.2 Mean average precision (mAP)

The mAP is the metric to measure the accuracy of object detectors like Faster R-CNN, SSD, etc. It is the average of all the average precision (AP) calculated for all the classes.

$$mAP = \frac{1}{n}\sum_{i=0}^{n} AP(i) \tag{3.2}$$

n represents the number of classes in the dataset.

Average precision is then calculated by taking the area under the precision-recall curve. This is done by segmenting the recalls evenly into 11 parts for different IOU values between 0

to 1.

$$AP = \frac{1}{11} \sum_{0.0}^{1} Pr \tag{3.3}$$

3.3.3 Precision and recall

Precision measures how accurate your predictions are. i.e., the percentage of your positive predictions are correct.

Recall measures how good you find all the positives. For example, we can find 80 percent of the possible positive cases in our top K predictions.

$$Precesion = \frac{TP}{(TP+FP)} \tag{3.4}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{3.5}$$

TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

3.4 Proposed framework

Proposed architecture for action detection uses automated drone for surveillance. More formally, the proposed architecture develops a dataset, an action detection model to automate the drone surveillance for search and rescue. The proposed model is inspired by the single shot object detection technique for human detection and their action recognition. In this chapter a dataset is developed and it has rich amount f variation for drone surveillance automation. The architecture of proposed model for for action detection is shown in Figure 3.4. The proposed model use the feature of multiple levels in parallel,instead of only the last level convolution fea-

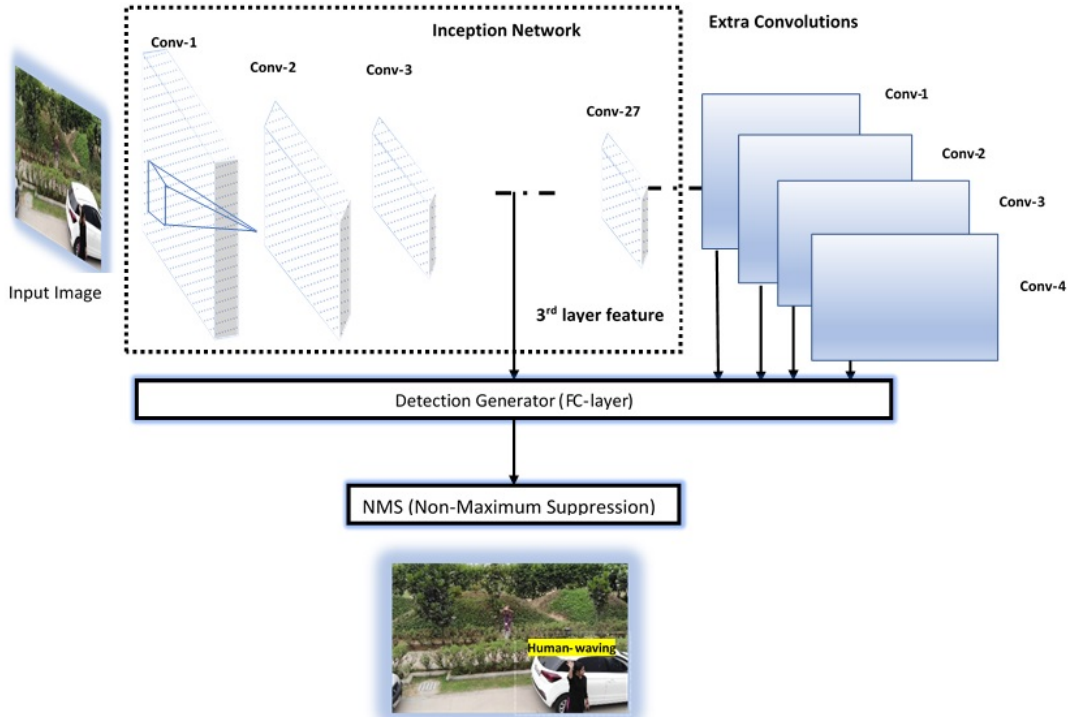


Figure 3.4: Architecture of the proposed action detection model

ture. Specially, the feature of the initial convolution layer contributes more to the classification of a small object. In the proposed action detection network, fully connected layer represents the dense layer of the network, which is used with softmax function for final feature classification. In the end, NMS is a function in computer vision that uses a threshold value and gives an output of single bounding box with input as multiple bounding boxes on the object. In the case of drone surveillance images were captured from the top, hence the object seems to be quite small, and its classification features less explored. The proposed model therefore uses the initial convolution layer feature, which enhances the outcome of the proposed model for drone surveillance action recognition.

In this study, have used the human detection and action detection approach to identify the place where help is required. For this we have developed a novel action detection model in drone surveillance videos. The proposed model for action detection is inspired by the pyramidal feature extraction [95] and utilization for classification and localization. As shown in Figure 3.4, the proposed model uses the feature of different convolution layer for the localization task. So, we have experimented with various convolution networks such as VGG16, Inception, and the feature of various layers for the localization. In the proposed architecture of the action detection model, a feature of the 3rd convolution layer is fed directly to the detection generation, which

Table 3.3: Details of hyper-parameters and their values in final trained model (Proposed model)

Hyper-parameters	Values
Number of classes	2 and 6
Activation	relu
Batch-normalization	Yes
IOU	0.5
Batch size	24
Optimizer	rmsprop
Momentum-optimizer value	0.9
Initial learning rate	0.004

is a key factor in performance improvement. Also, four extra convolution layer is used after inception network where the features are fed to the detection generation in a pyramidal way. The parameter value of these extra convolution networks is the same as the convolution layer in the Inception network. Details of hyper-parameters of the proposed model is given in Table 3.3. The proposed model is trained both dataset, with activation function as relu.

In our proposed approach, the drone will scan the disaster-affected area, and at the same time, our proposed model deployed on the drone will recognize the specific action as a help situation. Our proposed approach for SAR follows the intuition of human nature as they wave a hand in the aerial vehicle's direction as a symbol of asking for help. This specific gesture of waving hand has various actions such as human- standing and waving hand, human-laying and waving hand, human walking and running hand, human running, and waving hand. In other words, human detection and action recognition in a given drone image can be precisely done by identifying the feature at each layer of convolution. Its feature map is expected to consist of multiple structures and features of body parts and their pose, which corresponds to human action. The proposed implicit model exploits the feature of every CNN network in a hierarchical pyramidal way. NMS (Non-maximum suppression) is used to select the best bounding box based on its threshold score at the end of our trainable network. Our proposed model is applied to the detection of two-class and six-class actions. It is also applied to the publicly available Okutama dataset [9].

3.5 Experimental setup

Experiments were carried out on the NVIDIA DGX-1 V100 supercomputer, which has an FP64 computation power of 7.8 TFLOP/s. Initially, state-of-art object detection is applied to recognize human action. The models were trained and tested on the proposed dataset. In addition to this, experiments were performed on the publicly available Okutama dataset as well. Both datasets have frames that contain multiple people performing different actions simultaneously. As both datasets' image size is equal, i.e., 1920 * 1080 pixels, the various model results will give a good comparison between the actions performed in both the dataset. These models' results are compared based on the standard COCO performance metrics (mAP and IOU). The proposed dataset for two-class and six-class actions contains 6000 images. We split the dataset into training and testing, where 2000 images were filtered from 6000 frames to avoid redundancy. Out of 2000 images, 1800 images is for training, and the other 200 for testing. For evaluation of the trained model, randomly ten images were selected using random shuffle in each evaluation step. All three models were trained and tested in the TensorFlow environment on both datasets. The proposed model is tuned for various hyper-parameters such as batch size, step count, learning rate, and optimizer. The Faster R-CNN model uses the ResNet classification model, with the feature size 16, and the batch size varies from 32 to 8 with the learning rate in the range 0.003 to 0.002. Training starts with the initial learning rate as 0.003 for 11000 steps. The detail of all the hyper-parameters tuned and used for final training of model is listed in Table 3.3.

3.6 Results and analysis

Table 5.5 shows the performance of deep learning object detection models applied on the publicly available Okutama dataset. The performance is evaluated on a standard COCO evaluation metric (mAP). Our result shows that faster R-CNN is performing comparatively better on this dataset. In addition to this, Table 5.4 shows the results of models applied to the proposed dataset. Table 5.5 and Table 5.4 compares the result obtained with the proposed six class action dataset. The six-class dataset is developed for general surveillance applications. For the next set of experiments intended for SAR, the two class action dataset is used. Table 3.6 shows the

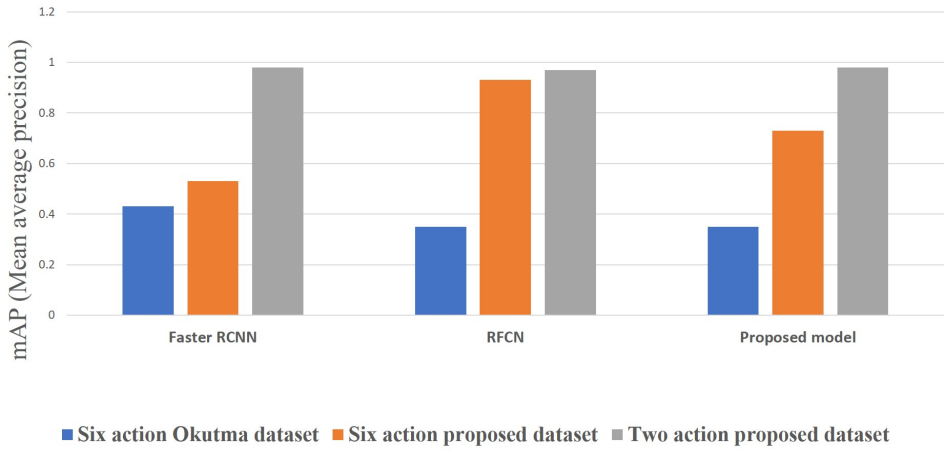


Figure 3.5: Comparison of deep learning models for action recognition in aerial images

result of deep learning models applied in a two-class action dataset, which is helpful for SAR applications. The experimental result of this chapter is summarized in Table 5.5, Table 5.4, and Table 3.6, and we have analysed the result in both ways ie. qualitative qualitative. Also, the qualitative analysis of the result is done based on the evaluation parameters mentioned here.

3.6.1 Qualitative analysis

The visual results of applying DL models on the proposed dataset is shown in Figure 3.2. In addition to this, models were used to publicly available Okutama dataset for 6-action. Figure 3.1 Shows the distribution of classes and samples available in our dataset. It shows that sufficient examples are available for models to learn the feature accurately for classification in our dataset. Figure 3.1 also shows the visuals of our dataset, where each action is shown; however, due to the presence of multiple activities in the same scene, the sub-images have captions that include various actions. In addition to this, Figure 3.2 shows the annotation of each action inside the image frame. In this, sub-images include captions, and inside that sub-image, each bounding boxes of the different classes have a different color. Our main objective is to identify the waving hand shown by the yellow color bounding box among all the classes. Other implementation results for models applied to this dataset are discussed and analyzed in the next section.

3.6.2 Quantitative analysis

Quantitative evaluation methods for action recognition depend on the approach we have followed. As we have applied the multi-human action recognition model, the standard evaluation metrics are average precision and mean average precision with different IOU values. COCO standard performance measuring parameter is mAP (mean average precision), which is the average of all the classes' best recalls. In this evaluation method, the detected bounding box is considered valid if it matches a corresponding ground truth bounding box. There should not be any other bounding box with the same ground truth. The second metric is used to evaluate the performance of modes average precision (AP).

Table 5.5 compares the performance of deep learning models applied to six class actions of the Okutama dataset. From these results, comparatively, SSD performs better, as the mAP value is relatively equivalent, while the inference time of SSD is very less than the other two models. Table 5.4 shows the result of deep learning models applied to six class actions of our proposed dataset. It shows the validity of our proposed dataset for the real-time application. The mAP value for all three models is higher than 90 percent. It can also be concluded from this that SSD is favorable, and our dataset is captured from a good enough height from where features are visible to be identified. Table 3.6 represents the experimental result of the selected models applied to our two-class action dataset. As mentioned in the dataset development section, this experiment is performed for SAR applications, where we have to identify the waving hand. The experimental result shows our approach's validity that waving had can be identified using these object detection models. Our trained model can be used for classifying the action as waving and non-waving. Among all those models that we have used here, SSD looks better as it gives comparatively equivalent with less inference time. We have also presented the result using the comparison graph represented in Figure 3.5. The graph shows the effect of model performance in all three datasets, including the publicly available Okutama dataset. It shows that the models perform better on our proposed dataset as the selected actions were different enough to be discriminated by the deep learning models. Also, the height from where images were captured is another factor of this performance.

Table 3.4: Performance of object detection models for action detection for six-classes of Okutma dataset

Models	22000 Steps mAP @ 0.50 IOU	200000 Steps mAP @ 0.50 IOU	500000 Steps mAP @0.50 IOU
Faster R-CNN	0.43	0.38	0.35
R-FCN	0.35	0.32	0.20
Proposed model	0.20	0.35	0.32

Table 3.5: Performance of deep learning models for action detection on proposed six-class aerial action dataset

Models	mAP @ 0.50 IOU
R-FCN	0.93
Faster R-CNN	0.53
Proposed model	0.73

Table 3.6: Performance of deep learning models for action detection on proposed two-class aerial action dataset

Models	mAP @ 0.50 IOU
Faster R-CNN	0.988
R-FCN	0.97
Proposed model	0.98

3.7 Conclusion

In this chapter, we have proposed a drone dataset for human action recognition. This dataset can also be used for human detection in drone surveillance applications. The proposed dataset has a rich amount of variety in color, height, actor, and background. Besides, our primary objective is to provide support for SAR using drone surveillance. We have presented an experimental comparison of the deep-learning action detection model applied on the proposed dataset and another publicly available dataset. This chapter also offers a novel detection model for action recognition. It achieves a 7% higher mAP value than the previously proposed SSD model when applied to a publicly available Okutama dataset. When the proposed model is used on our

two-class action detection dataset for SAR, it achieves a 0.98mAP value, which is a decent performance for real-time application.

Chapter 4

Video Classification Based Approach for SAR in Disaster

In this chapter, a complex scene classification problem is solved, which can be utilized for drone-based emergency situation classification in a natural disaster. The proposed model uses spatial and temporal features of the video to classify the scene as help or non-help in the natural disaster. Due to drone surveillance dataset's unavailability, the AI researchers have least explored the area of aerial and drone surveillance in the past. Therefore, it is essential to develop a scene classification dataset for search and rescue. This chapter proposes a dataset, and it is the first and unique dataset for scene classification using drone surveillance.

4.1 Introduction

In recent years a rising number of natural disasters hit several regions worldwide, causing thousands of human lives in danger. However, all government has some disaster management strategy, and by using them, they can save some lives. Out of all the disaster management strategies,

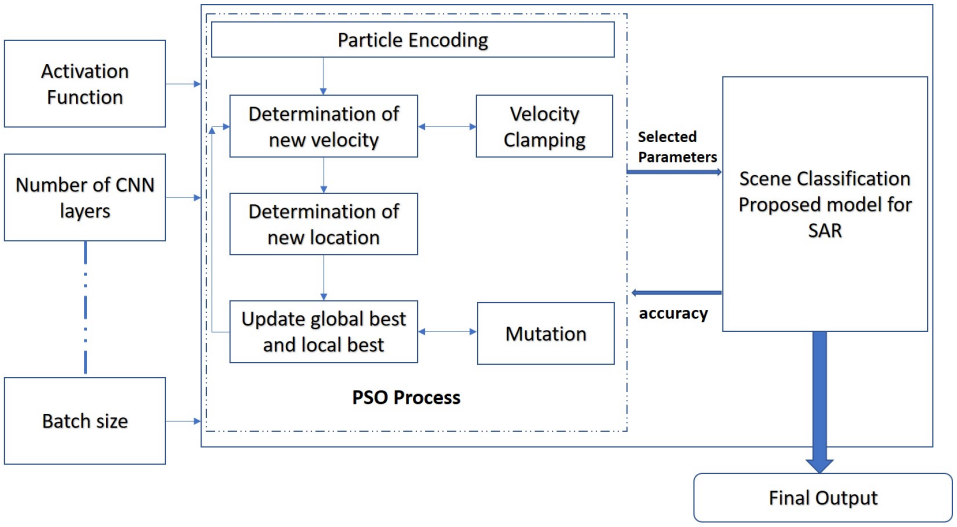


Figure 4.1: PSO based hyper-parameter tuning process

search and rescue (SAR) is a post-disaster operation to find the people and send them to a safer place. Till now, most of the SAR are volunteer intensive and cause human life in danger to find the people for whom rescue is required. Also, this is a time taking process and requires a faster SAR technique. The drone-based solution is suitable and versatile enough for outdoor disasters. The recent Flood of India and Nepal caused a massive loss of human lives and property. The loss of human lives is one of the prominent concerns here, and we focused on reducing it by providing the efficient search and rescue technique. This chapter aims to offer risk-free autonomous search and rescue at a lower cost, accessible for both public and private. The architecture and technology required for search and rescue vary according to the geographical area, type of disaster, and resource availability; for example, technique and solution will be different for in-house (Fire in the company of house) or outdoor disaster (earthquake, Flood). Quick recovery from disaster and sending help to the disaster victims is still a challenge and requires more attention. However, in the past, multiple methods were proposed, like Mini robots for indoor rescue; for example, [121] suggests a SAR technique based on ground and aviation robots. Recently, various SAR attempts were made by using these robots and drones to find the humans. However, all those approaches involve manual surveillance through drones. Figure.4.2 shows the block diagram of an automated surveillance system for SAR. By looking at the previous attempts to rescue humans, it can be more productive using the drone surveillance approach in Figure.4.2.

Drone-based search and rescue are valuable techniques, and similar operations have been

carried out using manual drone control in the past. Currently available drones for military and business applications are smart enough for SAR or other critical applications. The advanced deep learning technique can classify the action as a help or non-help situation for automation. Also, since we are working with the videos and 3D-models for feature selection and classification, the number of parameters in the network is comparatively high. It requires an automatic technique for the optimization of hyper-parameters. Figure 4.3 shows hybridization of optimization technique using PSO with the deep-learning model to solve the problem of help situation identification. PSO is a heuristic search technique that tries to imitate the travels of the flock of birds aiming at finding food [73]. In this, the population of particles flies through a multi-dimensional search space where each particle possesses a position and velocity [114]. Here, both variables are changed to emulate the social-psychological tendency to impersonate other individuals' success in the population. Using PSO to optimize hyper-parameters tuning, especially for the 3D models where the number of parameters is higher than any other deep-learning models, can speed-up the training and quickly find the best parameters values. Using an accurate optimization technique for the multiple hyper-parameters is important and can reduce the time required for the experiment, which is relatively high for the deep learning models works with videos. PSO is an important optimization method and is successfully used in [180] for hyper-parameters tuning of convolution neural network. Also, a modified version of PSO is proposed in [168], which can be applied with any kind of initial population like we start with the parameter values in the convolution neural network. The proposed methods in [127], [142], [104] uses different type of nature-inspired algorithms to optimize the hyper-parameters of classification algorithm like SVM. These all the techniques can help train other deep-learning models, especially for their hyper-parameter tuning, which is one of the most tedious but important for training deep learning models.

The desired output for the model is to identify the scene that contains single or multiple humans waving their hands in the direction of the drone as a help situation. Waving hand in drone surveillance from the top angle is an activity where humans' feature is not completely visible in a single frame. Hence it is required to make a decision based on multiple consecutive frames taken from different angles. Our developed dataset includes video-clips that cover the area from various heights, angles, and situations. The main contribution of this chapter is as follows:

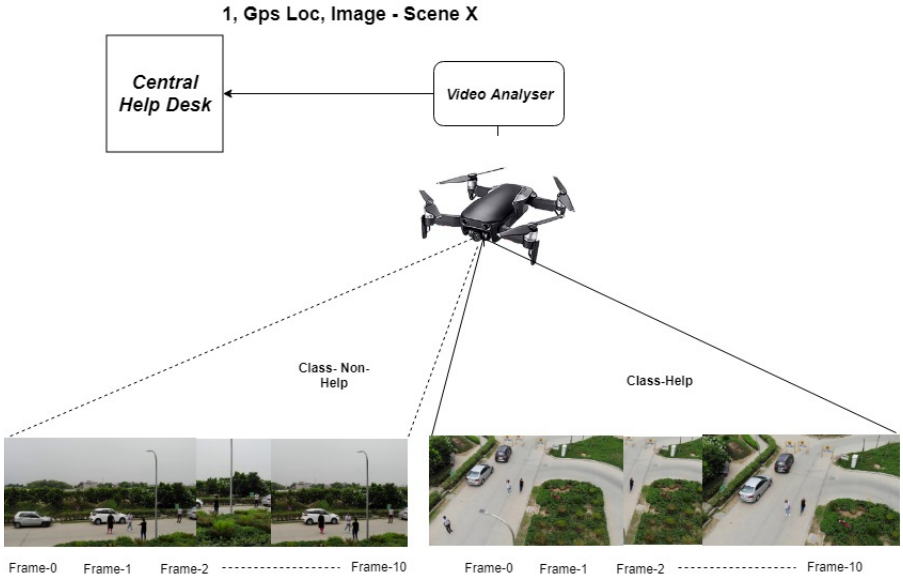


Figure 4.2: Automated support system for search operation in disaster

- Proposed a novel approach for SAR that uses the hybridization of PSO with 3D feature classification model (3D-CNN) for the training of model and its convergence.
- To automate the search operation, this chapter proposes a 3D-CNN model that uses spatial and temporal features of video to classify a situation as help-situation.
- In this chapter, we have introduced a unique video dataset having approximately 1000 clips in each class to classify the situations help-situation in the natural disaster. The proposed dataset is the first dataset captured through drones from aerial view for SAR. It applies to various other applications as well.
- This chapter also proposes video-augmentation techniques used here to increase the variety in our dataset.

4.2 Proposed framework

The proposed approach for help-situation identification consists of several steps: dataset collection, pre-processing, model development, training, testing, and hyper-parameter tuning. The proposed idea for help-situation recognition identifies a person with a waving hand. Waving a hand is a type of action that reflects a person requesting something, and it is helpful in the

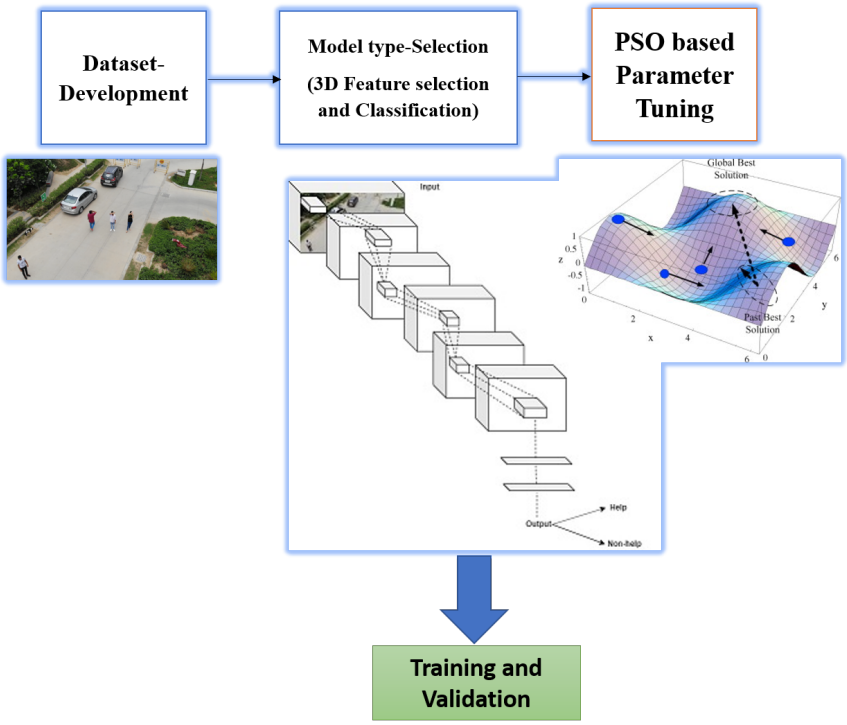


Figure 4.3: Flow of the proposed work for help situation identification

event of a disaster. So, the proposed drone surveillance scene classification dataset contains video clips in two classes: Help and Non-Help. In Help class, single or multiple humans were waving hand, while non-help class contains videos with actions other than waving hand. This chapter also proposes a novel hybrid approach for training the deep-learning model. The proposed model's hyper-parameters are tuned based on the PSO's recommendation for training. The PSO module shortens the time to find optimized global hyper-parameter values. Figure 4.1 and Figure 4.3 display two separate block diagrams of the proposed approach where, the first one explains the internal working of PSO based training while the other one describes the overall training process of the proposed architecture. The step-wise explanation of each phase of the proposed approach is as follows:

4.2.1 Dataset development

This portion provides a detailed description of the data set development process. Steps including selection and collection of action in the proposed dataset is explained here. Finally, we give a summary of the dataset.

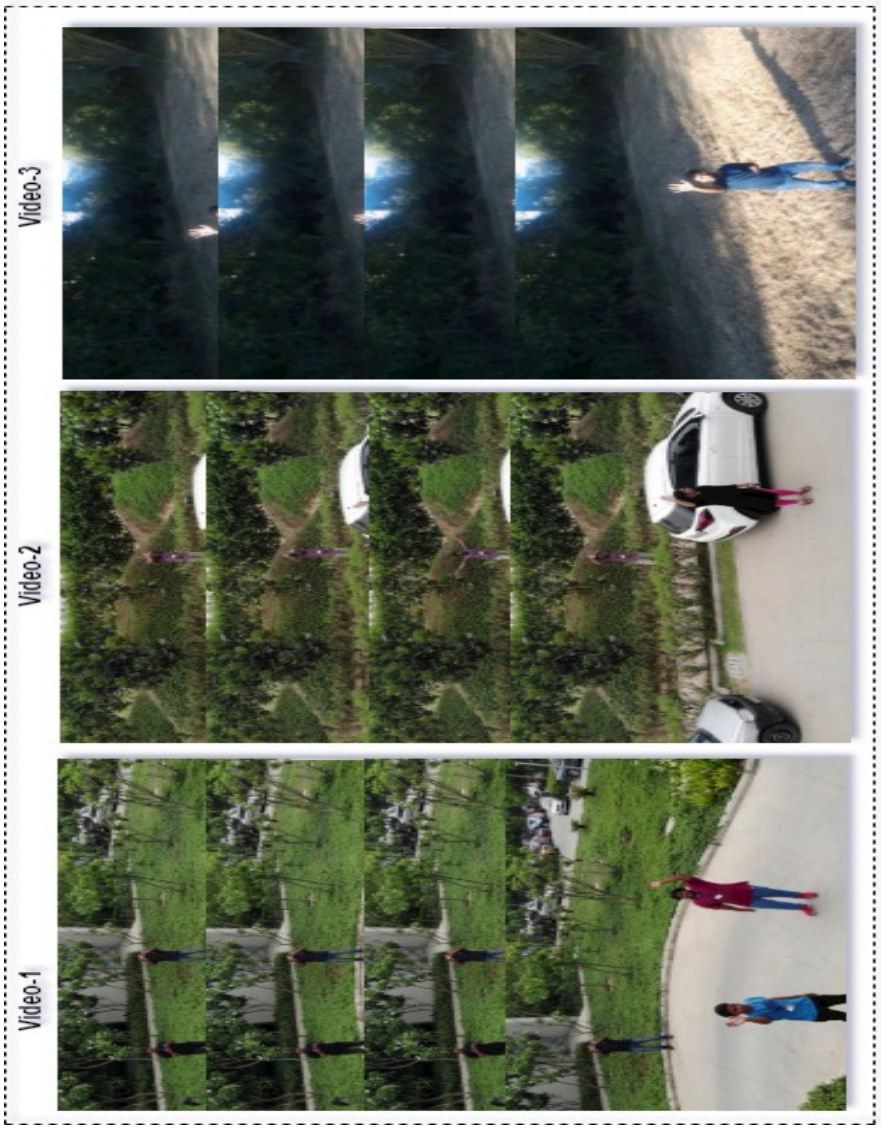


Figure 4.4: Sample videos shown in terms of few frames in the class representing help-situation

Figure 4.5: Sample videos shown in terms of few frames in the class representing help situation

4.2.1.1 Action selection

Due to the challenge of moving a camera, adjusting the landscape and changing drone height, identification of help scenarios by drone cameras can be more difficult. It has been assumed that if a human is asking for help then he or she will wave a hand in the drone's direction. Following variation of human hand gesture is considered to create a generalized dataset: standing and waving hand, standing and waving both hands, walking and waving hand, walking and waving both hands, running and waving hand, running and waving both hands.

4.2.1.2 Dataset collection

The data is gathered in the form of a video. It is recorded from different heights through drone fitted with a GoPro Hero 4 Black camera with a HD lens (5.4mm, 15MP, IR CUT) and a 3-axis solo gimbal. The videos were recorded with a 60 frame per second rate. Actors were picked randomly and they ate not professionals, hence, the videos were recorded with the natural actions and has sufficient variety. Here, we have labelled the video as a help-situation where at least one person is waving a single hand in the drone's direction. At the time of dataset collection, sometimes wind affected the drone and changed the height. Other than that, we have captured the video from height ranges between 10 to 40 meters.

4.2.1.3 Variation in dataset

To give the variety in the dataset, we have captured it on different days of the month, with actors. In addition, the videos were captured on different day of month, different daytime of a day,with different actors. Therefore, our dataset has a rich variety in terms of the camera motion, height, actors, action performing style, and clothing style.

4.2.1.4 Dataset summary

Video recorded for the total duration of 150 minutes from a varying height. After cropping, we have selected 100 video-clip in each class. When choosing these clips, we took special care to ensure that they had the minimum overlap and thus contribute most to scene classification. Each of our 1-sec video-clip has 60 different frames to experiment with the varying number of

depth of frame.

4.2.2 Dataset pre-processing

Our dataset is captured on different days of the month using the drone, with the actors acting naturally. It is captured in the form of video and pre-processed later to convert it into the required format. So, the following operation has been applied as the pre-processing step for the dataset creation.

4.2.2.1 Video-clipping

The video is captured by a drone and has a long sequence of scene where the models would be confused by plenty of repetitions and other effects. Therefore, we have used an online video-editing tool, and cropped all the frames giving no value for the learning to the model. In addition, we clipped the rest of the videos into small-small clips. Thereafter, the initial experiments were performed to check the validity of the proposed dataset. Furthermore, we decided which videos should be included in the content dataset, as certain video clips overlap and will not lead to confuse the model learning.

4.2.2.2 Video-augmentation

We had about 100 videos in each class after we had finalized the videos in our dataset. Following that, we began our experiments with the proposed model. Based on the results of our investigations and the PSO module's recommendations for hyper-parameter tuning, we discovered that there is a need for more variety in our dataset. Therefore, we have devised various video-augmentation techniques and used the expend the volume of the proposed dataset. In total, ten different transformation techniques were used in our video augmentation step. The following are some important transformation used in our video-augmentation steps:

- **Horizontal-flipping**: Horizontal flipping is an operation like a mirror image. It is beneficial for the action recognition dataset. For example, if a person is there waving their left hand, he waves his right hand in the transformed video after flipping.

Table 4.1: Summary of the dataset

Feature	Values
Number of Video in Class Help	100
Number of Videos in Class Non-Help	98
Frame-rate	60 fps
Rsolution	1920*2000
Camera-Motion	Yes, Slow and steady
Height of Camera	Varying (In range 10 to 30 meter)
Weather-Condition	Normal, Day-time
Number of Video after Preprocessing	1000

- **Down-sampling**: Down-sampling is an operation of video which samples the frame at a lower rate by applying the low-pass filter. In this, it reduces the rate-of frame per second by a factor M which can be thought of as a two-step process as follows:

  1. It passes every frame with a low-pass filter.

  2. It decimates the filtered frame by k, which is kept only for every kth sample.

- **Up-sampling**: Up-sampling is the process of inserting zero-valued frames between original frames to increase the sampling rate. It is also called zero-stuffing.

- **Super-pixel**: The term "super-pixel" refers to an image patch that is more aligned with intensity edges than a rectangular patch. The segmentation algorithms can extract the super-pixels of an input frame.

- **Gaussian Blur**: Gaussian blur in image processing is also known as Gaussian smoothing. The Gaussian function here minimizes the image noise.

4.2.3 Model development

A deep-learning model is suggested for feature collection and classification to create a novel search and rescue technique. The proposed model contains several layers of 3D-convolution

Table 4.2: PSO-tuned hyper-parameters for training proposed scene classification model for SAR

Particle	**Search space**
Activation Function	sigmoid, tanh, relu, leaky-relu
Padding	same, valid
Pooling	max-pooling, average-pooling
Optimizer	rmsprop, adam, sgd
Number-of-convolution-layers	2-100
Dataset-size	100-1000 videos

for feature extraction with a dense layer for the final classification of the extracted spatial and temporal features. Details of model is given in Table.4.3. Our experiment, on the other hand, started with a simple image classification convolution neural network. However, we improved our learning over time, and developed a scene classification model with the PSO based hyper-parameter tuning.

4.2.4 Hyper-parameter tuning

Hyper-parameter tuning is another critical step in the deep-learning model training. Deep-learning models consist of multiple hyper-parameter and need to be tuned for the best out-come.The particle and its significance as a swarm across its high dimensional feature space must be optimized and according to the step-wise learning, pooling, padding, learning rate, optimizer, and loss function are used here as hyper-parameter. The PSO based optimization technique works quite well and provides a small set of values to be tested with the proposed 3DCNN model. Table.4.2 shows the details of parameters used for automated tuning through PSO.

4.3 Implementation details

This section elaborates on the process of software implementation for the proposed framework to identify help-situation. The vital step of developing deep-learning models such as training and testing with its environments is discussed here.

4.3.1 Experimental set-up

The drone is utilized to collect the dataset in this experiment, and it is created according to the requirements, using an HP workstation with 6GB RAM and a 4GB dedicated NVIDIA GPU. This workstation was used for pre- and post-processing tasks such as video-cutting, video-augmentation, and so on. However, significant testing was carried out on the NVIDIA DGX-1 V100 supercomputer, which has a processing speed of 7.8 TFLOP/s. The supercomputer's vast processing capacity has aided our experiments, giving us plenty of time to run the experiment with multiple hyper-parameters. To set up our experiment, we'll need a large enough number of drone photographs to train the models, therefore we'll create a dataset. The dataset is described in depth in the preceding section.

4.3.2 Training

We have started experimenting with a small random learning rate with three different optimizers (RMSprop, SGD, and Adam). We have identified that the rmsprop optimizer has achieved significant performance with the minimum over-fitting problem even with less amount of data. The proposed model can learn the motion features, which is one of the deciding feature for action recognition and scene classification. Waving hand in the drone's direction is assumed as a prominent action. The proposed model has been trained on the developed scene classification dataset with the different frame patches. We have experimented with 10, 15, 20, 25, and 30 frame patch.

4.3.3 Testing

To train the model and test the performance, dataset is divided into 7:3 fractions, and the testing samples were drawn at random. The trained model was saved and used for prediction. The proposed model uses parameters such as validation loss and validation accuracy for evaluation.

4.3.4 Evaluation metric

We choose the standard criteria such as loss and precision for the proposed model's assessment, determined for both training and validation. Further, the models performance is compared based on the following performance evaluation metrics:

- Training accuracy, Validation accuracy, Training loss, Validation loss

4.4 Experiments summary

This section gives details of experiments and the result obtained. Overall, our experiments is divided into four different section based on the learning and models performance. The model is tested with various hyper-parameters with the initial dataset. Then, the model is applied to the proposed dataset. However, with the learning process through the experiment, we have used the PSO based hyper-parameter optimization and got the best parameter value of the proposed model. The last two experiments show the result of our proposed model with the final proposed dataset.

4.4.1 Experiment with 100 videos in each class (Experiment level-1)

At the initial stage of the experiment, we had 100 video clips in each class. In the first stage of the experiment with the proposed approach, our model consists of 9 convolution layers,

Table 4.3: Layered configuration of proposed model

Feature	**Specifications**
Conv3D	148, 148,10,32
Padding	same
Conv3D	148, 148,10,32
Padding	same
Pooling	3*3*3
Dropout	0.333
Conv3D	148, 148,10,32
Padding	same
Pooling	3*3*3
Dropout	0.333
Conv3D	148, 148,10,32
Padding	same
Pooling	3*3*3
Dropout	0.333
Conv3D	148, 148,10,32
Padding	same
Pooling	3*3*3
Dropout	0.333
Flattenn	....
Dense layer	512
Batch-Normalization	....
Dropout	0.333
Dense-Layer	....
Loss	Categorical-cross-entropy
Optimizer	Rmsprop
Metrics	Accuracy

followed by two dense layers. The last dense layer works with softmax activation, while all other layer's work with a relu activation function. Figure 4.6 and Figure 4.7 shows the result

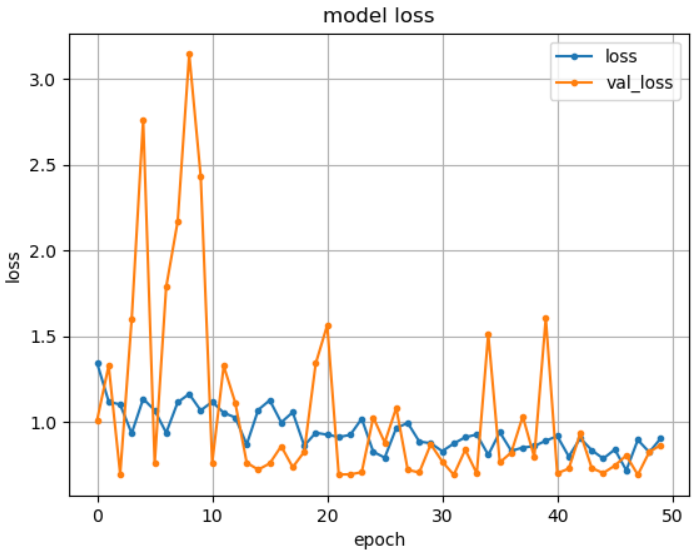


Figure 4.6: Loss comparison of training and validation for the proposed model

of proposed model with the initial dataset. It indicates an idea of over-fitting problem in the trained model, and required an adequate solution for this. In the next experiment we have tried to resolve this problem and uses different probable solution for this.

4.4.2 Effect of dropout and batch normalization (Experiment level-2)

In this experiment, we applied the dropout and batch normalization at the alternate layers of convolution. Although, the training of the network is same as in Experiment level 1, but the dropout function would make the network lighter and quicker. For this experiment, the dropout value was 0.33 at each layer. Overfitting is a major issue with neural networks, on the other hand, batch normalisation and dropout are two possible solution for this. Figure 4.8 and Figure 4.9 show the result of this experiment. We can see that the problem persists, so the previous experiments indicates a failure because the results still shows the problem of overfitting. In the following experiment, we tried various frames together instead of fixed tn frames.

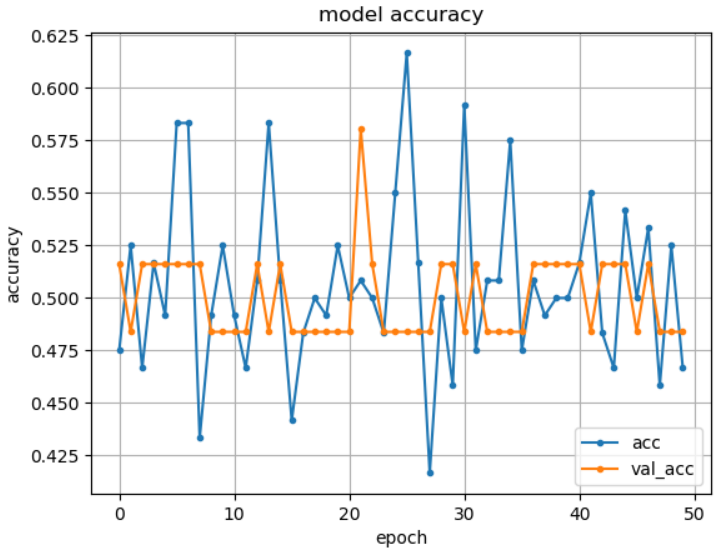


Figure 4.7: Loss comparison of training and validation for proposed model

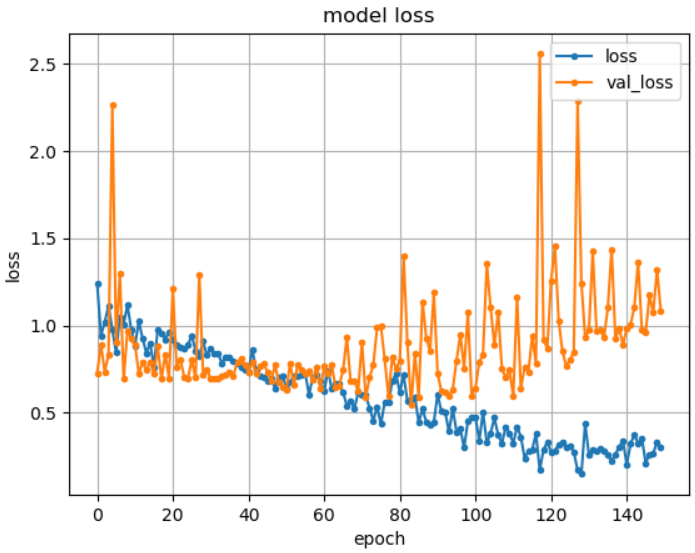


Figure 4.8: Loss comparison of training and validation for proposed model with batch norms

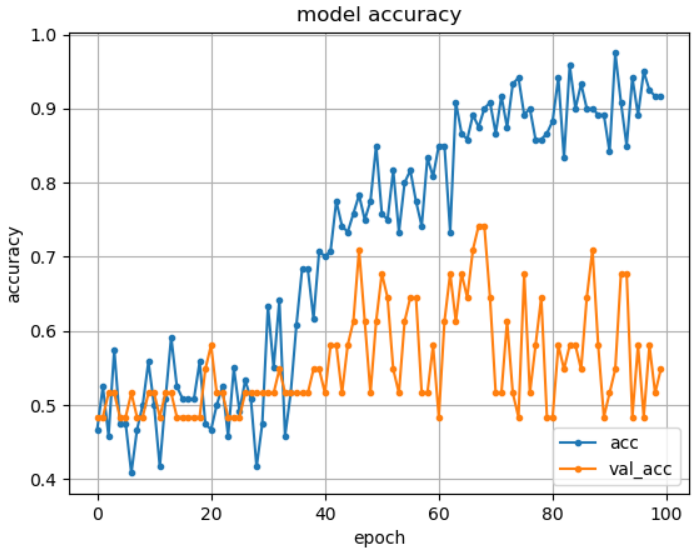


Figure 4.9: Accuracy comparison of training and validation of proposed model with depth of 20 frames and bath size 4 with batch norms

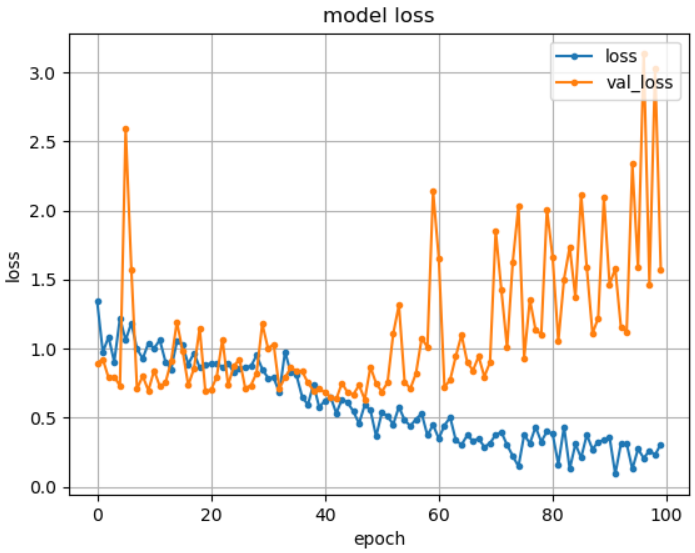


Figure 4.10: Loss comparison of training and validation of proposed model with depth of 20 frames and bath size 32

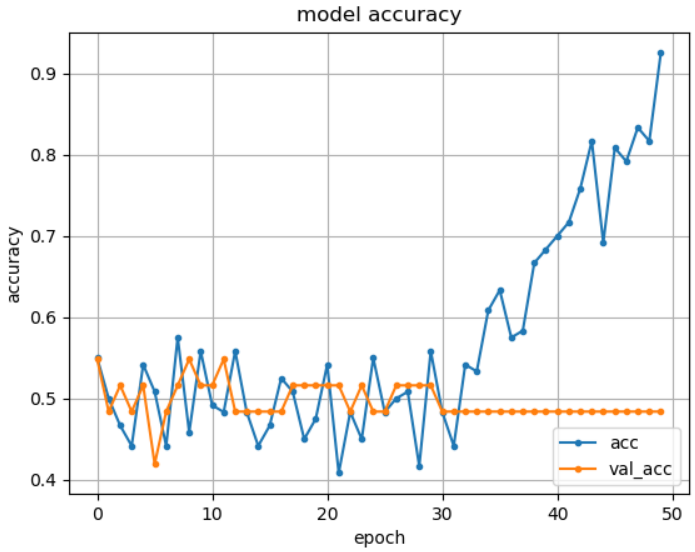


Figure 4.11: Accuracy comparison of training and validation of proposed model with depth of 30 frames and bath size 32

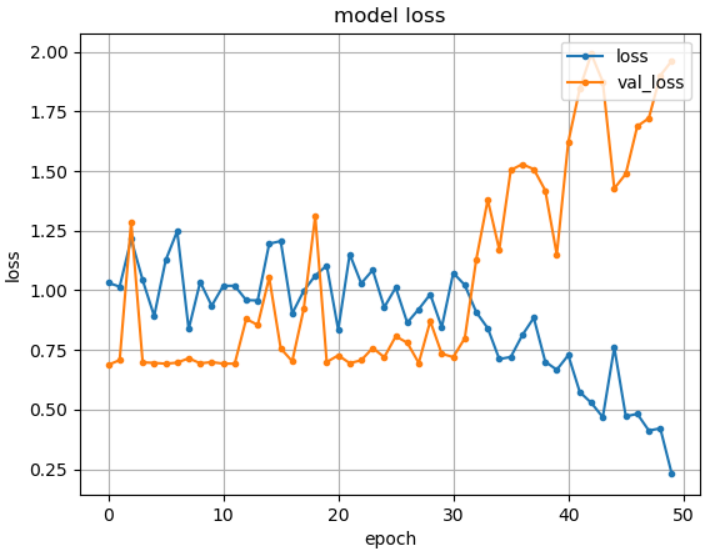


Figure 4.12: Loss comparison of training and validation of proposed model with depth of 30 frames and bath size 4

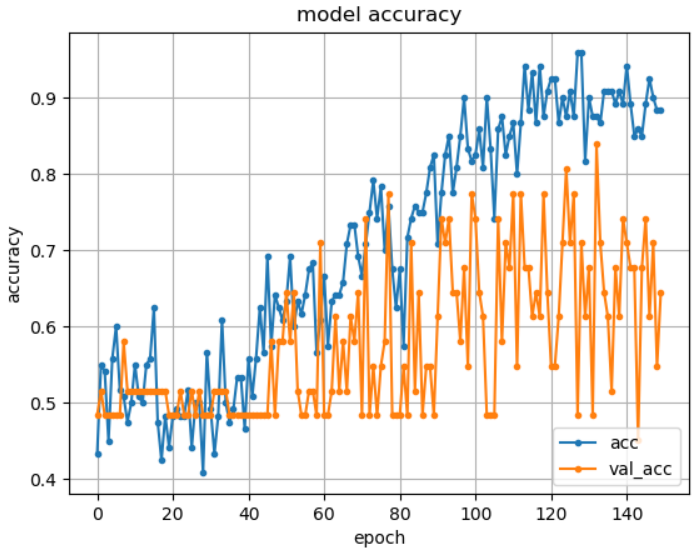


Figure 4.13: Loss comparison of training and validation for proposed model with depth of 30 frames and bath size 4

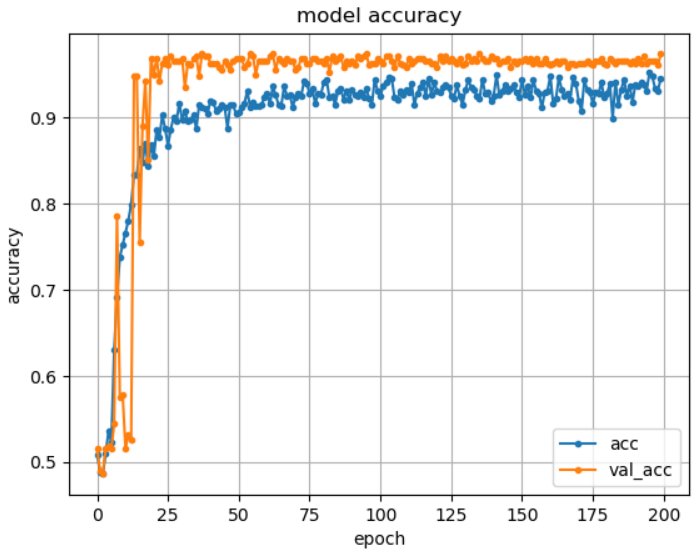


Figure 4.14: Loss comparison of training and validation for proposed model after data-augmentation

Figure 4.15: Loss comparison of training and validation for the proposed model after data-augmentation

4.4.3 Effect of depth of frame in video (Experiment level-3)

In this experiment, firstly, we varied the number of frames in each training video-clips and the number of frames in each training video-clip was 10. Thinking about avoiding the over-fitting, we have increased the size of the dataset.In this phase. our experiment started with the depth of frame as 10, with the learning we have moved to 15, 20, 25, and 30 frames as well. Figure 4.6, and Figure 4.7 shows the experimental result with the depth of frame as 15, while, Figure 4.8 and Figure 4.9 represents the result of 20 frames as depth. Figure 4.10, Figure 4.11, Figure 4.12, and Figure 4.13 shows the result with varying number of depth of frame and batch size. In these experiments, we have found a way to optimize these parameters and meanwhile, we have applied the PSO optimization technique for hyper-parameter tuning.

4.4.4 Experiments with final proposed dataset (Experiment level-4)

We have increased the dataset size after all the experiments in previous section which is another way to prevent the overfitting issue. The proposed dataset contains variety in terms of different actors, their acts, and their clothing. We have used the image augmentation technique and

Table 4.4: PSO-tuned hyper-parameters final value

Particle	**Final-output**
Activation Function	relu
Padding	same
Pooling	max-pooling
Optimizer	adam
Number-of-convolution-layers	5
Batch-size	8 videos

transformed this to be used for video augmentation. The dataset augmentation section contains more information on our augmentation. We have tuned the parameters and find the best value of parameters through a PSO based optimization, and the experimental result is shown in Figure 4.14 and Figure 4.16. Besides, Table.4.4 shows the global optimized value of tuned parameter through PSO, and by using the optimized hyper parameter values and augmented dataset, the proposed 3D-CNN scene classification model is trained.

4.5 Result and analysis

In this section, we have attempted to analyse the result obtained from all the experiments performed to validate the proposed approach for help situation identification. This section also displays the future direction of research for using technology and, in other words, the use of AI techniques for search and rescue.

The goal of all of the experiments shown in the preceding section is to identify a situation in which help is required by using computer vision and AI. Our experiment-1 shows the initial result of video classification in drone surveillance. It indicates the need for improvement in all aspect. so, we have started improving the dataset with the different techniques shown in the data-augmentation section. We have also tried different pre-trained models, however, all these models were developed and trained for the ground level human features. Further, we started experiment with the proposed 3DCNN model. In Experiment-2, we have tried various techniques to avoid over-fitting problems, with a similar network and dataset used in the last

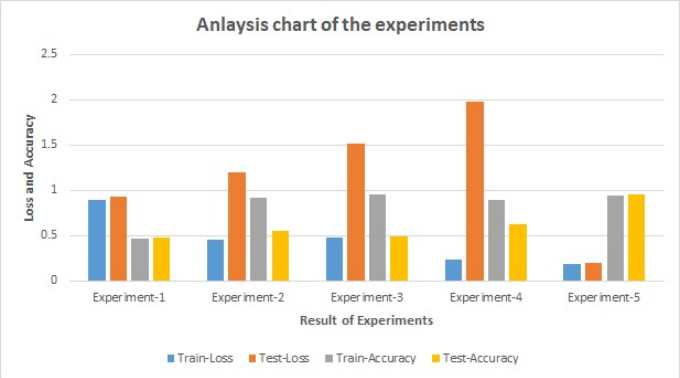


Figure 4.16: Analysis of our experiments

experiment. The result of these experiment shows that even with the most reliable network, all the techniques to avoid over-fitting technique can't help if we don't have enough variation in the dataset. Clearly, this experiment indicates the failure of this model trained on the initial dataset. We have tried the technique to avoid overfitting such as dropout, batch normalization, and reducing the number of layers, and it could not help. At this stage, we've done enough testing with various combinations of 3D convolution layers, parameter values, and techniques to prevent over-fitting. Our analysis for help-situation identification is as follows: either we don't have enough features to learn from this angle of surveillance or need more data in terms of various actions, background and actors. This point had opened the direction of our research in two different directions, i.e., change the direction of the camera and collect more data that should be more generic.

With the help of the results and analysis of the previous two experiments, we began experimenting with the other direction of the research. We started varying the number of frames in each video for training, validation, and inference. In this way, we increased the size of data and tried various combinations of the video's depth such as 10, 15, 20, 25, and 30. Experimenting with these varied numbers of frames for scene classification gives the freedom of using the number of frames at the time of inference, which tends to have more chance of recognizing particular action in a more extensive sequence. However, the smaller video will give faster processing at the time of training and inference. In this section, as we have used the initial data only, which had approximately 100 videos in each class. We can observe that still, the over-fitting problem exists. With the assistance of our previous experiment, and learning from it, the proposed 3D-based model has higher probability to achieve adequate performance as it uses both spatial as well as temporal features. However, the experimental result is not favourable,

and in some cases, accuracy was very low, and in the other two cases, our model was over-fit. We have tried two more techniques at this stage, i.e., one is optimizing the hyper-parameters of the network with the reliable PSO strategy and gets the final global optimized value, as shown in Table 4.4. With these optimized values, the proposed network is applied for training on the developed dataset after preprocessing(augmented). Final result in terms of accuracy (training and validation) and loss (training, validation) is shown in Figure 4.14 and Figure 4.16.

4.6 Conclusion

In this chapter, we introduced a novel idea for SAR based on a scene or video classification. The proposed hybrid approach of using the proposed 3D-CN model of deep-learning in association with PSO for parameter-tuning is a novel and fast technique for scene classification. The proposed architecture is helpful for scene classification in real-time rescue. For this, we have developed a dataset consisting of approximately 1000 video-clips in each class. The proposed model for scene classification is applied and tuned with the help of PSO on our developed dataset, and it gives a valuable real-time performance with 98% training and 99% validation accuracy.

Chapter 5

Emergency Text Classification based Approach for SAR in Disaster

This chapter proposes an approach to identify the exact places where rescue is needed using emergency text classification in drone surveillance. To determine the locations where rescue is required, we have assumed that people are writing emergency text such as SOS, HELP, and EMERGENCY on the ground, wall, or sand. With the availability of adequate datasets, deep learning algorithms can extract such text. However, no labelled dataset for emergency text categorization in aerial surveillance exists to our knowledge. Therefore, we first developed an image dataset for emergency text classification. The developed dataset contains more than 6000 images captured from varying height, angle, daytime and labeled them into Help and Non-Help classes. Furthermore, we have developed a lightweight CNN model with ten times less trainable parameters than previously proposed classification models like VGG16 and InceptionResneV2.

5.1 Introduction

Natural disasters such as the earthquakes in Nepal, Tohoku, and Haiyan, as well as floods in Europe and India, have demonstrated that local government officials and emergency services confront several problems in adequately managing a crisis [62]. The search for human survivors on the site is paramount for rescue services. Previous search and rescue (SAR) attempts are predominantly voluntary and, therefore, a time-consuming operation. The race to discover catastrophe victims as quickly as feasible is always a battle against the time. In contrast, automatic drone surveillance can save human lives in a catastrophe without taking the risk of a volunteer's life by quickly scanning the affected areas. Besides, the quality of automatic drone surveillance depends on the quality of the dataset and environmental hazards. In this thesis, previous chapters suggested some revolutionary techniques for reducing human lives' loss using action recognition and human detection in drone surveillance. Human detection and action recognition is beneficial when the people themselves can perform some action and visible at the time of drone surveillance.

In this research, a text-recognition-based approach is developed for the automation of drone surveillance. Text recognition has improved dramatically in recent years, largely due to the use of deep models[151], [23], [98],[97] and large datasets [174], [63]. Although many prominent 2D CNNs[22], [193], [172] have been designed for text classification and scene classification task, these 2D CNNs cannot directly utilized for aerial and drone surveillance. The inherent complexity of aerial and drone images in terms of features as per the height, angle, and size of the text makes it more challenging for recognition. Besides, the availability of appropriate data for training the network is an integral prerequisite of the deep learning algorithm. No such dataset was available in the literature for emergency text recognition. So, in this research, a dataset is developed to automate search operations in disaster. The generated dataset contains various images having the emergency text written on walls, roofs, and sand.

This chapter proposes a method for extracting emergency text from an input drone picture and classifying the picture as a Help or Non-Help situation. This chapter's primary contributions to the automation of drone monitoring for SAR are as follows: A. To our knowledge, this is the first paper to propose a convolution neural network for emergency scenario detection in

Table 5.1: Features of the proposed dataset

Class	**Image type**	**Background**	**No. of images before aug**	**Image size**	**No. of images after aug**
Help	HELP SOS EMERGENCY	Sea Hill Disaster	1000	150*150	3000
Non-Help	Random Images	Sea Hill Disaster	1000	150*150	3000

drone surveillance. B. Developed a dataset for the automation of text-based SAR using drone surveillance. C. Proposed a CNN model to recognize Help situations for the automation of SAR using drone surveillance.

5.2 Motivation

The main factors that motivate researchers for the development of drone-based surveillance applications are its capabilities, such as increasing payload and quick scanning of a wide area. A drone equipped with a camera and sensor can collect a large amount of information quickly. In addition to this, automated drone surveillance can benefit the various other surveillance activities such as traffic monitoring, suspicious activity recognition, and search operation in remote areas. Search and rescue is usually initiated by the army or through volunteers after the disaster happens. The purpose of search and rescue operations is to relocate catastrophe victims to a safer area, and the first step is to rapidly locate the victims. Therefore, a robust search and rescue technique with low response time can save more lives. By looking at the recent disasters and the loss caused by them, there is a need for an advanced SAR technique that can ease this process and reduce the disaster relief response time. This paper proposes a novel approach for searching for humans. The proposed system uses well-versed hardware with advanced AI algorithms for the automation of search operations.

The goal of text recognition is to take the image and identify the single word depicted

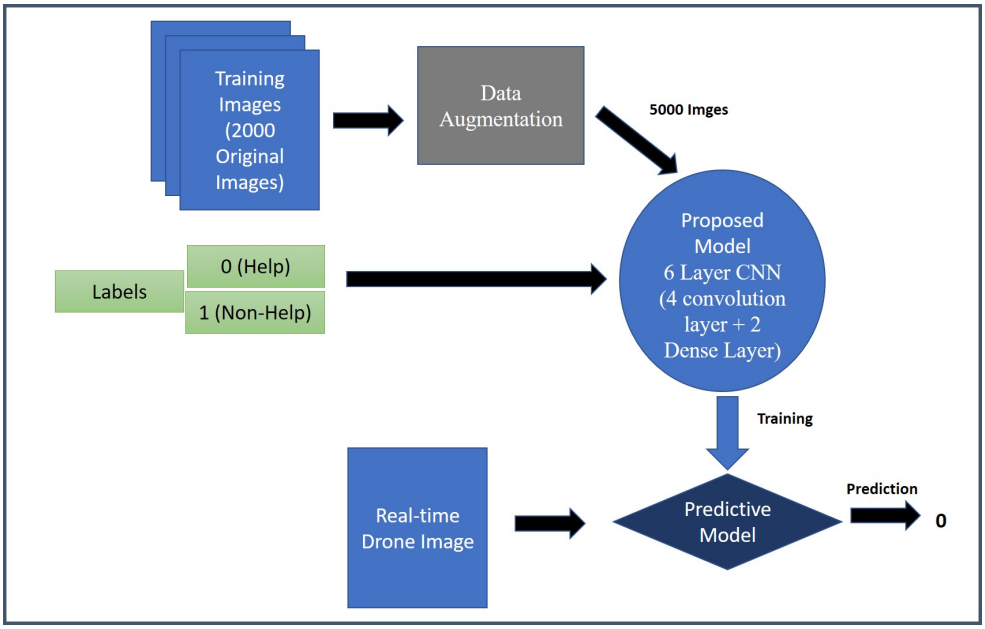
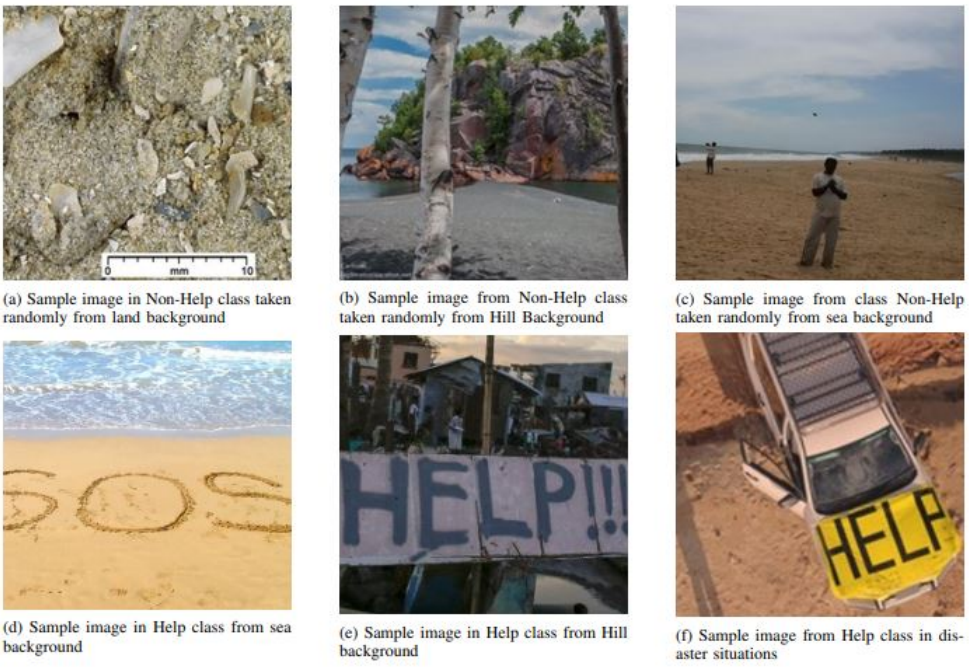


Figure 5.1: Flow-chart of proposed approach



(a) Sample image in Non-Help class taken randomly from land background

(b) Sample image from Non-Help class taken randomly from Hill Background

(c) Sample image from class Non-Help taken randomly from sea background

(d) Sample image in Help class from sea background

(e) Sample image in Help class from Hill background

(f) Sample image from Help class in disaster situations

Figure 5.2: Sample images in proposed dataset

inside it. There are few techniques available in literature of handwriting or historical document recognition[167, 49, 74, 7].

5.3 Methodology

This section explains the proposed approach for search and rescue, including the dataset development, preprocessing of the dataset, and the proposed model. We represent the image in the form of a function f(x,y), where at each coordinate, the value of function represents the pixel value of the image:

$$0 <= f(x,y) < \infty \tag{5.1}$$

5.3.1 Proposed system

The proposed system has three main phases, and the details of each phase is as follows:

5.3.1.1 Data collection and preprocessing

The first phase involves data collection from different sources, and then it is passed through the suitable image preprocessing techniques to make it appropriate for the proposed classification model. The dataset will be made available for public use on demand.

- Data collection: In automatic surveillance systems using a deep learning model, training and testing data is essential. It contains all the useful features on which the models have to be trained. Sample images of the proposed dataset is shown in Figure 6.3. The developed dataset has approximately 6000 images in two classes: Help and Non-Help. These images are collected from different backgrounds such as beach-side, hilly area, natural land and farming, and disasters such as earthquake and flood. Texts like "SOS", "HELP", "EMERGENCY" were written on walls, ground, or used a banner in each image of the Help class. However, the other class of our dataset, i.e., "Non-Help," contains a random

image from all those backgrounds considered in class "Help". The attributes of the images in the collection varied since they were acquired from various sources. As a result, we must resize, crop, enhance, and filter each image according to its specifications. The dataset's main details are listed in Table 5.1. This section also goes through the numerous additional preparation techniques that were used to increase the dataset's quality and quantity of photos.

- Data preprocessing: The images collected were in raw form, so we had to do a number of preprocessing procedures to make the dataset appropriate for training. Each image is first converted to a specified format (.jpg) before being fed directly into the deep learning model for training. Image translation and rotation are the most common preprocessing techniques used on our dataset. Preprocessing like translation of image is performed by shifting its pixel values by a vector dx and dy as given in equation 5.2 and equation 5.3.

$$x' = x + bx.y \tag{5.2}$$

$$y' = y + by.x \tag{5.3}$$

The rotation of an image is performed by an angle α, where each coordinates of the input image are shifted according to the formula given in equation 5.4 and equation 5.5.

$$x' = x.cos\alpha - y.sin\alpha \tag{5.4}$$

$$y' = x.sin\alpha + y.cos\alpha \tag{5.5}$$

Scaling, shrinking, and padding are the few image processing techniques required to transform all the images into single-sized image-datasets. An example of scaling is represented through equation 5.6.

$$f(x',y') = f(x,y) * p \tag{5.6}$$

Table 5.2: Layered configuration of proposed model

Layer	Output Shape	Filter size	Param
conv2d_11 (Conv2D)	148, 148, 32	3*3	896
conv2d_12 (Conv2D)	146, 146, 32	3*3	9248
max_pooling2d_6	73, 73, 32	3*3	0
dropout_5	73, 73, 32	....	0
conv2d_13	71, 71, 64	3*3	18496
conv2d_14	69, 69, 64	3*3	36928
max_pooling2d	34, 34, 64	3*3	0
dropout	34, 34, 64	...	0
flatten	73984	...	0
dense	256	...	18940160
dropout	256	...	0
dense	2	...	514

Table 5.3: Network architecture and their parameters

Network	Input image size	Optimizer	No. of Parameters
VGG16	150*150	sgd,adam	13,83,57,544
InceptionResnetV2	150*150	sgd,adam	5,43,36,736
Proposed model	150*150	sgd,adam	1,40,40,698

Where p is the factor by which the image is scaled.

5.3.1.2 Binary classification

After the collection and preprocessing of the dataset, the system is formulated as a binary classification problem. The labels of binary classes are Help as 0 and Non-Help as 1.

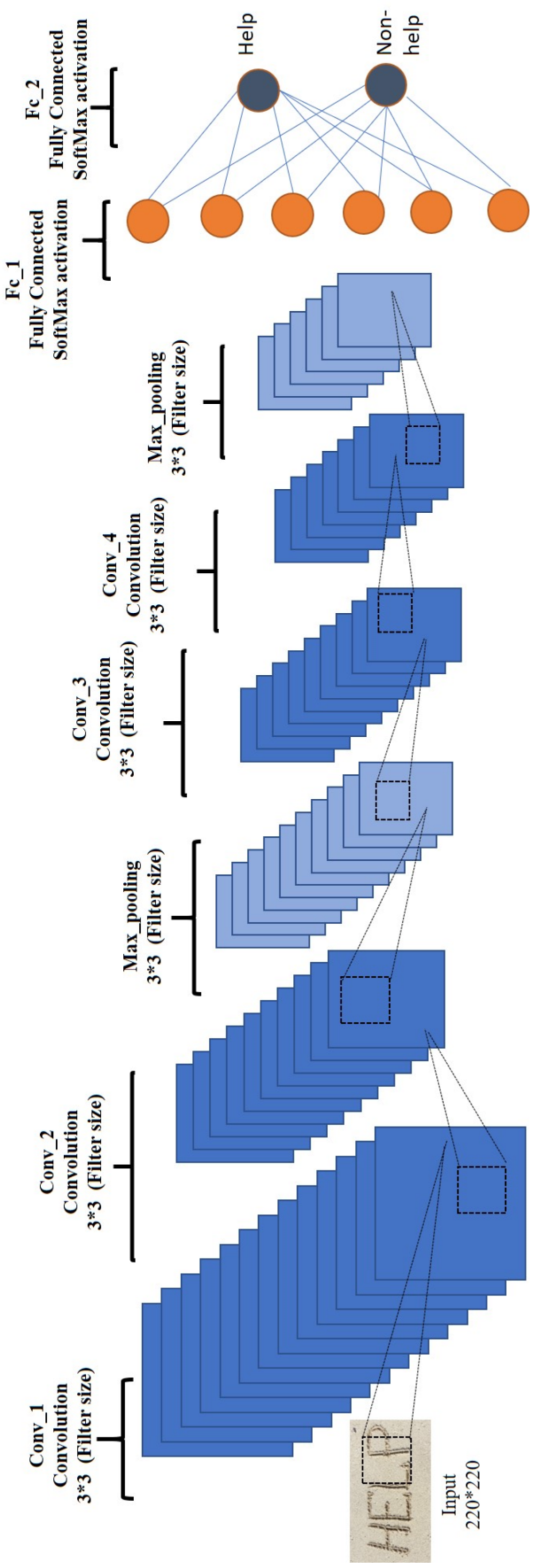


Figure 5.3: Architecture of proposed convolution neural network

$$L = -(z\log(p) + (1-z)\log(1-p)) \tag{5.7}$$

Where z_i = 0,1 represents the outcome of our designed classifier. In each iteration of the algorithm, it tries to minimize the loss shown in equation 5.7. The classifier's learning algorithm optimizes the parameters that can correctly predict the class *zi* for each input image $f(x,y)$ to learn from the input.

The convolution neural network is the best fit technique for an image-based problem in which a series of convolution operations is performed on each part of the image. Likewise, we have developed a binary classification model on the proposed dataset. The proposed model consists of multiple such convolution layers, including padding, pooling, and striding. Some important image and text classification networks are as follows:

- VGG16: This is 16 layers deep convolutional neural network containing 13 convolution layers and three dense layers[155]. In this network, initially, two convolutional layers have been used, followed by max-pooling and the three-pack of three-three convolution layers. In this, padding of size one is added after each convolution to avoid the spatial features. The Pretrained model of VGG16 is available online and is trained on 1000 classes of the Imagenet dataset. The first time proposed for large image dataset classification in 2014 became very popular for the image-net dataset.

- InceptionResnetV2: This is a convolutional neural network that has been pre-trained using the Imagenet dataset's 1000 classes. It is a deep convolutional neural network with 146 layers and a large feature representation. It works on the combination of the basic inception network with the residual connection. In residual connection, it combines multiple sized convolutional filters. It extracts the 1536 feature from the last fully connected dense layer after dropout for one input image. [166]. The residual connection here avoids the deep neural network's degradation problem, i.e., after a certain depth, the accuracy continues to decrease, which means more learning due to extra depth, and it creates further training error. Here, accuracy loss is not overfitting and is termed as a degradation problem in a deep neural network.

5.3.1.3 Proposed model

In this research work, a dataset is explicitly developed to train the emergency text classification. The proposed model is a convolution neural network-based binary classification model. The trained model is to be used to predict the input image between two different labels, Help and Non-Help. Figure5.2 represents the workflow of our research. When looking at the embedded devices' specifications integrated with a drone, a lightweight model is required. Therefore, our goal is to build a model that needs low computation at run-time and provides real-time accuracy of prediction for aerial emergency text recognition. The detail of the proposed model is as follows:

- Input image: For the proposed network's training, a dataset was created and fed into the network. Each image is resized into the 224*224 size before giving input to the network. The proposed network learns from each input image and the labels provided with it.
- Convolution: The proposed network uses four different convolution neural networks and is stacked in the form of a residual network of size 2. At each convolution layer, a 3*3 filter is applied for feature extraction. Two dense layers were placed to classify the text features into the labels after flattening the features coming from the convolution layer. The architecture of the proposed network is explained through Figure5.3. The proposed model is intended to build a lightweight model, so we tried to reduce the number of convolution layers and, with minimum trainable parameters, obtain the best precision value.
- Pooling: The convolutional layer's feature extraction is sensitive to the feature's position in the input. Pooling is used to down-sample the feature map in order to avoid this. Max-pooling, which uses a kernel, is one of the more contemporary approaches. After two layers of convolution, we employed one max-pooling of 3*3 kernel size.
- Dropout: It is a regularisation method used in neural networks to minimise overfitting. We tested with the values of dropout as 0.3, 0.33, and 0.5 in the beginning of our study. However, once our dataset is ready, the model stabilizes with a dropout value as zero.
- Flatten: The convolutional layer produces a multidimensional vector as the output feature map. To provide input to the completely linked dense layer, it must be in one dimension. The Flatten function allows you to turn a two-dimensional feature map into a one-

Table 5.4: Result experiments in terms of loss and accuracy with adam optimizer

Network	**Epochs**	**Accuracy**	**Loss**
VGG16	5	0.5	0.69
	15	0.4934	0.6936
	30	0.55	0.55
	50	0.56	0.5
InceptionResnetV2	5	0.86	0.51
	15	0.9568	0.115
	30	0.935	0.3
	50	0.95	0.33
Proposed model	5	0.5005	0.80074
	15	0.5	0.8007
	30	0.5	0.8007
	50	0.5	0.8007
	100	0.499	0.8007

dimensional feature map. Our suggested network turned a 34*34*64 feature map into a 73984 one-dimensional feature map, which it uses as an input to the fully linked layer.

- Activation function: At a fully connected dense layer in the proposed CNN, the softmax activation function is used, giving each label (class). In the end, using a threshold, the output class is estimated.

5.3.2 Experiments

We conducted different tests with the suggested model and the most relevant models available in the literature to ensure the validity of the suggested architecture. This section explains the set of experiments we carried out. We have also described the evaluation metric required to validate the proposed model.

Table 5.5: Result of experiments in terms of loss and accuracy with SGD optimizer

Network	**S.No**	**Accuracy**	**Loss**
VGG16	5	0.78	0.6937
	15	0.50	0.68
	30	0.49	0.69
	50	0.5	0.9
InceptionResnetV2	5	0.90	0.31
	15	0.9794	0.0061
	30	0.98	0.000023
	50	0.97	0.0000234
Proposed model	5	0.898	0.2481
	15	0.99	0.00016
	30	0.98	0.0011
	50	0.99	0.000001

5.3.2.1 Experimental setup

The suggested architecture is implemented on an 8 Tesla NVIDIA GPU system, with Tensorflow and Keras as deep learning frameworks in the backend. Other python and machine learning packages are used, such as OpenCV, Scikit learning, and Pandas. Deep learning models for text classification such as VGG16, Inception, Inceptionresnet are good with the image data, and we started the experiment with these models.

5.3.2.2 Evaluation metric

Each network's performance is assessed using five standard metrics: validation accuracy, validation loss, precision, recall, and f1 score. The network size is also measured in terms of the number of trainable parameters in each model. The number of trainable parameters represents the complexity of the model.

The loss for the classification model is shown in equation5.7. The classification loss used here is called binary cross-entropy loss, and it is for a two-class classification problem. Output class, y represents the binary indicator (0 or 1) for labels, and p represents the predicted probability.

The other parameter for evaluating the model is accuracy, also called validation accuracy. Classification accuracy is the rate of correct classifications made by our system out of complete classification applied. The proposed approach is a two-class binary classification model that needs to be analyzed using parameters such as precision, recall, and f1 score. Precision, recall, and f1 score is represented in equation5.9, equation5.10, and equation5.11.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{5.8}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{5.9}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{5.10}$$

$$f1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.11}$$

5.3.2.3 Experiment with Adam optimizer

Gradient descent optimization algorithms play a critical role in the optimization of neural networks. We have started our experiments with the most popular and recent optimization algorithm called adam. It computes the adaptive learning rate for each parameter. Each model has initially trained for 5 to 50 epochs for tuning the other hyper-parameters. Table 5.4 lists the experiments performed with the models. We have also trained the proposed model.

Table 5.6: Confusion matrix for VGG16 network for classes Help and Non-Help

Real / Predicted	**Help**	**Non-Help**
Help	172	28
Non-Help	25	75

Table 5.7: Confusion matrix for InceptionResnetV2 network for classes Help and Non-Help

Real / Predicted	**Help**	**Non-Help**
Help	175	25
Non-Help	28	72

5.3.2.4 Experiment with SGD optimizer

Another important optimization algorithm is a stochastic gradient descent algorithm (sgd). It performs the update of parameters for each input $f(x_i, y_i)$ and labels z_i. We have experimented with all three models with sgd. In this experiment, we used the number of epochs between 5 to 50 to train the models. We have recorded the performance of each experiment in Table 5.5. Accuracy and loss have been compared in this table for each of our experiments. Also, to validate our trained model, a separate experiment is performed individually with 300 different images. Precision, recall, and confusion matrix is recorded in Table5.8.

5.4 Results and analysis

Here, the results obtained after extensive experiments with the proposed and previously proposed models in literature are reported. This section further analyzes the data obtained and presents the most relevant parameters for the proposed neural convolution network.

Table 5.8: Proposed convolution neural network confusion matrix for classes Help and Non-Help

Real / Predicted	Help	Non-Help
Help	180	20
Non-Help	30	70

Table 5.9: Comparison of proposed model with VGG16 and InceptionResnetV2

Model	Accuracy	No. of param	Precision	Recall	f1 Score
VGG16	0.50	13,83,57,544	0.86	0.87	0.8649
InceptionResnetV2	0.99	5,43,36,736	0.875	0.862	0.864
Proposed model	0.999	1,40,40,698	0.90	0.857	0.877

5.4.1 Training and testing

Dataset is split into training and testing images in the ratio of 7:3 and is trained accordingly. In this paper, we have trained three models and the results are reported in Table5.4 Table5.5. Here, the results represent the validation accuracy and validation loss for the models trained between 5 to 50 epochs. Existing models such as VGG16 and InceptionResnetV2 were Originally models were trained on the Imagenet dataset and retrained on the proposed dataset using transfer learning. The proposed model is trained here on our developed dataset. Each model is trained and optimized for its best hyper-parameter value and particularly trained with two different optimizer: SGD and Adam.

5.4.2 Prediction with unseen data

We conducted various tests to evaluate our proposed architecture. However, the actual validity of the model depends on how the trained model behaves with unseen data. More than 300 images were collected to test the model's efficiency, with 200 from the Help class and 100 from

the Non-Help class. Prediction result with unseen data is reported in Table5.6, Table5.7, and Table5.8. The confusion matrix is used for the statistical analysis of the model, and here we can observe that for class Help, which is our target class, the proposed model gives a more accurate prediction.

5.4.3 Discussion

The comparison of all the models applied to the provided dataset is shown in this section. The proposed model is a lightweight model for natural scene recognition in the form of text recognition, and Table5.2 shows the proposed model's layer-wise configuration. The comparative result reported in this table represents that the convolution layer in the proposed model uses fewer trainable parameters and, hence, reduces the model's complexity. It can also be taken as the advantage, as it can predict faster with low computational devices. In Table5.3, total number of trainable parameters is reported. It shows that the overall number of the proposed model's trainable parameter is ten times less than the prior models. The number of trainable parameters directly indicates that the network complexity affects the training, testing, and prediction time. Result shown in Table5.6, Table5.7, and Table5.8, could be useful for the statistical analysis of applied models. Hence, we evaluated the model's performance with metrics such as precision, recall, and f1 score, and the results are summarized in Table6.5. The proposed model's precision value is higher than the other two models, and the recall value of the other two models is higher than the proposed model. In this case, the f1 score is a more accurate parameter for the proposed system's overall analysis. Here, we also calculated the f1 score value of all three models. It shows that the proposed model can classify the classes more accurately than the other two models as the f1 score of the proposed model is higher than both models.

5.5 Conclusion

In this chapter, a dataset is developed with approximately 6000 images. Images in Help class contains emergency words such as HELP, SOS, EMERGENCY, etc., these texts are written on walls, roofs, and ground. This paper proposes a novel approach to identify disaster victims

through emergency text in drone surveillance. It also proposed a convolution neural network model that uses the developed data for training and can predict the help situation through text classification. The proposed convolution neural network model achieves approximately 99 % accuracy with a precision value as 0.90. Compared to well-versed models such as VGG16 and InceptionResnetV2 trained on the developed dataset using transfer learning, the proposed model validates itself with a f1 score of 0.877.

Chapter 6

FMFM: Faster Motion Feature Modeling for Drone Action

Drone-based action recognition has got significant attention recently. However, the unavailability of particular human motion features for action classification is a critical drone surveillance challenges. For motion modelling and representation, most previously proposed methods rely heavily on optical flow which is a time-consuming operation. This chapter underlines individual human motion feature modeling by proposing a novel architecture that eliminates the dependency on Optical flow. The proposed architecture uses two sub-modules, FMFM (Faster motion feature modeling) and AAR (Accurate action recognition), to accurately classify the aerial surveillance action. The proposed architecture achieves a remarkable performance of 0.90 validation accuracy in aerial action recognition. Also, it could facilitate the invariant background training of models for surveillance.

6.1 Introduction

Video recognition has improved dramatically in the past years mainly because of the addition of models for deep models of learning actions [35, 165, 179, 88] and large-scale video databases [72, 86, 106]. In addition, many prominent convolution neural networks [178, 66, 188, 152] have been proposed for image recognition task. However, these CNN's can not model the motion feature of each individual effectively from crowd video. In particular, using these CNN's for aerial action recognition can provide a variety of real-life applications using the dataset proposed in [130, 9]. Further, aerial and drone surveillance can be applied to many real-time applications such as unusual activity detection in border area [76], violence and suspicious activity recognition in crowd [89], urban and rural scene understanding. However, the inherent complexity of aerial video still makes motion modelling a very difficult task in action recognition. One of the major concerns in aerial or drone surveillance is the varying nature of human features from different angles and heights.

The crucial need for a realistic approach to action detection works from different heights of live stream video, climate, and with single or multiple humans at the same time. The use of drone surveillance has the advantage of scanning a large remote area in less time. We have recently seen different surveillance tasks, such as police using a drone for traffic control and crowd monitoring during COVID 19. These activities demonstrate the potential and future of drone surveillance.Automation of such surveillance applications, on the other hand, is essential for taking it to the next level. On-device video analysis is required for this automation. In addition, drone hardware is improving in terms of battery life, storage capacity, and processing power. There is a critical need for sophisticated algorithms that can detect persons, as well as their activities and emotions.

In general, a two-staged procedure is used in most current action recognition procedures: first, the optical flow is estimated with the EPE loss. As an input for the action detection module, the measured optical flow works. However, the recent articles [149, 195] indicate that this co-relationship is not vital for the overall modeling of human action. Recently, in ground level images, the performance of the deep learning algorithm is exceptional. However, in aerial drone surveillance, deep learning algorithms are tried. Due to the lack of accurate data of individual

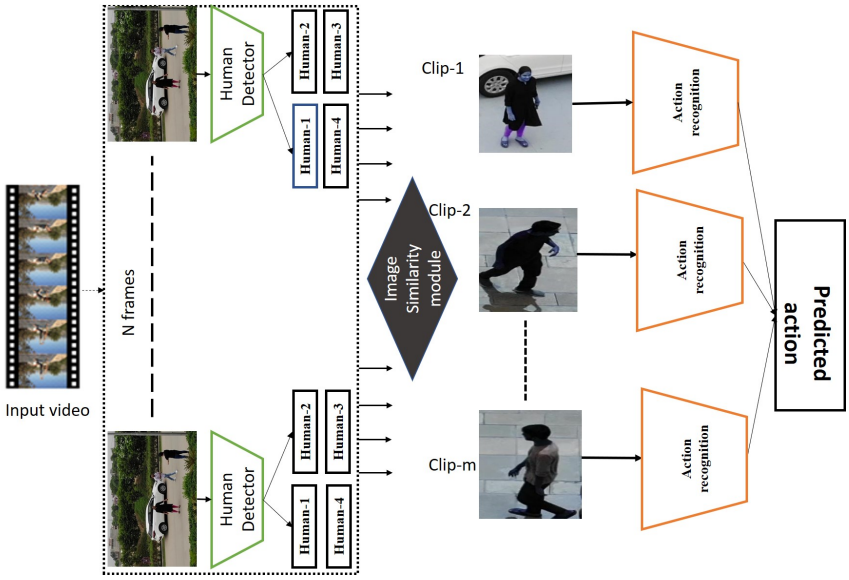


Figure 6.1: The proposed approach for background-invariant motion feature modeling

humans' temporal actions, the existing action recognition model's performance is not up to the mark.

In actual implementations, significant visual differences in features, such as occlusions, pose variations, and lighting adjustments, impose significant aerial surveillance challenges. For action detection, deep learning models use human features to understand the shape, texture, and size in the spatial-temporal domain. Such characteristics are not immediately apparent in drone-captured videos. Human action features can be exposed for activity classification using robust temporal modelling of each person. This chapter contributes to many ways by addressing these challenges in drone surveillance, including a novel architecture, a fast and accurate temporal motion modeling system, and an improved temporal network for action detection.

6.2 Literature

Drone surveillance search and rescue technology is one of the most futuristic and in recent years has received tremendous attention. It could be more effective with the help of profound learning methods such as activity recognition, object detection and object classification. Therefore, some crucial existing work is explored here, which is suitable for drone surveillance automation. In recent years recognition of human activities has been an important topic, with numerous

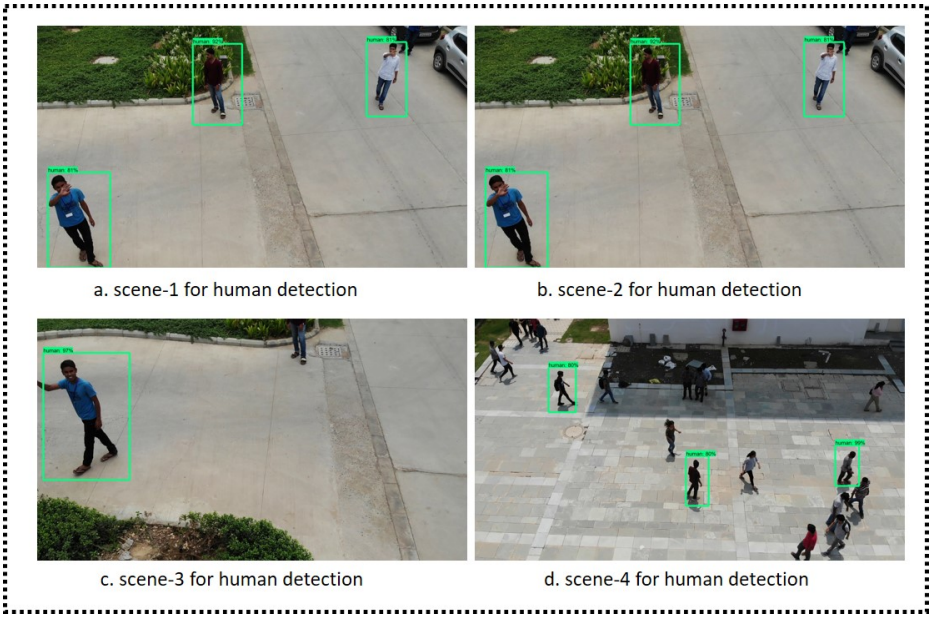


Figure 6.2: Sample images for human detected by human detection module

literature algorithms developed for intelligent surveillance systems [66, 4, 187, 92]. Traditional algorithms for action recognition uses manual feature extraction [27, 187, 125], and hence the performance is not up to the mark comparatively. Also, it requires lots of human effort unnecessarily. The ability of deep learning algorithms is demonstrated by the emergence of current deep learning algorithms for intelligent surveillance [13, 111, 118, 42]. Deep learning models are popular nowadays due to the dataset availability, which is critical for performance. [64] described the challenges facing the actual applications in the field of action recognition. Analysis of deep learning models shows the key to deep learning models' success are the dataset's availability, enabling the models to learn the features required for classification and prediction of human action. In literature, several such datasets is available freely, like UCF[161], Kinetics dataset[72]. However, each dataset has its limitation, which limits its use in different applications.

Deep learning techniques like T-CNN [60], two-stream convolution neural network [154], temporal segment network [178], and temporal pyramid network [188] have performed well with this dataset. Also, a variant of 3D convolutional network with deformation attention model is proposed for modeling the motion and appearance feature in action recognition[88]. Action recognition models with Spatio-temporal feature extraction have got tremendous success in recent years, such as in[107], a composition model is proposed based on the Spatio-temporal feature. Besides, [25] suggests a transferable function technique using a two-stream convolution

neural network for action recognition. Hand-gesture recognition is a field of action recognition, and [87] proposes a two-stage convolution neural network for this.

Other methods of action recognition use the body key points for classifying action into different classes such as in [99], both 2d and 3d features of body key points are used. [94] proposes an action recognition model which uses pose estimation method on pose estimated map. It uses NTU rgb+d [150], UTD-MHAD[21], and PennAction datasets for training of models. These action recognition models outperformed for ground-level action recognition. However, due to the angle and height from which it is recorded, these algorithms do not function well in aerial or drone surveillance. Similarly, a dataset was created and different deep learning algorithms were tested in a work on aerial action recognition [9]. Nonetheless, the methods used on this dataset perform poorly. The problem with such a dataset is, the features are not explored, and hence, the deep learning algorithm could not learn appropriately for the classification of action. Recently, a few datasets for aerial and drone surveillance have been published that is useful for action recognition, such as in [130, 113, 132]. Drone use has increased in recent years, and it now encompasses practically, including traffic analysis for road safety [146], crop-land monitoring [115]. Drone surveillance employing action recognition model for search and rescue has recently been published in this field [112]. [163] proposed a discontinuous multitasking strategy for activity detection in drone films in the area of aerial action recognition. However, the training sample is insufficient and cannot be used as a universal foundation.

These applications using the drone for review and analysis require models trained on the drone dataset for practical performance. Pre-trained models for object detection (OD) and action recognition are not useful in drone surveillance. For the ground level object recognition various object detection models such as RCNN [171], SPP-Net [135], Fast R-CNN [44], and Faster R-CNN [140] have become popular recently. These techniques have been used to identify actions as well in some of the articles in litearture. However, the inference time of the OD techniques outlined above is a significant disadvantage. SSD [95] is one of the models in the OD family that has relatable performance with the most popular Faster RCNN model with having less inference time. However, as shown in the work [9], these algorithms are iefficient for long-distance and aerial or drone footage. Some of the work [17] employs a distinct module for human body detection and its extension to action recognition. Due to the size and angle of the features, this idea of employing a separate module for both tasks could be effective for aerial

surveillance.

6.3 Dataset

This segment discusses the dataset used for experimentation. Besides, it gives the detail of the developed dataset for aerial action recognition. In addition, the newly created dataset is compared to previously published drone based image and video dataset.

6.3.1 Human detection dataset

Human detection is an essential sub-module of the proposed architecture. The proposed architecture of drone surveillance for action recognition requires humans to be detected precisely and requires an accurate human detection dataset. For this, we have utilized the dataset proposed in [113], where the images are captured from the required height and angle. Dataset has more than 5000 humans annotated from different angles and heights. It has a good variety in terms of actors, weather conditions, and the the time of the day.

6.3.2 Proposed action recognition dataset

The second important module in the proposed architecture is action recognition. For this, an accurate dataset is required. A recently published dataset for aerial action recognition has mostly captured for multi-view action performed by more than one human. For this, we have developed a dataset required for training with five different classes. The process of the development of the dataset is as follows:

6.3.2.1 Dataset collection

The dataset is collected in the form of video clips, with over 30 actors performing various actions. The dataset is collected at separate daytime of varying days of the month. Actors dressed in different outfits on the different day of the month gives another variety in our dataset. The captured video has actors of different ages and gender. The videos are captured in various parts of India and mainly in a rural background, making them useful for multiple real-life drone surveillance applications.

6.3.2.2 Dataset preprocessing

All video captured are then resized into the single-size video format without disturbing the aspect ratio of frames. Each video is then labeled into five different classes as required, especially for drone action recognition in remote areas. Data is also converted in the image form as we planned to test 2D ConvNet for action recognition. We have picked the frame in the ratio of 1:5 in the image dataset for action recognition. However, the image dataset has only spatial features, which is sometimes not helpful in classifying the overlapping action such as human walking and human standing could not be classified with only spatial features.

6.3.2.3 The proposed dataset summary

The dataset contains 480 clips and in total 14400 frames. The dataset is labeled in five different classes, and the number of videos are equally divided between all five classes. The important details of proposed dataset is summarized in Table6.1. We can see in this table, UCF rooftop dataset has almost 50 video clips in each class. However, it was released with more than five classes. For drone action recognition, the proposed five actions are sufficient for identifying the emergency in a remote area. The phase, orientation, camera motion, perspective, the number of people in action, the size of humans,and the background are the basic parameters which gives our dataset enough variety.

Table 6.1: Features of action classification dataset

Class	**Video type**	**Background**	**No. of videos in proposed dataset**	**No. of videos in UCF rooftop dataset**
Fight	.MP4	Rural area, College campus	85	48
Lying	.MP4	Rural area, College campus	85	48
Shaking hand	.MP4	Rural area, College campus	85	48
Waving Hand	.MP4	Rural area, College campus	85	48
Waving both hand	.MP4	Rural area College campus	85	0

6.4 Methodology

The proposed technique uses drones for remote monitoring and for various monitoring tasks. In this chapter a complete training method is designed which can model the individual feature of human movement and helps to classify drone actions accurately. Faster motion feature (FMFM) and accurate action classification (AAR) are the two blocks of the proposed system. The following are some necessary notations in this paper: $Xi = I^{c*l*h*w}$ represents a single video with feature c, l, h, and w. Where $I = f(x,y)$ shows a single frame in video, c is the number of channels, l is a frame number, h and w are the height and width of each frame in video clip.

6.4.1 Faster motion feature modeling (FMFM)

For the accurate and faster motion modeling of temporal features, multiple modules are interconnected and trained together. It combines deep learning model object detection and traditional computer vision algorithms to extract detected bounding boxes. This module extracts the temporal feature that appears in consecutive frames of original crowd monitoring surveillance

Figure 6.3: Sample images in our dataset and generated by the proposed FMFM module

videos. For this, the initial step is detecting the human precisely. After detection, the output image is passed to the second part of FMFM, which extracts the detected human, calculates image similarity, and patches most similar extracted images together in the video clip. This stage is further divided into two parts:

- **Step 1** Accurately detecting humans in the surveillance video is the primary step. For this, object detection techniques could help. The benchmark results of object detection techniques motivate us to apply for aerial human detection. It performs well after training the models on the aerial surveillance image dataset using transfer learning.
- **Step 2** The second and most important step in FMFM is extracting the detected humans from the object detection inference image and combining them with the other extracted human from the different frame with the highest similarity.

Therefore the proposed FMFM module transforms the crowd monitoring video dataset into individual human action videos. Each video will have a sufficient human feature and valuable for accurate action recognition. Another advantage of using FMFM is that it will remove the background's effect, affecting the action recognition algorithm's performance. The training of models is usually performed in a structured environment, but when used in a real-time environment or when the background shifts, they struggle.

Overall, the FMFM method begins with precisely detecting humans, so here are some

cutting-edge object detection techniques for detecting humans in aerial surveillance.

- **Faster R-CNN** Originally, the faster RCNN is proposed in [140], comparatively, this model has very less inference time. It is a kind of region-based object proposal generation for object detection. This algorithm generates fewer region proposals than other region-based object detection algorithms. The VGG-16 module is first used as a feature extractor in the Faster RCNN network. However, several advanced models, such as Inception and Mobilenet, have been employed to implement Faster R-CNN over time. As a function to implement feature extractor in our approach, we have used inception network in Faster R CNN.

- **SSD (Single-shot detection)** It was first introduced in [95], which detects objects using multiple-layer features. For the feature extraction, it uses VGG16 Classification network. This method combines classification and localization into a single network. Although we have evaluated the base network with Inception and Mobilenet in our proposed model for object detection.

Basically, object detection is a combination of object localization and object classification in an image $I = f(h,w)$. Applying object detection on a video stream or video clip $X_i = I^{c*l*h*w}$, require to apply the same process on each frame of the video clip.

$$Y_i = O(I) \tag{6.1}$$

Where Y_i is the output label, i.e., human, car, truck, bus. For a good quality object detection model, two different type of loss, localization loss and classification loss need to be optimized. We define the localization loss as $L_{loc}(y,y')$ and classification loss as $L_{class}(y,y')$, in which y is ground truth object class and y' denotes the predicted object class.

$$L(y,y') = L_{loc}(y,y') + L_{class}(y,y') \tag{6.2}$$

6.4.2 Accurate action recognition (AAR)

Here, the primary objective is classifying drone actions, and it is a multi-class activity recognition problem. Usually, in drone surveillance, the height and angle of surveillance change with time, making it more difficult. Identifying a specific action from an image captured by a drone in which multiple humans are present is difficult and necessitates detecting each action individually. Therefore, the proposed architecture uses two other modules FMFM and AAR, which can detect each action accurately. Only spatial features can perform the action recognition in drone surveillance; however, it restricts only upto few actions. We can combine spatial features with motion features and accurate action classification. We performed several experiments with existing model and in combinations of different architectures such as 2D CONVNet with RNN and 2D CONVNet with Time-Distributed RNN, and the proposed architecture a 3D convolution neural network. However, for recognizing few actions required for emergency identification in disaster, motion features are needed, and hence we will stick with the models using spatial and motion features together. For example, in our case, if we have to identify the waving hand, it requires spatial and motion features.

Following are some equation to represent the mathematical modeling of the action recognition module.

$$z_i = A\,(\,I\,) \tag{6.3}$$

$$z_i = A\,(I^{c*l*h*w}) \tag{6.4}$$

Where z_i represents the output class of action recognition model, I represent the image, and $I^{c*l*h*w}$ represents the input video for action recognition module A.

Equation6.4 is further derived into another equation separately using spatial feature and

motion feature for action classification.

$$z_i = S(\sum_{r=1}^{w} A\ (\ I_r\ )) + T(\ A\ (I^{c*l*h*w})) \tag{6.5}$$

Where function S represents the action recognition based on 2d spatial features, and T function represents the action classification feature extraction based on motion features. Equation6.5 represents the working of CNN+RNN, and CNN+Time distributed RNN models used in our experimentation. Two different networks are employed here to extract temporal and spatial characteristics, and in the end both characteristics are combined for classification of action.

6.4.2.1 Action classification with spatial features

Human actions such as lying, sitting, fighting, shaking hands could be easily recognized by the image's spatial features. Image classification models are capable of recognizing such features. Some important 2D ConvNet used in this paper is as follows:

- Resnet50: It uses a series of residual function with 50 layer deep convolution network. It starts from 7*7 convolution with stride two and max pool kernel of size 3*3 with stride 2. After that, at four different stages, it uses the residual function of sizes 3, 4, 6, and 3. Each residual function contains three convolution network of size 1*1,3*3,and 1*1 [55].
- Resnet152: In this, the network depth was 152 layers with the residual function. It also starts with a convolutional layer of 7*7, and all other convolution layers are patched in the form of residual function as in Resnet50 [55].
- VGG16: It is a type of convolution layer with 16 layers. In this 13 convolution layers are present with three fully connected layers[155].

6.4.2.2 Action classification with temporal features

Multiple architectures were designed and trained for action recognition with Spatio-temporal features. Spatio-temporal features are captured through different networks such as 2D ConvNet,

RNN, and time-distributed RNN. The architecture of the models was as follows:

- 2D ConvNet with RNN: In this, we have experimented with a five-layer 2D ConvNet working in parallel with three layers RNN combined with three dense layers in the end for feature classification. 2D ConvNet extracts spatial features, and RNN networks extract the motion features from the sequence of frames. Extracted features are then passed to the network's dense layer after flattening it into a single dimension vector. Both networks working in parallel are merged into a single network with the concatenated feature available in Keras.

- 2D ConvNet with time distributed RNN: This network consists of 5 layer 2D ConvNet with three layers of time distributed RNN layer and one RNN layer working in parallel for spatial and motion feature extraction. Both the network are then merged for overall feature classification with three dense layers in the end.

- Proposed architecture of 3D ConvNet: Proposed network for action recognition uses the modified architecture of VGG16 with 12 3D ConvNet and two dense layers in the end. Convolution layers are stacked in the fashion as it appears in VGG16 with 3*3*3 convolution filter in the first layer having 64 kernels. Detail structure of the proposed network is represented in Table 6.2. Each layer employs the relu activation function, with the exception of the final dense layer, which employs the softmax activation function.

6.4.3 Proposed architecture

The proposed architecture contains two different stages, ie. FMFM and AAR as explained above. In this, in the beginning, humans are detected in surveillance videos, and the output of the human detection module is then passed to the function written for cropping the bounding boxes predicted. It uses humans' coordinates detected by human detection module for single or multiple humans in a surveillance image. Further, the cropped image function's output is passed to another module, which combines the cropped images based on the similarity. For similarity of images, multiple techniques were tested and based on the best outcome with the mean square error result, and the proposed module uses MSE for stacking.

Table 6.2: Detail of proposed convolution neural network for action recognition

Layer	**Kernel size**	**Output**
3DConv	3*3*3	(None, 32, 32, 12, 64)
3DConv	3*3*3	(None, 32, 32, 12, 128)
3DConv	3*3*3	(None, 32, 32, 12, 128)
Maxpool	2*2*2	(None, 16, 16, 6, 128)
3DConv	3*3*3	(None, 16, 16, 6, 256)
3DConv	3*3*3	(None, 16, 16, 6, 256)
3DConv	3*3*3	(None, 16, 16, 6, 256)
Maxpool	2*2*2	(None, 8, 8, 3, 256)
3DConv	3*3*3	(None, 8, 8, 3, 512)
3DConv	3*3*3	(None, 8, 8, 3, 512)
3DConv	3*3*3	(None, 8, 8, 3, 512)
Maxpool	2*2*2	(None, 4, 4, 2, 512)
3DConv	3*3*3	(None, 4, 4, 2, 512)
3DConv	3*3*3	(None, 4, 4, 2, 512)
3DConv	3*3*3	(None, 4, 4, 2, 512)
Maxpool	2*2*2	(None,2, 2, 1, 512)
Flatten	None	(Non, 2048)
Fully Connected	None	(None, 512)
Fully Connected	None	(None, 5)

$$V_i = MSE\ (i_1, i_2, i_3, ...i_n) \tag{6.6}$$

Where V_i is the video clip formed with MSE (mean square error) parameter. MSE is used here to group most similar image together as per the pixel values of images.

$$(i_1, i_2, i_3, ...i_n) = C_{rop}\ (I_1, I_2, I_3, ...I_{60}) \tag{6.7}$$

Where $(I_1, I_2, I_3, ...I_{60})$ are the individual output image from object detection module. Each detected human is cropped by the function C_{rop}, and $(i_1, i_2, i_3, ...i_n)$ are the different instance of humans cropped.

The proposed architecture uses a patch of human detection output of 60 consecutive frames to stack it into individual action recognition video clips. The output individual action video clips is passed through the proposed action recognition sub-module for final action recognition.

The proposed approach's overall goal is to minimize the unified loss *LBackI*, a combination of object detection loss $L(y, y')$, patching mean square error (MSE), and action classification loss $LA - class$. The detailed equation of loss is explained through equation 6.8. In this, y and y' represent the output class for human detection, ie. human and non-human is class labels for this. Overall, our final output is based on the five action classes, where z represents the ground truth labels and z' is for the predicted class.

$$L_{BackI}(z, z') = L(y, y') + MSE(f_1(h, w), f_2(h, w) + L_{A-class} \tag{6.8}$$

Table 6.3: Accuracy & Inference time comparison of object detection models for human detection trained on our human detection dataset on NVIDIA Qudro K1200

Object Detection Techniques	Mean average precision	FPS
Faster RCNNC with Innception	0.833	19
SSD with Inception	0.848	23
SSD Mobilenet Model	0.796	33

Table 6.4: Performance of 2D convolution model applied on our dataset for action recognition

Network	Validation accuracy
Resnet50	0.65
Resnet152	0.61
VGG16	0.91

6.5 Experiments and result

We assessed our dataset using the action classification accuracy measure. We ran a series of tests with several 2D and 3D models that deal with the spatial and temporal aspects of input video in the dataset. This section goes over all of the tests in depth and provides the results in the most appropriate way.

6.5.1 Experimental set-up

The experiment is running on an HP workstation with 6 GB of RAM and a 4 GB NVIDIA Quadro K1200 GPU. The preprocessing and post-processing work of the dataset is also performed on python and its OpenCV package. For our experiments we have used a tensorflow environment on Python 3. Various other Python packages were used during the entire study, including OpenCV, Keras, tensorRT and others.

6.5.2 Experiments

The tests were carried out using the dataset given in section 6.3, which comprises of two distinct datasets, one for human detection and the other for the action recognition presented in this chapter. At the primary level, experiments were performed with object detection models for recognizing the action in drone videos. It is converted into five-class action recognition with the dataset for human detection after preprocessing, and three different object detection models were trained using transfer learning. Experimental results with object detection model, Faster RCNN, SSD Inception, and SSD Mobilenet for action recognition motivate us to develop an architecture where the features are exposed to classification. Therefore, we come up with a novel architecture in this chapter. The proposed architecture is tested with the help of two different trainable networks FMFM and AAR. For the first module, three different human detection modules are trained with more than 20000 steps and its best-tuned hyper-parameters. In this phase, cropping Haman and stacking the maximum matched extracted human, which is also be termed as motion feature modeling of the individual human, is performed to optimize different image matching algorithms.

The second stage of our experimentation is designing and testing a novel action recognition model that can recognize drone surveillance videos' human action. For this, with the help of transfer learning, three different models are trained on our dataset. Besides, as the learning of 2D models are not sufficient for accurately classifying the action, two different advance existing models are also trained on our dataset (CNN + RNN, CNN + TimeDLSTM). These experiments have motivated us to design a little complex but accurate model for classifying the models based on the video's spatio-temporal features. All these models are trained and tuned with its best hyperparameter values.

6.5.2.1 Training and testing

Here, different models were trained and tested and for this,three models for human detection is tried in the first phase of the proposed architecture. Besides, For the second module, AAR, three 2D CNNs are trained on our dataset, also, including proposed network and advanced existing models were trained and optimized for its bets hyper-parameter values. Twelve different

Table 6.5: Comparison of aerial action recognition

Network	Validation Accuracy
Okutama SSD [9]	0.18 mAP
CNN for multi-label PD[157]	0.28 mAP
3D-Resnet[188]	0.65
MOD-20[131]	0.74
CNN+RNN (Our dataset)	0.45
CNN+TimeDLSTM (Our dataset)	0.53
Proposed model (Our dataset)	0.85

Table 6.6: Detailed result comparison of models applied on proposed dataset using transfer learning with proposed model performance

Network	**Accuracy**	**loss**	**Val Accuracy**	**Val loss**
3D-RES-Net[188]	0.75	0.33	0.65	1.45
CNN+RNN	0.69	0.33	0.45	1.75
CNN+TimeDLSTM	0.65	0.33	0.53	1.33
Proposed model	0.89	0.25	0.85	0.75

Table 6.7: Incremental analysis of modules directly applied for action recognition in proposed action dataset

Network	**Validation Accuracy**
Faster RCNN Inception	0.39 mAP
SSD Inception	0.23 mAP
Proposed model for action recognition	0.85
Combined architecture proposed for action recognition	0.90

networks were trained and compared based on the evaluation parameters such as mAP (mean average precision) for object detection models and validation accuracy for action classification models. Video and images in our dataset is split into a 7:3 ratio while training and testing models. Every model is trained until the point of saturation.

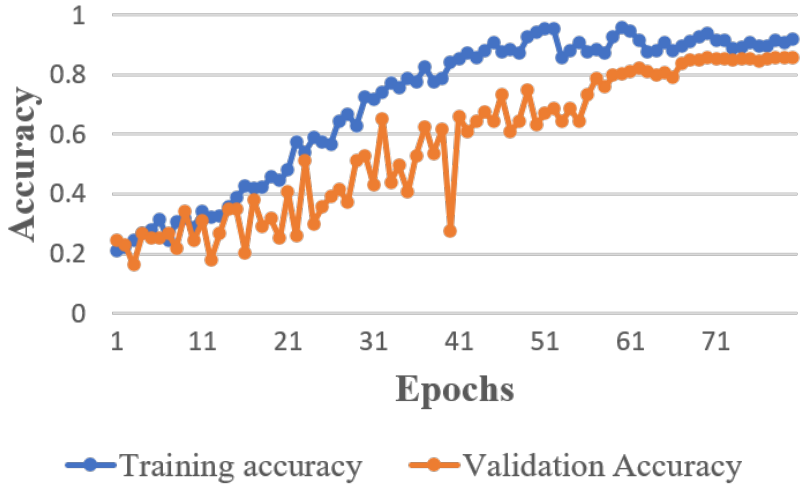


Figure 6.4: Result of proposed model for training and validation accuracy with proposed dataset

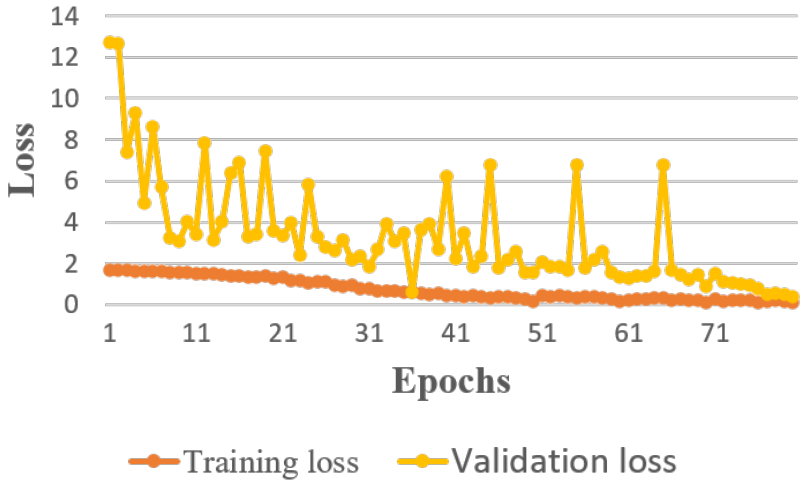


Figure 6.5: Result of proposed model for training and validation loss with proposed dataset

6.6 Result and analysis

For feature extraction in a video frame, the proposed action recognition module employs a CNN. The recurrent neural network, on the other hand, uses several video frames to extract and use the temporal feature to categorize video. For action recognition models trained on our dataset, and result of each model is reported in Table 6.4 and Table 6.6. Table 6.4 summarizes the experimentation result with the 2D ConvNet trained on our dataset using transfer learning. In contrast, Table 6.6 represents the experiments performed with the spatio-temporal models for action recognition, including the proposed model. The result obtained in this paper is compared

with the previously proposed models performance and reported in Table 6.5. Besides, Table 6.6 reports the result of all the models applied on our dataset based on parameters training accuracy, training loss, validation accuracy, and validation loss. Figure 6.4 and Figure 6.5 represents the loss and accuracy of proposed action recognition model trained on our developed dataset. In another level of testing, the proposed architecture's incremental testing is performed, and each module is tested for the outcome as action recognition. The result of incremental module testing is reported in Table 6.7.

6.6.1 Discussion

The proposed architecture is tested at various levels, and one of the most critical analysis we have performed is through incremental module testing. Object recognition models have been used directly to recognise action in some previous work, and we have also used it directly to recognize action.

6.6.1.1 Evaluation metric

The proposed architecture uses two different performance evaluation metrics as follows:

- mAP(mean average precision): It is a standard coco evaluation metric and is used for object detection models. In a few paper[9] for action recognition, object detection models have been used, and our experiment result is compared with their models.
- Validation accuracy: Accuracy of validation is a key parameter where recognition of deep learning action is tested and we have also compiled the results on these parameters with the advanced existing model performance.

6.6.2 Comparative analysis

The performance of the object detector models utilised for action recognition is shown in Table 6.7. With the obejct detection RCNN model, it only reaches 0.39 mAP. However, when

it is used to identify humans, it detects with the accuracy value of 0.833 mAP, the table 6.3 presents comparisons of the several models of human detection trained and evaluated on the suggested data set. The second module offers an action recognition model and the suggested video recognition dataset is also evaluated separately. The maximum accuracy we have got with the proposed model is 0.85, which is validation accuracy with a training accuracy as 0.89. The detail comparative result is represented in Table 6.6 and Table 6.7. In this research, we have proposed an accurate and faster 3D temporal motion modeling architecture and is applied to the input video. While testing the architecture with validation video, the architecture has achieved maximum accuracy with 0.90 validation accuracy and is represented in Table 6.7. Through our experimental result and the incremental analysis of each module, it has been observed that using FMFM and AAR modules together could achieve higher accuracy than any other combination.

6.7 Conclusion

In this chapter, a novel architecture for motion feature modelling is provided for quicker and more accurate action identification. By reducing the influence of backdrop and making it background invariant, the suggested design makes the action recognition algorithm easier to use. A new dataset with five action classifications gathered from moving aerial cameras is also suggested. Furthermore, a precise model for temporal action recognition is proposed, which outperforms in comparison with the existing advanced action recognition models applied to aerial datasets. The incremental analysis of the proposed architecture is also presented in this chapter and it shows that the overall architecture's performance is higher than the individual action recognition models applied on the aerial datasets. Overall performance of proposed architecture is more than 90%, which sets a new benchmark in aerial action recognition.

Chapter 7

Summary and Conclusion

In this thesis, deep-learning based SAR algorithms are developed which can be applied in drone for surveillance. The developed algorithm can operate at a height of up to 50 metres above the ground. In search and rescue, finding the precise location where help is needed takes the most of the time. Hence, the proposed automated monitoring schemes have the potential to save millions of lives by providing the precise location of disaster victims. Timely information collection about the humans in need of relief and rescue in a disaster will help the rescue team in saving their lives. In case of disaster, a well known quote is "If we can reduce one minute of time inn disaster, it could save millions of lives". However, the aerial and drone surveillance puts a certain challenges for computer vision and deep learning techniques in automation. One of the major challenges is the changing height of the drone during surveillance makes human faces opaque at times for prediction. Furthermore, top angle surveillance for human identification and action recognition is one of the most challenging tasks even for the human eyes. Environmental hazard also make the challenge more difficult in drone surveillance.

In this thesis, we have worked on challenges highlighted above for a more accurate search and rescue technique. For this, we have developed few benchmark dataset and proposed various algorithms which are suitable for aerial and drone surveillance in different terrain. In natural

disaster different type of condition occurs such as:

- Human stuck in disaster and need instant help like food packets and medicine.
- Human stuck in disaster require instant help and can give signal in terms of action at the time of surveillance.
- Human stuck in disaster and require help, however, at the time of surveillance, he could not give any physical signal due to environmental hazards or their physical condition.
- Human stuck in the disaster however, he can survive for some time and don't need instant rescue.

We have considered all the cases mentioned above and proposed various algorithm for SAR. For the first scenario, we created a human detection model and verified through experiments that humans can be identified automatically in aerial and drone surveillance. Using the trained human detection and action recognition models, food packets and other instant relief items could be distributed in the disaster prone area for instant help. For second case, where, human could perform certain action, the developed action recognition algorithms could help. Proposed action recognition is trained on aerial and drone surveillance dataset which could easily detect the action such as waving a hand in the direction of drone. However, for the third case, where human could not perform some specific action at the time of surveillance, we have considered emergency text extraction from drone images. In the proposed approach, it has been assumed that human's are writing emergency text such as "HELP", "SOS", and "Emergency" on ground wall or roof. Here, the proposed convolution neural network is able to extract such emergency text and estimate the location of human's stuck in the affected area.

In this thesis, four major algorithms were presented for the development of a faster search and rescue process using drone surveillance. The algorithms have been proposed and validated using SAR data developed for human detection and action recognition. The following are some major aspects covered in this thesis:

- A new methodology for handling the search and rescue is proposed in this thesis. For this set of datasets have been developed to validated the proposed algorithms.

- A new novel algorithm has been proposed to detect the humans and their action in images. The proposed model was applied to the dataset developed for human detection and action recognition in images.

- The exact human action could be identified through the consecutive frames of a scene and hence a novel method is developed for classifying the action. Developed method is based on the 3D convolution neural network and work on the basis of scene classification approach.

- At the time of disaster, it is possible sometime that human is not visible outside or he can not perform particular action in outdoor environment and hence an emergency text classification based approach is developed. The proposed emergency scene classification approach uses the emergency words such as "HELP", "SOS", and "Emergency" to classify the situation as help situation.

- It has been observed that the human detection and action recognition approaches applied directly to the distant drone surveillance video usually fails. Here, we have identified the issues and proposed a unified end to end trainable approach for the faster motion modeling of human actions and accurate action recognition in drone surveillance.

7.1 Contributions

During this course of research work, four major contributions were made in-terms of new algorithms development for the automation of drone based search and rescue.

- A dataset is developed for human detection and action recognition in aerial and drone surveillance for search and rescue.

- Another dataset is developed for action classification in the form of video clips for accurately classifying the action.

- Also, a dataset for emergency text classification is developed for classifying the emergency texts into the help and non help situations.

- A novel architecture is proposed for search and rescue which uses human detection and their action classification in images.
- A new novel algorithm has been developed to to classify the human action as help action using SAR action classification video data.
- A lightweight convolution neural network is developed to classify the text is help situation text in drone surveillance.
- A unified end-to-end trainable approach is proposed to faster motion modeling of human actions and accurate classification of human action. The proposed approach remove the effect of background and hence, background invariant for identifying the help situation.

7.2 Scope for future research

- The proposed methods for action recognition is applied on a low height aerial dataset, however, in future, our target will be to develop such algorithms which can work with more than 50 meter height videos.
- Specifically, for the real-time search and rescue, the developed algorithms should work all together, and integrated in drone. In future, our aim is to develop a unified drone fitted with much stronger algorithm, and work together for a real-time search operation in disaster.
- For the real-time surveillance application such as border area surveillance, city unusual activity surveillance, and remote area surveillance, developed and tested algorithms in this thesis is very much useful.
- In industry, currently we can clearly see the trend of developing different surveillance application related to age and gender classification, crowd surveillance and monitoring, crowd counting,etc, these surveillance algorithms developed in this thesis could help at various level.
- In future, these algorithms should be tested with more diverse background images and videos to provide more generalization to these surveillance applications.

- Looking at the real-time surveillance challenges in disaster prone area, wind, fog, and other environmental hazard could also affect the result and hence need to be taken care while developing a unified application for drone surveillance.
- Another aspect of developing a unified application of drone surveillance in disaster, night vision cameras and thermal camera should also be included for a better and accurate result in different daytime.
- In this thesis, we have included only human search for rescue, which can be extended for other living being also in future.

Appendix A

Additional Project related to this Study

.

Drone surveillance is one of the most emerging technology and it can be used in various application areas. Using the current trends of artificial intelligence techniques, drone based system could be automatize and hence used for various life saving applications. To achieve the desired functionality in the developed SAR technique, various experiments has been performed, the key portion is listed in the different chapters of this thesis, and some other such experiment we summarize here. Some other research has been conducted and projects has been developed which gives us learning and research direction to complete this thesis. The summary of those projects are described in the following section of this Appendix.

A.1 Human detection for search and rescue

Human detection is one of the basic functionality of this thesis work, by detecting human at the time of surveillance, food packet and instant medicine delivery could be served. Instant relief item is very important for the survival in disaster and it could be a life saving idea. In this,

extensive experimentation has been performed on the developed dataset and dataset converted from the Okutama dataset.

A.2 Action recognition for search and rescue

Other important functionality of this thesis is to identify the actual person who is asking for help. In this, it has been assumed that we have detected multiple persons in the disaster prone area, and model is in dilemma for taking decision weather which particular person is asking for rescue and who need some instant relief item. To address this challenge, action recognition based system was proposed in this thesis, and some particular action such as human waving their hands in the direction of drone is considered as the indication of human action for asking help. For this, different type of images has been captured from drone and annotated for two-class, and six class action dataset.

A.3 Help situation identification in drone surveillance using action recognition

This project has been conducted with a team of four different intern students from different college of India. In this internship project, various models have been developed and tested, such as 3DCNN, 2DCNN + LSTM, TSN on the proposed action classification datasets, to demonstrate the recognition of action to identify assistance situations in drone surveillance.

A.4 Pose estimation based action recognition for help situation identification

This is another project conducted with the help of 4 different internship students to assist our experiments of help situation identification. In this, we have focused on actions that are neces-

sary to detect emergency situation among other actions such as single hand wave, waving with both hands, shaking hands, fighting, lying on the ground etc. In this, two models OpenPose and HRNet has been used to detect the aforementioned actions.

A.5 Real-time anomaly detection using drone surveillance

This is another related project we have conducted with the help of two different internship students to recognize actions in drone surveillance videos. From all the available application detecting anomaly is a key problem that has been studied within research domains. The purpose is to assists with recognizing individual actions and detecting whether it is ananomaly or normal activity. To address this challenge, 3d CNN algorithm is used in which, the initial stage of the model extract features from each person in the video and represents the data. Analysis of each extracted sequence to detect the associated actions.

A.6 Pose estimation in drone videos for identifying hand gestures

This is another project conducted with the help of four different internship student to assist this thesis work. In this, pose estimation algorithms has been tested on our dataset for identifying different hand gestures. In this, a complex human pose classification problem is solved using the proposed OpenPose model. The proposed model uses spatial as well as temporal features of the video for the classification of the pose of any number of human bodies, to recognize what the subject is trying to signal the safety personnel using its body pose. The model has been trained on many videos and photos, comprising different number of people with varying body shapes, size, and locations in the frame. The proposed dataset is a MPII Human Pose Dataset which is a state-of-the-art benchmark for evaluation of articulated human pose estimation. The major contribution of this paper is (1) a novel 2DCNN powered model for key-point classification, (2) identification of the pose of the human body after joining all the key-points and determining the correct pose. The model accurately predicts all the key-points present on all the human

bodies present in the frame and then, predicts the best pose identified, among the following poses: jump, kick, punch, run, etc. The proposed model gives an impressive performance to help situation identification with 72 percent training and 77 percent validation accuracy.

A.7 Armed, injured and other suspicious activity recognition using drone surveillance

One of the important application of surveillance activity is to recognize suspicious activity for border-area. In this aspect, w have conducted an application project with the help of four different internship students for armed, injured and other suspicious activity recognition in drone surveillance videos. In this, we propose to create an efficient and less resource-intensive way to loading video data for training and an attempt at creating a viable model that can identify suspicious activities and distinguish them from normal events. Our model can be used to detect suspicious activity in the forests. It captures spatial semantics and motion simultaneously for recognition and classification It uses SlowFast networks for capturing the action. SlowFast networks can be described as a single stream architecture that operates at two different frame rates, but we use the concept of pathways to reflect analogy with the biological Parvo and Magno cellular counterparts. The dataset used consists of 1900 long and untrimmed real-world surveillance videos, with 13 realistic anomalies such as fighting, burglary, robbery etc., as well as normal activities. We achieve 84 percent test accuracy with our model while also training with large number of frames and image sizes.

Appendix B

Additional Images Related to this Study

Figure B.1: Glimpse of images captured and annotated for action recognition data

Figure B.2: Sample images in the original Okutama dataset

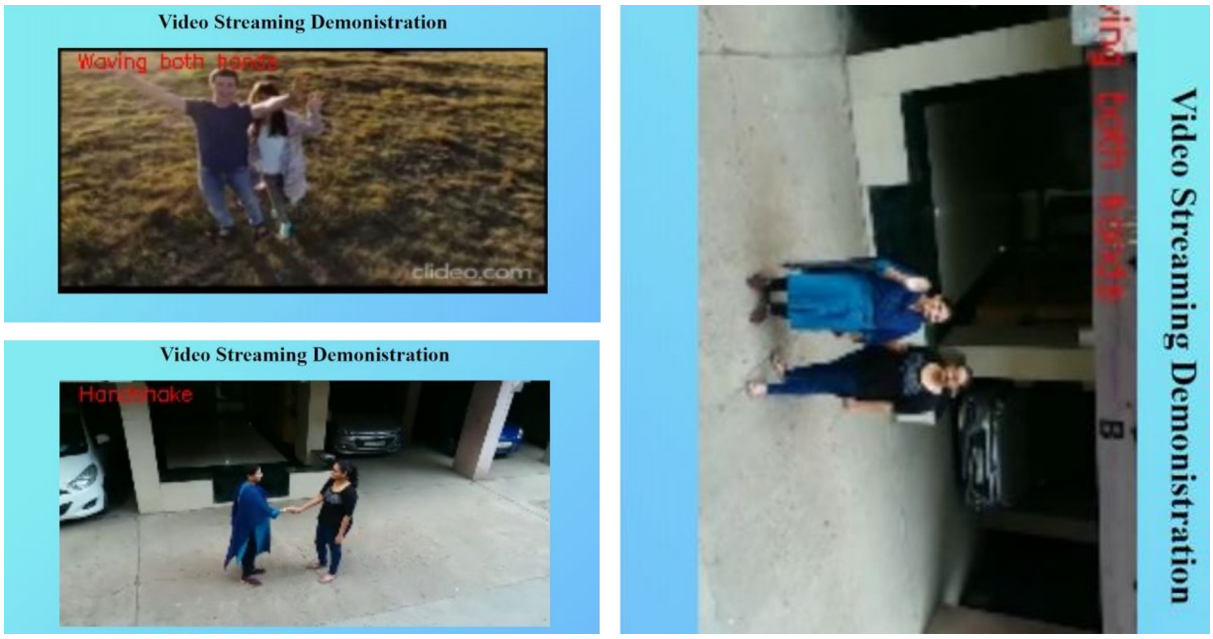


Figure B.3: Inference example of the developed application for action recognition in drone surveillance

Figure B.4: Sample visual example detected human pose for action recognition through Hrnet model

Figure B.5: Sample Image of NVIDIA DGX supercomputer available in campus and has been highly used in this thesis for video and image processing and AI models training

References

[1] Q. K. Al-Shayea. Artificial neural networks in medical diagnosis. *International Journal of Computer Science Issues*, 8(2):150–154, 2011.

[2] J. Albusac, J. J. Castro-Schez, L. M. López-López, D. Vallejo, and L. Jimenez-Linares. A supervised learning approach to automate the acquisition of knowledge in surveillance systems. *Signal Processing*, 89(12):2400–2414, 2009.

[3] O. Alsharif and J. Pineau. End-to-end text recognition with hybrid hmm maxout models. *arXiv preprint arXiv:1310.1811*, 2013.

[4] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *International workshop on human behavior understanding*, pages 29–39. Springer, 2011.

[5] A. Bachrach, R. He, and N. Roy. Autonomous flight in unknown indoor environments. *International Journal of Micro Air Vehicles*, 1(4):217–228, 2009.

[6] J. Bai, Z. Chen, B. Feng, and B. Xu. Image character recognition using deep convolutional neural network learned from different languages. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 2560–2564. IEEE, 2014.

[7] A. Baldominos, Y. Saez, and P. Isasi. Evolutionary convolutional neural networks: An application to handwriting recognition. *Neurocomputing*, 283:38–52, 2018.

[8] B. Banerjee and V. Murino. Efficient pooling of image based cnn features for action recognition in videos. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2637–2641. IEEE, 2017.

[9] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35, 2017.

[10] A. L. Beam and I. S. Kohane. Big data and machine learning in health care. *Jama*, 319(13):1317–1318, 2018.

[11] R. Begg, J. Kamruzzaman, and R. Sarker. *Neural networks in healthcare: Potential and challenges: Potential and challenges*. Igi Global, 2006.

[12] B. Bera, A. K. Das, S. Garg, M. J. Piran, and M. S. Hossain. Access control protocol for battlefield surveillance in drone-assisted iot environment. *IEEE Internet of Things Journal*, 2021.

[13] V. Bloom, D. Makris, and V. Argyriou. G3d: A gaming action dataset and real time action recognition evaluation framework. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 7–12. IEEE, 2012.

[14] M. Bonetto, P. Korshunov, G. Ramponi, and T. Ebrahimi. Privacy in mini-drone based video surveillance. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 4, pages 1–6. IEEE, 2015.

[15] A. Burney and T. Q. Syed. Crowd video classification using convolutional neural networks. In *2016 International Conference on Frontiers of Information Technology (FIT)*, pages 247–251. IEEE, 2016.

[16] R. Canlas. Data mining in healthcare: Current applications and issues. *School of Information Systems & Management, Carnegie Mellon University, Australia*, 2009.

[17] B. Chakraborty, O. Rudovic, and J. Gonzalez. View-invariant human-body detection with extension to human action recognition using component-wise hmm of body parts. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE, 2008.

[18] A. A. Chandio and M. Pickering. Convolutional feature fusion for multi-language text detection in natural scene images. In *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–6. IEEE, 2019.

[19] D. S. Char, N. H. Shah, and D. Magnus. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, 378(11):981, 2018.

[20] S. Chaturvedi, T. A. Faruquie, L. V. Subramaniam, and M. K. Mohania. Estimating accuracy for text classification tasks on large unlabeled data. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 889–898, 2010.

[21] C. Chen, R. Jafari, and N. Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, pages 168–172. IEEE, 2015.

[22] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*, pages 5076–5084, 2017.

[23] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou. Aon: Towards arbitrarily-oriented text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5571–5579, 2018.

[24] D. P. Coppola. *Introduction to international disaster management*. Elsevier, 2006.

[25] C. Dai, X. Liu, and J. Lai. Human action recognition using two-stream attention based lstm networks. *Applied soft computing*, 86:105820, 2020.

[26] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. *arXiv preprint arXiv:1605.06409*, 2016.

[27] S. Danafar and N. Gheissari. Action recognition for surveillance applications using optic flow and svm. In *Asian Conference on Computer Vision*, pages 457–466. Springer, 2007.

[28] D. de Oliveira and M. Wehrmeister. Using deep learning and low-cost rgb and thermal cameras to detect pedestrians in aerial images captured by multirotor uav. *Sensors*, 18(7):2244, 2018.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[30] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou. Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks.

IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(8):3652–3664, 2017.

[31] A. Dr. Satendra Dr. K. J. Anandha Kumar Maj. Gen. Dr. V. K. Naik, KC. India disaster report, 2013. *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine*, 37(3):185, 2013.

[32] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 370–386, 2018.

[33] C. L. Dunis, J. Laws, and G. Sermpinis. Higher order and recurrent neural architectures for trading the eur/usd exchange rate. *Quantitative Finance*, 11(4):615–629, 2011.

[34] M. ElMikaty and T. Stathaki. Car detection in aerial images of dense urban areas. *IEEE Transactions on Aerospace and Electronic Systems*, 54(1):51–63, 2017.

[35] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4768–4777, 2017.

[36] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

[37] C. Fogarty and A. Region. Hurricane michael, 17–20 october 2000. *Part I: Summary report and storm impact on Canada. Meteorological Service of Canada*, 39, 2002.

[38] C. Forces. B–ga–209–001/fp–001 dfo 5449 national sar manual. *PDF). Archived from the original on*, pages 08–03, 2008.

[39] J.-Y. Forcier. *The Canadian Navy and the Canadian Coast Guard: Cooperating Sea Services or Co-existing Federal Fleets*. Citeseer, 2011.

[40] L. M. Francis and N. Sreenath. Live detection of text in the natural environment using convolutional neural network. *Future Generation Computer Systems*, 98:444–455, 2019.

[41] O. Frunza, D. Inkpen, and T. Tran. A machine learning approach for identifying disease-treatment relations in short texts. *IEEE transactions on knowledge and data engineering*, 23(6):801–814, 2010.

[42] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 33(11):2188–2202, 2011.

[43] Z. Ghahramani. Unsupervised learning. In *Summer School on Machine Learning*, pages 72–112. Springer, 2003.

[44] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[45] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[46] T. Gneiting and A. E. Raftery. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005.

[47] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[48] M. A. Goodrich, B. S. Morse, D. Gerhardt, J. L. Cooper, M. Quigley, J. A. Adams, and C. Humphrey. Supporting wilderness search and rescue using a camera-equipped mini uav. *Journal of Field Robotics*, 25(1-2):89–110, 2008.

[49] A. Graves and J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in neural information processing systems*, pages 545–552, 2009.

[50] A. Guillen-Perez, R. Sanchez-Iborra, M.-D. Cano, J. C. Sanchez-Aarnoutse, and J. Garcia-Haro. Wifi networks on drones. In *2016 ITU Kaleidoscope: ICTs for a Sustainable World (ITU WT)*, pages 1–8. IEEE, 2016.

[51] R. G. Guimaraes, R. L. Rosa, D. De Gaetano, D. Z. Rodriguez, and G. Bressan. Age groups classification in social network using deep learning. *IEEE Access*, 5:10805–10816, 2017.

[52] Q. Guo, D. Tu, J. Lei, and G. Li. Hybrid cnn-hmm model for street view house number recognition. In *Asian Conference on Computer Vision*, pages 303–315. Springer, 2014.

[53] P. Gupta, A. Khanna, and S. Majumdar. Disaster management in flash floods in leh (ladakh): A case study. *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine*, 37(3):185, 2012.

[54] Y. Gurwicz, R. Yehezkel, and B. Lachover. Multiclass object classification for real-time video surveillance systems. *Pattern Recognition Letters*, 32(6):805–815, 2011.

[55] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[56] D. Hein, S. Bayer, R. Berger, T. Kraft, and D. Lesmeister. An integrated rapid mapping system for disaster management. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:499, 2017.

[57] G. E. Hinton and R. S. Zemel. Autoencoders, minimum description length, and helmholtz free energy. *Advances in neural information processing systems*, 6:3–10, 1994.

[58] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[59] J. Hou, H. Gao, Q. Xia, and N. Qi. Feature combination and the knn framework in object classification. *IEEE transactions on neural networks and learning systems*, 27(6):1368–1378, 2015.

[60] R. Hou, C. Chen, and M. Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 5822–5831, 2017.

[61] K. M. Hunt and A. Menon. The 2018 kerala floods: a climate change perspective. *Climate Dynamics*, 54(3):2433–2446, 2020.

[62] IFRC. Resilience: Saving lives today, investing for tomorrow. *World Disasters Report*, 2016.

[63] S. Inunganbi, P. Choudhary, and K. Manglem. Meitei mayek handwritten dataset: compilation, segmentation, and character recognition. *The Visual Computer*, pages 1–15, 2020.

[64] I. Jegham, A. B. Khalifa, I. Alouani, and M. A. Mahjoub. Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation*, 32:200901, 2020.

[65] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013.

[66] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[67] A. Jindal, N. Aggarwal, and S. Gupta. An obstacle detection method for visually impaired persons by ground plane removal using speeded-up robust features and gray level co-occurrence matrix. *Pattern Recognition and Image Analysis*, 28(2):288–300, 2018.

[68] M. Y. Kabir, S. Gruzdev, and S. Madria. Stimulate: A system for real-time information acquisition and learning for disaster management. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 186–193. IEEE, 2020.

[69] I. Kalra, M. Singh, S. Nagpal, R. Singh, M. Vatsa, and P. Sujit. Dronesurf: Benchmark dataset for drone-based face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019.

[70] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for document image classification. In *2014 22nd International Conference on Pattern Recognition*, pages 3168–3172. IEEE, 2014.

[71] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[72] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[73] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE, 1995.

[74] D. Keysers, T. Deselaers, H. A. Rowley, L.-L. Wang, and V. Carbune. Multi-language online handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1180–1194, 2016.

[75] S. G. Khan, G. Herrmann, F. L. Lewis, T. Pipe, and C. Melhuish. Reinforcement learning and optimal adaptive control: An overview and implementation examples. *Annual reviews in control*, 36(1):42–59, 2012.

[76] S. J. Kim and G. J. Lim. Drone-aided border surveillance with an electrification line battery charging system. *Journal of Intelligent & Robotic Systems*, 92(3):657–670, 2018.

[77] D. Kollias and S. Zafeiriou. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *arXiv preprint arXiv:1910.01417*, 2019.

[78] I. Kononenko. Inductive and bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal*, 7(4):317–337, 1993.

[79] I. Kononenko, I. Bratko, and M. Kukar. Application of machine learning to medical diagnosis. *Machine Learning and Data Mining: Methods and Applications*, 389:408, 1997.

[80] D. Koundal, R. Vishraj, S. Gupta, and S. Singh. An automatic roi extraction technique for thyroid ultrasound image. In *2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS)*, pages 1–5. IEEE, 2015.

[81] J. R. Koza, F. H. Bennett, D. Andre, and M. A. Keane. Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In *Artificial Intelligence in Design'96*, pages 151–170. Springer, 1996.

[82] S. Kushwaha, S. Bahl, A. K. Bagha, K. S. Parmar, M. Javaid, A. Haleem, and R. P. Singh. Significant applications of machine learning for covid-19 pandemic. *Journal of Industrial Integration and Management*, 5(4), 2020.

[83] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[84] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[85] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, et al. Comparison of learning algorithms for handwritten digit recognition. In *International conference on artificial neural networks*, volume 60, pages 53–60. Perth, Australia, 1995.

[86] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov, and A. Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020.

[87] C. Li, S. Li, Y. Gao, X. Zhang, and W. Li. A two-stream neural network for pose-based hand gesture recognition. *arXiv preprint arXiv:2101.08926*, 2021.

[88] J. Li, X. Liu, M. Zhang, and D. Wang. Spatio-temporal deformable 3d convnets with attention for action recognition. *Pattern Recognition*, 98:107037, 2020.

[89] M. Li, M. O'Grady, X. Gu, M. A. Alawlaqi, G. O'Hare, et al. Time-bounded activity recognition for ambient assisted living. *IEEE transactions on emerging topics in computing*, 2018.

[90] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. *arXiv preprint arXiv:2104.00946*, 2021.

[91] L.-J. Lin. Reinforcement learning for robots using neural networks. Technical report, Carnegie-Mellon Univ Pittsburgh PA School of Computer Science, 1993.

[92] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016.

[93] K. Liu and G. Mattyus. Fast multiclass vehicle detection on aerial images. *IEEE Geoscience and Remote Sensing Letters*, 12(9):1938–1942, 2015.

[94] M. Liu and J. Yuan. Recognizing human actions as the evolution of pose estimation maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1159–1168, 2018.

[95] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[96] Y. Liu, P. Sun, M. R. Highsmith, N. M. Wergeles, J. Sartwell, A. Raedeke, M. Mitchell, H. Hagy, A. D. Gilbert, B. Lubinski, et al. Performance comparison of deep learning techniques for recognizing birds in aerial images. In *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pages 317–324. IEEE, 2018.

[97] Y. Lu, J. Lu, S. Zhang, and P. Hall. Traffic signal detection and classification in street views using an attention model. *Computational Visual Media*, 4(3):253–266, 2018.

[98] C. Luo, L. Jin, and Z. Sun. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118, 2019.

[99] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5137–5146, 2018.

[100] P. Lynch. The origins of computer weather prediction and climate modeling. *Journal of computational physics*, 227(7):3431–3444, 2008.

[101] C. Machinery. Computing machinery and intelligence-am turing. *Mind*, 59(236):433, 1950.

[102] R. H. Major. *Early Voyages to Terra Australis, now called Australia: A Collection of Documents, and Extracts from Early Manuscript Maps, Illustrative of the History of Discovery on the Coasts of that vast island, from the beginning of the sixteenth century to the time of Captain Cook*, volume 25. Hakluyt society, 1859.

[103] D. Makris and T. Ellis. Learning semantic scene models from observing activity in visual surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(3):397–408, 2005.

[104] R. G. Mantovani, T. Horváth, R. Cerri, J. Vanschoren, and A. C. de Carvalho. Hyperparameter tuning of a decision tree induction algorithm. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 37–42. IEEE, 2016.

[105] Ž. Marušić, D. Božić-Štulić, S. Gotovac, and T. Marušić. Region proposal approach for human detection on aerial imagery. In *2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech)*, pages 1–6. IEEE, 2018.

[106] J. Materzynska, G. Berger, I. Bax, and R. Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[107] J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang, and T. Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020.

[108] A. E. Maxwell, T. A. Warner, and F. Fang. Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, 39(9):2784–2817, 2018.

[109] L. R. Medsker and L. Jain. Recurrent neural networks. *Design and Applications*, 5, 2001.

[110] A. R. Mekala, S. Madria, and M. Linderman. Aerial vehicle trajectory design for spatio-temporal task aggregation. In *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 1085–1094. IEEE, 2016.

[111] H. Meng, N. Pears, and C. Bailey. A human action recognition system for embedded computer vision application. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6. IEEE, 2007.

[112] B. Mishra, D. Garg, P. Narang, and V. Mishra. A hybrid approach for search and rescue using 3dcnn and pso.

[113] B. Mishra, D. Garg, P. Narang, and V. Mishra. Drone-surveillance for search and rescue in natural disaster. *Computer Communications*, 156:1–10, 2020.

[114] V. K. Mishra and A. Sengupta. Mo-pse: Adaptive multi-objective particle swarm optimization based design space exploration in architectural synthesis for application specific processor design. *Advances in Engineering Software*, 67:111–124, 2014.

[115] U. R. Mogili and B. Deepak. Review on application of drone systems in precision agriculture. *Procedia computer science*, 133:502–509, 2018.

[116] L. Moncla, M. Gaio, E. Egorova, and C. Claramunt. An automatic extraction method of static and dynamic spatial contexts from texts. 2018.

[117] D. A. Monroe. Multimedia surveillance and monitoring system including network configuration, Nov. 29 2005. US Patent 6,970,183.

[118] S. V. Mora and W. J. Knottenbelt. Deep learning for domain-specific action recognition in tennis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 170–178. IEEE, 2017.

[119] L. Mou, Y. Hua, P. Jin, and X. X. Zhu. Era: A dataset and deep learning benchmark for event recognition in aerial videos. *arXiv preprint arXiv:2001.11394*, 2020.

[120] I. Muhammad and Z. Yan. Supervised machine learning approaches: A survey. *ICTACT Journal on Soft Computing*, 5(3), 2015.

[121] R. R. Murphy, S. Tadokoro, D. Nardi, A. Jacoff, P. Fiorini, H. Choset, and A. M. Erkmen. Search and rescue robotics. *Springer handbook of robotics*, pages 1151–1173, 2008.

[122] M. Nieuwenhuisen, D. Droeschel, M. Beul, and S. Behnke. Autonomous navigation for micro aerial vehicles in complex gnss-denied environments. *Journal of Intelligent & Robotic Systems*, 84(1-4):199–216, 2016.

[123] U. of Central Florida. Ucf-arg dataset, 2020 (accessed March 3, 2019) https://www.crcv.ucf.edu/data/UCF-ARG.php.

[124] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160. IEEE, 2011.

[125] E. Ohn-Bar and M. Trivedi. Joint angles similarities and hog2 for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 465–470, 2013.

[126] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

[127] L. C. Padierna, M. Carpio, A. Rojas, H. Puga, R. Baltazar, and H. Fraire. Hyper-parameter tuning for support vector machines by estimation of distribution algorithms. In *Nature-Inspired Design of Hybrid Intelligent Systems*, pages 787–800. Springer, 2017.

[128] E. J. Palomo, J. North, D. Elizondo, R. M. Luque, and T. Watson. Application of growing hierarchical som for visualisation of network forensics traffic data. *Neural Networks*, 32:275–284, 2012.

[129] X. Pang, Y. Zhou, P. Wang, W. Lin, and V. Chang. An innovative neural network approach for stock market prediction. *The Journal of Supercomputing*, 76(3):2098–2118, 2020.

[130] A. G. Perera, Y. W. Law, and J. Chahl. Drone-action: An outdoor recorded drone video dataset for action recognition. *Drones*, 3(4):82, 2019.

[131] A. G. Perera, Y. W. Law, T. T. Ogunwa, and J. Chahl. A multiviewpoint outdoor dataset for human action recognition. *IEEE Transactions on Human-Machine Systems*, 50(5):405–413, 2020.

[132] A. G. Perera, Y. Wei Law, and J. Chahl. Uav-gesture: a dataset for uav control and gesture recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[133] H.-H. Pham, T.-L. Le, and N. Vuillerme. Real-time obstacle detection system in indoor environment for the visually impaired using microsoft kinect sensor. *Journal of Sensors*, 2016, 2016.

[134] K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter, and T. Unterthiner. Interpretable deep learning in drug discovery. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 331–345. Springer, 2019.

[135] P. Purkait, C. Zhao, and C. Zach. Spp-net: Deep absolute pose regression with synthetic views. *arXiv preprint arXiv:1712.03452*, 2017.

[136] T. Qu, Q. Zhang, and S. Sun. Vehicle detection from high-resolution aerial images using spatial pyramid pooling-based deep convolutional neural networks. *Multimedia Tools and Applications*, 76(20):21651–21663, 2017.

[137] M. Radovic, O. Adarkwa, and Q. Wang. Object recognition in aerial images using convolutional neural networks. *Journal of Imaging*, 3(2):21, 2017.

[138] A. Rajkomar, J. Dean, and I. Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[139] 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*, pages 779–788, 2016.

[140] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[141] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016.

[142] A. Rojas-Domínguez, L. C. Padierna, J. M. C. Valadez, H. J. Puga-Soberanes, and H. J. Fraire. Optimal hyper-parameter tuning of svm classifiers with application to medical diagnosis. *IEEE Access*, 6:7164–7176, 2017.

[143] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[144] S. Russell and P. Norvig. Ai a modern approach. *Learning*, 2(3):4, 2005.

[145] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In *ICCV*, volume 1, page 2. Citeseer, 2009.

[146] G. Salvo, L. Caruso, and A. Scordo. Urban traffic analysis through an uav. *Procedia-Social and Behavioral Sciences*, 111:1083–1091, 2014.

[147] T. J. Sejnowski. *The deep learning revolution*. Mit Press, 2018.

[148] Z. Selmi, M. B. Halima, A. Wali, and A. M. Alimi. A framework of text detection and recognition from natural images for mobile device. In *Ninth International Conference on Machine Vision (ICMV 2016)*, volume 10341, page 1034127. International Society for Optics and Photonics, 2017.

[149] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black. On the integration of optical flow and action recognition. In *German Conference on Pattern Recognition*, pages 281–297. Springer, 2018.

[150] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[151] B. Shi, X. Bai, and C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016.

[152] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020.

[153] M. Silvagni, A. Tonoli, E. Zenerino, and M. Chiaberge. Multipurpose uav for search and rescue operations in mountain avalanche events. *Geomatics, Natural Hazards and Risk*, 8(1):18–33, 2017.

[154] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[155] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[156] M. Singh, S. Singh, and S. Gupta. Investigations on roi selection for liver classification. In *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–6. IEEE, 2014.

[157] A. Soleimani and N. M. Nasrabadi. Convolutional neural networks for aerial multi-label pedestrian detection. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1005–1010. IEEE, 2018.

[158] L. Sommer, T. Schuchert, and J. Beyerer. Comprehensive analysis of deep learning based vehicle detection in aerial images. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[159] L. W. Sommer, T. Schuchert, and J. Beyerer. Fast deep vehicle detection in aerial images. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 311–319. IEEE, 2017.

[160] Y. Song, D. Demirdjian, and R. Davis. Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In *Face and Gesture 2011*, pages 500–506. IEEE, 2011.

[161] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[162] A. Srivastava, A. Kundu, S. Sural, and A. Majumdar. Credit card fraud detection using hidden markov model. *IEEE Transactions on dependable and secure computing*, 5(1):37–48, 2008.

[163] W. Sultani and M. Shah. Human action recognition in drone videos using a few aerial training examples. *arXiv preprint arXiv:1910.10027*, 2019.

[164] W. Sultani and M. Shah. Human action recognition in drone videos using a few aerial training examples. *Computer Vision and Image Understanding*, 206:103186, 2021.

[165] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018.

[166] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[167] C. C. Tappert, C. Y. Suen, and T. Wakahara. The state of the art in online handwriting recognition. *IEEE Transactions on pattern analysis and machine intelligence*, 12(8):787–808, 1990.

[168] D. Tian and Z. Shi. Mpso: Modified particle swarm optimization and its applications. *Swarm and evolutionary computation*, 41:49–68, 2018.

[169] T. Tomic, K. Schmid, P. Lutz, A. Domel, M. Kassecker, E. Mair, I. L. Grixa, F. Ruess, M. Suppa, and D. Burschka. Toward a fully autonomous uav: Research platform for indoor and outdoor urban search and rescue. *IEEE robotics & automation magazine*, 19(3):46–56, 2012.

[170] D. Tosato, M. Farenzena, M. Spera, V. Murino, and M. Cristani. Multi-class classification on riemannian manifolds for video surveillance. In *European conference on computer vision*, pages 378–391. Springer, 2010.

[171] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[172] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen. A convolutional neural network approach for acoustic scene classification. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1547–1554. IEEE, 2017.

[173] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, 2019.

[174] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

[175] S. Viaene, G. Dedene, and R. A. Derrig. Auto claim fraud detection using bayesian learning neural networks. *Expert Systems with Applications*, 29(3):653–666, 2005.

[176] S. Waharte and N. Trigoni. Supporting search and rescue operations with uavs. In *2010 International Conference on Emerging Security Technologies*, pages 142–147. IEEE, 2010.

[177] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In *European conference on computer vision*, pages 825–841. Springer, 2016.

[178] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[179] Y. Wang, M. Long, J. Wang, and P. S. Yu. Spatiotemporal pyramid network for video action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1529–1538, 2017.

[180] Y. Wang, H. Zhang, and G. Zhang. cpso-cnn: An efficient pso-based algorithm for fine-tuning hyper-parameters of convolutional neural networks. *Swarm and Evolutionary Computation*, 49:114–123, 2019.

[181] M. Wazid, B. Bera, A. Mitra, A. K. Das, and R. Ali. Private blockchain-envisioned security framework for ai-enabled iot-based drone-aided healthcare services. In *Proceedings of the 2nd ACM MobiCom Workshop on Drone Assisted Wireless Communications for 5G and Beyond*, pages 37–42, 2020.

[182] H. White et al. *Artificial neural networks*. Blackwell Cambridge, Mass., 1992.

[183] B. Wiese and C. Omlin. Credit card transactions, fraud detection, and machine learning: Modelling time with lstm recurrent neural networks. In *Innovations in neural information paradigms and applications*, pages 231–268. Springer, 2009.

[184] Wikipedia. April 2015 nepal earthquake, 2019 (Accessed: 2019-09-30) ://en.wikipedia.org/wiki/April$_2$015$_N epal_e arthquake$.

[185] Wikipedia. Natural disasters in japan, 2019 (accessed March 3, 2019) https://en.wikipedia.org/wiki/Natural$_d isasters_i n_J apan$.

[186] Z. Z. Wint, Y. Manabe, and M. Aritsugi. Deep learning based sentiment classification in social network services datasets. In *2018 IEEE international conference on big data, cloud computing, data science & engineering (BCD)*, pages 91–96. IEEE, 2018.

[187] L. Xia, C.-C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27. IEEE, 2012.

[188] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 591–600, 2020.

[189] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang. Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical reviews*, 119(18):10520–10594, 2019.

[190] C. Yao, X. Bai, and W. Liu. A unified framework for multioriented text detection and recognition. *IEEE Transactions on Image Processing*, 23(11):4737–4749, 2014.

[191] B. Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

[192] H.-J. Yoo. Deep convolution neural networks in computer vision: a review. *IEIE Transactions on Smart Processing and Computing*, 4(1):35–43, 2015.

[193] F. Zhan and S. Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2059–2068, 2019.

[194] Y. Zhao and Y. Peng. Saliency-guided video classification via adaptively weighted learning. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 847–852. IEEE, 2017.

[195] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann. Hidden two-stream convolutional networks for action recognition. In *Asian conference on computer vision*, pages 363–378. Springer, 2018.

List of Publications

International Journals

1. Mishra, Balmukund, Deepak Garg, Pratik Narang, and Vipul Mishra. "Drone-surveillance for search and rescue in natural disaster." Computer Communications 156 (2020): 1-10.

2. Mishra, Balmukund, Deepak Garg, Pratik Narang, and Vipul Mishra. "A hybrid approach for search and rescue using 3DCNN and PSO." Neural Computing and Applications (2020): 1-15.

3. Mishra, Balmukund, Deepak Garg, Pratik Narang, and Vipul Mishra. "Background Invariant Search and Rescue using an End-to-End Joint Trainable Network for Fast and Accurate Action Recognition in Drone Surveillance." Submitted in a Elsevier Journal, Computer Vision and Image Understanding.

International Conferences

1. Mishra, Balmukund, Deepak Garg, Pratik Narang, and Vipul Mishra. "Enhancing Emergency Text Classification in Drone Surveillance using Deep Learning." Submitted in 6th International Conference on Computer Vision and Image Processing, IIT Ropar.

Other publication during Ph.D

1. Ghosh S.B., Muddalkar K., Mishra B., Garg D. (2021) Amur Tiger Detection for Wildlife Monitoring and Security. In: Garg D., Wong K., Sarangapani J., Gupta S.K. (eds) Ad-

vanced Computing. IACC 2020. Communications in Computer and Information Science, vol 1368. Springer, Singapore.

2. Vyas, P., Mishra, B., Garg, D., Narang, P.. (2021). Human Detection for Search and Rescue using Automatic Drone Surveillance. In National conference on Deep Learning and Applications, IITM Kerala.

Acknowledgments

I would like to thank my supervisors **Dr. Deepak Garg (Professor and Head, CSE Department, Bennett University, Greater Noida)** and **Dr. Pratik Narang (Asst. Professor, Dept. of CSIS, BITS Pilani)** for their continuous encouragement, support, guidance, and valuable suggestions throughout this research work. I am thankful to Dr. Vipul Mishra (Asst. Professor, Bennett University, Greater Noida) for all his support, motivation, and guidance throughout this research. I would like to express my deepest thanks to the almighty (The Omnipresent, The Omnipotent, and The Most Knowledgeable), whose continuous blessings kept me on the right path and enabled me to complete this research work. I am thankful to the NVIDIA, Bennett Research Center on AI, who has provided excellent environment and infrastructure for research and developments.

I would like to thank my RAC Chairman, Prof. Shivani Goel, for her valuable suggestion throughout this research. I take this opportunity to thank Dr. Sridhar Swaminathan (Asst. Professor, Bennett University), who has always motivated me to think positive and helped me throughout this journey. I would also like to thank Dr. Gaurav Singhal and Dr. Anurag Goswami for sharing their experience and guiding me from time to time whenever needed. I am incredibly grateful to Dr. R. K. Shevgaonkar (Ex Vice-Chancellor, Bennett University, Greater Noida, India) and Dr Prabhu Aggarwal (Honorable Vice-chancellor, Bennett University) for creating an academic and research environment at Bennett University. I wish to thank Dr. Ved. P. Mishra, my elder brother and uncle Dr. K. N. Mishra, for all their support and motivation they have given to me throughout the journey of research. I am also grateful to my elder brother Rajani Kant Mishra, who always stood with me in up and downs and supported me every time I felt low throughout this research journey. I would like to thank my wife, Pooja Mishra, for her continuous support and encouragement throughout this research. I would like to dedicate this thesis to my parents Mr. Sudhakar Mishra and Smt. Vidya Devi who have constantly encouraged me to achieve new height. Lastly, I would like to acknowledge all the people who have been of help and assisted me throughout my research.

Synopsis

on

Automation of Drone Surveillance for Search and Rescue

Submitted By

Balmukund Mishra

(Admission No. E17SOE801)

Under the Supervision of

Prof. Deepak Garg and **Dr. Pratik Narang**

Department of Computer Science Engineering
Bennett University, (The Times Group)
Greater Noida, INDIA 201310

Contents

1. Introduction

Search and rescue (SAR) is an important application of disaster management aims to provide assistance to humans stuck in calamities. Recent natural calamities such as the earthquake in Nepal's, Tohoku, and Haiyan, or floods in Europe and India have shown that local civil officials and emergency services are having difficulty managing the crisis properly[1]. The search for human survivors on the site is paramount for rescue services. Previous search and rescue (SAR) attempts are mainly volunteered specific and hence a time consuming process. A Human-operated SAR mission is a complicated and harmful job that may even cause the human crisis executives to lose their lives.

The effort to identify the disaster victims as early as possible is always a race against time in disaster response. Rapid scanning of impacted areas through drones equipped with high definition camera and navigation system can save human lives in disaster without taking the risk of volunteers' lives. The drone could be a powerful tool for the SAR because its capability is increasing day by day. The drones are now also accessible for civilian uses. More investments are coming in drone technology these days, and research is going on to develop the drone-based application, which can help humans ease their life [2, 3, 4]. Different SAR methods are available in the literature, ranging from volunteers-based SAR to fully automated drone-based SAR. However, the type of solution required is solely based on the geographical location and disaster type. For example, the robot-based SAR can only benefit in case of indoor disaster. The type of SAR technique needed depends on the disaster's geographical position, and for example, the SAR robots are less efficient in hills. In the category of surveillance and rescue through an aerial vehicle, the first rescue operation was performed on 29 November 1945 by Sikorsky's chief pilot Dmitry "Jimmy" Viner, in the cockpit[5].

1.1. Research directions and applications

Drone-based surveillance application is helpful for almost every type of disaster and in different terrain also. It can cover a wide remote area quickly and provide instant help to disaster victims. In developing the drone-based application, training the machines on such images and videos could be more realistic. It can help for identifying those places where rescue is required. Here, our research direction is to find the computer vision solution by providing the surveillance system with a basic understanding of natural human actions and their instances in a disaster-prone area. Our research focuses on using the basic AI technique such as object classification, object detection, human action classification, and emergency text classification.

1.2. Motivation and research goal

One of India's biggest challenge is Natural Disaster, where thousands of lives come to an end every year by the floods and heavy snowfall. Most of these places where such natural disasters happen are remote and less accessible, where providing help is a difficult task. To access those areas quickly and providing help to stuck people should be a primary goal to

save their lives. The first and most critical step in assisting disaster victims is to determine where support is needed.

The main factors that motivate researchers to develop drone-based surveillance applications are its capabilities, such as increasing payload and quick scanning of a wide area. A drone equipped with a camera and sensor can collect a large amount of information quickly. Search and rescue are usually initiated by the army or through volunteers after the disaster happens. Once the places are identified where rescue is required, usually they shift the stuck people to a safer place. Giving a faster response to the people stuck in the affected area can save more lives. By looking at the recent disasters and the loss caused by them, there is a need for an advanced SAR technique that can ease this process and reduce the disaster relief response time.

2. Related work and research methodology

Various strategies in the area of search and rescue have been tried in the past. In this section, we provide a brief overview of related SAR approaches as well as strategies for automating drone-based catastrophe surveillance. This section also gives an overview of challenges identified in literature and the objectives set for our research. The research methodology is also discussed in the end.

2.1. Existing research work

Due to the hasty process of evacuation and rescue, ordinary life becomes stressful and painful. Recently some article has shown the path to quickly acquire the information about disaster, and [7] uses the social media analysis such as tweets for quickly information gathering. Multiple methods are accessible and used for post-disaster rescue. The robot is one of the critical tools for this. A fully independent indoor and outdoor rescue UAV robot is offered by [8]. For fire prevention and safety in a home environment, an IOT based model is developed in[9]. In addition, there is a wide range of studies on rescue and navigation in complicated GPS and unknown environment exists such as [10] and [11]. In addition to robotics, significant study attempts were made to decrease the adverse effects on human society and the environment of natural disasters. To strengthen disaster management techniques, to address extreme environmental circumstances, and perform quick rescue work [12], UAV's were implemented to disaster management. A communication protocol proposed in [13] for supporting the communication and identifying the needy people though the network communication happening in the affected area. Here a drone-based IoT environment, called ACPBS-IoT is proposed. It could also be utilized for the remote area surveillance for identifying the help situations, and could be research direction for search and rescue.

In [15], a wireless network on UAV is suggested to extend wireless devices coverage. In this, a wireless sensor network (WSN) node deployed in the drone is suggested to extend the network's coverage for remote areas. In addition to this, [16] proposed a technique to design the trajectory of aerial vehicle, which can help to design customize drone with camera fitted on UAV drone. The author in [17] analyzed the impacts of variables influencing UAV-based SAR systems efficiency and and its optimization requirements is identified. In [18], a multipurpose UAV was proposed for rescue operations in the mountain. Developed UAV is have the facility to capture high definition visual and thermal image. This multi-rotor drone is intended to satisfy environmental demands such as low temperature, high altitude, and powerful winds for mountaineers terrain. This less than 5 kg capacity drone is capable of fully automatic landing and take-off operation.

In the event of a crisis, the drone can assist the rescue team, which is critical for a successful disaster management plan. An efficient disaster management have the basic steps such as collecting the disaster data, identify the victims location, and rescue the victims earliest[19]. The drone is useful in the whole emergency relief process. The medium and small-scale UAVs can be deployed and integrated for the police departments' rescue work, the fire brigades, and catastrophe response. Small scale UAVs like a drone is now very popular and valuable for these kinds of tasks. Involving the technical aspects with the

drones, small-sized devices like Raspberry Pi and NVIDIA Nano Jetson [20] [21]are helpful. It uses the latest computer vision technique for automatically identifying the rescue places. Developing drone based solution for surveillance leads to various potential attacks also, as the nature of communication in drone is wireless and specially in remote areas makes it easier to attack for attackers. To protect the drone from such attacks, [14] proposed a private block-chain based mechanism for secure communication in IoT based UAV devices.

Text recognition aims to take the image and identify the single word depicted inside it. In the initial days of text extraction, usually, rule based system is followed. [22] presents the comparison of ruel based system and a supervised ML model trained on large dataset for document classification and postal address classification. There are few techniques available in literature of handwriting or historical document recognition [23, 24, 25, 26]. Usually, these models are not generalized and cannot give the best result for problems like text extraction or generic scene text extraction [27, 28, 29, 30]. The main challenge of text extraction from the generic scene is the variable foreground and background texture. For scene text recognition, these methods can be categorized into mainly two groups which is character-based recognition and word-based recognition. Character-based models utilize the classification method per character to produce complete word recognition throughout the word picture. In another work [31], by clustering sub-patches of characters, it can learn a collection of mid-level characteristics, where characters are identified by random forest classifiers strokeletal and HOG characteristics.

2.2. Research gaps

Following gaps are identified in the existing research based on the literature review:

- Deficiency of Drone-based datasets:
- Having very little work in the field of combining technology with disaster management techniques:
- Less automation work in the field of aerial drone surveillance:

2.3. Research objectives

This research work addresses the novel work related to disaster management and saving humans life in disaster using technology. After a comprehensive literature review, the research gaps is identified. Our ultimate aim is to build an accurate system for identifying victims in disaster for rescue. To achieve this goal, the following are the objectives of our research:

- To analyze existing search and rescue techniques.
- Acquisition and creation of data for search and rescue.
- Designing the action classification based searching of human in drone surveillance.
- Suggesting a different solution for the search operation in disaster using computer vision and AI techniques.

2.4. Research methodology

To achieve the research objectives discussed earlier, the following research methodology is adopted:

- Study of existing techniques: Comprehensive survey of existing disaster management techniques. Also, an extensive survey of drone-based applications and its capability for performing search and rescue.
- Data collection: Acquisition and creation of surveillance data for search and rescue in terms of emergency text classification dataset, human detection dataset and human action in emergency-situation of disaster.
- Installation and study of the tool(s): Python toolkit with NumPy, OpenCV, packages, and TensorFlow and Keras packages for back-end implementation of the existing/proposed techniques.
- Propose a novel framework for action classification based search and rescue:
- Propose a novel framework based on human detection and their action classification for background-invariant search and rescue application:

3. Drone-Surveillance for Search and Rescue in Natural Disaster

The idea of drone surveillance for SAR is to use the drone to scan the affected area with a camera and model deployed on the drone to identify the exact places where help is required. The recent success of deep-learning approaches for object detection and action recognition motivates us to apply them in the drone-surveillance. The essential part of a deep-learning approach is that a significant amount of data is needed for training. Since most of the literature data are for ground-level surveillance, such as UCF[32], which does not help train the deep-learning model of aerial surveillance. Hence, it is our primary objective to develop a dataset of aerial action recognition for SAR. Besides, deep-learning models use these datasets for training in different tasks such as classification and localization. Deep-learning models used for these tasks can automatically extract the feature. Out of all other neural networks used for classification or localization, convolution neural networks (CNN) suits more for image-based feature extraction. The classification problem of images is mainly to classify the image into a different category (labels). The detection objective is to identify the object's label and determine the exact position of classified labels in that image.

As the dataset plays a crucial role in the model's performance, we have proposed a unique aerial action recognition dataset for SAR. Also, in aerial surveillance, since human appears very small and existing algorithms cannot identify the action performed by them, we propose a modified action detection model for aerial action detection.

3.1. Proposed framework

Proposed architecture with the developed dataset for action detection in drone surveillance can identify situations where humans ask for help. The proposed dataset is generalized and has enough variation to be used to automate such an application. The proposed model use the feature of multiple levels of convolution for the classification of the object. It has been observed that the lower layer feature contribute more to the classification of the small object. In the case of drone surveillance images taken from the top, the object appears very small, and their features are less explored for classification. The proposed model utilizes the initial convolution layer feature, which improves the proposed model for action recognition in drone surveillance.

The proposed model for action detection is inspired by the pyramidal feature extraction[33] and utilization for classification and localization. We have experimented with various convolution networks such as VGG16, Inception, and various layer features for the localization. After analyzing these experiments, we found that the detection network with the Inception network as a classifier performs better. In the proposed architecture of the action detection model, a feature of the 3rd convolution layer is fed directly to the detection generation, which is a key factor in performance improvement. Also, four extra convolution layer is used after inception network where the features are fed to the detection generation in a pyramidal way. These different convolution networks' parameter value is the same as the convolution layer in the Inception network. Details of hyper-parameters of the proposed model are given in Table 1. Our proposed model is trained for two class and six class (both) dataset, with activation function as relu. Other parameters are used after optimization as given in Table 1 and details of results are given in section3.

Table 1: Details of hyper-parameters and their values in final trained model (Proposed model)

Hyper-parameters	**Values**
Number of classes	2 and 6
Activation	relu
Batch-normalization	Yes
IOU	0.5
Batch size	24
Optimizer	rmsprop
Momentum-optimizer value	0.9
Initial learning rate	0.004

Table 2: Summary of proposed six-class action dataset

Feature	**Values**
Number of Actions	6
Number of Images	7000
Average instance per class	1000
Frame rate	60 fps
Resolution	1920*1080
Camera motion	Yes, slow and steady
Annotation and its format	Yes, Bounding box, .xml format

Table 3: Performance of deep learning models for action detection on proposed two-class aerial action dataset

Models	**mAP @ 0.50 IOU**
Faster R-CNN	0.988
R-FCN	0.97
Proposed model	0.98

3.2. Dataset summary

There are many datasets available for aerial object detection and action recognition in the literature. The best-suited data for our drone-based action recognition is the Okutama action dataset [34], however, is not helpful for real-time application. Existing dataset for human and their action detection is very complicated and take from more than 65-meter height. Results are not satisfactory with the dataset, as shown in our experiment and state of the art. So, for SAR and other surveillance applications using a drone camera, this dataset is not applicable. Here, we have collected the drone surveillance dataset for search and rescue and annotated into tow class, six class, and human detection dataset.

3.3. Results and analysis

Different existing object detection model for human detection and action recognition was applied on the Okutama dataset and our proposed dataset. Table 3 shows few results of our experiments performed to validate the proposed data and model in this work. The performance is evaluated on a standard coco evaluation metric (mAP). The proposed dataset for six class-action annotation is represented in Table 2. Our experimental result on this dataset shows the validity of the proposed model, where it achieves a comparable performance to the prior object detection model Faster RCNN. Comparing the inference time of Faster RCNN and the proposed model, the proposed model performs better as it is from the family of one-stage object detection techniques having very little inference time comparatively.

4. A Hybrid Approach for Search and Rescue using 3DCNN and PSO

At the time of drone surveillance, some critical human action can be seen as an indication of emergency, and recognizing these actions could help the automation of search and rescue. The vision-based system, high definition camera, and advanced deep-learning algorithm are three essential pillars in a robust drone-based surveillance system. Here, we use a video-based scene classification approach for emergency scene classification. Also, since we are working with the videos, 3D models are required for feature extraction and classification. However, in 3DModels, the number of trainable parameters is comparatively high, and it takes extra effort to train models and tune the hyperparameters. Tuning of the proposed network becomes more difficult as the number of hyperparameter in the network increases. Here, we tried a heuristic-based approach for hyperparameter tuning. In our research, we used a PSO based heuristic search technique that tries to imitate the travels of the flock of birds aiming at finding food [35]. It is based on a population of particles flying through a multi-dimensional search space in which each particle possesses a position, and velocity [36]. Both variables are changed to emulate the social-psychological tendency to impersonate other individuals' success in the population. Using PSO to optimize hyper-parameters tuning, especially for the 3D models where the number of parameters is higher than any other deep-learning model, can speed up the training and quickly find the best parameter values. The proposed approach aims to identify the scene containing single or multiple humans waving their hands in the drone's direction as a help situation. Waving hand in Drone surveillance from the top angle is an activity where humans' feature is not completely visible in a single frame, hence it is required to make a decision based on multiple consecutive frames taken from different angles. Our developed dataset includes video clips that cover the area from multiple heights, angles, and situations.

4.1. Proposed framework

The proposed approach for help-situation identification uses the idea of drone surveillance in disaster. It uses the proposed deep learning model for scene classification to recognize the help situation using the proposed model. The proposed model combines 3D convolution and a fully connected neural network. In this, we have used five 3D-convolution layers along with the two fully connected layers in the end. Simultaneously, the onboard device sends the GPS location to the central help desk for humans' real-time rescue. Framework for developing such a model to automate the drone surveillance application consists of several steps, such as dataset collection, pre-processing, model development, training, testing, and hyper-parameter tuning. However, all these steps inherently applied in association with PSO. The proposed idea for help-situation identification identifies a person with a waving hand in the drone's direction. So, for the automation of drone surveillance in disaster, a dataset of two classes is developed. In Help class video clips, single or multiple humans were waving hand, while in other classes, waving hand action is not available.

4.2. Dataset summary

We have recorded the dataset on four different days. The total duration of the video recorded was 150 minutes. Out of that 150-minute video captured on a different day and

Table 4: PSO-tuned hyper-parameters final value

Particle	Final-output
Activation Function	relu
Padding	same
Pooling	max-pooling
Optimizer	adam
Number-of-convolution-layers	5
Batch-size	8 videos

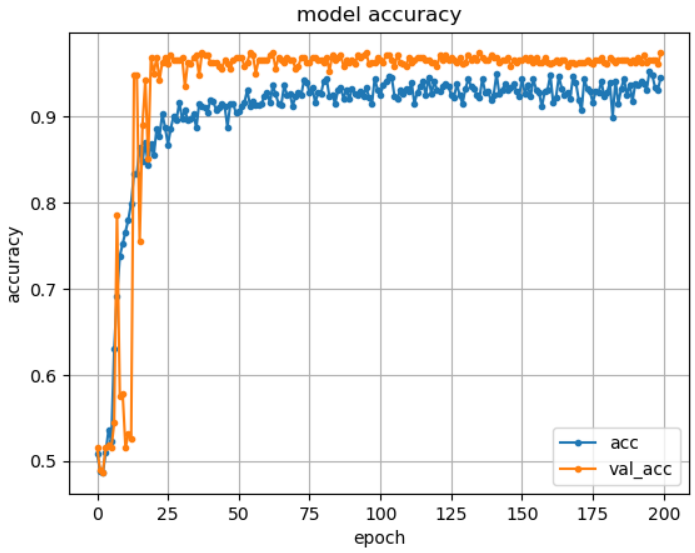


Figure 1: Loss comparison of training and validation for proposed model after data-augmentation

from a varying height. After cropping, we have selected 100 video-clip in each class. When selecting these clips, it has been taken care that clips should have minimum overlapping, and thus their feature should contribute maximum in the scene classification. Furthermore, each of our 1-second video clips comprises 60 separate frames, allowing us to play with each frame's differing number of frames to train the models.

4.3. Result and analysis

The objective of all the experiments shown in the previous section is to find a situation where help is required using computer vision and AI techniques in drone surveillance. Our initial experiment shows the need for improvement in terms of dataset and models. The proposed model's hyperparameter is then used with the aid of the heuristic search technique PSO, and the dataset size is expanded with the help of augmentation. The final experiment shows the validity of the proposed model and achieves an outstanding performance of 98% validation accuracy with the developed scene classification dataset. Figure 1 shows the performance of the proposed model.

5. Enhancing Emergency Text Classification in Drone Surveillance using Deep Learning

In this research, a text recognition based approach is developed for the automation of drone surveillance. Text recognition has improved dramatically in recent years, thanks in large part to the use of deep models [37], [38], [39],[40] and large datasets [41], [42]. In addition, a variety of 2D CNNs[43], [44], [45] have been designed for text classification and scene classification task, these 2D CNNs cannot directly utilized for aerial and drone surveillance. The inherent complexity of aerial and drone images in terms of features as per the height, angle, and size of the text makes it more challenging for recognition. Besides, the availability of appropriate data for training the network is an integral prerequisite of the deep learning algorithm. No such dataset was available in the literature for emergency text recognition. So, in this research, a dataset is developed to automate search operations in disaster. The generated dataset contains various images having the emergency text written on walls, roof, and sand. In this research work, we have suggested a solution for extracting emergency text from the input drone image and using the proposed convolution neural network to identify the scene as a Help or Non-Help scenario.

5.1. Proposed framework

The proposed system is conceptualized as a binary classification system, where a text dataset is developed, and a binary text classification model is developed. In this research work, a dataset is explicitly developed to train the emergency text classification. The proposed model is a convolution neural network-based binary classification model. The trained model is to be used to predict the input image between two different labels, Help and Non-Help. A lightweight model is expected due to the requirements of the embedded devices integrated with a drone. Our goal is to construct a model that requires little computation at runtime and provides real-time prediction precision for aerial emergency text recognition.

5.2. Dataset summary

For this research, a dataset is developed of two classes, Help and Non-help. The proposed dataset is an image dataset, and in Help class, each image contains text such as "HELP", "SOS", and "Emergency" written on the wall, roof, and ground. The proposed dataset has approximately 6000 images collected from different backgrounds such as beachside, hilly area, natural land and farming, and disasters such as earthquake and flood.

Table 5: Comparison of proposed model with VGG16 and InceptionResnetV2

Model	**Accuracy**	**No. of param**	**Precision**	**Recall**	**f1 Score**
VGG16	0.50	13,83,57,544	0.86	0.87	0.8649
InceptionResnetV2	0.99	5,43,36,736	0.875	0.862	0.864
Proposed model	0.999	1,40,40,698	0.90	0.857	0.877

5.3. Result and analysis

Several experiments were conducted on the proposed dataset and pre-trained text classification models using transfer learning. However, based on the results, we have selected a few best-suited versions, and the outcome of our final experiment is seen in Table 5. VGG16 and InceptionresnetV2 model is one of the best classification models and gives previous results in approximately every classification application type. Our experimental result shows the validity of the proposed model with comparable performance in terms of validation accuracy having approximately ten times faster inference speed as the number of the proposed network's trainable parameter is comparatively low.

6. Background Invariant Faster Motion Modelling for Drone Action Recognition

Video action recognition have got tremendous performance improvement in recent year and is mainly due to the addition of deep learning action recognition models [46],[47],[48],[49] and large-scale video databases [50],[51],[52]. In addition, many prominent convolution neural networks[53],[54],[55],[56] have been proposed for image recognition task. However, these CNN's can not model the motion feature of each individual effectively from crowd video. In particular, using these CNN's for aerial action recognition can provide a variety of real-life applications using the dataset proposed in [57],[34]. Further, aerial and drone surveillance can be applied to many real-time applications such as unusual activity detection in border area[58], violence and suspicious activity recognition in crowd[59], urban and rural scene understanding. However, due to the intrinsic complexity of aerial footage, motion modelling in action detection remains a difficult task. One of the major concerns in aerial or drone surveillance is the varying nature of human features from different angles and heights.

In actual implementations, significant visual differences in features, such as occlusions, pose variations and lighting adjustments, impose significant aerial surveillance challenges. The deep learning models use human features to learn the shape, texture, and size in the spatial-temporal domain for action classification. In the videos directly captured from the drone, such features are not readily evident. For the total exposure of human action features, robust temporal modelling of each human is required. This research work contributes to multiple ways to addressing these difficulties in drone surveillance, including a novel architecture, a quick and precise temporal motion modelling system, and an improved temporal network for action recognition.

6.1. Proposed framework

The proposed approach uses drones to monitor remote areas and be capable of being applied for various surveillance tasks. In this research, an unified trainable system is developed for faster motion feature modelling and action recognition from crowd surveillance videos. This approach combines two different modules. Hence, the architecture and its work are presented with the help of each module's mathematical background. Some necessary notations used here is as follows: each video clip is represented by X with the formula $X_i = I^{c*l*h*w}$. Where $I = f(x, y)$ represents the frames of the video clip, c stands for the number of channels, and l represents the frame number in each video clip. Here, h and w represents the height and width of each frame.

6.1.1. (Faster motion feature modelling (FMFM)

For the accurate and faster motion modelling of temporal features, multiple modules are interconnected and trained together. It combines deep learning model object detection and traditional computer vision algorithms to extract detected bounding boxes. Based on the image similarity, an individual human's temporal feature model is extracted from the original crowd monitoring video. For this, the initial step is detecting the human precisely. After the humans are detected accurately, the output image is passed to the second part of FMFM, which extracts the detected human, calculates image similarity, and patches most similar extracted images together in the form of video-clip.

Table 6: Incremental analysis of modules directly applied for action recognition in proposed action dataset

Network	Validation Accuracy
Faster RCNN Inception	0.39 mAP
SSD Inception	0.23 mAP
Proposed model for action recognition	0.85
Combined architecture proposed for action recognition	0.90

6.1.2. Accurate action recognition (AAR)

Here, the primary objective is classifying the drone action, and it is a multi-class activity recognition problem. Usually, in drone surveillance, the height and angle of surveillance change with time, making it more difficult. Identifying a specific action from the image, captured from a drone where multiple humans are present, and all may be doing different activities is quite challenging and requires detecting each action individually. Therefore, the proposed architecture uses two different modules FMFM and AAR, which can detect each action accurately. However, in drone monitoring, action recognition may also be accomplished using only spatial features. We can combine spatial feature with motion feature and accurate action classification. We have performed various experiments with the existing models in both categories and the combination of different architecture such as 2D CONVNet with RNN and 2D CONVNet with Time-Distributed RNN and with the proposed architecture using 3D CONVNet. However, for recognizing few actions required for emergency identification in disaster, motion features are required, and hence we will stick with the models using spatial and motion features together. For example, in our case, if we have to identify the waving hand, it requires spatial and temporal features.

6.2. Result and analysis

Table.6 shows the performance of object detection models applied for action recognition. However, it achieves only 0.39 mAP with the existing model faster RCNN. However, when applied for human detection, it can detect the human with a 0.833 mAP value. The second module proposes an action recognition model and is also tested independently on the proposed action recognition video dataset. The maximum accuracy we have got with the proposed model is 0.85, which is validation accuracy with a training accuracy as 0.89. The detailed comparative result is represented in Table6. While testing the architecture with validation video, the architecture has achieved maximum accuracy with 0.90 validation accuracy and is represented in Table6. Through our experimental result and the incremental analysis of each module, it has been observed that using both proposed modules (FMFM and AAR) together could achieve higher accuracy than any other combination.

7. Summary and Conclusions

In this research, deep-learning-based SAR algorithms are developed to be applied in drones for surveillance in a natural disaster. The developed algorithm can work from up to 50 meters of height from the ground. In search and rescue, searching for exact places takes most of the time. Hence, the proposed automated surveillance can save millions of lives. In case of disaster, a well-known quote is, "If we can reduce one minute in disaster, it could save millions of lives". However, aerial and drone surveillance puts a particular challenge for computer vision and deep learning automation techniques. One of the critical challenges is that the varying height of drone at the time of surveillance makes the human features invisible sometimes for prediction. This research has addressed these challenges and proposed various architecture, dataset, and computer vision algorithms. Our experimental result validates the proposed models and architecture. Overall, this research proposes three techniques to identify the exact places where rescue and relief are required based on human action recognition, scene classification, and emergency text classification.

7.1. Scope for future research

- The proposed methods for action recognition is applied on a low height aerial dataset. However, our target will be to develop such algorithms that can work with more than 50-meter height videos in the future.
- Specifically, for the real-time search and rescue, the developed algorithms should work altogether and integrated into the drone. In future, we aim to develop a unified drone fitted with a much stronger algorithm and work together for a real-time search operation in disaster.
- In this research, we have included the only human search for rescue, which can be extended for other living beings also in future.

References

[1] IFRC, "Resilience: Saving lives today, investing for tomorrow," *World Disasters Report*, 2016.

[2] S. Siebert and J. Teizer, "Mobile 3d mapping for surveying earthwork projects using an unmanned aerial vehicle (uav) system," *Automation in construction*, vol. 41, pp. 1–14, 2014.

[3] A. L. Fall, *Assistive Drone Technology: Using Drones to Enhance Building Access for the Physically Disabled*. PhD thesis, University of Cincinnati, 2018.

[4] J. T. K. Ping, A. E. Ling, T. J. Quan, and C. Y. Dat, "Generic unmanned aerial vehicle (uav) for civilian application-a feasibility assessment and market survey on civilian application for aerial imaging," in *2012 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT)*, pp. 289–294, IEEE, 2012.

[5] S. P. HISTORY, "SIKORSKY PRODUCT HISTORY."

[6] B. V. Salazar, "Murphy and robots featured on TED Talk."

[7] M. Y. Kabir, S. Gruzdev, and S. Madria, "Stimulate: A system for real-time information acquisition and learning for disaster management," in *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pp. 186–193, IEEE, 2020.

[8] T. Tomic, K. Schmid, P. Lutz, A. Domel, M. Kassecker, E. Mair, I. L. Grixa, F. Ruess, M. Suppa, and D. Burschka, "Toward a fully autonomous uav: Research platform for indoor and outdoor urban search and rescue," *IEEE robotics & automation magazine*, vol. 19, no. 3, pp. 46–56, 2012.

[9] F. Saeed, A. Paul, A. Rehman, W. H. Hong, and H. Seo, "Iot-based intelligent modeling of smart home environment for fire prevention and safety," *Journal of Sensor and Actuator Networks*, vol. 7, no. 1, p. 11, 2018.

[10] M. Nieuwenhuisen, D. Droeschel, M. Beul, and S. Behnke, "Autonomous navigation for micro aerial vehicles in complex gnss-denied environments," *Journal of Intelligent & Robotic Systems*, vol. 84, no. 1-4, pp. 199–216, 2016.

[11] A. Bachrach, R. He, and N. Roy, "Autonomous flight in unknown indoor environments," *International Journal of Micro Air Vehicles*, vol. 1, no. 4, pp. 217–228, 2009.

[12] D. Hein, S. Bayer, R. Berger, T. Kraft, and D. Lesmeister, "An integrated rapid mapping system for disaster management," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, p. 499, 2017.

[13] B. Bera, A. K. Das, S. Garg, M. J. Piran, and M. S. Hossain, "Access control protocol for battlefield surveillance in drone-assisted iot environment," *IEEE Internet of Things Journal*, 2021.

[14] M. Wazid, B. Bera, A. Mitra, A. K. Das, and R. Ali, "Private blockchain-envisioned security framework for ai-enabled iot-based drone-aided healthcare services," in *Proceedings of the 2nd ACM MobiCom Workshop on Drone Assisted Wireless Communications for 5G and Beyond*, pp. 37–42, 2020.

[15] A. Guillen-Perez, R. Sanchez-Iborra, M.-D. Cano, J. C. Sanchez-Aarnoutse, and J. Garcia-Haro, "Wifi networks on drones," in *2016 ITU Kaleidoscope: ICTs for a Sustainable World (ITU WT)*, pp. 1–8, IEEE, 2016.

[16] A. R. Mekala, S. Madria, and M. Linderman, "Aerial vehicle trajectory design for spatio-temporal task aggregation," in *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 1085–1094, IEEE, 2016.

[17] S. Waharte and N. Trigoni, "Supporting search and rescue operations with uavs," in *2010 International Conference on Emerging Security Technologies*, pp. 142–147, IEEE, 2010.

[18] M. Silvagni, A. Tonoli, E. Zenerino, and M. Chiaberge, "Multipurpose uav for search and rescue operations in mountain avalanche events," *Geomatics, Natural Hazards and Risk*, vol. 8, no. 1, pp. 18–33, 2017.

[19] S. Karma, E. Zorba, G. Pallis, G. Statheropoulos, I. Balta, K. Mikedi, J. Vamvakari, A. Pappa, M. Chalaris, G. Xanthopoulos, *et al.*, "Use of unmanned vehicles in search and rescue operations in forest fires: Advantages and limitations observed in a field trial," *International journal of disaster risk reduction*, vol. 13, pp. 307–312, 2015.

[20] E. Li, Z. Zhou, and X. Chen, "Edge intelligence: On-demand deep learning model co-inference with device-edge synergy," in *Proceedings of the 2018 Workshop on Mobile Edge Communications*, pp. 31–36, ACM, 2018.

[21] Y. Ukidave, D. Kaeli, U. Gupta, and K. Keville, "Performance of the nvidia jetson tk1 in hpc," in *2015 IEEE International Conference on Cluster Computing*, pp. 533–534, IEEE, 2015.

[22] S. Chaturvedi, T. A. Faruquie, L. V. Subramaniam, and M. K. Mohania, "Estimating accuracy for text classification tasks on large unlabeled data," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 889–898, 2010.

[23] C. C. Tappert, C. Y. Suen, and T. Wakahara, "The state of the art in online handwriting recognition," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 12, no. 8, pp. 787–808, 1990.

[24] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in neural information processing systems*, pp. 545–552, 2009.

[25] D. Keysers, T. Deselaers, H. A. Rowley, L.-L. Wang, and V. Carbune, "Multi-language online handwriting recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1180–1194, 2016.

[26] A. Baldominos, Y. Saez, and P. Isasi, "Evolutionary convolutional neural networks: An application to handwriting recognition," *Neurocomputing*, vol. 283, pp. 38–52, 2018.

[27] Z. Selmi, M. B. Halima, A. Wali, and A. M. Alimi, "A framework of text detection and recognition from natural images for mobile device," in *Ninth International Conference on Machine Vision (ICMV 2016)*, vol. 10341, p. 1034127, International Society for Optics and Photonics, 2017.
[28] L. M. Francis and N. Sreenath, "Live detection of text in the natural environment using convolutional neural network," *Future Generation Computer Systems*, vol. 98, pp. 444–455, 2019.
[29] L. Moncla, M. Gaio, E. Egorova, and C. Claramunt, "An automatic extraction method of static and dynamic spatial contexts from texts," 2018.
[30] A. A. Chandio and M. Pickering, "Convolutional feature fusion for multi-language text detection in natural scene images," in *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pp. 1–6, IEEE, 2019.
[31] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4737–4749, 2014.
[32] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.
[33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.
[34] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 28–35, 2017.
[35] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-International Conference on Neural Networks*, vol. 4, pp. 1942–1948, IEEE, 1995.
[36] V. K. Mishra and A. Sengupta, "Mo-pse: Adaptive multi-objective particle swarm optimization based design space exploration in architectural synthesis for application specific processor design," *Advances in Engineering Software*, vol. 67, pp. 111–124, 2014.
[37] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.
[38] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "Aon: Towards arbitrarily-oriented text recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5571–5579, 2018.
[39] C. Luo, L. Jin, and Z. Sun, "Moran: A multi-object rectified attention network for scene text recognition," *Pattern Recognition*, vol. 90, pp. 109–118, 2019.
[40] Y. Lu, J. Lu, S. Zhang, and P. Hall, "Traffic signal detection and classification in street views using an attention model," *Computational Visual Media*, vol. 4, no. 3, pp. 253–266, 2018.
[41] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*, 2016.
[42] S. Inunganbi, P. Choudhary, and K. Manglem, "Meitei mayek handwritten dataset: compilation, segmentation, and character recognition," *The Visual Computer*, pp. 1–15, 2020.
[43] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *Proceedings of the IEEE international conference on computer vision*, pp. 5076–5084, 2017.
[44] F. Zhan and S. Lu, "Esir: End-to-end scene text recognition via iterative image rectification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2059–2068, 2019.
[45] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1547–1554, IEEE, 2017.
[46] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4768–4777, 2017.
[47] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and

robust motion representation for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1390–1399, 2018.
[48] Y. Wang, M. Long, J. Wang, and P. S. Yu, "Spatiotemporal pyramid network for video action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1529–1538, 2017.
[49] J. Li, X. Liu, M. Zhang, and D. Wang, "Spatio-temporal deformable 3d convnets with attention for action recognition," *Pattern Recognition*, vol. 98, p. 107037, 2020.
[50] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
[51] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov, and A. Zisserman, "The ava-kinetics localized human actions video dataset," *arXiv preprint arXiv:2005.00214*, 2020.
[52] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.
[53] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*, pp. 20–36, Springer, 2016.
[54] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.
[55] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 591–600, 2020.
[56] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with multi-stream adaptive graph convolutional networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 9532–9545, 2020.
[57] A. G. Perera, Y. W. Law, and J. Chahl, "Drone-action: An outdoor recorded drone video dataset for action recognition," *Drones*, vol. 3, no. 4, p. 82, 2019.
[58] S. J. Kim and G. J. Lim, "Drone-aided border surveillance with an electrification line battery charging system," *Journal of Intelligent & Robotic Systems*, vol. 92, no. 3, pp. 657–670, 2018.
[59] M. Li, M. O'Grady, X. Gu, M. A. Alawlaqi, G. O'Hare, *et al.*, "Time-bounded activity recognition for ambient assisted living," *IEEE transactions on emerging topics in computing*, 2018.
[60] B. Mishra, D. Garg, P. Narang, and V. Mishra, "Drone-surveillance for search and rescue in natural disaster," *Computer Communications*, vol. 156, pp. 1–10, 2020.

Biography

Balmukund Mishra was born on 20-July-1992 in Kushinagar, a district of state Uttar Pradesh, India. He received his bachelor's degree (B-Tech) from A.P.J, Abdul Kalam University, Lucknow, formerly known as Uttar Pradesh Technical University and master's degree (M-Tech) from Jaypee University of Information Technology, Waknaghat, Solan, HP, India with 9 CGPA.

Currently, he is a PhD Scholar in Bennett University since July 2017. Before coming to Bennett University, Mr. Mishra was Assistant Professor in Computer Science and Engineering Department, PIET Haryana affiliated under Kurukshetra University from July 2015 to July 2017. His research interest includes Aerial and Drone Surveillance, Computer Vision, Deep Learning, Image and Video Processing.