Satbayev University

UDC 621.397:004.932.72’1(043)

SEIDALIYEVA ULZHALGAS OMIRTAEVNA

Research of effective UAV detection using smart sensors

6D071900 – Radio engineering, Electronics and Telecommunications

Thesis for the Degree of
Doctor of Philosophy (PhD)

Supervisors
candidate of technical science,
associate professor
L.B. Ilipbayeva

doctor PhD,
Professor
E.T. Matson
(Purdue University)

Republic of Kazakhstan
Almaty, 2023

CONTENTS

NORMATIVE REFERENCES

This thesis uses references to the following standards:

Instructions for the preparation of a dissertation and author's abstract. External Attestation Committee of the Ministry of Education and Science of the Republic of Kazakhstan, 377-3 Zh.

GOST 7.32-2001. Report on research work. Structure and design rules.

GOST 7.1-2003. Bibliographic record. Bibliographic description. General requirements and compilation rules.

GOST 7.32-2017. System of standards of information, librarianship, and publishing. Research report. Structure and design rule.

DEFINITIONS

In this dissertation, the following terms are used with the corresponding definitions:

Artificial neural networks (ANN) are computing machines built on the basis of biological neurons for complex data processing and simulation of analytical operations.

Artificial intelligence (AI) is the replication of human intellectual functions by machines, particularly computer systems.

Background subtraction (BS) is a popular computer vision-based moving object detection technique that removes extraneous background from the image and takes just the essential details of the foreground object.

Computer vision (CV) is a branch of artificial intelligence, which allows machines to extract useful information from visual data such video sequences, images etc.

Deep learning is one type of machine learning that uses neural networks consisting of two and more layers for modeling and solving complicated tasks.

Drone is an unmanned flying object that moves autonomously in the air and does not require pilot control.

Image classification is another computer vision task of classifying and labeling sets of pixels or vectors in an image in accordance with established criteria.

Machine learning is a subfield of artificial intelligence that aims to build systems that can learn from the data they are fed.

Object detection is computer vision approach for finding and locating regions of interest in videos and images.

Radar is a reliable sensing instrument that determine the presence, distance, speed, and direction of any object.

RF sensor is an effective detection technique that detect objects through energy emitted by radio transmitters.

Sensor fusion is a technique of combining or integrating data from multiple sensors.

Transfer learning is the process of reusing pertinent elements of a machine learning model that has already been trained to address a new but related problem.

DESIGNATIONS AND ABBREVIATIONS

AI – Artificial Intelligence
ANN – Artificial neural network
AP – Average Precision
BN – Batch Normalization
BS – Background Subtraction
CNN – Convolutional Neural Network
CPU – Central Processing Unit
CS – Confidence score
CV – Computer vision
C-UAS – Counter Unmanned Aerial System
DL – Deep Learning
FAA – Federal Aviation Administration
FC – Fully connected
FD – Fourier descriptor
FN – False Negative
FP – False Positive
FPS – Frame Per Second
GAN – Generative adversarial network
GFD – Generic Fourier Descriptor
GMM – Gaussian Mixture Model
GPU – Graphics Processing Unit
GSM – Global Positioning System
HOG – Histogram of oriented gradients
HSV – Hue Saturation Value
IoU – Intersection over Union
KCF – Kernelized Correlation Filter
KSQ – Korean Square
LOS – Line of sight
ML – Machine Learning
MLP – gMultilayer Perceptron
NN – Neural Network
PBAS – Pixel Based Adaptive Segmentator
PBRT – Physical Based Rendering Toolkit
PTZ – Pan-Tilt-Zoom
RCS – Radar cross-section
ReLU – Rectified Linear Unit
ResNet – Residual network
RF – Radio Frequency
ROC-curve – Receiver operating characteristic curve
RGB – Red, Blue, Green
RPN – Region Proposal Neural Network
TN – True Negative

TP – True Positive
SGD – Stochastic Gradient Descent
SSD – Single-shot detector
SVM – Support vector machine
UAV – Unmanned aerial vehicle
UV – Ultra violet
ViBe – Visual Background Extractor
VGG – Visual Geometry Group
VS – Visible spectrum
YOLO – You Only Look Once
ZFNet – Zeiler and Fergus Network

INTRODUCTION

General characteristics of research. This work is aimed at research and development of an unmanned aerial vehicle (UAV) detection system based on smart sensors.

Relevance of the research topic. The best drone manufacturers on the market, like DJI, Parrot, and 3D Robotics, are constantly producing affordable and simple-to-use models of unmanned aerial vehicles (UAVs), also referred to as "drones," that can be used for a variety of legal commercial applications, including photography, first aid, agriculture, delivering packages, monitoring crowded places etc. However, the use of drones for illegal purposes, such as *smuggling* (transporting illegal substances at borders, in restricted areas, and prisons), *espionage* (illegal video surveillance of people, businesses, and government organizations), *collisions* with aircraft, drones loaded with explosives or chemicals, the use of drones for *attack* purposes [1] and other situations can cause serious problems for society. Prohibition of an unauthorized drone flight over the building of the Ministry of Defense of the Republic of Kazakhstan by the operational response group of the Military Police in March 31 of 2019 [2], confiscation of illegal transportation of the psychotropic drug “Tramadol” from the border of Kazakhstan to the border of Uzbekistan by Uzbek border guards in September 14, 2019 [3], for the first time in the history of the prison system of Kazakhstan transportation of prohibited items to the colony by an unmanned aerial vehicle (hereinafter UAV) in September 4 of 2020 [4], the arrest of a resident who launched an unregistered UAV over a military unit in Aktobe in June 13, 2022 [5] and other events that took place indicate that the careless or deliberate use of UAVs can pose a serious threat to the airspace of airports, power plants, civilians, organizations, and even the entire state.

The infrastructure may experience risky incidents such as information privacy violations, aircraft collisions, attacks on significant objects, allowing the transportation of illegal substances, etc. if the intrusion of drones into specially protected areas is not identified early on and stopped in time. In order to prevent such dangerous incidents, it is important to establish a reliable detection system that will detect drones in real time in the territory of important infrastructures. Usually, conventional radar and radio frequency technologies are frequently utilized in the preparation of UAV target detection and tracking systems; however, the accuracy of these sensors reduces when the UAV flies in the area where the signal is obstructed or the received signal is blocked. Due to their accessibility and relative accuracy in object detection from a sufficient distance, optical camera sensors are useful in the development of effective detection systems that identify the UAV as soon as it approaches the specially protected area and present the visual output result (bounding box) to security personnel in real time.

In the field of communications, some possible UAV incidents may also occur, such as:

- Failure of telecommunication systems due to interference from drones;

- Violation of safety rules when using drones can lead to accidents that can damage telecommunications equipment, such as communication towers or cables, leading to communication interruptions and network failures, as well as damage to equipment and endanger the safety of personnel.

- Unauthorized use of drones to perform espionage operations or store classified information transmitted over communication networks.

Thus, the development of technologies for detecting and preventing the use of UAVs is becoming relevant due to the increase in the number of dangerous incidents associated with drones in various fields. The above situations require a deeper study of drone detection and avoidance systems, which leads to the development of this research field.

Research problem. The task of real-time UAV object detection in accordance with camera system requirements while maintaining a balance in accuracy and speed is challenging due to the territorial size and location of the specially protected area and the fact that UAVs are moving objects and move quickly in frames, which makes the detection task more difficult. In order to ensure effective detection, an important requirement is that the model must be able to identify the drone from a distance, as it approaches the area of a specially protected infrastructure, respectively, the dimensions of the drone in the images are very small in terms of pixels.

Research aim. Research and development of a real-time UAV detection system using smart camera sensors.

Research objectives. The following tasks must be completed in order to achieve research aim:

1. In-depth literary review of UAV detection methods based on smart sensors.
2. Choosing a camera sensor, taking into account the territory of the special protected area and the camera parameters, fixing the position of the camera sensor to detect the object with sufficient accuracy.
3. Data preparation and pre-processing, which allows the neural network to identify drones more accurately.
4. Theoretical description of the proposed detection system.
5. Research and development of a real-time and accurate drone detection system with a static background.
6. Research and development of a multiple sensor fusion system to avoid blind spots and reduce drone confusion with birds.

The object of research is UAV detection system.

The subject of research are primary visual data preparation and video signal processing methods, the structure and algorithms of neural networks used in object detection and classification, sensor fusion methods.

Research methods. To solve the research tasks the following methods were solved: Digital signal processing methods, machine learning theory, object detection methods, image classification methods, as well as sensor fusion techniques etc.

The scientific novelty of the work. The scientific novelty of the research lies in the development of a smart sensor fusion system using voting method for multi-angle detection of UAVs.

The following scientific statements are to be defined:

1. Data acquisition, processing and preparation methods.
2. Moving object detection methods and algorithms.
3. Moving object classification methods and algorithms.
4. Sensor fusion methods and algorithms.
5. Experiments, results and discussions.

The theoretical significance of the research results. This research can be used as a methodological guide for camera sensor selection, data collecting and preprocessing, neural network model selection, and training by anyone wishing to do visual camera sensor-based UAV detection.

The practical significance of the research results. The proposed sensor fusion-based unmanned aerial vehicle recognition model serves as the basis for the future work of the researcher under the Zhas Galym project AP14971031 «Research and implementation of a bimodal system for real-time detection of unmanned aerial vehicles» (priority direction «9. National security and defense»). That is, it is the basis for the development of a bimodal system that combines LiDAR sensors and cameras in real time to detect unauthorized penetration of flying objects into specially protected infrastructure (Appendix A).

Personal contribution of the author. The dissertation is the original work of the author, all the results of scientific research are obtained by the author herself. Approval of tasks to achieve the goal of the study, analysis of research methods and implementation of the proposed system, analysis of the results of scientific research were carried out by the author herself and under the guidance of a domestic supervisor and a foreign scientific consultant.

The validity and reliability of scientific provisions, conclusions and recommendations are confirmed by publications in journals included in the list of scientific publications recommended by the Committee for Control in Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan, and the Web of Science and Scopus database; approbation in a foreign international scientific and practical conference.

Approbation of research results. The main scientific results of the dissertation research outlined in the dissertation are presented at the international conference «The Fourth IEEE International Conference on Robotic Computing (IRC)», and the results are published in the IEEE Xplore proceedings:

1. Detection of loaded and unloaded UAV using deep neural network // 2020 Fourth IEEE International Conference on Robotic Computing (IRC), Taichung, Taiwan, 2020, pp. 490-494, DOI: 10.1109/IRC.2020.00093. Conference Paper.

2. Deep residual neural network-based classification of loaded and unloaded UAV images // 2020 Fourth IEEE International Conference on Robotic Computing (IRC), Taichung, Taiwan, 2020, pp. 465-469, DOI: 10.1109/IRC.2020.00088. Conference Paper.

Publications. Based on the main scientific results of the dissertation work, 6 publications were published, including 1 journal article in publications indexed in the Scopus and Web of Science databases, 3 articles in publications recommended by the

Committee for Quality Assurance in Science and Higher Education of the Ministry of Science and higher education of the Republic of Kazakhstan and 2 papers in the proceedings of international conferences.

Acknowledgements. I would like to express my gratitude to professors Eric Matson, Lyazzat Ilipbayeva and Akhan Almagambetov for giving academic guidance and strong support throughout the PhD dissertation work. It was a real delight to work at KSQ lab with supportive colleagues and other fellow PhD students under Professor Eric Matson's guidance. My great appreciation to the SafeShore project for supporting Drone-vs-Bird challenge dataset. Finally, I would like to thank my best friend-motivator Dana Utebayeva and my family for supporting me morally throughout this long period.

Structure and scope of the dissertation. The thesis consists of 88 pages of typewritten text, including normative references, definitions, designations and abbreviations, an introduction, 5 primary chapters, a conclusion and the list of references. The primary chapters contain 63 figures and 10 tables, as well as 75 cited references. The work begins with an introduction part, where the author gives a general description of the work, the relevance of the study, the problems and provisions on defense. In **Chapter 1**, UAV object detection technologies are explained in detail, and a huge number of literature reviews have been conducted on UAV detection based on visual data recognition. **Chapter 2** describes the steps for preparing UAV visual data including video signal acquisition and digital image processing techniques. **Chapter 3** describes moving object detection and classification methods using for visual UAV detection and classification tasks. **Chapter 4** presents the proposed real-time drone detection system in the scene with a static background. **Chapter 5** presents the developed model and algorithm of a smart sensor fusion system using voting method for multi-angle UAV detection and classification. **The conclusion** discusses the analysis and outcomes of research work, its future directions.

1 BACKGROUND

1.1 Security threats of unmanned aerial vehicles

The flight of UAVs over the defense objects is a violation of the law "On the use of airspace and aviation activities of the Republic of Kazakhstan", as well as the registration of unmanned aerial vehicles in the state register is mandatory. Owners of UAVs in the Republic of Kazakhstan can obtain a written permission to use the airspace from the Civil Aviation Committee only after registering the UAV and indicating the direction of flight [4; 5]. Nevertheless, licensing of UAVs is unlikely to prevent the actions of terrorists. After all, although law-abiding citizens register and license their drones in the state register, it is quite possible that terrorists and other criminals buy UAVs abroad and continue to use them freely inside the country. Therefore, recently, the research on the UAV detection system has been developing rapidly.

In recent years, due to the continuous development of technology the technical capabilities of unmanned aerial vehicles (UAVs), publicly known as drones are being improved, as well as the scope of their application is expanding. Long-range flight, payload options, mobility and compactness have increased the potential use of drones from personal to military use [6]. Due to their low cost and ease of use, drones play an important role in modern life, becoming an assistant in delivering packages and medicines, filming various video content, monitoring crowded places, providing first aid, etc. Nevertheless, the problem occurs when drones are used in Illegal purposes posing a threat to society, such as smuggling (transporting illegal substances across borders, restricted areas, prisons), illegal video surveillance, drone collisions with flying aircraft, etc., which requires timely detection of UAVs entering protected areas for illegal purposes.

Considerate use of UAVs. Not all the UAVs pose a danger to society. For example, small-sized and shorter-range drones are often bought by young people and amateurs, as well as for entertainment purposes, such as shooting images and video content for social networks. Usually such short-range drones are controlled by smartphones via Wi-Fi, and this can limit their service radius to around tens of meters. As well as their small size limits the ability to transport the payload. And the poor wind resistance of some small-sized drones allows them to be used exclusively indoors.

Although such drones for image and video filming pose an issue in terms of confidentiality, they do not pose large-scale problems such as attacks, collisions or smuggling [7].

Malicious use of UAVs. UAVs are usually equipped with a video camera and an an image transmission system, and drones costing about $ 1,000 can stay in the air for half an hour and fly several kilometers away from the operator. This, in turn, allows terrorist groups to use the drone as a means of surveillance, collecting information from the intended object, as well as a vehicle for the delivery of explosives or poisonous substances.

In the middle of the 20th century, the United States first began using reconnaissance drones in Vietnam. However, only in the 21st century UAVs have

become full-fledged weapons. After the September 11 attacks, the United States began a large-scale fight against terrorists around the world. For terrorist groups, the tracking function is an important component as they first send a drone to collect information and then carry out an attack to create attack coordinates and organize an attack plan before the attack. Terrorists can destroy targets not by special military drones, but by loading a colorless toxic liquid such as sarin, which narrows the pupil and paralyzes the nerves, instead of a field chemical on an agricultural multicopter intended for the chemical treatment of ordinary fields, or by installing a firearm on a multicopter and firing accurate shots at the target [7].

Parameters of dangerous drones. The main parameters of dangerous drones used carelessly or intentionally (figure1):

– the ability to fly long-range distances;
– carrying capacity of heavy payloads;
– resistance to different weather conditions and wind;
– unsafe design of UAVs (UAVs with open propellers).

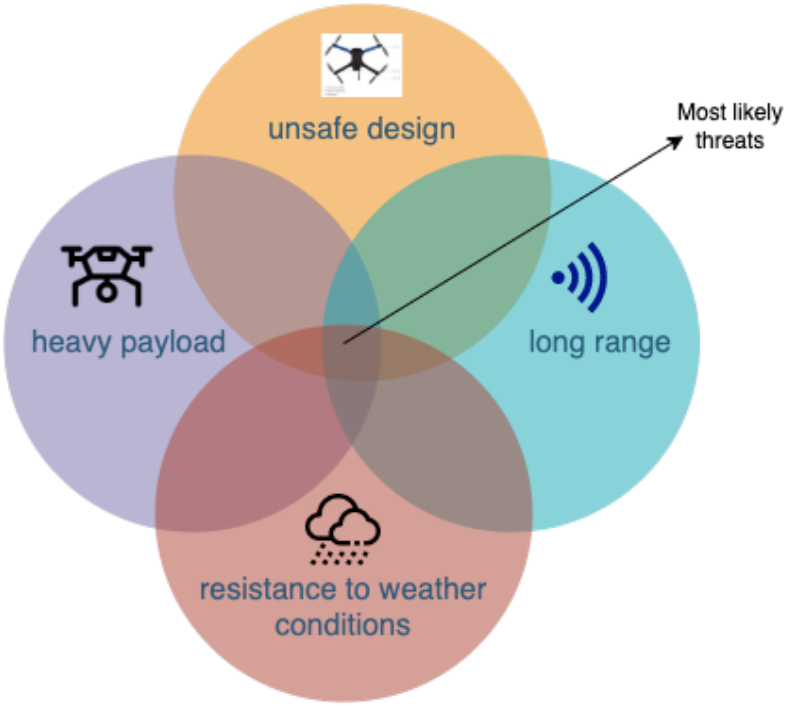


Figure 1 – Parameters of dangerous UAVs

UAV threats. The deep emphasis on the UAV threat categories allows us to better understand the factors that make up the unmanned aerial vehicles that pose the most danger to society. The threats posed by drones can be divided into the following important categories (figure 2).

The first category of potential threats is a drone attack. Since explosives, as well as biological and chemical weapons can be carried by unmanned aerial vehicles. Such explosives can be used to attack a number of targets, including individuals, organizations, and even nations. However, for such attacks, criminals would require

drones, which can accurately reach targets, fly long distances and with heavy payload capabilities to carry the necessary weapons [7].

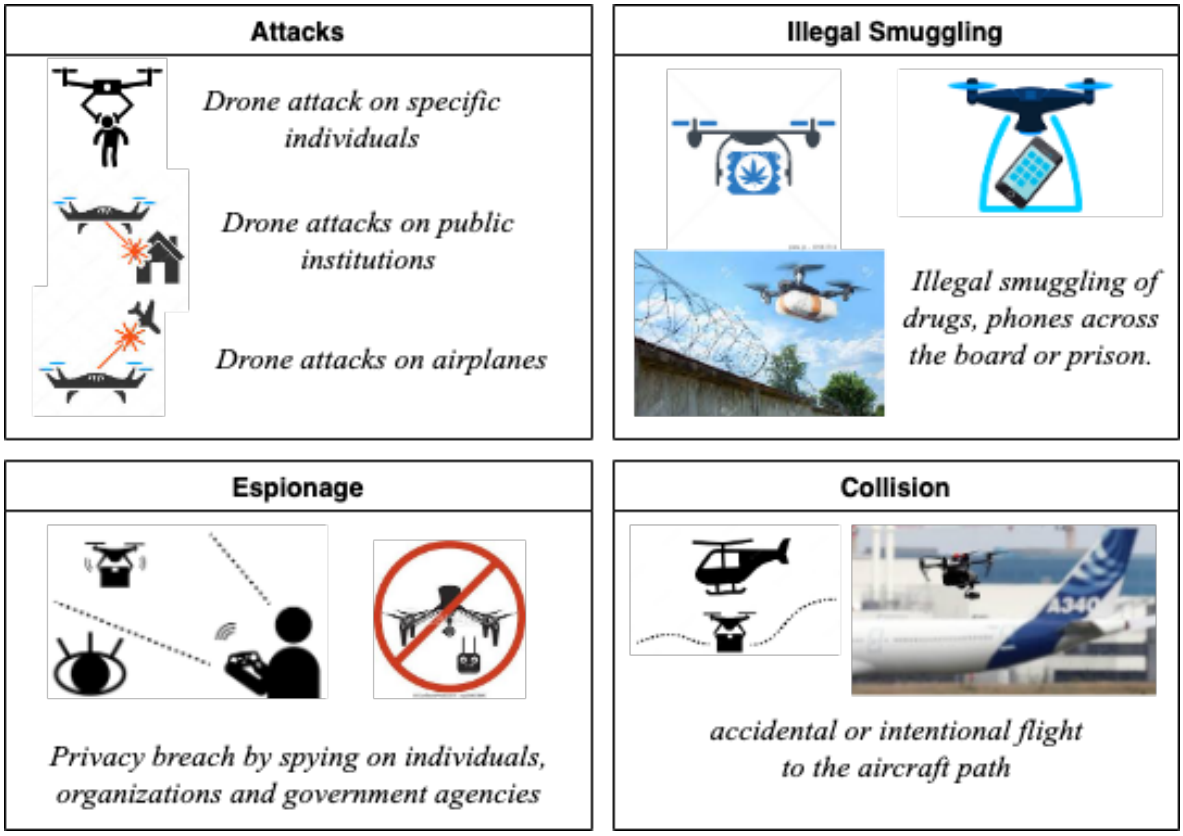


Figure 2 – The main UAV threat categories

Launching a remote-controlled drone, either accidentally or intentionally, in close proximity an aircraft or in its flight path may endanger the safety of passengers and crew, as well as cause property damage [7; 8]. In turn, this is the next category of threat from unmanned aerial vehicles, which is an example of a *collision*. It is also clear that the launch of UAVs directly into public places, such as a crowded stadium, is also dangerous, especially if the drone is large in size and heavy, or the drone is loaded with heavy explosives, posing a serious threat to society.

Drones equipped with powerful cameras, in turn, can be used to remotely spy on individuals, enterprise companies and government agencies. Contrary to privacy claims, this concern could be an example of drone threat such as *invasion of privacy* or *espionage*. To carry out such illegal acts, drones have to be able to spy into a window or gather information about an important object, as well as illegally film sports events. All of these activities require drones with embedded cameras, long-range and weather-resistant abilities.

Frequent incidents of drug smuggling using drones by creating a huge problem for prison officials and border patrols, as well as the smuggling of weapons or other illegal items without attracting the attention of security services can serve as an example of the next category of threats from unmanned aerial vehicles such as *smuggling*. Drones used for these purposes must respond to parameters such as the ability to carry heavy loads to their destination, fly long distances, and operate in

extreme conditions regardless of different weather conditions in the border territory [9].

1.2 Drone detection technologies

Drones have found innovative applications and use cases for everyone from children and enthusiasts to police officers and firemen. Given the widespread use of drones and their considerate and malicious potential, it is necessary to accurately detect and classify them as permitted or not, as well as track their route. The UAV detection system is usually deployed close to the region of interest. The detection system can monitor the drone as it enters or is expected to enter the no-fly zone and determine if it is a friendly drone or an unauthorized intruder. The system then has the option of alerting an operator or applying an automated policy. The frequent use of unmanned aerial vehicles (or drones) for commercial and entertainment purposes has recently stimulated research in the field of UAV detection systems. According to the industrial and academic studies carried out so far, UAV detection system uses non-optical technologies using radar, microphones for detecting acoustic vibrations, radio frequency (RF) sensors and optical technologies (visible spectrum and thermal infrared camera sensors) that detect drones based on features obtained from images and videos as shown (figure 3). As well as these drone detection sensors are divided into primary and secondary categories by the Federal Aviation Administration (FAA).

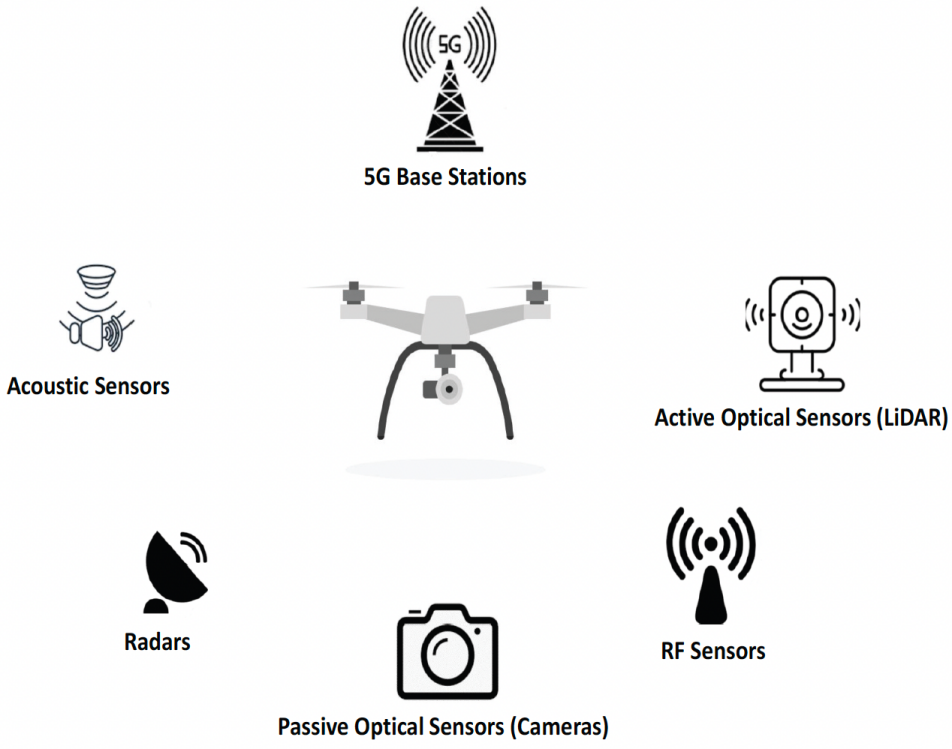


Figure 3 – Different UAV detection technologies

Note – Compiled according to the source [10]

Primary sensors, as a rule, are capable to detect drones with sufficiently high accuracy and low false positives to function as standalone (autonomous) solutions. That is, they will not need data from other types of sensors to confirm detection.

There are two types of primary sensors:

1. Radar.
2. Radio frequency (RF).

To reliably identify drone threats with a low false positive rate, secondary sensors require additional data from other kinds of sensors. These additional types of sensors can be used to improve detection accuracy or provide more detailed information about drone hazards, however, they cannot be used as standalone detection systems. Examples of secondary sensors are the following types of sensors:

1) acoustic;
2) optics (camera)/infrared.

The advantages and disadvantages of each of the listed drone detection systems will be listed below.

1.2.1 Non-optical drone detection technologies

Detection of flying objects based on Radar. In both military and civilian applications (such as aviation), radars (Radio Detection and Ranging) have historically been used for aircraft detection, as a result, they are frequently seen as a reliable sensing instrument that comes to mind when discussing UAV detection. The components of any radar system are a transmitter that generates EM (electromagnetic waves) in the radio or microwave spectrum, transmitting and receiving antennas, a receiver and processor that determine the objects' properties [9]. In a radar, a radio wave is transmitted (usually in the microwave frequency range), and the radiolocation system captures reflected radio wave from the physical object (i.e., aircraft or UAV). To determine the presence, distance (range), radial velocity and direction of object, the receiver examines the Doppler shift brought on by moving objects [9; 10-12]. The schematic structure of radar-based drone detection method is illustrated below (figure 4) [12, p. 42635-42658].

There are two types of radars: active and passive radars. To find objects, active radar broadcasts a signal and receives the reflected signal. In contrast, passive radar relies on other signal sources, including cellular signals, FM radio, analog television, and digital audio and video broadcasting and does not transmit any kind of signals itself. The main benefit of radar-based detection sensors over other techniques are its superior detection and localization accuracy [12, p. 42635-42658]. Radars are able to identify UAVs with noise suppression, unlike acoustic or RF detection. Visual circumstances including rain, fog, dust, etc. generally have no effect on the detection accuracy. Hence, radars are frequently regarded as the best option when great precision and reliability are sought. Yet, there are several practical limitations and financial concerns to keep in mind while developing and implementing a radar that is appropriate for detecting drones. First of all, as drones have a smaller radar cross section (RCS) than airplanes, traditional radars are less effective in detection of tiny objects. Secondly, drones that are hovering or moving slowly could be mistaken for stationary reflecting objects by radars. Thirdly, in terms of classification ability, radar sensors frequently may not accurately determine whether the detected object is a drone or another flying object when performing the task of distinguishing between small objects

such as birds and drones [11, p. 250-15]. Although Radar is a sensor that is less affected by weather and can simultaneously detect and track multiple objects at long distances, they do a poor job of recognizing a swarm of drones flying erratically and at different speeds. As well as, they can only determine the general direction of an object, and cannot provide detailed information about its shape or size. Finally, the high cost of radar sensors and the complexity of the installation process are ineffective in preparing an affordable anti-radar system [10; 11, p. 250-3; 12, p. 42639].

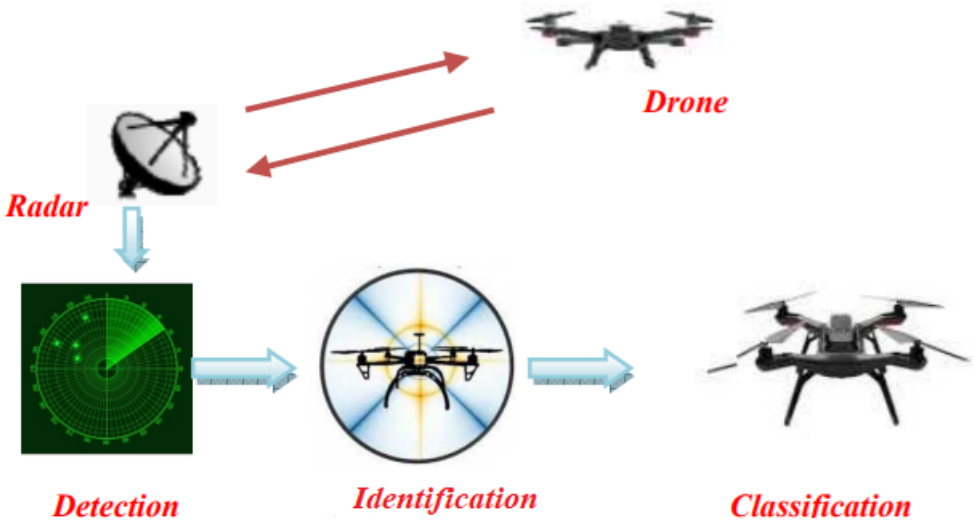


Figure 4 – Schematic structure of radar-based drone detection method

Note – Compiled according to the source [12, p. 42641]

Detection of flying objects based on acoustic sensors. A fairly simple technique for detecting and tracking tiny UAVs is acoustic processing techniques. After all, we are all too acquainted with the recognizable buzzing sound that makes a tiny UAV flying nearby. Nevertheless, there is more to these acoustic techniques than meets the eye. Acoustic antennas are able to recognize the distinct noise made by the drone's propellers and detect the UAV's presence in the region of interest. Examples of current studies and publications that focus on UAV detection using audio sensors may be found in [12, p. 42640; 13-21]. The distinctive acoustic signals (sound waves) produced by a drone's rotating blades can be utilized to locate and recognize a particular UAV model [12, p. 42640]. Several drone parts, such as the engine, wind, or propeller blades generally produce the noise. Propeller blade sound, on the other hand, has a significantly larger amplitude and is frequently employed for detection. Many studies have recommended the use of a microphone array to recognize UAVs by analyzing the noise of the rotors since acoustic detection methods are not reliant on either the target UAV's size or LOS (line of sight). Several studies described ways to detect drones by comparing a drone's recorded acoustic signature with other signatures kept in a database of previously gathered sound signatures. A drone's engine and propellers generate acoustic waves in the frequency range of 20 Hz to 20 kHz, creating the vehicle's acoustic signature. This information may be captured by a single microphone,

and when compared to a database of audio characteristics, it can differentiate the drone from other objects. Therefore, the UAV sound collected using acoustic sensors such as microphones is processed to match with the UAVs ID in a database, as well as this procedure is called as *fingerprinting* [14]. The database normally holds the acoustic fingerprints (also known as drone IDs) of every drone. Either traditional methods or cutting-edge ML algorithms are used to process the acoustic signals. The schematic structure of fingerprinting-based UAV sound detection method is illustrated below (figure 5) [14].

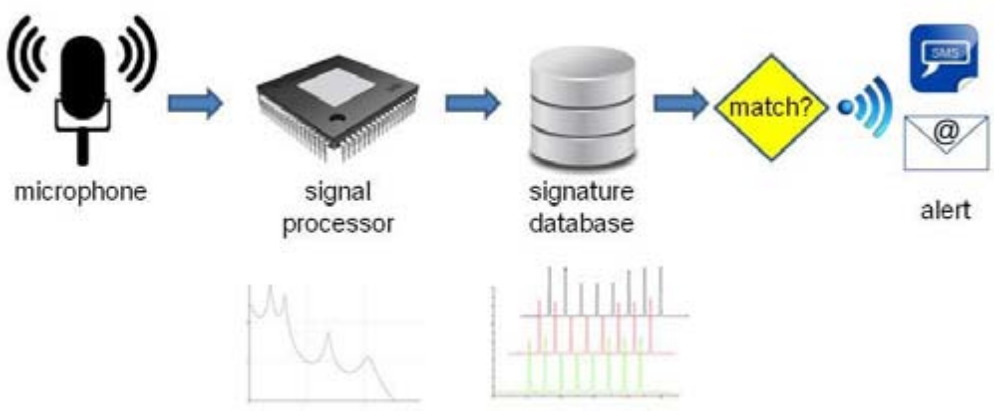


Figure 5 – Acoustic sensor-based drone detection system

Note – Compiled according to the source [14]

Using acoustic signals to detect UAVs can be somewhat less expensive (but the cost rises as accuracy requirements rise). It has the benefit of functioning normally in dim light and a variety of weather conditions, such as rain, fog or dust etc. Moreover, it does not require LOS to the target UAV Acoustic systems work best in isolated locations with minimal background noise. However, it is sensitive to ambient noise and climatic conditions due to wind or temperature, and usually has a range determined by the size of the microphone grid. When there is background noise, this method's accuracy drastically decreases and it becomes impossible to tell if a drone is using a sound-suppression device. The detecting range varies depending on the surroundings and is often just about 200 meters. Therefore, the decrease in the accuracy of detection of acoustic sensors in noisy environments and their dependence on wind and noise indicate the inefficiency of using these sensors as the main detection system.

Radio Frequency detection. Another effective method of drone detection is based on radio frequency (RF) detection. Radio frequency sensors can detect energy emitted by electronic devices on drones, such as radio transmitters or GPS receivers. The presence of UAVs in the target region may be ascertained using RF scanner technologies that capture wireless signals. Sensing the RF transmission between a drone and its ground operators is one of the most popular methods for UAV detection in no-fly zones**.** This technique uses RF sensors acting as receivers to search for RF communication channel signals. Recent research and publications that concentrate on UAV detection and classification utilizing RF sensors may be found in [22-26]. The RF sensors are made to recognize the RF frequency bands that UAVs use to

communicate with their ground controllers and control them. To connect with the controller, drones commonly employ RF signals between 2.4 and 5 GHz [22, p. 114575]. Wi-Fi-enabled drones operate at a frequency of 5.4 GHz, while 5G drones utilize the range of 3.5 GHz. Other less popular bands include the range between 1.2 GHz and 1.3 GHz [23]. Hence it is possible to passively listen to the signals sent between a drone and its controller using an RF sensor (also known as a scanner). RF signal scanning may be done in two different ways. While using the first technique, the receiver listens to and decodes the packets being sent between the UAV and its controller. As the entire packets are decoded, this approach provides more thorough information including the identity, position, and speed of the drone as well as data included inside the packets, such as video feeds. Nevertheless, this technique demands network snooping, which is illegal in many countries. In the second technique, the RF receiver does not perform decoding the packets, but just captures the RF signals consisting of amplitude and phase, as well as looks for certain patterns in the captured data. By detecting the UAV presence, predicting its model, and estimating its position, this technique delivers only a limited amount of information. The second approach is mostly utilized to UAV detection due to regulatory restrictions.

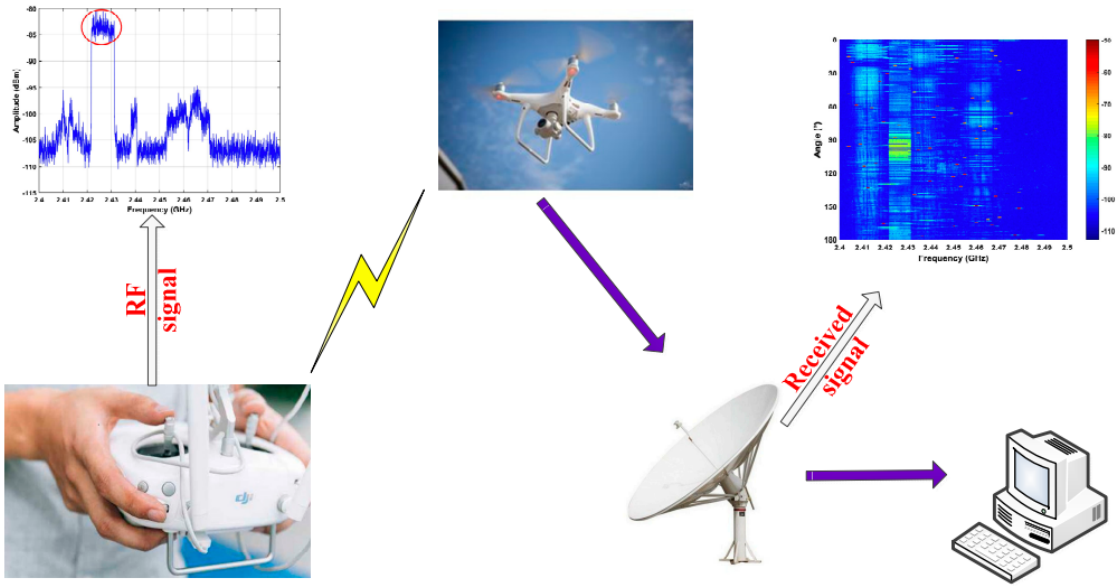


Figure 6 – RF sensor-based system module for UAV detection

Note – Compiled according to the source [24, p. 274-3]

Below in (figure 6) is illustrated the schematic structure of RF sensor-based UAV detection system consisting of the receiver antenna, UAV, computer and controller [24, p. 274-3]. RF signals are used by a UAV to communicate with the ground controller so that these signals can be utilized for detection.

Below is illustrated the infrastructure monitoring using an RF-based UAV detection system under Bluetooth and WiFi wireless interference (figure 7) [26]. In order to detect UAVs, the UAV controller signals are utilized. Bluetooth, WiFi, and UAV controller function at 2.4 GHz frequency spectrum. RF signals from WiFi, UAV

controller and Bluetooth device are intercepted using an antenna. Different wavelet transform categories were used to extract the features from the processed captured signals, as well as these processed signals were used to detect and classify UAV targets based on CNN classifier SqueezeNet [26, p. 101569-27].

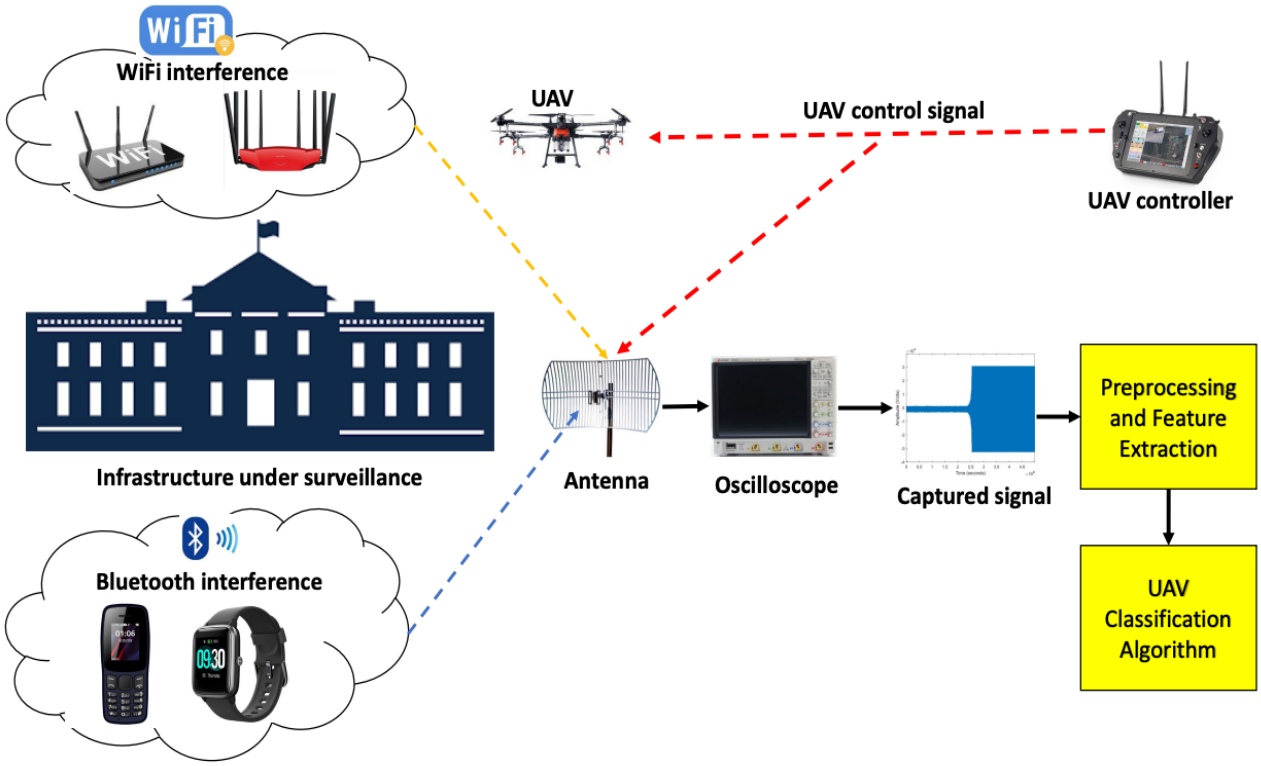


Figure 7 – Infrastructure monitoring using an RF-based UAV detection system under Bluetooth and WiFi wireless interference

Note – Compiled according to the source [26, p. 101569-8]

The ability of detection and localization of drones is made possible by the RF signal, which is a key feature of drones. Nevertheless, whether the drone is used in a partially or totally autonomous mode, RF-based solutions are ineffective. Moreover, all of the currently used RF-based detection methods have poor performance for low signal-to-noise ratios [26, p. 101569-4]. As well as, the detection range of drones can be limited by RF technology, and RF sensors are susceptible to interference from other electronic devices such as cell phones, radios, or televisions. This can lead to false positives and make drone detection difficult.

1.2.2 Optical drone detection technologies

Detection of flight targets using infrared camera sensors. Certain conditions such as night time, poor visibility, or urban environments poses difficulties in detecting flying UAVs. Flying objects detection based on the heat emitted by the drone's internal components using thermal imaging infrared cameras has been conducted in several studies [27-28]. Thermal infrared cameras are able to detect small temperature

fluctuations at the level of tens of μK [27, p. 183]. Thermal infrared sensors produce images like conventional cameras, but use light in the 14 μm wavelength range instead of visible light in the 450-750 nm wavelength range. Physical components such as batteries, motors emit a significant amount of heat and can be recognized/detected using thermal infrared cameras. In [27, p. 184] the authors tested how low-cost and mobile the FLIR Lepton micro-camera, which provides 80×60 pixels video, installed on a Raspberry Pi, can be used to monitor air traffic or prevent collisions of multi-rotor unmanned aerial vehicles at night time. During the experiment, different-sized drones such as DJI Phantom 4, Parrot AR. Drone 2.0 and a custom hexacopter were launched at a test site with a length of more than 100 m. All three drones were launched at a constant speed of 2 m/s at an altitude of about 10 m above ground level in the absence of wind on a warm summer night. As a result of the study, the authors found that batteries are the main sources of heat, unlike engines, which consume the most energy. This is due to the fact that the engines cooled quickly due to the rapid air circulation and decreased in the thermal spectrum. The study did not use a specific machine learning algorithm, and drone detection and classification was carried out using human observation of output video streams in real time. In [28, p. 4450] the authors used a convolutional classifier consisting of ResNet-18 backbone network, pre-trained in the ImageNet classification database and two 1×1 and 4×4 neural layers trained online, which allows them to distinguish objects from the background for reliable tracking of infrared drones in real time. To test the effectiveness of the proposed method, the authors conducted experiments with several well-known algorithms on the anti-UAV infrared dataset consisting of 100 thermal infrared video sequences covering several of multi-scale UAV cases. Additional application of feature attention mechanism and expansion search strategy to the classifier has shown that the infrared tracking algorithm proposed by the authors is more resistant to real-time problems in the infrared scenes compared to the SiamFC and ATOM algorithms.

UAV detection using camera sensors. As explained above various sensors based on radar, acoustic, visual, Radio Frequency (RF) signals are used for detection and classification of UAVs. Among these technologies, despite its traditional success in object detection and tracking, radar is considered a highly professional and expensive technology that requires qualified professionals capable of interpreting the visual results of a radar system. This complexity of radar technology and rapid progress in the field of Computer vision have prompted some researchers to consider drone detection and classification using visual data (images or videos) [29]. Computer vision has long been proven as an effective method for detecting and tracking UAVs in order to prevent air accidents [30]. Today, researchers are investigating how computer vision can be used as an optimal method for detecting and tracking malicious UAVs. Because cameras are relatively inexpensive, light weight and alternative to other sensors that detect objects from long distances. Another advantage of cameras is low power consumption. In the case of placing a sensor with a large power consumption is placed on board an unmanned aerial vehicle, the available flight time can be significantly reduced [31]. In addition, the OpenCV open source software for computer vision and machine learning offers more than 2,500 optimal computer vision algorithms that can

be easily implemented in many applications. For these reasons, Computer vision is one of the main methods used to detect and prevent the use of UAVs.

Although Computer vision is considered one of the most commonly used methods in object detection based on image data from the camera sensor, various lighting and weather conditions such as fog and rain greatly reduce the detection accuracy of the computer vision method [31, p. 15]. As well as the fact that most drones are small in size and fly at low speeds and at different altitudes has made the UAV detection task challenging, which prompts to use sensors of the same type or different sensor systems when preparing an anti-drone system.

1.3 Related works on visual detection of flying objects based on optical camera sensors

Visual detection using hand-crafted features. In order to describe binary shapes (e.g., silhouettes) of drones and birds Unlu et al. [32] developed vision-based features called the Generic Fourier Descriptor (GFD), which is scale, invariant to displacement and rotation changes, able to define shapes with high accuracy. According to the system proposed by the authors, first, a silhouette of a moving object is obtained using a fixed wide-angle camera and a background subtraction algorithm, a special Region Growing image segmentation algorithm is used to separate the pixels of the object from the background, morphological operations are not applied after image segmentation phase in order not to lose any information about the shape, GFD is calculated after normalizing and centering the silhouette, and finally, GFD features are classified into birds and drones using a neural network of about 10,000 neurons. The authors collected 410 drone and 930 bird images from public sources to construct a dataset for training their system. On the test data set, CNN showed a classification accuracy of 85.08%, while the proposed GFD approach showed an accuracy of 93.10%, and by calculating the CNN architecture and inputting the GFD feature vector before the neural network classification, it significantly improved the classification performance of the small dataset.

Qiang Dong et al. [33] proposed a new UAV visual detection method that combines *foreground detection* and *online feature classification* to solve the problem of detecting different types of UAVs against different backgrounds. Since micro- and mini-UAVs have completely different external manifestations, the problem of detecting UAVs cannot be considered as the problem of detecting an object using a trained classifier. The first step in the UAV detection method proposed by the authors is to obtain foreground detection results from images captured by static cameras, foreground detection results may include dynamic background pixels, which leads to an increase in the level of false detection. In this case, an online classification of objects was proposed to obtain detection results with a high probability of the presence of UAVs. The histogram of oriented gradients (HOG) was constructed based on the results of foreground detection to obtain information about the edge and local shape. To evaluate their method, the authors conducted experiments with the previous ViBe (Visual Background Extractor) and PBAS (Pixel Based Adaptive Segmentator) background detection algorithms for visual UAV detection. The results of the

experiment showed that the method proposed by the authors provides significantly better detection accuracy than the other two algorithms for detecting micro- and small unmanned aerial vehicles.

Boddhu et al. [34] presented a collaborative sensor platform consisting of an intelligent smartphone application that uses the appropriate sensors on the device to capture drone attributes such as flight direction, shape, color. The proposed platform was developed by combining an intelligent processing module based on a sensor cloud and a probabilistic model that can evaluate and predict the probabilistic flight paths of a dangerous drone based on a set of geographically distributed data points. As an alternative to the RADAR airspace surveillance system, this system has been tested outdoors and shown to be effective in providing a real-time alerting mechanism to prevent or eliminate potential damage from detected hostile drones, however, flight path estimation by showing a low statistical error with the actual flight path requires improvements and additional testing to be comparable.

Wang et al. [35] proposed a simple, fast and efficient system for detecting unmanned aerial vehicles based on video images captured by static cameras that cover a vast area and are very cost-effective. The method of temporal median background was used to detect moving objects in a video sequence captured by a static camera, and then Global Fourier Descriptors and local HOG features were obtained from images of moving objects. As a result, the combined FD+HOG features were sent to the SVM classifier, which performs classification and recognition. To prepare the data set, the authors converted 10 videos of the Dajiang Phantom 4 quadcopter drone, taken under various conditions, into a series of images and manually annotated drones as a positive sample, and the leaves, buildings, etc. other objects as negative samples. FD (Fourier Descriptor), HOG (Histogram of Oriented Gradients) and the proposed FD+HOG algorithms were used to recognize "drone" and "no drone" objects, and as a result of the experiment, the overall recognition accuracy of the proposed method was 98%. As well as, based on the experimental results the authors proved that the proposed FD+HOG algorithm performs birds and drones classification task with higher accuracy than the GFD (Generic Fourier Descriptor) algorithm even with a small dataset.

In [36], the authors proposed a system for detecting and tracking unmanned aerial vehicles using a stationary and Pan-Tilt-Zoom (PTZ) camera. The authors considered several background modeling algorithms. According to the results of the experiment, the GMM (Gaussian Mixture Models) algorithm showed the best result in terms of accuracy. Object features were extracted by combining the drone predictive models trained by GMM, adaptive model, and principal component analysis to model the object and extract relevant features. In future research, the authors will consider the problem of tracking the violation of very bright colors due to reflection on sunny days and the fact that the system cannot separate the object from the background.

Amy R. et al. [37] analyzed many methods of detecting and monitoring first-class UAVs based on computer vision. According to the research paper, the detection of UAVs using computer vision is carried out using the categories of object detection and feature detection. The authors individually explained these two categories and conducted a literary review of each of them. Object detection is often carried out using

morphological filters, such as dilation and erosion. Detection of UAVs at long distances can give better results using morphological filters than feature detection algorithms. On the other hand, feature detection algorithms can provide a more detailed analysis of detected UAVs. Therefore, the authors plan to achieve better results in their future studies using a combination of morphological filters and feature detection algorithms.

In [38], the authors proposed a method underlying the creation of a cooperative UAV autonomous landing system by detecting and locating the UAV from video sequences obtained from an RGB sky-directed camera. The system proposed by the authors consists of a multi-rotor UAV with vertical takeoff and landing capabilities and a helicopter mounted above a mobile robot or on a static platform. According to the proposed system, moving objects are first detected from an image taken from a static camera using the background subtraction method, the foreground is separated from the background, and morphological operations called opening and dilation are used to reduce pixels isolated from the final movements of clouds from the main foreground i.e. to reduce noise, since the result of background segmentation is often a set of separated BLOB-objects, the clustering method is used to combine each set so that it belongs to the same object. Since the UAV is a more chaotic pattern of movement than other objects (for example, airplanes and birds), the analysis of movements was carried out using the method of distinguishing UAVs from other objects using motion signatures. According to the results of the motion analysis, it was determined whether the object corresponded to the UAV or not. Based on all the data obtained, the previous positions of the UAV were tracked and the current positions were calculated. To test the proposed model, 12 video sets with an image resolution of 320x240, consisting of 915 individual UAV frames, were prepared and analyzed. Each frame in the dataset was classified as "Moving", "Hovering", "Partially Out". The first category was detected when the UAV was moving at a significantly high speed, the second category was detected when the UAV was oscillating or moving at a significantly slow speed, and the third category was detected when the UAV was partially out of the camera's field of view. The process was repeated for various outputs: raw data obtained directly from single frame analysis without any tracking algorithm; using the Kalman filter; averaged position tracking algorithm using the output of the Kalman filter at N=5 and N=8. As a result of the experiment, the method proposed by the authors showed particularly good results in frames where the UAV fluctuates/oscillates in one place, and the error increases at the moments when the UAV partially disappears from the frame, because the exact calculation of its center is difficult. The results at N = 8 deteriorated compared to the results at N = 5 in moving frames. While the Kalman filter performs well on moving frames, the average position tracking algorithm using the output of the Kalman filter at N = 5 proposed by the authors performs better on still frames.

Visual detection using trained features or based on deep learning. Researchers who applied deep learning to visual data to identify UAVs fulfilled the following objectives:

a) classification of drones and birds: two labels are used to annotate these data: «drone» and «bird». Arne Schumann and et al. [39] proposed UAV detection framework that can accurately detect flying objects from long distances based on video images and separate UAVs and birds into appropriate classes. In the system proposed by the authors, depending on whether video images are captured by static or moving cameras, the region that may contain the object was determined by the methods of median background subtraction or deep learning-based RPN (Region Proposal Neural Network). The use of RPN has reduced the number of false positive results and made it possible to achieve an early warning detection system. The detected regions were classified into UAVs or birds using a convolutional neural network (CNN) classifier. To train a reliable CNN classifier, the authors prepared a dataset consisting of 3,386 drones, 3,500 birds, and 3,500 background images. The proposed UAV detection system was tested and evaluated on 6 complex video sequences consisting of UAVs, birds and background movements at different distances. In future work, the authors will use temporal information to improve accuracy.

Due to the lack of data used in training, Saqib et al. [40] conducted experiments to detect moving objects in video data based on pre-trained Zeiler and Fergus (ZF), Visual Geometry Group (VGG16) and other CNN networks Since the aim of the work was to identify drones and avoid confusion with birds, the authors used a Bird-Vs-Drone dataset consisting of 5 MPEG4-encoded videos captured at different times. The videos were divided into 2,727 image frames on which the drones were annotated. The authors analyzed the performance of each network architecture in different iterations. As a result of the experiment, the VGG-16 architecture in combination with Faster R-CNN showed good performance with an average accuracy of 0.66. In conclusion, the authors noted that by considering and annotating birds as a separate class, false positive factors can be minimized, allowing the trained model to accurately distinguish birds and drones, and improves performance.

Aker et al. [41] solved the tasks of predicting the drone's position in video frames and distinguishing them from birds by adapting and adjusting the YOLOv2 algorithm, a single-stage object detection method based on deep learning. To solve the problem of lack of data during training the network, the authors proposed an algorithm for creating an extended artificial dataset by combining background-subtracted real images. The authors combined the Bird-Vs-Drone dataset and artificial dataset and divided them into training (85%) and validation (15%) parts. Precision-recall curves were used to evaluate the network. In the proposed adjusted version of YOLOv2, the Precision and Recall values reached about 0.9 simultaneously.

Celine Cray et al. [42] proposed a spatio-temporal semantic segmentation method using two convolutional neural networks that localize potential objects using a U-Net semantic segmentation network and classify the detected object using a ResNet instead of SSD detector to solve problems detecting very small objects. The difference of the proposed system from the Faster R-CNN object detection algorithm is that the object detection and recognition paths are carried out in two separate networks to correctly manage the network for small objects, and the tracking algorithm is used to filter false positives and false negatives based on continuity. time. The dataset used

was 11 MPEG4 encoded videos from the Bird-Vs-Drone dataset, plus many additional drone and bird videos. Birds in all videos were semi-manually annotated and areas annotated with bounding boxes were converted to segmentation masks using background subtraction. One-, three-, and seven-channel configurations were considered, and experimental results showed that the greater the number of input channels, the better the network can handle temporal aspects and distinguish very small drones from the background or birds. The difference between the proposed system and the Faster R-CNN object detection algorithm is that the object detection and recognition paths are carried out in two separate networks for the correct management of the network for small objects, and the tracking algorithm is used to filter false positives and false negatives based on time continuity. The dataset used was 11 MPEG4 encoded videos from the Bird-Vs-Drone dataset, as well as many additional videos of drones and birds. Birds in all videos were annotated semi-manually, and regions annotated with bounding boxes were converted to segmentation masks using background subtraction. One-, three-, and seven-channel configurations were considered, and experimental results showed that the higher the number of channels in the input, the better the network can handle temporal aspects and distinguish very small drones from the background or birds.

In [43] the authors proposed adding a deep Super-Resolution technique to the UAV detector, which gives the possibility of image enlarging. That is, small drones flying in the sky from a far will be magnified and further enhanced so that they can be detected by the DNN detector. To test how the Super-Resolution technique can improve the detection result of remote UAVs by affecting the weights and the detector learning module, and not just the pre-processing phase of the object detection model, the authors examined the full DCSCN (Deep Residual CNN with Skip Connection and Network in Network) consisting of feature extraction and reconstruction networks and its compact version c-DCSCN. The Faster-RCNN detector was trained using a large dataset with an appropriate number of annotated birds from short and long-range drone footage provided by the organizers of the WOSDETC Drone-vs-Bird Detection Challenge, as well as other publicly available datasets. The proposed detector was trained to predict three classes classified as "UAV", "bird" and "other objects". 2814 randomly selected frames were used as a validation set to tune the detector's hyperparameters. Experimental results show that although DCSCN is only slightly better than c-DCSCN in terms of Recall results for all clip sequences, the reduced c-DCSCN model is effective for real-time applications with less computational time.

In [44] the authors considered the problem of detecting small drones in a remote video surveillance system using popular and advanced deep learning based object detection methods, such as Faster R-CNN based on Inception v2 and ResNet-101 base architectures and SSD based on Inception v2 base architecture. A total of 8771 frames were extracted from 11 MPEG4 encoded videos in the Bird-Vs-Drone dataset. The time when the drone was close to the camera, i.e. large size, and cases when the drone was far from the camera, i.e., small size, were considered for each algorithm, and according to the results of the experiment, in the first case, all algorithms were able to detect the drone, and in the second case, when two drones simultaneously appeared in

the frame at a long distance, Faster R-CNN based on ResNet-101 was able to successfully detect both drones, while Faster R-CNN based on Inception v2 was able to detect only one of the drones. SSD, on the other hand, had a poor ability to detect both long-range drones and tiny objects. Due to the fact that the authors performed the detection offline, the detection time was not taken into account, however, in their future work, they will consider the detection time as a key indicator for evaluating the effectiveness of the proposed model in real-time detection of the drone. The difference between the proposed system and the Faster R-CNN object detection algorithm is that the object detection and recognition paths must be performed in two separate networks for proper network management with respect to small objects, as well as in the use of a tracking algorithm to filter false positives and false negative results based on time continuity. 11 MPEG4 encoded videos in the Bird-Vs-Drone dataset were used as data sets, as well as many additional videos about drones and birds. The birds in all the videos were partially annotated manually, and the areas annulled by the bounding box were transformed into segmentation masks by the background subtraction method. Configurations with one, three and seven channels were considered, and experimental results showed that the more channels at the input, the better the network handles temporal aspects and can distinguish very small drones from background or avian ones.

In [45], the authors proposed a multi-level UAV detection algorithm based on the CNN model, which distinguishes UAVs from birds and background obstacles in a very wide range of static and moving video cameras. The proposed model was trained using frames from the Purdue UAV and DronevsBird Detection Challenge 2019 datasets, which include drones both above and below the horizon, short and long range drones, drones taken from moving and static cameras. To evaluate the performance of the proposed method in different observation situations, 5 video sets of the Drone vs Bird Detection Challenge 2019 dataset (model evaluation in static camera scenarios) and 5 video sets of the Purdue UAV dataset (model evaluation in moving camera scenarios) were used. As a metric for evaluating the proposed method on two different datasets, the F-1 score was calculated separately for cases with and without frame difference in the proposed system. According to the results of the experiment, it was shown that using the frame difference for the proposed method for videos taken from a static camera gives an effective result, and for videos taken from a moving camera, a good result can be obtained without using the frame difference for the proposed system;

b) drone detection: two labels are used to annotate this data: «drone» and «no drone». Manjia Wu et al. [46] developed a video-based real-time drone detector using deep learning. Since training a reliable detector requires a large number of training images, during the research, the authors created a dataset that was semi-automatically labeled with KCF (Kernelized Correlation Filters) tracker instead of manual labeling. The KCF tracker-based semi-automatic data set labeling method allowed to speed up the pre-processing process of the trained images. The authors improved the YOLOv2 (You Look Only Once) deep learning model by changing the resolution structure of the input images and adjusting the anchor box size parameters. To obtain the detection network, the authors removed the last convolutional layer of Darknet19, which was pre-trained with a set of standard ImageNet 1000 classifier, and added three 3x3

convolutional layers with 1024 filters and one 1x1 convolutional layer with 30 filters at the end of the network. The USC drone dataset and the KCF-tracker-labeled Anti-Drone dataset were used to train the network. 2 Gb and 4 Gb GPU-RAM configurations were used to verify the real-time performance of the detector and low power consumption. Processing speed with GPU-RAM 2 GB reached 19 FPS, while GPU-RAM 4 GB showed processing speed 33 FPS. By conducting various experiments, the authors achieved a good result of real-time detection with the proposed detector at an affordable price of the system.

Jihun Park et al. [47] considered six state-of-the-art convolutional object detectors for real-time UAV detection and tracking system in terms of accuracy and speed using using a Pan-Tilt-Zoom (PTZ) camera. According to the authors, the main challenges of the system are the small size of drones and the creation of a real-time detection model to track drones through panning, tilting and zooming actions. The drone dataset captured video of 11 different multi-rotor drone models in different views, at different distances and under different background conditions, from captured videos were taken ten image frames per second and manually labeled a total of 9525 multi-rotor drone images. The results of the experiment were analyzed separately for accuracy and speed. Since the vast majority of drones in the dataset are small in size, SSD performed the lowest in terms of accuracy, as SSD models give poor performance on small objects. According to the F-measure that considers Precision and Recall together, Faster R-CNN based on Inception Resnet has the highest result (74.3%), R-FCN based on ResNet-101 (73.2%) and YOLOv2 (72.8%) showed. In terms of speed, the MobileNet-based SSD showed the fastest detection by detecting 20.8 frames per second, while YOLOv2 (13.0 fps) and Inception V2-based SSD managed to detect 12 frames per second. Given the real-time constraints, the authors believe that these three fast models can be used in their system. Also, the training time of the model should not be too long, since the object detection model is constantly updated due to the addition of datasets with new models of UAV and new background conditions. The authors investigated the time taken to train the model for 100,000 iterations and found that the RFCN model based on ResNet-101 took the shortest time to train, and Faster R-CNN based on ResNet-101 and SSD models based on ResNet-101 took 20 hours to train 100,000 iterations. found that little time was spent. In the experiment, Faster R-CNN based on Inception Resnet shows the highest accuracy, but it is the slowest in detection and training. Considering the trade-off between speed and accuracy, the authors conclude that the YOLOv2 model shows comparable accuracy to the Faster R-CNN and R-FCN models and is faster than these models. In future research, the authors will consider how different configurations, such as number of object proposals, affect speed and accuracy in R-FCN and Faster R-CNN models;

c) detection, tracking and classification of flying objects (drone, bird, aircraft, etc.). According to Peng et al. [48] solved the problem of limited visual data by creating photorealistic images of unmanned aerial vehicles using the Physical Based Rendering Toolkit (PBRT). The authors prepared a large-scale training set consisting of 60,480 rendered images by selecting various UAV positions and orientations, 3D models,

extrinsic materials, intrinsic and extrinsic camera parameters, environment maps, and post-processing methods of rendered images.

This rendered dataset includes not only the anchor box of the NAC, but also the position of some important parts of the UAV used in complex applications such as mask detection and keypoint detection, and the location of all the pixels contained in the UAV. For drone detection, the Faster R-CNN network was fine-tuned with Detectron provided by Facebook AI Research using ResNet-101 base model weights. Faster R-CNN trained with rendered images shows an Average Precision of 80.69% on the test set of manually annotated UAV images, AP 43.03% when pre-trained only with COCO 2014 dataset and AP 43.36% on PASCAL VOC 2012 dataset, while only in the rendered training set showed an average accuracy of 56.28%. According to the experimental results, the average accuracy of Faster R-CNN detection network trained with rendered images was relatively higher than other methods.

Yoshihashi et al. [49] proposed a new joint detection and tracking system using information about the movement of small flying objects. This system, called a recurrent convolutional network (RCN), consists of four modules, each of which performs a specific task: conv layer, ConvLSTM, cross-correlation layer, and fully connected layers. The authors used the system training technique by tuning AlexNet and VGG16 without training the system from scratch. To evaluate the system, the system was first tested on a bird detection dataset around a wind farm and then tested on a UAV dataset of 20 hand-captured data to ensure that the system could not be applied to other flying objects. Experimental results, presented as ROC curves, show that the proposed system performs better than previous solutions.

Artem Rozantsev et al. [50] proposed two approaches to detect the presence and danger of a flying object by classifying three-dimensional descriptors from spatio-temporal image cubes captured by a single camera. Both approaches are based on three-dimensional histograms of gradients (HoG3D) and CNN model. The proposed system starts by dividing the video frames into time segments that overlap each other by 50%. A multiscale sliding window is then placed to construct the st-cubes. To create stabilized st-cubes, a regression-based motion compensation algorithm is applied to each st-cube patch. Each st-cube is then classified as presence or absence of an object of interest (a UAV or an aircraft). To this end, the authors trained two different augmented tree regressors to predict the desired transition to the input patch based on the HoG features and two separate CNNs for the regression task based on the trained features. After training, the regressors are used to generate st-cubes, which are given as input data for motion compensation and classification. To test the performance of the system, the authors prepared a dataset consisting of 20 video sequences of unmanned aerial vehicles and 20 publicly available videos of radio-controlled aircraft. In testing and comparing several motion compensation methods on the test dataset, the CNN regression showed a detection accuracy of 0.849 on the UAV dataset and 0.864 on the aircraft dataset. As a final step, the regressor is trained to place different image scales, that is, to place the detected object accurately.

Sobue et al. [51] proposed a system for detecting and tracking objects flying over the city using several 4K video cameras. According to the proposed system, detection

of flying and non-flying objects is carried out in three stages: background subtraction method, comparison of KCF control method and background subtraction methods, classification of detected objects into six different classes using deep learning-based CNN image classifier. Also, SFM (Structure from Motion) was used to calculate the three-dimensional trajectory. Thanks to the automatic classification of flying objects, the authors used deep learning to classify images of birds, helicopters, airplanes, drones and background (noise) of the same size with about 80% accuracy. Since this result is not sufficient for a real system, the authors pointed out that the classification accuracy of flying objects can be improved by using distance information.

Yuanyuan Hu et al. [52] proposed an improved YOLOv3 system for detecting unmanned aerial vehicles. The improved YOLOv3 system uses the last four scales instead of the last three scales of feature maps to predict the bounding box of objects, which allows more information about the structure and contours to detect small objects. In addition, to reduce the calculation, the size of the UAV in the four-scale feature maps is calculated from the input data, and then the number of anchor boxes is also adjusted. The proposed system was pre-trained on the ImageNet dataset and tuned with the UAV dataset. The authors prepared a dataset consisting of images of four-, six-, and eight-rotor UAVs from the Internet and from their own cameras. According to the experiment results, the mean average precision (mAP) and the number of frames per second (FPS) were determined, which evaluate the detection efficiency for the YOLOv3 and advanced YOLOv3 methods. mAP for the YOLOv3 method was 33.25, and mAP for the improved YOLOv3 method proposed by the authors was 37.41. YOLOv3 had 60.8 frames per second, while improved YOLOv3 had 56.3. The overall accuracy was 84% in the YOLOv3 method and 89% in the proposed system. Experimental results show that the proposed system provides the best detection accuracy and allows obtaining the most accurate bounding boxes for high-speed UAVs. In future work, the authors plan to improve the proposed system by increasing the network speed, using a more efficient k-means substitution method and applying it to various objects.

In [53], an autonomous drone detection and tracking system using a wide-angle and low-angle camera mounted on a rotating tower is proposed. The proposed system describes an integrated deep learning multi-frame detection method in which a frame from a zoom tower camera is superimposed on a frame from a wide-angle static camera for efficient use of memory and time.

In [54], the authors proposed a complex system for detecting an unknown drone based on machine learning using a camera installed on a surveillance drone. The proposed system was trained by adapting the OpenCV library Haar cascade classifier. 2,088 positive (drone) and 3,019 negative (non-drone) samples were used to prepare the dataset, and the number of positive samples was increased to 7,000 using image manipulation. Based on these positive patterns, the 2D region where the drone is located in the frame is determined by the trained Haar cascade classifier, then the region is cut and sent to a simple CNN identifier consisting of two convolutional (Conv), two fully connected (Fully connected) layers to identify drone models. Due to the very long time spent on training for drone detection and identification with deep learning models

such as ResNet or Faster R-CNN, and the complexity of building the system, the authors separated the system into a classifier and an identifier CNN. According to the results of the experiment, the detection accuracy showed 89% even in several iterations, and the identification accuracy showed 91.6%. This system does not require much training data and it is very effective to use Deep CNN in classifying drone models. In future work, the authors plan to develop a remote assessment module to supplement the existing system.

1.4 UAV detection using Sensor fusion techniques

Combining sensor data from many sources is a technique known as sensor fusion. Compared to using individual sources, these data sources help to lessen the uncertainty of the information. Recent studies on UAV detection using sensor fusion methods may be found in [55-60]. Below different sensor fusion methods are presented:

1. Fusion of data from sensors of the same type: two sensors and more than two sensors.

2. Fusion of data from different types of sensors: two sensors and more than two sensors. By combining the data acquired from different types of sensors, it is possible to recognize not only the drone and its payload, but also indirect parameters such as the position, the direction of movement and the model of a drone [55, p. 9448699-3].

Identity and tracking fusion has been facilitated by three general sensor fusion frameworks. They include postindividual sensor processing or sensor-level fusion, preindividual sensor processing or central-level fusion, as well as a hybrid fusion, which is the combination of above two methods [60, p. 317].

The method used here is known as "postindividual sensor-processing fusion," in which each individual sensor processes data separately before sending the results to the fusion processor. In order to detect, classify, and maybe track the objects, they are integrated here using a sensor fusion technique. Preindividual sensor-processing fusion involves feeding minimally processed data from a number of sensors to a central processor, where they are merged, either pixel by pixel or feature by feature, and assessed for the occurrence of targets and their tracks using a fusion algorithm [60, p. 317].

The third strategy uses a combination of postindividual and prcindividual sensor-processing. Single sensor failures, which could have an adverse effect on the performance of other fusion architectures, are tolerated in this case because the detection and tracking operations of the unaffected sensors continue through their individual data processing pathways and, potentially, sensor-level fusion.

Preindividual sensor processing or central-level fusion – fusing the raw data (also known as early fusion). Sensor fusion is performed by fusing the raw data coming from multiple sensors, therefore the outputs of unimodal analyses are fused before a concept is learned [60, p. 317]. The general scheme of preindividual sensor processing or central-level fusion is illustrated in (figure 8).

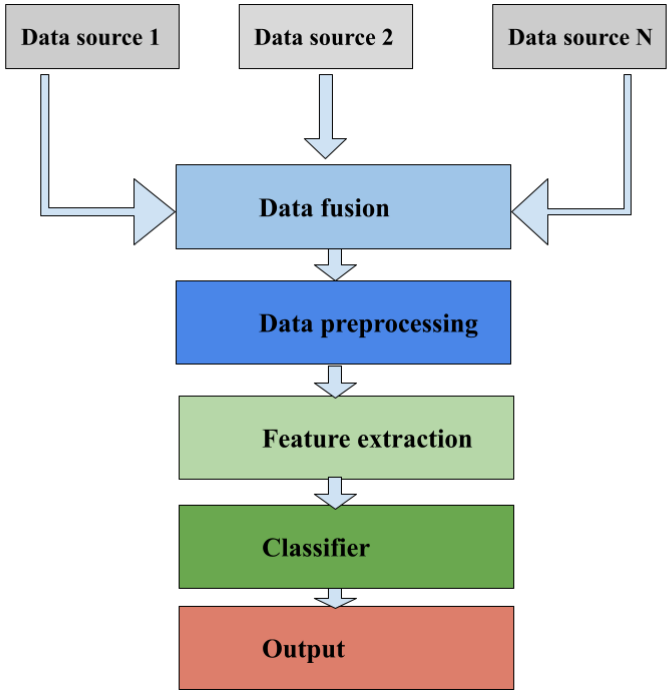


Figure 8 – General scheme for preindividual sensor processing or central-level fusion

Postindividual sensor processing or sensor-level fusion – fusion the detections. Sensor fusion is performed by fusing the objects detected independently on sensor data, therefore the outputs of unimodal analyses are fused to learn separate scores for a concept. A final score for the idea is learned after the fusion process [60, p. 318]. The general scheme of postindividual sensor processing or sensor-level fusion is illustrated in (figure 9).

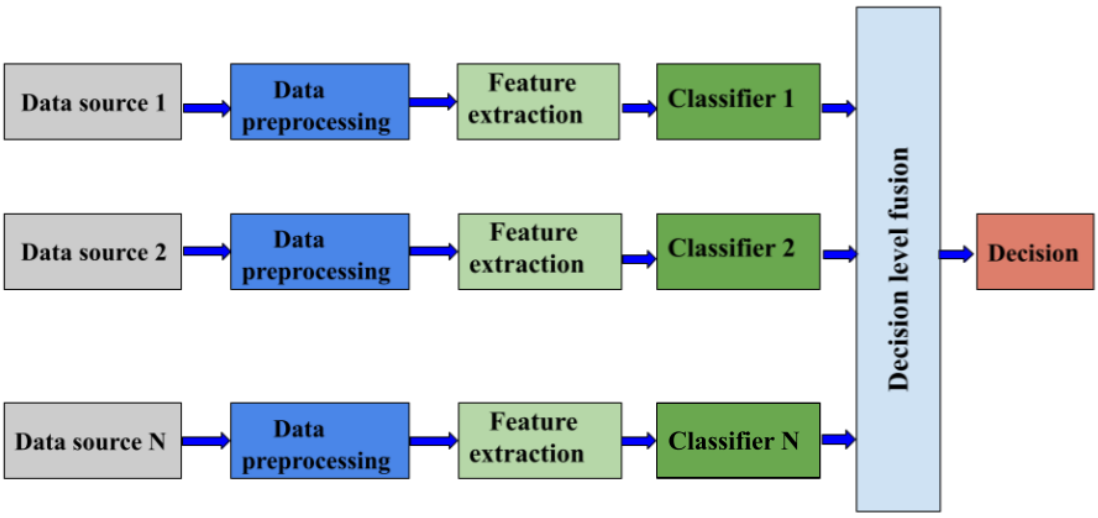


Figure 9 – General scheme postindividual sensor processing or sensor-level fusion

Hybrid fusion or high-level fusion. Sensor fusion is performed by fusing both objects and their trajectories, therefore the output results are relied not only on

detections, as well as on predictions and tracking. The general scheme of hybrid fusion or high-level fusion is illustrated in (figure 10).

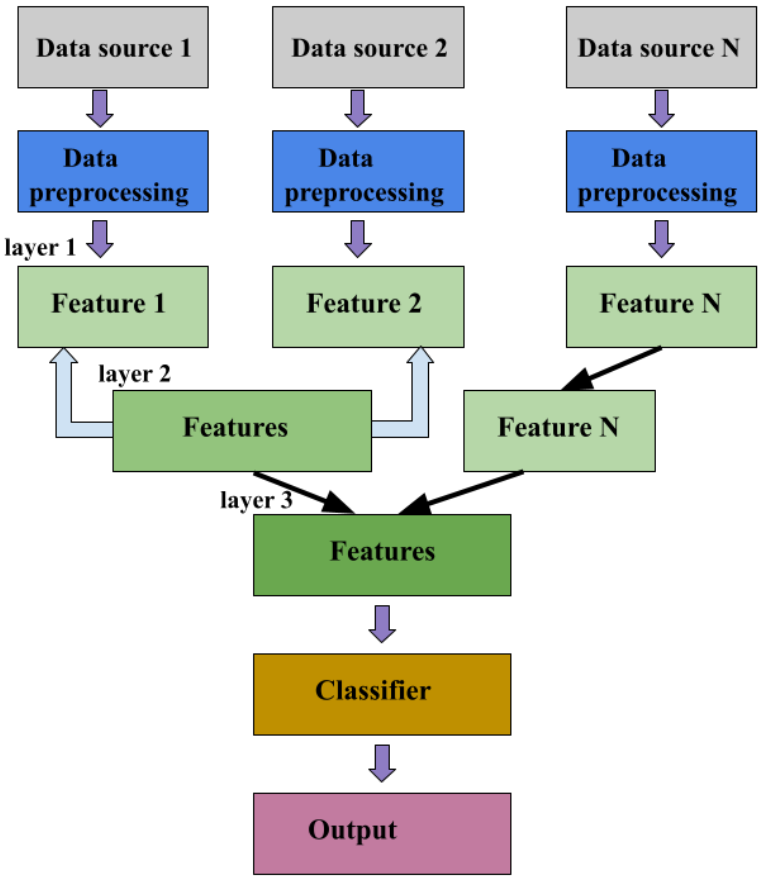


Figure 10 – General scheme for high-level fusion

Chapter's conclusion. This chapter is dedicated on detailed theoretical analysis started from security UAV threats to literature review of related works on UAV detection and classification methods based on different drone detection technologies. Particular attention is paid to methods for detecting and classifying UAVs based on visual data, since this work is devoted to the study of effective UAV detection using image sensors. To sum up, based on conducted state-of-the-art research, it was found out that the vast majority of studies aimed at detecting and classifying unmanned aerial vehicles based on visual data have been performed using features trained using various models and methods of deep learning. However, the lack of publicly available large-scale UAV datasets is a major obstacle to research in this area, as large labeled datasets are required to build robust models of deep learning approaches. Due to this situation, researchers tried to solve the problem of limited visual data by using transfer learning without building the network from scratch through the methods of creating photorealistic UAV images, using image distortion and data augmentation techniques, as well as creating extended artificial datasets. In future research, some authors have noted that they use generative models such as GAN (Generative Adversarial Networks) to create artificial data similar to real data. Most of the visual data studies do not provide detailed information about the type of data collection device, UAV model, the

detection range, and the dataset that would validate the work and allow comparison with other works. Most of the studies have focused only on the detection efficiency of the UAV, and no work has focused on the classification efficiency based on the distance to the UAV. This issue may be an interesting area of research in the future. Research related to deep learning-based CNN detection is mostly based on two-stage Faster R-CNN and one-stage YOLO architectures. The second chapter considers video signal acquisition and processing methods that are crucial in the process of visual data preparation for neural network training.

2 VIDEO SIGNAL PROCESSING AND DATA PREPARATION TECHNIQUES

2.1 Image acquisition and processing techniques

Image processing is a technique for converting a physical image into a digital format and applying various operations on it to create an enhances image or extract relevant information from it. It is a sort of signal distribution where the input is an image, such as a video frame or a photograph, and the output might be another image or some characteristics related to the image. Figure 11 illustrates the block diagram of general Image Processing system.

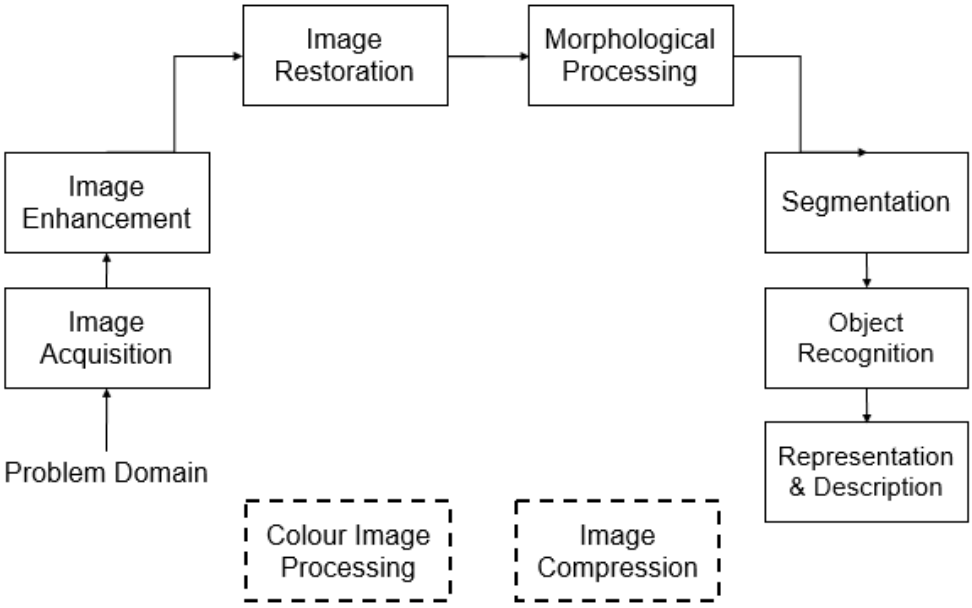


Figure 11 – The block diagram of general Image Processing system

As it is seen from the above block diagram Image acquisition is the first step of digital image processing. An image must first be captured by a camera and transformed into a controllable entity before any video or image processing is started. This is the procedure referred to as image acquisition. Image acquisition process is performed through these three steps: firstly, energy is reflected from the object of interest, then the energy is focused by an optical system, finally the energy amount is measured using a camera sensor. Figure 12 illustrates overview of the image acquisition process [61].

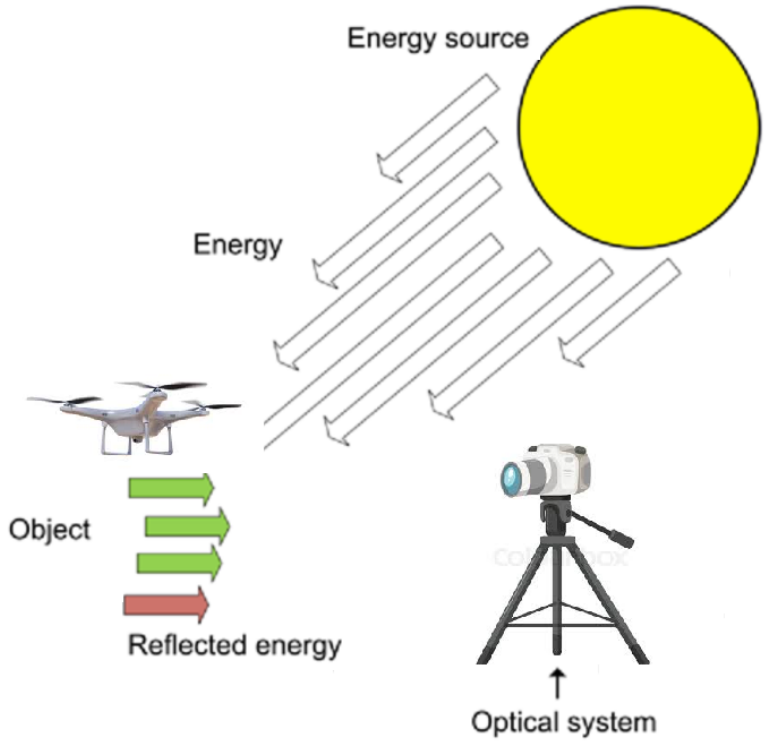


Figure 12 – The main steps of image acquisition process

The visual light spectrum is defined as the region between 400 and 700 nm. Therefore, human eye, along with the majority of cameras, can detect EM waves in this frequency range. As a result, the signal used to broadcast TV, radio, mobile phones, etc. is fundamentally the same as the light from the sun.

Digital image processing methods. The following is how digital signal processing (DSP) operates: an analog-to-digital converter converts the analog signal into a digital signal. The received signals are then processed by the digital computer. The DSP systems also make use of computer peripherals with built-in signal processors that enable real-time signal processing. It is occasionally essential to convert the signal back to analog (e.g., to control a device). Digital-to-analog converters are employed for this purpose. There are several uses for digital signal processing. It may be utilized for image processing, speech recognition, or sound processing.

The term «image processing» in imaging science refers to any type of signal processing where the input is an image, such as a video frame and a photograph. The output of image processing can either be another image or a set of attributes or parameters that are associated with the image. The majority of image processing methods consider the image as a two-dimensional signal and using common signal processing methods on it. Although analog (or optical) image processing is equally feasible, digital image processing is considered as the most frequently used signal processing method.

Imaging is the process of acquiring pictures, which initially produces the input image. Image processing is a technique used to improve unprocessed or raw images from cameras or sensors mounted on airplanes, spacecraft, satellites, and other objects for various uses. Throughout the last four to five decades, several approaches have

been created in the field of image processing. A set of square pixels organized in columns and rows is called an image. Each picture element in a (8-bit) grayscale image includes assigned intensity that varies from 0 to 255. The term «grey scale image» is typically referred to as a black and white image, but the term stresses that there will also be several shades of grey in the image. A typical grayscale image includes 256 grayscales (8-bit color depth). RGB and CMYK are the two primary color spaces used in science communication. The way that humans experience color via the R, G, and B receptors in their retinas is quite similar to the RGB color paradigm. RGB, which employs additive color mixing, is the fundamental color model used in television and other visual medium that projects color through light [62].

For an 8-bit true color image, any colors can be determined by the values of red (R), green (G) and blue (B). As a general rule, any RGB color ranges from 0 (least saturated) to 255 (most saturated). In the RGB color model, a wide range of colors can be obtained by mixing three colors in different ways. With this system 256x256x256 discrete color combinations can be obtained. To convert a 0-255 system to a 0-1 system, the RGB colormap values can be divided by a maximum value of 255. The three-color components are stored in a 3D matrix of numbers. A 2D matrix can be warped using with various edge detection filters. As shown in Figure 1, each of the filters is used to drive one channel and three outputs, respectively [61; 62]. Figure 13 shows a color image represented as a matrix of three-dimensional numbers.

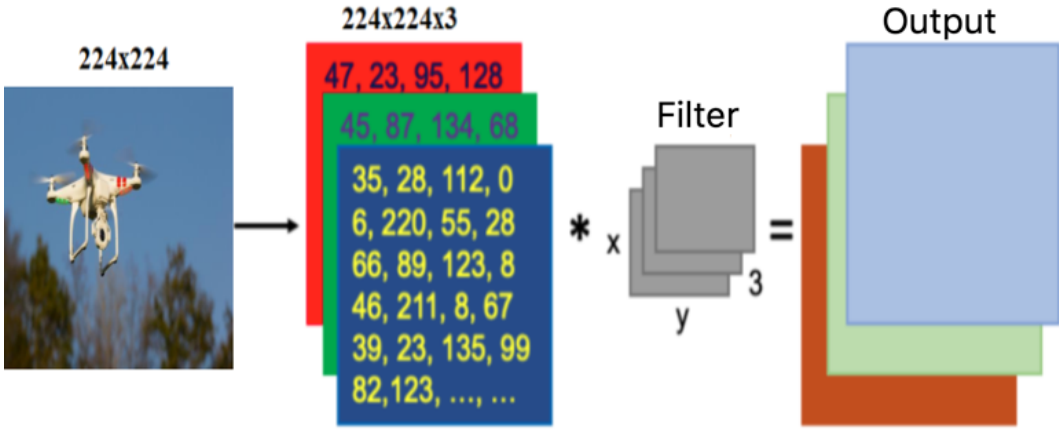


Figure 13 – A color image represented as a matrix of three-dimensional numbers

For image preprocessing, a single image is represented as an array of pixels using grayscale or RGB values. To increase the learning speed, the image data should be scaled with minimum-maximum normalization. A categorical sign color can be converted into a vector of three digital values using one-hot encoding.

Putting it simply, digital image processing involves converting the input image into the output image. Choosing the most crucial information (such shape) and removing the rest (e.g., noise) is the goal of this method. There are several distinct image operations included in the digital image processing, such as filtering, segmentation, thresholding, compression and geometry transformation.

Filtering. There are particular tools for acting on signals in both the one-dimensional domain (for audio signals) and the two-dimensional domain, in this case

on images. Filtering is one such instrument. It entails some pixel-based mathematical operations that produce a new image. It is standard practice to apply filters to images in order to enhance their quality or extract key image features.

Data augmentation. Data augmentation is one of the most important preprocessing techniques that can be carried out online or offline [63]. Offline techniques are used to increase the size of small data sets, while online techniques are used to increase the size of large data sets. Image data augmentation methods generate more training data from the original data and do not require additional memory to store. Common methods for creating new images include horizontal or vertical translation, random rotation, scaling, etc. Additionally, there are advanced data augmentation techniques that use conditional generative adversarial networks (GANs). The examples of data augmentation techniques are shown in (figure 14).

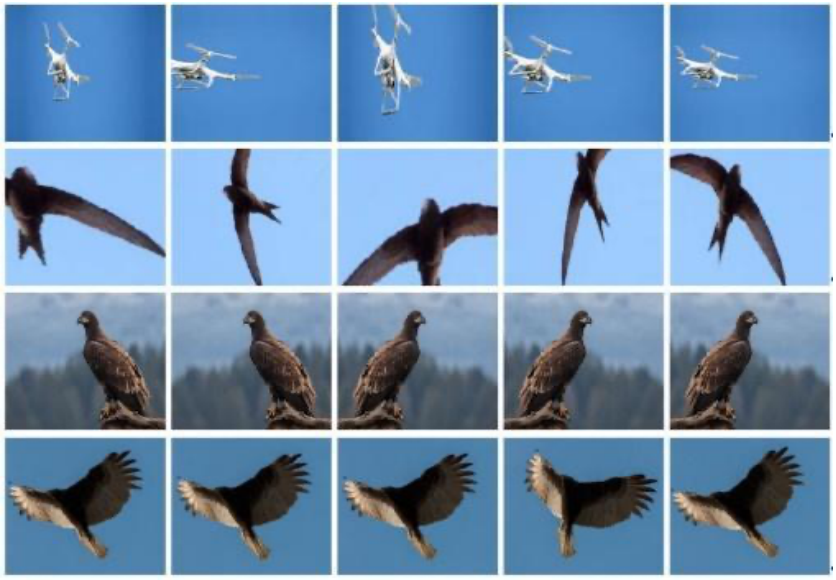


Figure 14 – Examples of image data augmentation

2.2 Data preparation and labeling

Input data preparation process includes drawing a ground truth bounding box around the certain object and then converting it to the standard format between 0 and 1. This process is called data labeling or annotation. Based on this process, the converted format the labeled data consists of the class number, the bounding box's center coordinates such as (x, y), as well as its width and height value (w, h). Therefore x, y – are the bounding box's center coordinates, while w, h – are width and height of the bounding box, [(x1, y1), (x2, y2), (x3, y3), (x4, y4)] – is the ground truth bounding box, respectively. Below in (figure 15) an example of image labeling process is illustrated. As well as (figure 16) shows the bounding box values.

Figure 15 – Example of image labeling

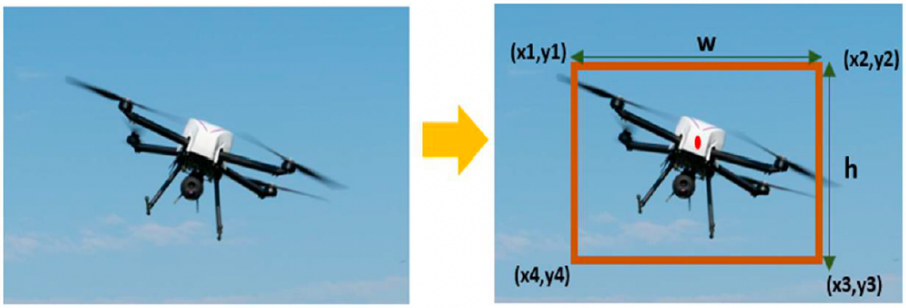


Figure 16 – Labeled image with bounding box values

The image below shows the moment when two flying objects, a bird and a drone, flew into the field of view of the camera at the same time. The ground truth bounding boxes of the two flight objects are circled in green and red, and each is labeled with the respective drone and bird classes. Below is shown an example of labeling an image with two flying objects (figure 17).

Figure 17 – Labeling an image with two flying objects

Data preparation is the meaningful part of any object detection task. This chapter presented a brief description of video signal acquisition and processing steps, as well as the importance of such steps as data collection, preprocessing, and labeling in object detection has been described.

3 OBJECT DETECTION AND CLASSIFICATION USING NEURAL NETWORKS

3.1 Deep Learning algorithms for image classification

3.1.1 Convolutional neural networks

Typically, a neural network is made up of several layers of linked neurons. Its design was influenced by the way the visual cortex is organized and is comparable to the way neurons communicate in the human brain. There are some primary components inside a neuron that calculate the measured sum of inputs. A layer is referred as a storage of neurons. CNNs are a well-known family of deep neural networks frequently employed in visual image processing, as well as it includes three main layers such as the convolution layer (Conv), the pooling layer (Pool), and the fully connected layer (FC). The general structure of CNN is illustrated (figure 18) [64].

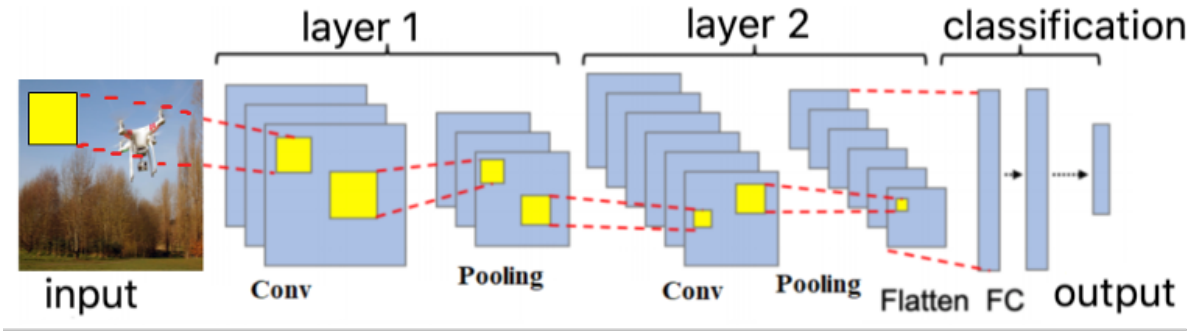


Figure 18 – Outline of CNN architecture

Convolution layer. The fundamental blocks of a CNN model are convolutional layers. A mathematical operation called convolution combines two sets of data. Filters and feature maps are the main components of convolution layers. Convolution is performed by sliding the filter along the input. Each place does the product of matrices by element and then sums the result. This sum is entered in the feature map. In other words, trained filters are utilized to extract significant features from the input image and the output of the filter that applied to the previous layer. Convolution operation is always performed via an odd number size filter (such as 1x1, 3x3, 5x5, 7x7, or 11x11). The size of the resulting image decreases with increasing filter size. This process causes the loss of information. That is why, CNN models frequently employ 3x3 filters during the convolution operation. Moreover, using small-sized filters is not always effective. Large-size filters might be useful in some situations. Filters are used for feature extraction. As a result, the filter size varies based on the used network and performed operation [65]. The fact that only the most crucial features are extracted during convolution operations causes the feature map's size to be less than the original input image’s size. In the illustration below, the original input image's size is 5x5; the filter's size is 3x3; and the size of the feature map produced by the convolution process is 3x3. Two crucial characteristics of the convolution operation are stride and padding. Stride, which is frequently chosen to be equal to s=1, denotes the step by which the filter

should be moved. The output feature map size is likewise reduced when the stride value is greater than 1, and significant information may be slightly lost (figures 19, 20).

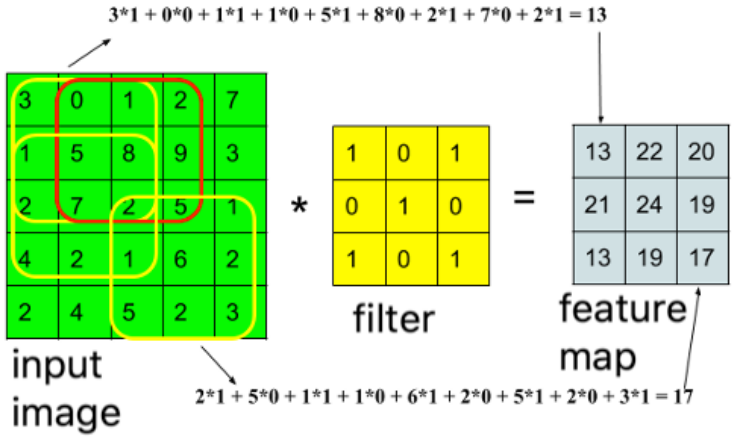


Figure 19 – The sequence of execution of the convolution operation at s=1, p=0

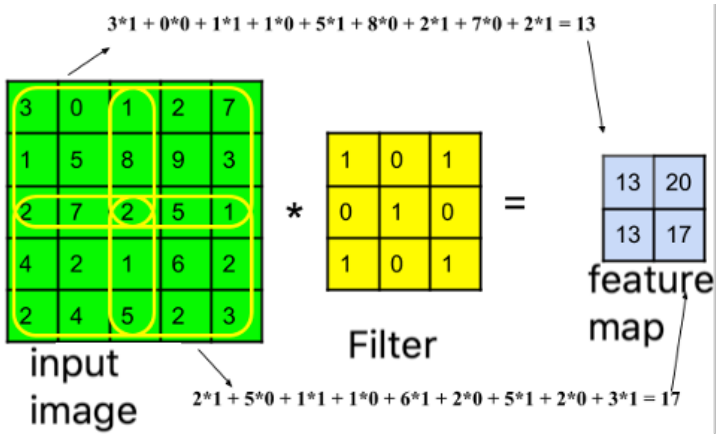


Figure 20 – The sequence of execution of the convolution operation at s=2, p=2

The size of the feature map can be determined using the following formula (1):

$$y_{size} = \frac{n+2\times p-f}{s} + 1 \quad (1)$$

where n – is the input size, if the size of an input image is 5x5 then n=5;

p – is the padding value, for the above example the padding type is "valid", i.e., $p = 0$; f is the filter size, $f = 3$ for a 3x3 filter; s is the striding value, s = 1.

$$y_{size} = \frac{n + 2 \times p - f}{s} + 1 = \frac{5 + 2 \times 0 - 3}{1} + 1 = 3$$

Hence the size of the feature map in the output is equal to y_{size} x y_{size} = 3x3.

In the example above, no padding was used, so we assumed p = 0. If the output needs to be the same size as the input, then zero-padding is used, and the value of p is determined by the following formula (2):

$$p = \frac{f-1}{2} \quad (2)$$

When using a 3x3 filter, p=1 is taken so that the input size is 5x5 and the output is the same 5x5, and a border consisting of one row and one column is added to the input image. This type of padding is called "*same*".

As a result, the 5x5 size input becomes 7x7 size. When convolution operation is performed, 5x5-size feature map is obtained at the output (figure 21).

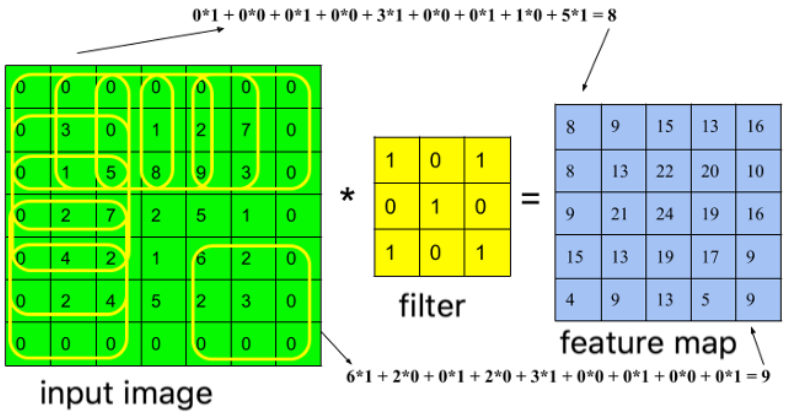


Figure 21 – The sequence of execution of the convolution operation at s=1, p=1

Hyperparameters of the convolution layer. The convolution layer's four most crucial hyperparameters are as follows:

1. Filter size: 3x3 filters are frequently used. Yet, there are other cases when 5x5, 7x7, or even 1x1 filters are utilized, depending on the type of application.

2. Number of filters: it is the parameter that mostly varies between 32 and 1024, equal to the power of 2. The model gets more powerful when more filters are utilized, although overfitting in the network is possible because of the increased number of parameters. Typically, in the first layers the number of filters is small, then the number of filters is gradually increased as the network becomes deepens.

3. Stride: The value is assumed to be 1.

4. Padding: Padding is frequently utilized.

Activation function. Deep learning is frequently employed to resolve non-linear problems. This is because deep learning has outperformed other approaches in addressing non-linear problems. The results of the convolution layer's matrix product are linear values.

After each convolution layer, a nonlinear activation function (elu, selu, relu, tanh, sigmoid, hardsigmoid, softplus, softsign, linear, and exponential) is typically used to transform these values into nonlinear. For deep neural networks, the non-linear rectified linear block relu has been chosen as the primary activation function due to its ease of use and speedy training.

Pooling. Pooling layer is regularly embedded into the CNN architecture. Its main job is to compress each feature map and extract the greatest number of pixel values

from the grid in order to shrink the size of the image. By using max pooling, the largest elements of the input feature map are extracted using a 2x2 window with a stride of 2, and the output feature map that is twice as tiny as the input is created. While using average pooling, the average sum of all values in the same window is calculated [65, p. 17]. In CNN designs, pooling is often carried out using a 2x2 window with stride 2 and no padding (figure 22).

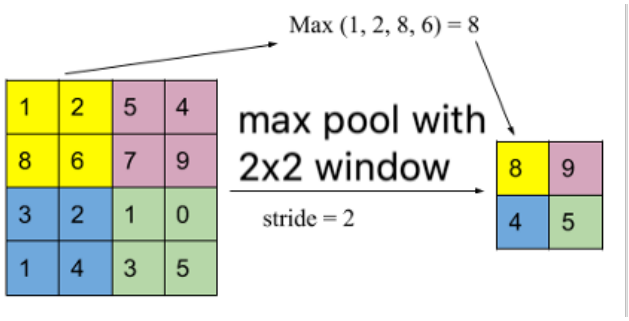


Figure 22 – Max pooling operation

Flattening and fully connected (FC) layer. The model will be able to comprehend the features once it has gone through the aforementioned procedures. The fully connected layer (FC) follows the convolution, activation, and pooling layers. This layer depends on all the neurons of the previous layer. A multilayer perceptron is comparable to a fully connected layer (FC). While the fully connected layer (FC) anticipates a 1D vector of integers, the output of the Conv and Pooling layers is consistently a 3D volume [65, p. 19].

The output of the last pooling layer is then flattened into a vector and used as the input for the fully connected layer (FC). Moreover, it is supplied into the neural network nodes that carry out the classification. A convolutional network and a neural network are connected using full connection, which then assembles the network (figure 23).

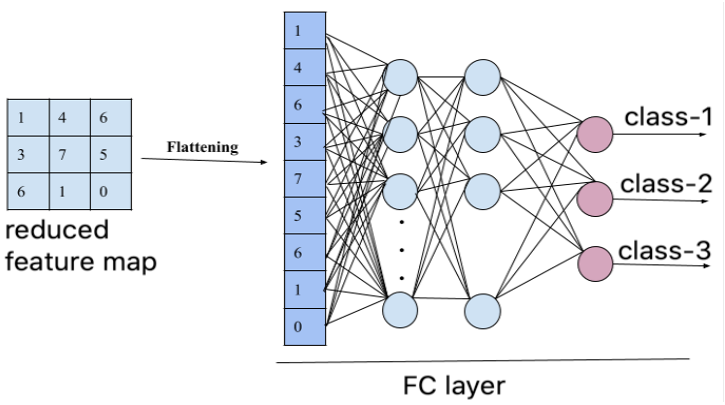


Figure 23 – Fully connected (FC) layer

3.1.2 Different CNN model architectures for image classification

LeNet. Although the LeNet network was proposed by LeCuN in 1998, limited computing capabilities and memory volumes made it difficult to implement the algorithm until 2010. LeCun et al. proposed a CNN with a back propagation algorithm

and experimented on the MNIST dataset of handwritten numbers to achieve state-of-the-art accuracy. The base configuration of LeNet-5 consists of the following components: two convolution (Conv) layers, two Average pooling layers, two Fully Connected (FC) layers and a Softmax classifier in the output layer. Input is a 32x32x1 grayscale image. In the first step, a 28x28x6 feature map is obtained by applying six 5x5 filters with stride=1 without applying any padding. Using Average pooling with a 2x2 filter with stride=2, the size is reduced by a factor of 2, and the result is 14x14x6. Then a 10x10x16 feature map is obtained by applying a second convolution layer (Conv) with sixteen 5x5 filters. This feature map passes through the pooling layer and has a size of 5x5x16. The next layer is the Fully connected layer (FC) with 120 neurons. 400 (5*5*16=400) neurons in the previous layer are connected to these 120 neurons. These 120 neurons are connected to 84 neurons of the next fully connected layer (FC) and send this connection to a softmax (previously tanh) classifier to recognize numbers from 0 to 9 [65, p. 26]. LeNet network structure is shown on the following (table 1).

Table 1 – LeNet network architecture

Layer		Size	Filter size	Stride	Activation function
Input	Image	32x32x1	-	-	-
1	Convolution	28x28x6	5x5	1	tanh
	Average Pooling	14x14x6	2x2	2	
2	Convolution	10x10x16	5x5	1	tanh
	Average Pooling	5x5x16	2x2	2	
3	Dense	1x1	5x5	1	
4	FC	84	-	-	
5 (output)	FC	10	-	-	Softmax

AlexNet. Although LeNet began the history of deep CNN, at that time CNN was limited to handwritten number recognition tasks and could not do a good job for all image classes. The most difficult ImageNet competition for visual object identification, known as the ImageNet Large Scale Visual Recognition Competition, was won in 2012 by Alex Krizhevesky et al. who offered a deeper and more comprehensive CNN model compared to LeNet (ILSVRC) [65, p. 26]. Compared to all existing machine learning and computer vision techniques, AlexNet has attained state-of-the-art recognition accuracy. It was an important breakthrough in the field of computer vision and machine learning to perform visual recognition and classification tasks, and a point in history where interest in deep learning grew rapidly.

The depth of the AlexNet architecture was increased from 5 (LeNet) to 8 layers to apply CNN to different image categories. The model consists of 5 convolution layers in the feature extraction part and 3 fully connected layers in the classifier part. The use of smaller size filters such as 5×5 and 3×3 is now normal. Input images are fixed to 224x224 size with three color channels. As in LeNet, the pattern of increasing the number of filters in each layer was preserved: 96, 256, 384, 384, and 256. Also, in each layer, the filter size decreased: in the initial layers, the size decreased from 11×11 to 5×5, and further 3×3 size filters were used in deep layers. Max pooling operation is

performed using 3×3 dimensional filters with stride=2. The use of smaller size filters such as 5×5 and 3×3 is now normal. Table 2 shows the full AlexNet network architecture.

Table 2 – AlexNet network architecture

Layer		Size	Filter size	Stride	Activation function
Input	Image	224x224x3	-	-	-
1	Convolution	55x55x96	11x11	4	ReLU
	Max Pooling	27x27x96	3x3	2	
2	Convolution	27x27x256	5x5	1	ReLU
	Max Pooling	13x13x256	3x3	2	
3	Convolution	13x13x384	3x3	1	ReLU
4	Convolution	13x13x384	3x3	1	ReLU
5	Convolution	13x13x384	3x3	1	ReLU
	Max Pooling	6x6x384	3x3	2	
6	FC	4096	-	-	
7	FC	4096	-	-	
8 (output)	FC	1000	-	-	Softmax

Below in (table 3) is given the difference between LeNet and AlexNet networks.

Table 3 – Difference between LeNet and AlexNet

Parameters	LeNet	AlexNet
Activation function	tanh	ReLU
Pooling	Average	Max
Dropout regularization	-	+
Data augmentation	-	+

ZFNet/Clarifai. In 2013, Matthew Zeiler and Rob Fergue won the 2013 ILSVRC with CNN architecture, which was an extension of AlexNet. The network was named ZFNet [65, p. 29] in honor of the authors. Since CNN is computationally expensive, optimal use of parameters is required in terms of complexity of models. The ZFNet architecture is an improvement of AlexNet, with a modified version of its network parameters. Here, instead of 11x11 filters, 7x7 filters are used to significantly reduce the number of weights. This significantly reduces the number of network parameters and increases the overall accuracy of recognition.

VGG (Visual Geometry Group). The development of deep neural networks for Computer vision tasks was a little blurry after AlexNet. In 2014, Karen Simonyan and Andrew Zisserman presented a much deeper VGG network [65, p.30]. Their model was developed for the ILSVRC 2014 competition. In their work, the authors showed that the depth of the network plays an important role to achieve better recognition or classification results in CNN. The first important difference of the network is the use of many smaller size filters, in particular stride=3×3 and 1×1 with stride=1 step. VGG networks use examples of two, three, and even four convolution layers before applying

the max pooling layer. The reason for this is that several stacked layers with small filters approximate the effect of a single convolution layer with a large-sized filter. Another important difference is the sheer number of filters used. The number of filters increases with the depth of the model, it starts from 64 and increases to 128, 256, at the end of feature extraction part it reaches to 512. Several versions of the architecture have been developed and evaluated, although two of them are often mentioned in terms of performance and depth. They are named by the number of layers: VGG-16 and VGG-19. Table 4 shows the full network structure of VGG network.

Table 4 – VGG network architecture

Layer	VGG-16	VGG-19
input 224x224x3		
1	Conv-64	Conv-64
2	Conv-64	Conv-64
3	Conv-128	Conv-128
4	Conv-128	Conv-128
5	Conv-256	Conv-256
6	Conv-256	Conv-256
7	Conv-256	Conv-256
8	Conv-512	Conv-256
9	Conv-512	Conv-512
10	Conv-512	Conv-512
11	Conv-512	Conv-512
12	Conv-512	Conv-512
13	Conv-512	Conv-512
	maxpool	
14	FC-4096	Conv-512
15	FC-4096	Conv-512
16	Fc-1000 (softmax)	Conv-512
		Maxpool
17		FC-4096
18		FC-4096
19		Fc-1000 (softmax)

GoogleNet (Inception v1). Inception – it has filters of different sizes (for example 1×1, 3×3, 5×5) a concatenate block consisting of parallel pile layers and a 3×3 max pooling layer, and their results are combined.

Winner of the GoogleNet 2014-ILSVRC competition. The main goal of the GoogleNet architecture was based on achieving high accuracy at reduced computational costs [65, p. 31]. It proposed a new concept for the Inception blog at CNN, so this network architecture is sometimes called Inception v1. Inception is a concatenate block that consists of parallel convolution layers with filters of different sizes (eg 1×1, 3×3, 5×5) and a 3×3 max pooling layer (figure 24).

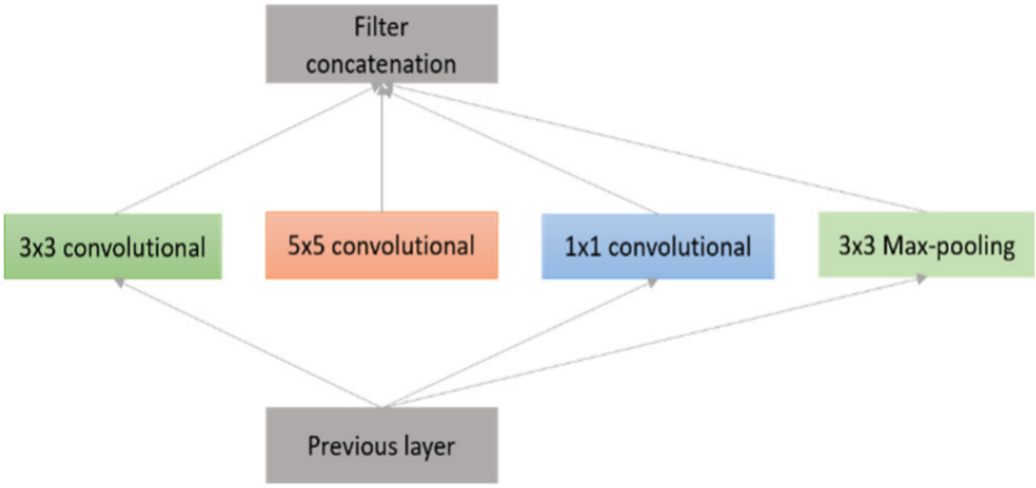


Figure 24 – Inception block: Naive version

The problem with the implementation of the Naive version of the Inception model is that the number of filters (depth or channels) starts to increase rapidly when the Inception modules are merged. Performing a convolution of filters of large sizes (for example, 3 and 5) can be computationally expensive when the number of filters is large. In order to solve this problem, 1x1 convolution layers are used to reduce the number of filters of the Inception model. In particular, before the 3×3 and 5×5 convolution layers and after the pooling layer.

The difference between the Naive Inception layer and the last Inception layer lies in the addition of 1×1 convolution kernels. These kernels made it possible to reduce the size to computationally expensive layers. GoogLeNet has a total of 22 layers, far higher than any network before it. However, the number of GoogLeNet network parameters used was much smaller than the previous AlexNet or VGG. While GoogLeNet has 7M network parameters, AlexNet has 60M and VGG has 138M network parameters. While GoogLeNet has 7m network settings, AlexNet has 60m and VGG 138m network settings (figure 25).

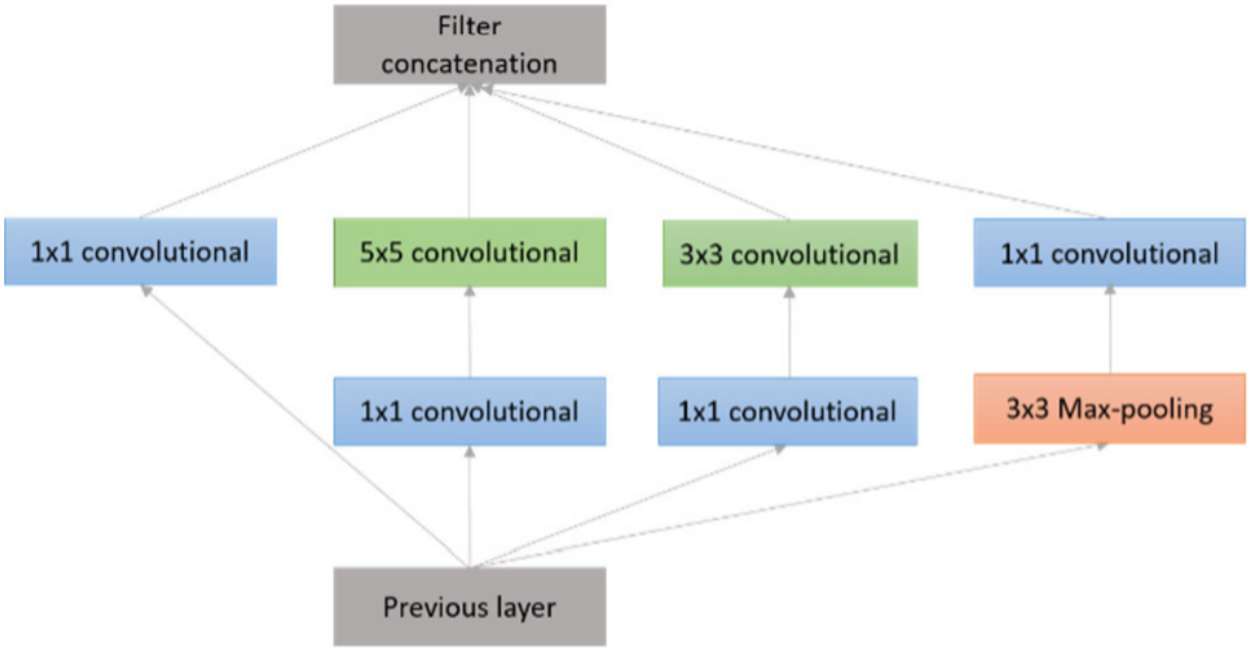


Figure 25 – Inception block with reduced size

Residual Network (ResNet 2015). The winner of the ILSVRC 2015 challenge was the ResNet residual network. Kaiming He developed the Resnet model in order to create very deep networks that do not suffer from the vanishing gradient problem. The main point of the model is the idea of residual blocks using shortcut connection [65, p. 33]. That is, these are simple connections in a network architecture, where the input is stored as it is (not measured), and then goes to a deeper layer, for example, by skipping the next layer (skip connection). ResNet is made with different number layers: 34, 50, 101, 152 and even 1202. The popular ResNet-50 consists of 49 convolution layers and 1 fully connected layer at the end of the network. Traditional feed-forward networks with residual connections are called ResNet. The output of the residual layer is determined based on the *(l-1)*-th output from the previous x_{l-1} layer. $F(x_{l-1})$ – is the output obtained after performing various operations, Batch normalization, which is accompanied by the ReLU activation function in x_{l-1}.

$$x_l = F(x_{l-1}) + x_{l-1}$$

A residual block is a model of two convolution layers with ReLU activation, where the output of the block is combined with the input of the block, i.e., a shortcut connection is performed (figures 26, 27).

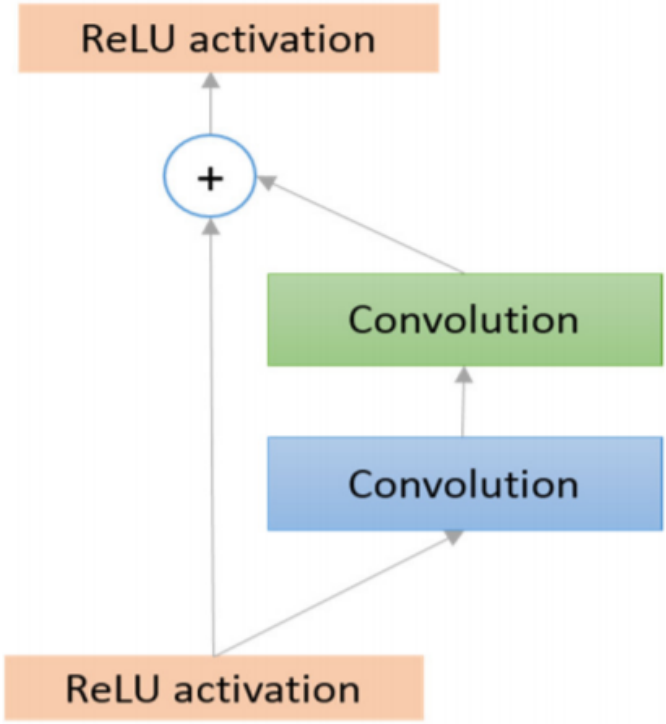


Figure 26 – Basic drawing of the residual block

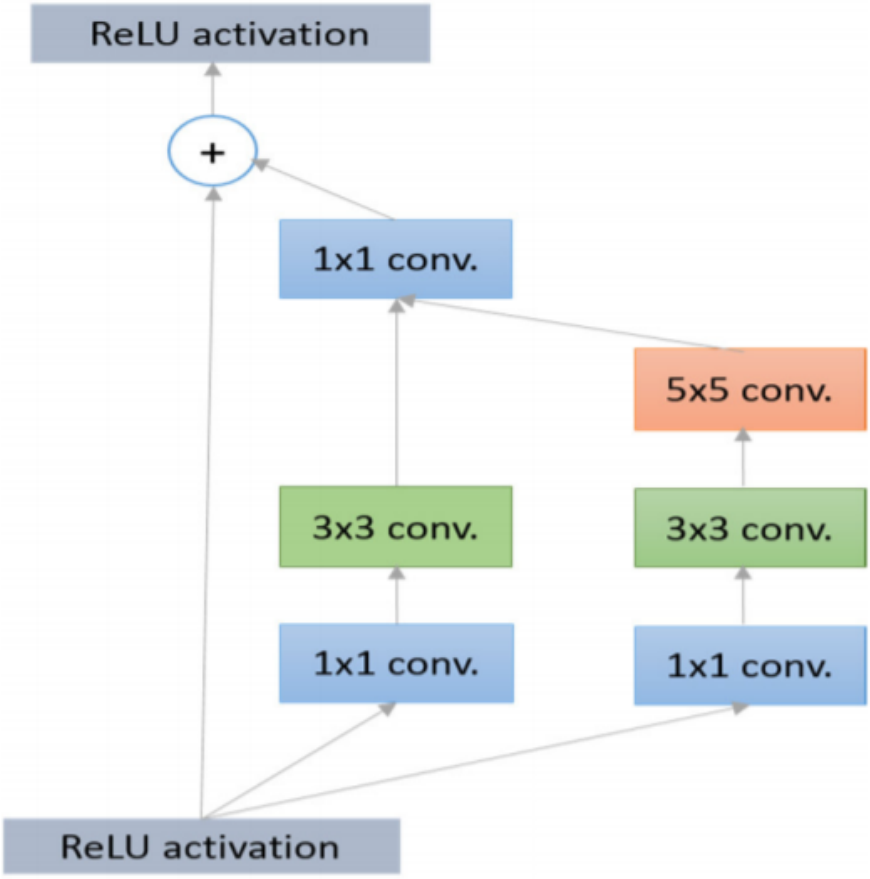


Figure 27 – Basic block diagram of the Inception residual block

3.2 Object detection algorithm

3.2.1 Background subtraction algorithm

The detection of flying or moving objects from a video sequence may be done using a variety of techniques, including background subtraction, the optical flow approach, edge detection, and frame differencing. Optical flow uses the relative velocities of the items in the picture to detect moving objects and estimate motion in a video. The optical flow approach cannot be used for real-time detection applications due to its complex computation. A frame differencing method separates the moving objects from a video sequence by computing the difference between the current and previous frames. Although while frame differencing has several benefits, such as rapid implementation, adaptability to dynamic changes in the environment, and relatively cheap computational requirements, it is typically ineffective for collecting all the pertinent pixels from moving regions. Two-point background subtraction has been implemented in preparatio of the proposed real-time drone detection system. This video processing method was used to identify drones in the scenario with a static background [66].

The following procedures have been used for background subtraction:

1. Acquisition of two distinct flow variables.
2. Determining the difference between the two frames' subtraction to determine which pixels have changed.
3. Using the cv2.COLOR BGR2GRAY filter to convert the video frames' RGB color matrices into unidimensional matrices that match to grayscale images.
4. Applying the noise-removal and edge-blurry GaussianBlur filter to the images.
5. Conversion of the blurred grayscale image into a binary image based on a predetermined threshold.
6. Extending the detected region more evenly by enlarging the white pixels using the morphological operations such as dilation method.
7. Updating an older frame with the current frame.

For pixels that change their values in the subsequent frame, the binary image transformation enables set the value 255, which means white color, while the unchanged pixels are made to be 0, which is black color. The final processing is depicted (figure 28), where the essential processing steps are visible. The original video frame may be seen in the upper right corner. The frame subtraction is reported there, to its right. The threshold application is shown in the bottom right corner, and the final processing – where the altered pixels have been dilated to better identify the drone – is visible to its right.

Figure 28 – The main steps of the background subtraction process

Chapter's conclusion. Classification, detection, and segmentation are the major three problems in the field of computer vision research. In order to draw a bounding box around a drone in a frame, we need to know what and where the item in the frame is. This information is needed for object detection task.

The background subtraction method is used to separate a foreground object from a video sequence's background. Because it can be used for real-time detection and is quick and accurate, the background subtraction approach is one of the most used detection techniques. It is also simple to implement. This method's primary flaw is that moving cameras cannot utilize it since the background of each shot changes. All of the brief films in the dataset for the Drone-vs-Bird identification challenge were captured by a static camera at a great distance when we utilized a video with a static background. Therefore, a background subtraction technique served as the foundation for our motion detector.

4 THE PROPOSED REAL-TIME DRONE DETECTION SYSTEM IN THE SCENE WITH A STATIC BACKGROUND

The first experiment is concentrated on the real-time drone detection task in a scenario with a static background. This task was split into two distinct steps: the first step deals with moving object detection task, as well as the second task is classification the detected object as drone, bird, and background. Therefore, the proposed drone detection method comprises of two modules as shown (figure 29) moving objects detector and a drone-bird-background classifier.

Figure 29 – The proposed drone detection pipeline

Background subtraction (BS) method is served as a motion detector, as well as all the moving objects in the scene are the outputs of this module. In order to distinguish drones from other the flying objects, all the detected moving objects are fed to CNN classifier. CNN classifier was trained on the entire dataset of image frames taken from different video signals. In the third chapter, the theoretical description of both methods is separately explained in detail and below are the steps of the experiment.

4.1 Moving Objects Detection

The detection of any objects in a scene that are moving are found based on motion detector. This module's effectiveness was evaluated using its Recall value. Using the dataset from the Drone-vs-Bird challenge, experimental studies of several motion detectors were carried out. The motion detector based on the two-point background subtraction technique [67] managed to attain the highest Recall. The result of background subtraction method is shown (figure 30).

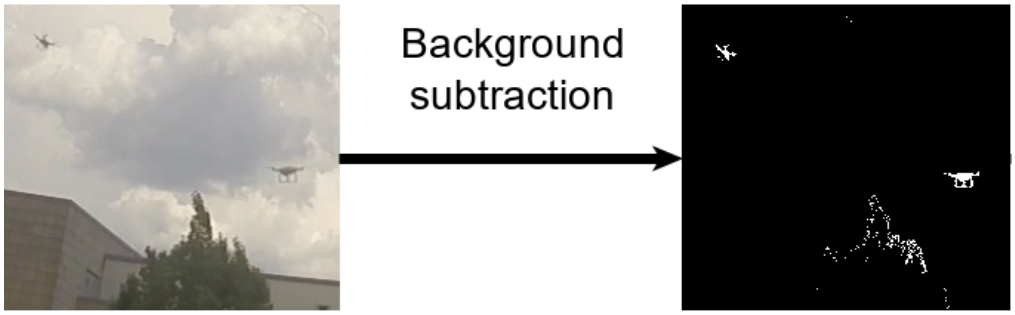


Figure 30 – The result of Background subtraction algorithm

The common background subtraction technique's output is a binary image, where pixels that change their values in the following frame are assigned the value 1 while the pixels that remain unaltered are set to 0.

The output image of BS algorithm includes single pixels that are dispersed throughout the image along with moving objects. These single pixels cause noise, as well as to get rid on the noise the output binary image is filtered using various filters. Figure 31 displays an illustration of a filtered binary image.

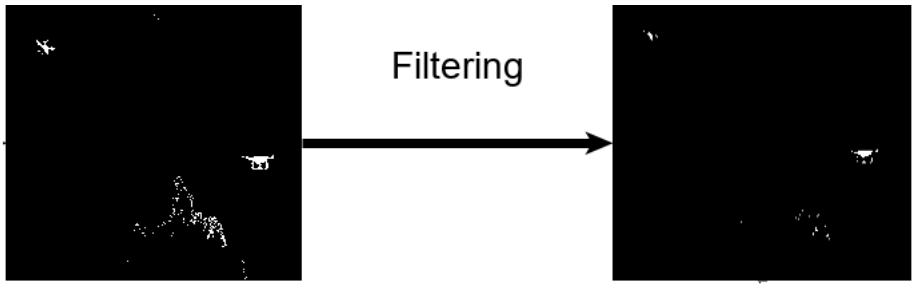


Figure 31 – Filtered binary image

The next step is to apply morphological operation called dilation to join closely spaced pixels. This operation helps to increase the detector's processing speed by reducing the number of distinct regions, which are checked by CNN classifier. The result of dilation operation is illustrated (figure 32).

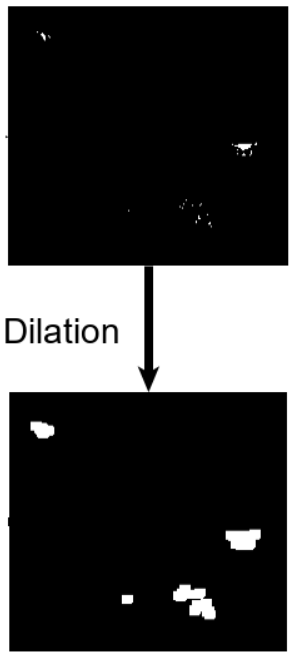


Figure 32 – The output image of dilation operation

Finding the bounding boxes that cover the regions identified in the previous step is the final stage on the moving object detector. Finally, all detected bounding boxes are forwarded to the drone-bird-background classifier. The detected bounding boxes is illustrated (figure 33).

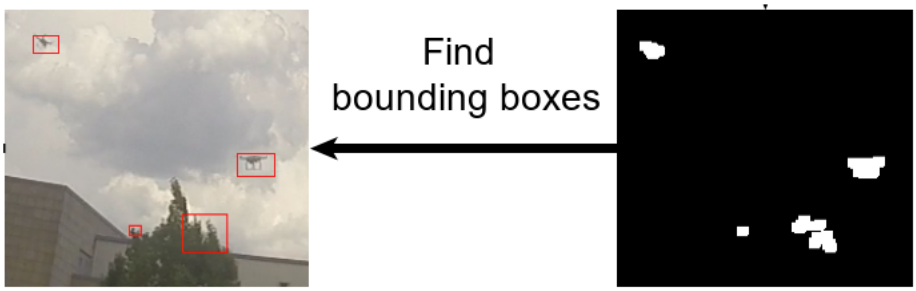


Figure 33 – The output image of dilation operation

All above mentioned steps together form the entire model of proposed drone detection system (figures 34, 35).

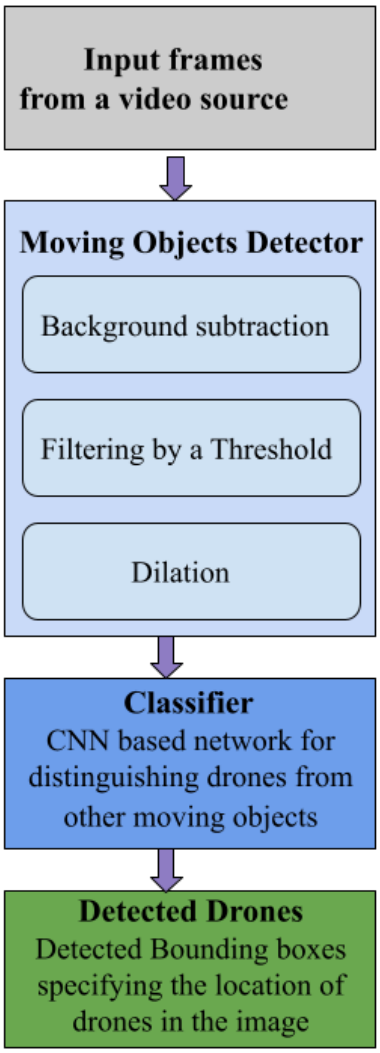


Figure 34 – The steps of the proposed drone detection algorithm

Figure 35 – Proposed drone detection system structure

4.2 Moving Objects Classification

The classification of detected or found objects is one of the crucial parts of the proposed approach. Drones, birds, aircraft, insects, and moving scene parts are among the moving objects that are detected in real-world scenarios. Consequently, we made the decision to use a classifier that categorizes all the detected objects into three classes: drones, birds, and background respectively. The moving objects classification task was performed using CNN classifiers such as VGG-16 and MobileNetV2 [68]. The CNN was chosen because of its great accuracy and short inference time. [69] claims that a detector using the MobileNet [70] backbone network has the highest detection speed. With a considerable increase in accuracy, MobileNetV2 is well-known as an upgraded version of MobileNet. The detailed theoretical description of both CNN classifiers was explained in Chapter 3. There are 19 original basic blocks, referred to as bottleneck residual blocks, in the MobileNetV2 network design. After these blocks, a 1x1 convolution layer with an average pooling layer was added. The network design ends up with a classification layer. We compared two classifiers such as VGG-16 and MobileNetV2, both of which were modified from the original network design, as well as based on the experiment results, we chose MobileNetV2 classifier due to the highest classification accuracy. By adjusting the stride, padding, and filter size, the modified network became more appropriate for small images. As well as, we made changes to the classification layer, as a result the network was focused to classify only above three classes.

To perform above mentioned tasks the experiment consists of the following parts:

– data preparation;
– training;
– evaluation Metrics;
– results;
– discussion.

Data Preparation. In order to train CNN, the amount of data is essential. The network's generalization ability is impacted by insufficient data. As a result, whenever the network receives new data, the classification accuracy is decreased. The Drone-vs-Bird challenge dataset consisting of 11 videos shot with a static camera, served as the training data. The videos may include a drone as well as birds and other moving objects. Therefore, all moving objects that appear in the frames of each video are annotated. The ground truth bounding boxes' coordinates and sizes are used as annotations. We collected 10,155 drone images from the videos and their annotations. We used the moving object detector discussed in the preceding chapter to the whole dataset in order to extract images of birds and background from the videos. Afterwords, the image frames of each detected object are manually annotated (or labeled) using LabelImg Tool. As a consequence, 9348 background images and 1921 bird images were collected. We further included additional 2651 bird images from Wild Birds in a Wind Farm: Image Dataset for Bird Detection [71] because the amount of bird images was less than other two classes. Hence, the entire dataset consisted of 24,075 image frames. Since the MobileNetV2 network's input size was 32x32x3, all annotated images resized to fit the networks' input layer. Figure 36 illustrates a few examples of the resized images.

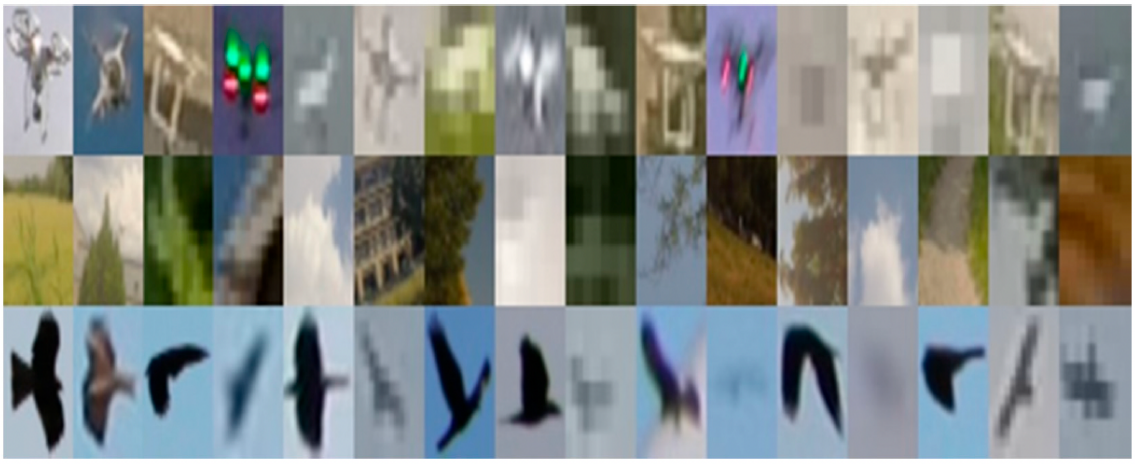


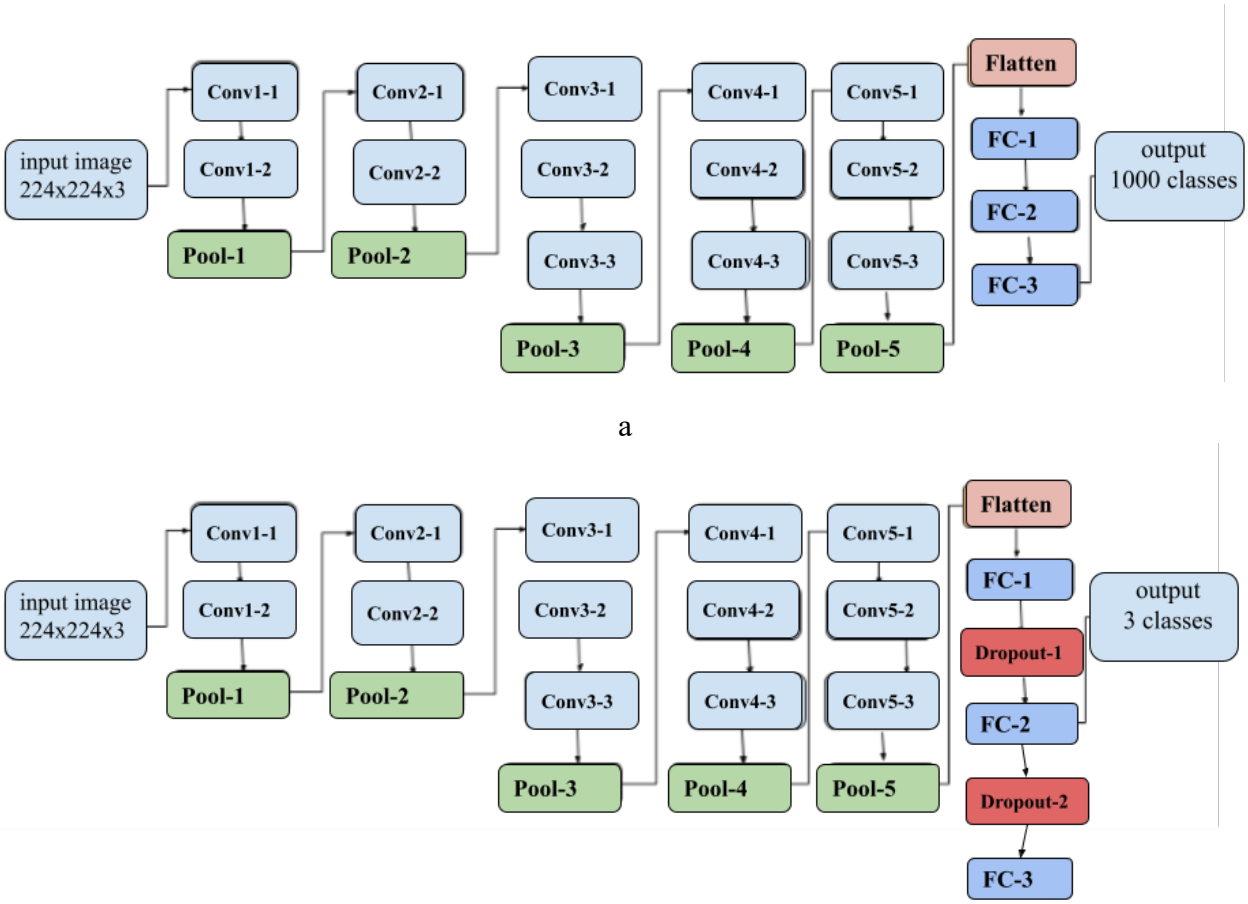
Figure 36 – Some image examples gathered for training

Note – The top row shows drones, the second row includes background images, while bird images are featured the third row

4.3 The architecture of proposed CNN classifiers

Modified VGG-16 CNN classifier. Figure 37a shows the basic original VGG-16 model with 13 convolutional (conv) and 3 fully connected (FC), i.e., 16 layers in total. The reason it is called VGG-16 is because of the number of layers. The size of the input image is 224x224, and the output softmax activation function uses the ImageNet dataset to classify it into 1000 classes. That is, if we use the original model directly, we can classify up to 1000 classes of objects. However, our study only needs to detect three classes, which are called drones, bird and background, so we modify the CNN model to adapt to only 3 classes. Also, in order to reduce the complexity of the

proposed model and prevent the model from learning excessive details, i.e., to prevent overfitting, after fully connected layers FC-1, FC-2, Dropout-1 and Dropout-2 layers with a probability of 0.5 and 0.3 added. That is, by running a dropout layer, we distinguish or ignore certain neurons with a probability of 0.5 and 0.3 when training the model. Dropout is only used during model training. The original VGG-16 model consists of 138 million trained parameters. That is, it is very cumbersome and takes a very long time to train the model if we train it from scratch. Therefore, we train only fully connected layers using pre-trained ImageNet weights. As a result, only 2.5 million training parameters are used to train the model. The modified version of VGG-16 CNN is illustrated (figure 37b).



a – the structure of original VGG-16; b – modified VGG-16 (b) models

Figure 37 – The model structures of original VGG-16 and its modified version

4.4 Evaluation Metrics

Evaluation is needed to check the performance of object recognition model and compare it with other models. ROC curves, Precision and Recall, F-scores, and False positives per image are some statistical and machine learning metrics that may be used to assess any object detection method. To assess the effectiveness of the object detection model, the Intersection of Union (IoU) idea was developed. Typically, first of all, the output of any object detector, which is a predicted bounding box is compared

to a list of manually annotated ground-truth bounding boxes. Most investigations on object detection have employed the overlap criteria, which was developed by Everingham et al. [71] for the Pascal VOC challenge, to address the question of when a detection may be regarded as accurate. As previously mentioned, the detections are allocated to ground truth objects, and their true or false positive status is determined by computing the bounding box overlap. According to [72], the overlap ratio between the predicted and ground truth boxes must be more than 0.5 (50%) in order to be regarded as a correct detection. IoU is determined by dividing the intersection by the union of the two bounding boxes: the predicted and ground truth bounding boxes (figure 38). The intersection over union (IoU), which is used to establish the Pascal VOC overlap criteria, is calculated as follows:

$$IoU = a_0 = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{3.1}$$

where Bp and Bgt stand for predicted and ground truth bounding boxes; IoU means intersection over union, while a_0 stands for an overlap ratio; $\text{area}(B_p \cup B_{gt})$ refers to the area of union of predicted and ground truth bounding boxes, whereas $\text{area}(B_p \cap B_{gt})$ refers to the overlap or intersection of these two bounding boxes. Once detections and ground truth have been matched the number of properly detected objects, also known as true positives (TPs), inaccurate detections, or false positives (FPs), and ground truth objects that the detector missed, also known as false negatives (FNs), can be calculated. Many assessment metrics might be calculated using the total number of TPs, FPs, and FNs.

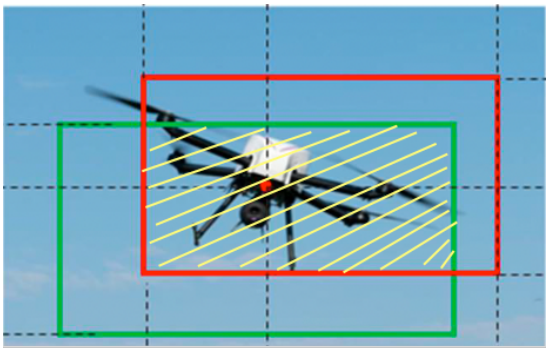


Figure 38 – An example of IoU

The higher the IoU, the closer the predicted bounding box and the ground truth bounding box are to one another. To evaluate whether the detected object is valid, a threshold value is set. The threshold value in this thesis is fixed at 0.5. If so, then:

1. If IoU ≥ 0.5, then the detected object is considered valid and is classified as True Positive (TP).

2. If IoU ≤ 0.5, then the detected object is considered invalid and is classified as False Positive (FP).

3. If the model is unable to recognize the ground truth in the image, then the classification outcome is False Negative (FN).

4. True Negative (TN) is the classification result for any part that lacks ground truth and recognized objects.

The Drone-vs-Bird challenge's metrics were used to assess our approach. Three test videos bearing the names gopro 001, gopro 004, and gopro 006 were provided by the challenge for evaluation. Frames from the first video included two drones and a moving background. The second video had a static background and a drone that was quite modest in size. The third video had multiple birds on the frames in addition to the drone.

We used our drone detector on the above-mentioned test videos in order to determine Precision and Recall by counting the total number of TPs, FPs, and FNs.

$$Precision = \frac{TP}{TP+FP} \tag{4.2}$$

$$Recall = \frac{TP}{TP+FN} \tag{4.3}$$

Recall indicates if the detection is capable to detect all the objects, whereas Precision indicates whether the detection is accurate. Precision and Recall of the drone tracking system should ideally be high in order to prevent tracking of incorrect objects or drone misses. In certain circumstances, customers might want a model with extremely high Precision and not care about the Recall since they would not want any false alarms. A model with high Recall and the ability to tolerate moderate false alarms would be desired by other customers who wish to track every suspect target. Therefore, the majority of information extraction methods may be evaluated using the metrics of Precision and Recall. They can occasionally be used independently or as the basis for derived metrics like F-score and Precision-Recall curves [73]. So, on the basis of these two metrics, we may calculate the F1-score metric, which combines Precision and Recall data. Hence, the primary evaluation metric employed in the Drone-vs-Bird challenge's is determines as below:

$$F_1 = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{4.4}$$

4.5 Experiment Results

Using the dataset described in the preceding section, we trained the modified VGG-16 and proposed MobileNetV2 CNN classifiers from scratch. 70% of the image frames taken from the video signal of the whole dataset were used for training, 15% for validation, and 15% for testing. The stochastic gradient descent (SGD) optimization approach with a starting learning rate of 0.05, a momentum of 0.9, and a weight decay of 0.001 was used to train the network. Using a 2 GB NVIDIA GeForce GT 1030 GPU, the training was carried out with the batch size of 88. Every 50 epochs throughout the training, the starting learning rate was reduced by a factor of 10.

Training the moving objects classifier based on modified VGG-16 and proposed MobileNetv2 CNN. The modified VGG-16 CNN model was trained on 70% of the entire dataset and showed a training accuracy of 99.37% for the three classes called background, drone and bird. Whereas the proposed MobileNetv2 CNN model showed a training accuracy of 99.83% for above three classes.

To use the classification metric, the test data is converted to another number format, i.e., numpy array. Then this converted code is run through the built-in classification metrics to get accurate results. Then class names are added as background, drone, and bird. Below in tables 1 and 2 were illustrated training classification metrics for both classifiers. Precision value in the table takes into account only positive predictions, while Recall value takes into account both positive and negative predictions. Training classification metrics both VGG-16 and MobileNetv2 CNN networks are given below in (tables 5, 6).

Table 5 – Training classification metrics for modified MobileNetv2 CNN classifier

Class names	Precision	Recall	F-1 score
Background	0.99	0.99	0.99
Bird	1.00	0.99	100
Drone	0.99	0.99	0.99

Table 6 – Training classification metrics for modified VGG-16 CNN classifier

Class names	Precision	Recall	F-1 score
Background	0.98	0.99	0.99
Bird	1.00	0.99	100
Drone	0.99	0.99	0.99

An epoch represents how many times the model has been trained on the entire dataset. In our case, the number of epochs is equal to 50. From the dependence graph of training and validation classification accuracies on the number of epochs shown in (figure 39), it can be seen that even 30 epochs are sufficient for training the model. Training a model with a very small number of epochs may under-learn the features and characteristics of the data and, as a result, perform very poorly when classifying new data. In machine learning, this situation is called *underfitting*, that is, when the amount of training data is small, or when the data is very small, it is known that the training is insufficient. On the other hand, if the model is trained with a very large number of epochs, the model may become unreliable when the model learns the training data in detail down to the noise and is tested against new data. This situation in machine learning is called *overfitting*. Also, the lower value of the loss function indicates the reliability of the model. According to the result of our experiment, the value of the verification loss function in the last epoch is equal to 0.0507.

It is not enough to evaluate the classification model by this result alone. To assess the reliability of a model, it is necessary to test how well it correctly predicts untrained or unseen data. In our case, the MobileNetv2 classification accuracy during testing

showed 99.83% for, while VGG-16 CNN classification accuracy was reached to 99.317% and the value of the loss function, in turn, was equal to 0.06.

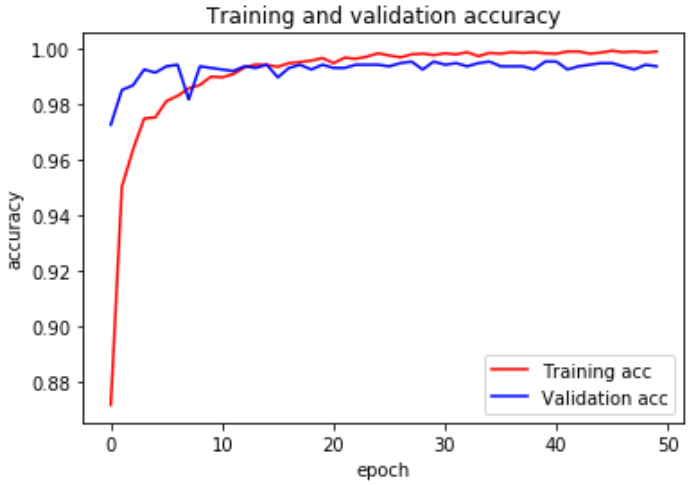


Figure 39 – Training and validation classification accuracies over 50 epochs run

However, the evaluation of how well the proposed model works should not be limited to this. There are two other ways to show how well our model predicts or classifies untrained/unseen, new data. One of them is the confusion matrix, and the other one is the classification metrics.

One of the ways to evaluate the model is to build a confusion matrix. The numpy array we created is placed inside the data frame, because the confusion matrix works well in the data frame. To evaluate the quality of the output data in the dataset of the proposed classifier, we constructed two types of confusion matrix here: normalized and unnormalized confusion matrices.

In our case, due to class imbalance, i.e., the number of data in each class is different, it is useful to construct an unnormalized confusion matrix to clearly show which class is misclassified. Here, the diagonal elements show the number of matches of the predicted class with the actual class, and the off-diagonal elements show the number of elements incorrectly predicted by the classifier. The higher the number of diagonal elements, the better the model predicts the test data. For example, out of 9991 UAV images, 9980 are correctly predicted and 11 are misclassified as background. Normalized and unnormalized confusion matrices of moving object classifiers are illustrated in (figures 40, 41).

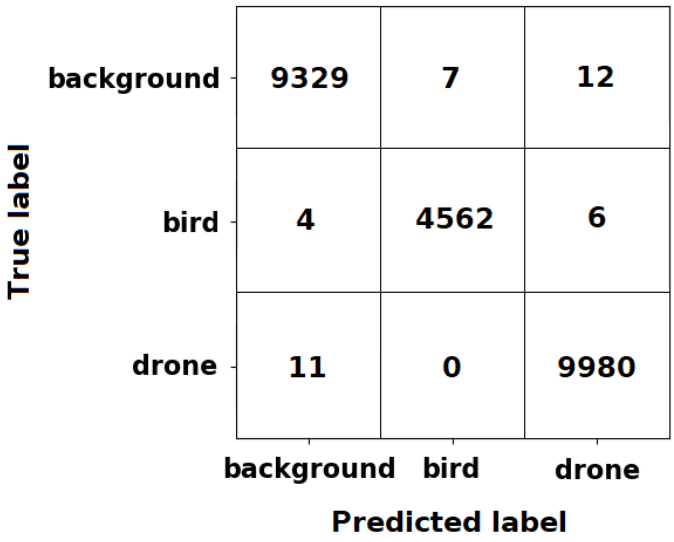


Figure 40 – Unnormalized confusion matrix of the trained convolutional neural network (CNN)

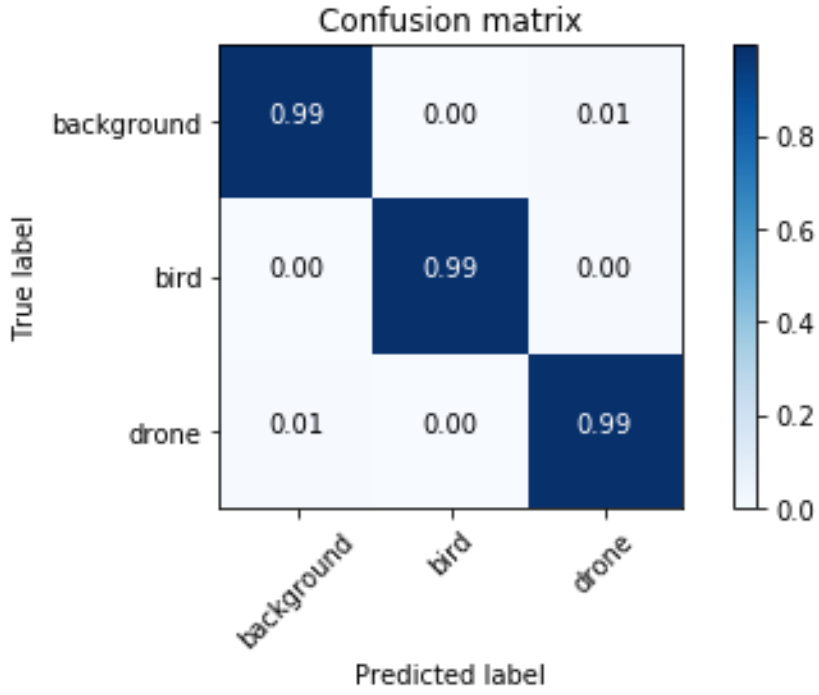


Figure 41 – Normalized confusion matrix of the trained convolutional neural network (CNN)

Figure 42 displays the experiment results attained by using proposed detector on all test videos. With an IoU of 0.5, the true positives and false positives values were counted. Depending on the size of the drone, the results were separated into three ranges.

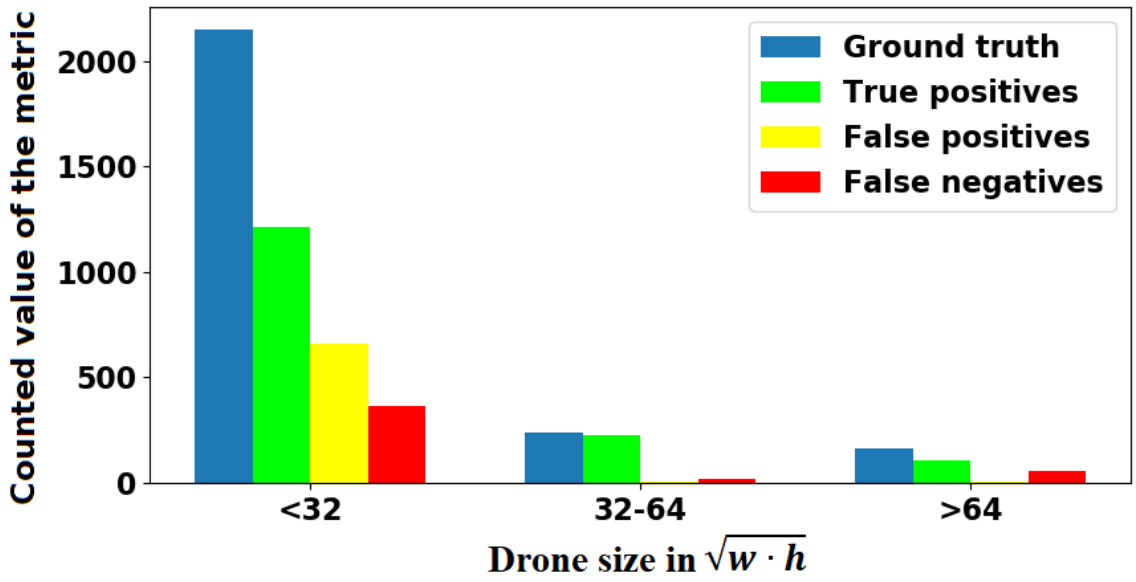


Figure 42 – Experiment results for different drone sizes

The ground truth bounding box's width and height are shown in Figure 43 as w and h, respectively, in pixels. $\sqrt{w * h}$ is the drone's size as seen in the image This parameter decreases as the distance between the drone and camera increases. Precision and Recall values were calculated based on these data. The F1-score was then determined using Equation (4.4) and appended to the final row of (table 7). Each video was separately carried out the same sequence.

Table 7 – The evaluation's results for IoU value of 0.5

Video name	Precision	Recall	F1-score
gopro_001	0.786	0.817	0.801
gopro_004	0.554	0.910	0.689
gopro_006	0.735	0.691	0.712
Overall	0.701	0.788	0.742

We performed experiments for various IoU values in order to analyze the detector in detail. As illustrated in (figure 43), the curves were drawn based on the Recall, Precision, and F1-score data that were obtained.

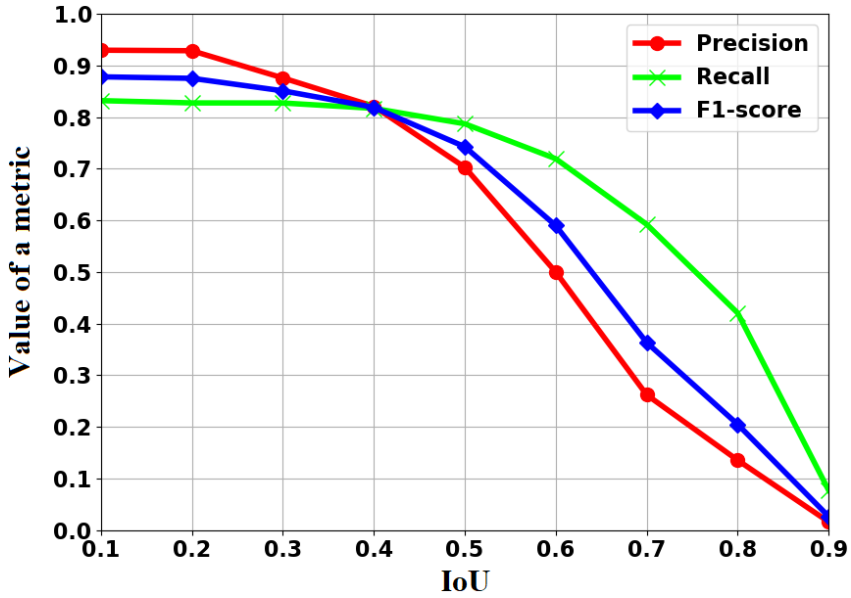


Figure 43 – Values of evaluation metrics for various IoU levels

Figure 44 shows qualitative detection results.

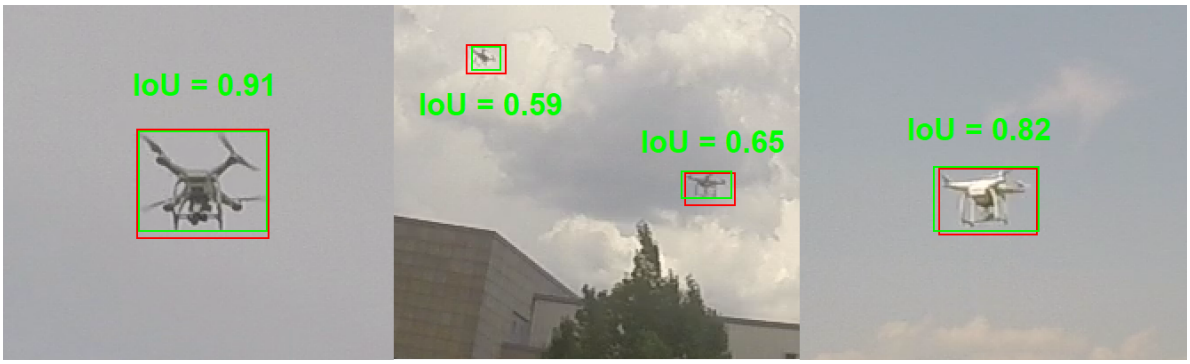


Figure 44 – Qualitative detection results

Note – Ground truth boundary boxes are highlighted in green. The results of applied detector are red bounding boxes

Inaccurate bounding box calculations were to blame for 85% of all false positives, which led to an estimated IoU value of less than 0.5. The remaining 15% were classification errors that led to the misclassification of drones as other moving objects. Figure 45 provides examples of incorrect classification-based false detections.

Figure 45 – Incorrect classification-based false detections

The detector classified the images shown in figure 45 as drones. Birds, clouds, swaying tree branches, and grass were the most often misclassified objects. The detector processed nine 1920x1080 frames per second on average. Moving object classification took up one-third of the processing time, while their detection took up the rest time. It has been discovered that the moving object detection speed was dependent on the background change rate, which was increased with increasing number of bounding boxes supplied into the classifier. Figure 46 demonstrates this dependence in further detail.

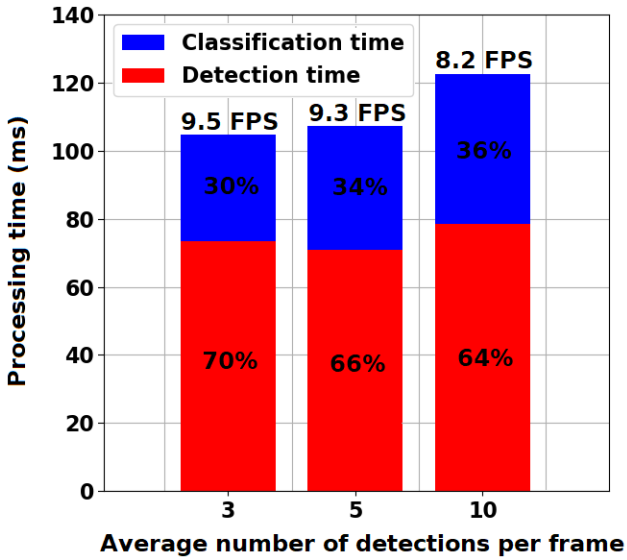


Figure 46 – The results of the detection speed evaluation

4.6 Discussion

According to conducted research, dividing drone detection process into two parts such as moving object detection and classification of detected moving objects is efficient for precise and fast drone identification. Yet there are some limitations with using motion information to detect moving objects. Firstly, as illustrated in (figure 46) moving background increased the number of detected objects, which in turn increased the classification time and the quantity of false positives. Second, as shown in (figure 47), it became difficult to distinguish the drone from other objects when it was flying nearby moving objects.

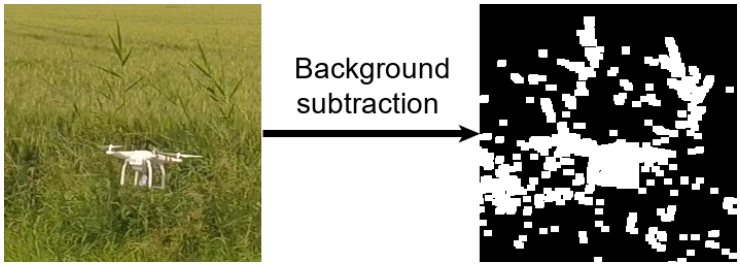


Figure 47 – The result of background subtraction method applied to the video segment when a drone was flying close to swaying grass

As a consequence, the drone was not detected which increased the false negatives. As well as, due to the fact that more images were supplied to the classifier the false positive detection also increased in number. An increased number of classification errors resulted from the classifier's accuracy was not equal to 100%. Using the metrics from the Drone-vs-Bird detection competition [74] allowed to compare research results with other teams' results, who took part in the Drone-vs-Bird detection challenge. Table 8 displays the comparison results for prior works and proposed approach.

Table 8 – Comparison results for prior works and proposed approach

Methods	Precision	Recall	F1-score
Method used*	0.756	0.713	0.734
Method used**	0.795	0.591	0.678
Method used***	0.103	0.146	0.121
Method used****	0.524	0.342	0.414
Proposed approach	0.701	0.788	0.742
* – in [42, p. 8909854-1-8909854-4]; ** – in [43, p. 8909865-1-860865-4]; *** – in [44, p. 8909830-1-890830-4]; **** – in [45, p. 8909856-1-890856-4]			

The experiment's results indicated that proposed approach's accuracy was comparable to that of the approaches put forward in [42, p. 8909854-5; 43, p. p.8909865-5]. Proposed detector had a much faster detection speed compared to [43, p. 8909865-4], where just the super resolution application was carried out at a speed of 0.58 FPS. Recall and F1-score values of the proposed approach were higher than other methods used in above mentioned works.

4.7 Chapter's conclusion

This chapter considers a real-time approach for drone detection that is comparable to current algorithms. It was clarified that dividing the drone detection task into the detection and classification steps can solve the given task effectively. The results of the experiment demonstrated both the benefits and drawbacks of the proposed approach. The main limitation of the proposed detector lies on its heavily influenced performance on the moving background existence. In order to solve this limitation and reach higher detection accuracy the next chapter is dedicated to a sensor fusion approach combining multi-angle visual information from several camera sensors.

5 PROPOSED DRONE DETECTION SENSOR VOTING SYSTEM BASED ON VISUAL DATA FROM MULTIPLE CAMERAS

Multi-sensor fusion refers to techniques for data or decision fusion from many sensors, sometimes even from distinct ones, in order to make one sensor make up for the shortcomings of another or to increase the overall accuracy or robustness of a decision-making process [75].

5.1 A three-sensor system

This research work focused on using camera sensors for UAV detection and classification task. As the fusion is performed by fusing data from sensors of the same type, all sensors have the same input data, which are RGB images. Therefore, sensors A, B, C can be referred as camera-1, camera-2 and camera-3 (figure 48).

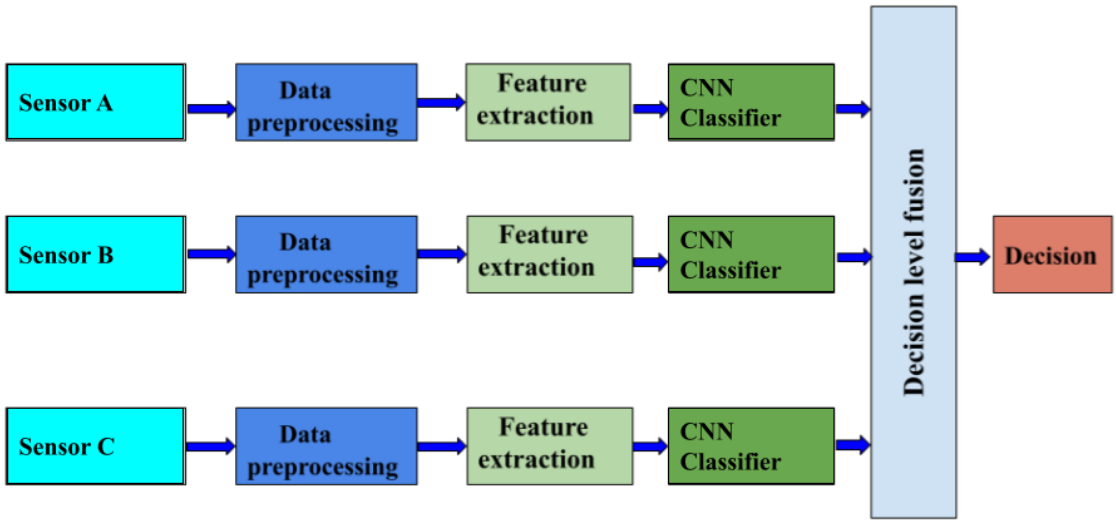
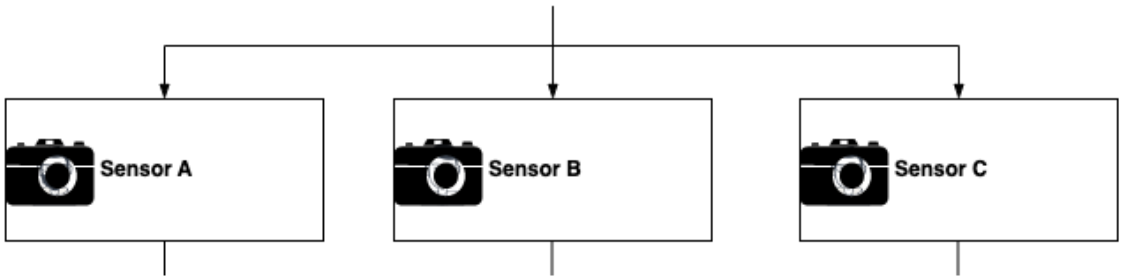


Figure 48 – Decision level fusion of three camera sensor system

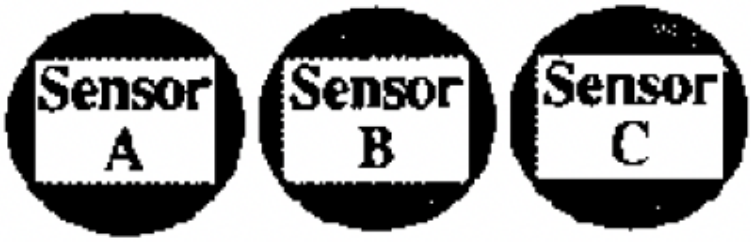
Figure 48 illustrates the late sensor fusion process, where fusion is performed in the decision level stage. As we can see the data source from each sensor is trained separately, detection and classification are made for each sensor, and in the final stage the decisions from each sensor are fused based on voting method.

Several sensor output data combinations. There are three types of data output combinations of a three-sensor system: parallel, series and series/parallel. The topologies and detection spaces of each type are illustrated in following below figures:

1. Parallel. In this topology sensors operate independently of one another. The example of parallel topology of a three-sensor fusion system and its Venn diagram with shaded sensor detection spaces are illustrated in (figure 49a, 49b).



a



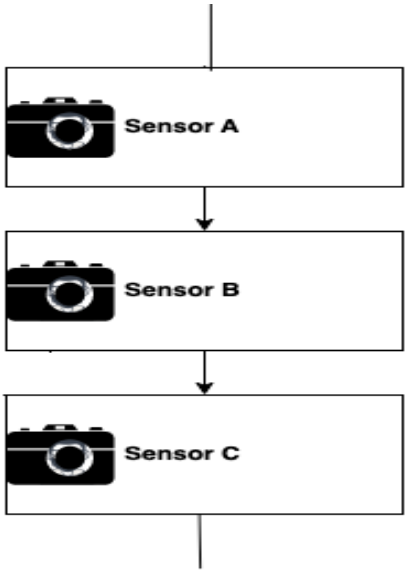
b

a – parallel topology of a three-sensor fusion system; b – Venn diagram of parallel topology. Sensor detection space is shaded

Figure 49 – Parallel sensor data output combination

2. Series

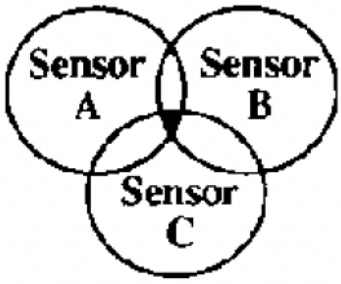
Each sensor's output influences the system's output. The example of series topology of a three-sensor fusion system and its Venn diagram with shaded sensor detection spaces are illustrated in (figure 50a, 50b).



a

a – series topology of a three-sensor fusion system

Figure 50 – Series sensor data output combination, sheet 1

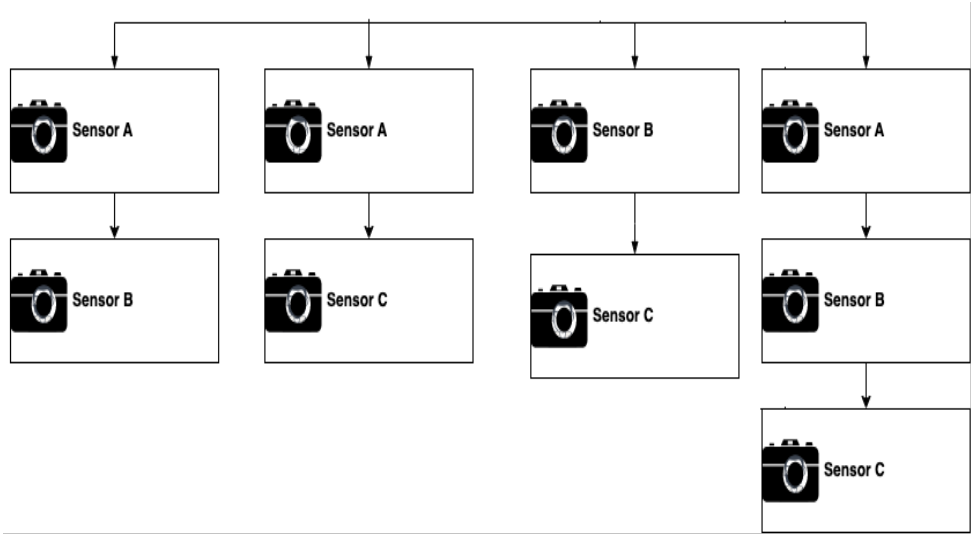


b

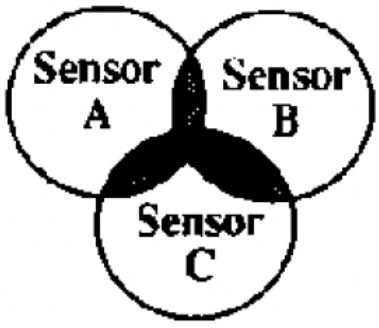
b – Venn diagram of series topology. Sensor detection space is shaded

Figure 50, sheet 2

3. Series/Parallel. The combination of several sensor outputs determines system output. The example of series/parallel topology of a three-sensor fusion system and its Venn diagram with shaded sensor detection spaces are illustrated in (figure 51a, 51b).



a



b

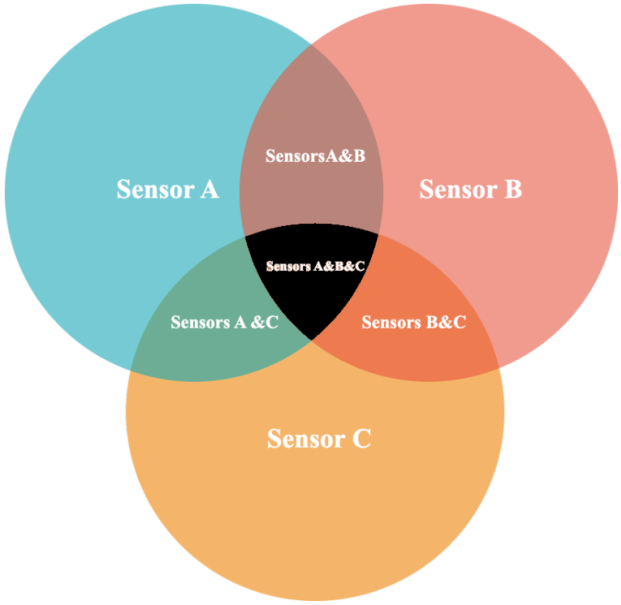
a – series/parallel topology of a three-sensor fusion system; b – venn diagram of series/parallel topology

Figure 51 – Series/parallel sensor data output combination

5.2 The proposed decision-level fusion system

The proposed decision-level fusion system is based on voting method, the series/parallel data combination was chosen as an example of voting fusion because it has the ability to detect targets that are suppressed and reject both naturally occurring false alarms from clutter and artificial decoys. It does, however, need a fusion algorithm that can handle decisions from sensors with various levels of confidence.

Venn diagrams are a useful tool for illustrating the detection space (or classification space) of several sensors. Figure 52 depicts the detection space for a three-sensor system with Sensors A, B, and C. Labeled areas indicate those with one sensor, two sensors, or three sensors interacting.



– detection modes for three-sensor system: 1 – sensors A&B&C; 2 – sensors A&B; 3 – sensors A&C; 4 – sensors B&C

Figure 52 – The detection space for a three-sensor system with Sensors A, B, and C

As Figure 52 shows, three camera sensors are connected based on the series/parallel topology, as well as in the center their detection space is shaded with black color. The main fusion combination is ABC mode, because it is the central connection all of the three sensors. Other modes are combinations of two sensors' fusion such as AB, BC and AC, respectively. These detection modes are given below in (table 9).

Table 9 – Multi-sensor detection modes for a three-sensor system

Mode	Sensor and Confidence level		
	A	B	C
ABC	A1	B1	C1
AC	A2	-	C2
BC	-	B2	C2
AB	A3	B3	-

System detection probability. Once the detecting modes have been established, Boolean algebra can be utilized to construct an expression for the detection probability and false alarm probability of the sensor system. The system detection probability equation has the following form for the example above with one three-sensor mode and three two-sensor modes.

$$SystemP_d = P_d\{A_1B_1C_1 \; or \; A_2C_2 \; or \; B_2C_2 \; or \; A_3B_3\}. \qquad (5.1)$$

Probability Axioms: 0 < p(x) < 1
P(true) = 1
P(false) = 0
By repeated application of the Boolean algebra expression given by

$$P\,(X \vee Y) = P(X) + P(Y\,) - P(X \wedge Y\,) \qquad (5.2)$$

(5.1) can be transformed into the expressions for difference and sum as [60, p. 321]:

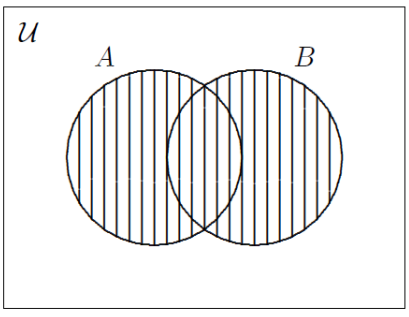
$$\begin{aligned} SystemP_d = {} & P_d\{A_1B_1C_1\} + P_d\,\{A_2C_2\,\} + \; P_d\,\{B_2C_2\,\} + \; P_d\,\{A_3B_3\,\} - P_d\,\{B_2C_2 \\ & * A_3B_3\} - P_d\,\{A_2C_2B_2\} - P_d\,\{A_2C_2 * A_3B_3\} + P_d\,\{A_2C_2B_2 * A_3B_3\} \\ & - P_d\{A_1B_1C_1 * A_2C_2\,\} - \; P_d\,\{A_1B_1C_1 * B_2C_2\,\} - P_d\,\{A_1B_1C_1 * A_3B_3\,\} \\ & + P_d\{A_1B_1C_1 * B_2C_2 * A_3B_3\} + P_d\{A_1B_1C_1 * A_2C_2B_2\} + P_d\{A_1B_1C_1 \\ & * A_2C_2 * A_3B_3\} + P_d\{A_1B_1C_1 * A_2C_2B_2 * A_3B_3\}. \end{aligned}$$

Since the confidence levels for each sensor are independent of one another (by the nonnested or disjoint assumption), the applicable union and intersection relations are.

The union of A and B is defined as

$$A \cup B = \{x \in \, | x \in A \vee x \in B\}$$

as the name suggests, the set combining all the elements from A and B.
$A \cup B \rightarrow$ A or B, which corresponds to OR logic function (figure 53).



$A \cup B$

Figure 53 – The union of two sensors

Given two sets A and B, define their intersection to be the set

$$A \cap B = \{x \in \circledcirc \mid x \in A \wedge x \in B\}$$

it means that $A \cap B$ contains elements common to both A and B, which corresponds to AND logic function (figure 54).

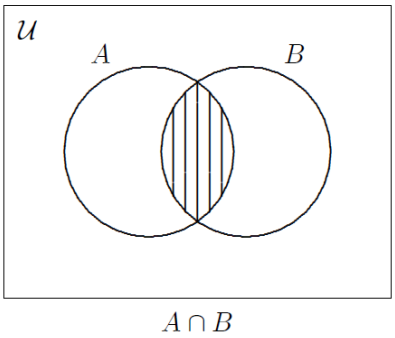


Figure 54 – The intersection of two sensors

$$P_d\{A_1 \cup A_2\} = P_d\{A_1\} + P_d\{A_2\}$$

$$P_d\{A_1 \cap A_2\} = 0$$

respectively. Analogous statements apply for the other sensors.

The above relations allow (3) to be simplified to

$$SystemP_d = P_d\{A_1B_1C_1\} + P_d\{A_2C_2\} + P_d\{B_2C_2\} + P_d\{A_3B_3\} - P_d\{A_2C_2B_2\}.$$

The four positive terms correspond to each of the detection modes, while the one negative term eliminates double counting of the {A2 B2 C2} intersection that occurs in {A2 C2} and {B2 C2}.

Each camera sensor can have output different confidence score values. In general, there are three types of confidence score values:

– high confidence (equal to or greater than 90%);

– medium confidence (between 70 and 89%);

– low confidence (less than 70%).

Below in (figure 55) three-sensor voting logic function's hardware implementation is structured based on the confidence levels of each sensor. As in the ABC mode all three sensors participate in the voting, this mode gives the correct result even in low confidence value. For the case of AB mode two high confidence values from A and B sensors are taken. As well as for AC and BC modes medium confidences of A and B sensors and high confidence from C sensors are considered.

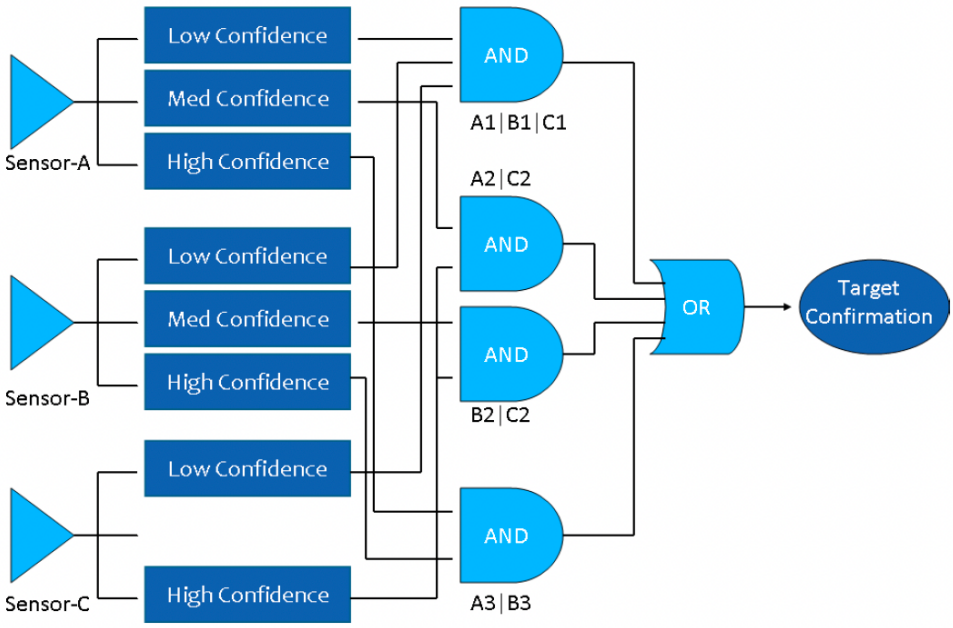


Figure 55 – Three-sensor voting logic fusion

Note – Compiled according to the source [60, p. 326]

Since the research work uses camera sensor in UAV detection, all three sensors used here are visual data based camera sensors. Two of the three sensors are panoramic cameras – Panasonic WV-SF448E (WV-SF448E), and the third one is a Sony HDR CX-405. The camera sensors planning and their angles are illustrated below (figure 56).

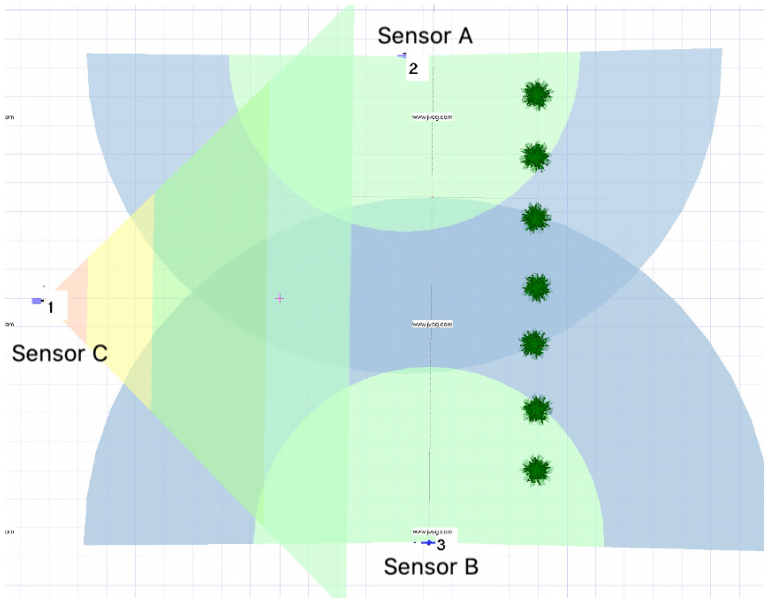


Figure 56 – Camera sensors planning

Below the parameters of each camera and its scene are illustrated (figure 57).

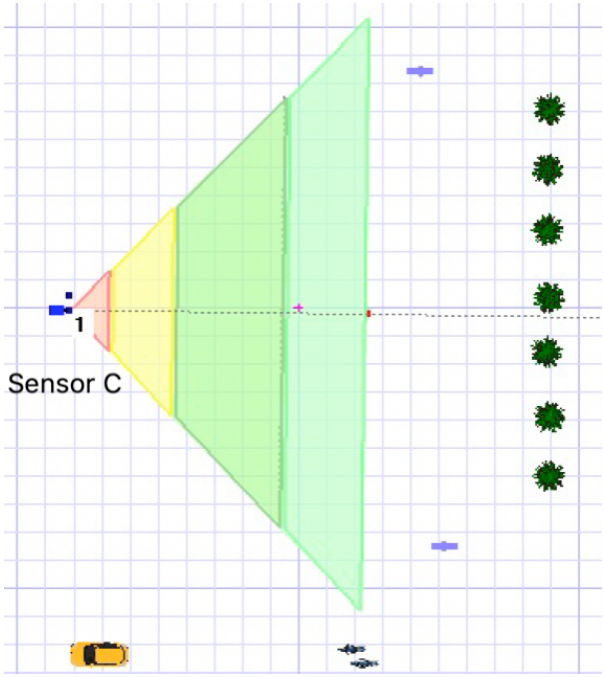


Figure 57 – The angle of camera C

Note – Camera resolution 1920x1080, focal length 2.4 meters, camera height 9 meters, camera pixel density 45 pixels/m

The view from camera C is depicted in (figure 58). The camera's resolution is 1920x1080. The camera is installed in 9 meters.

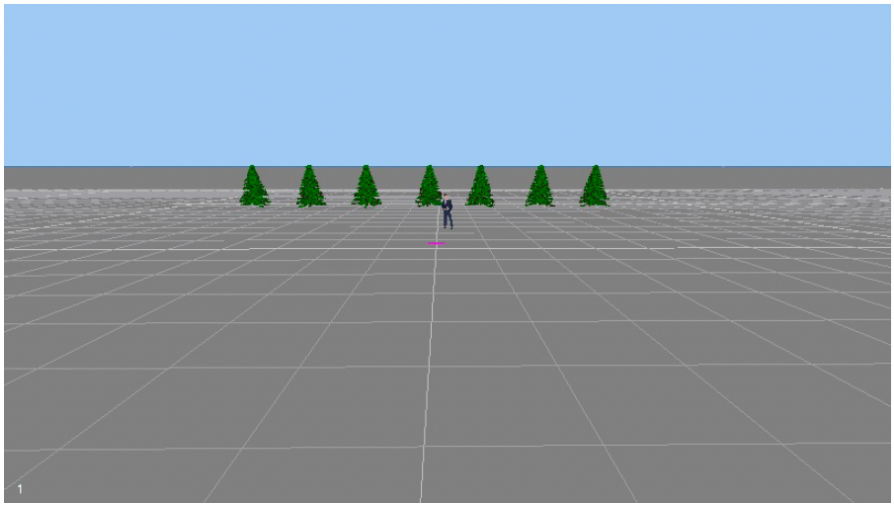


Figure 58 – View from camera C

Sensor A (figure 59).

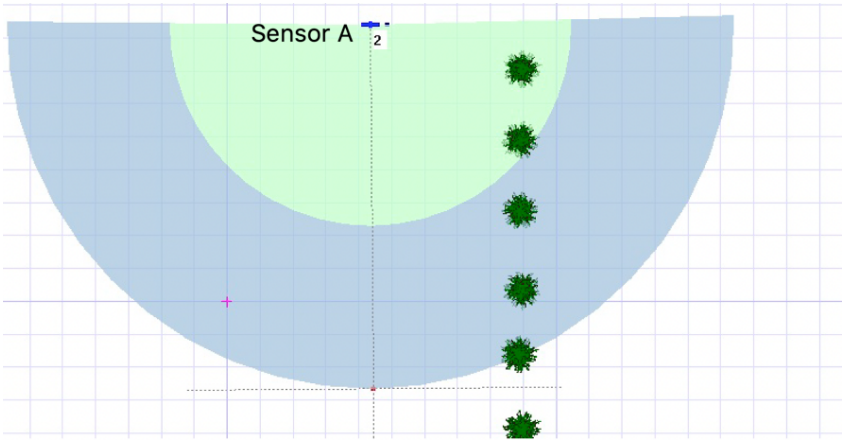


Figure 59 – The angle of camera A

Note – Camera resolution 1920x1080, focal length 0.84 meters, camera height 9 meters, camera pixel density 27x15 pixels/m

The view from camera A is depicted in (figure 60). The camera's resolution is 1920x1080. The camera is installed in 9 meters.

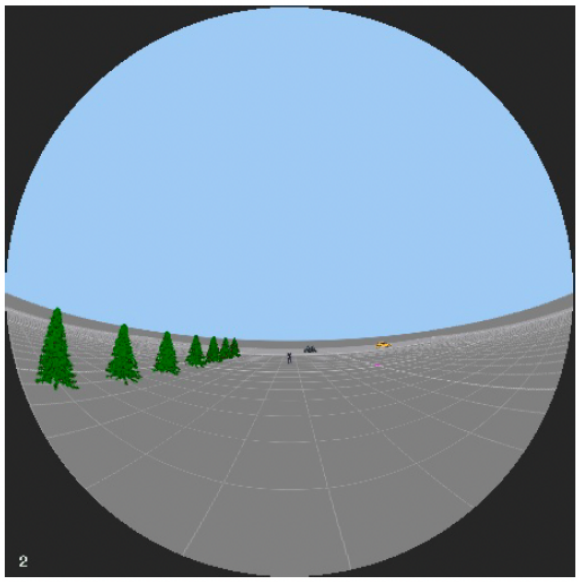


Figure 60 – View from camera A

Sensor B (figure 61).

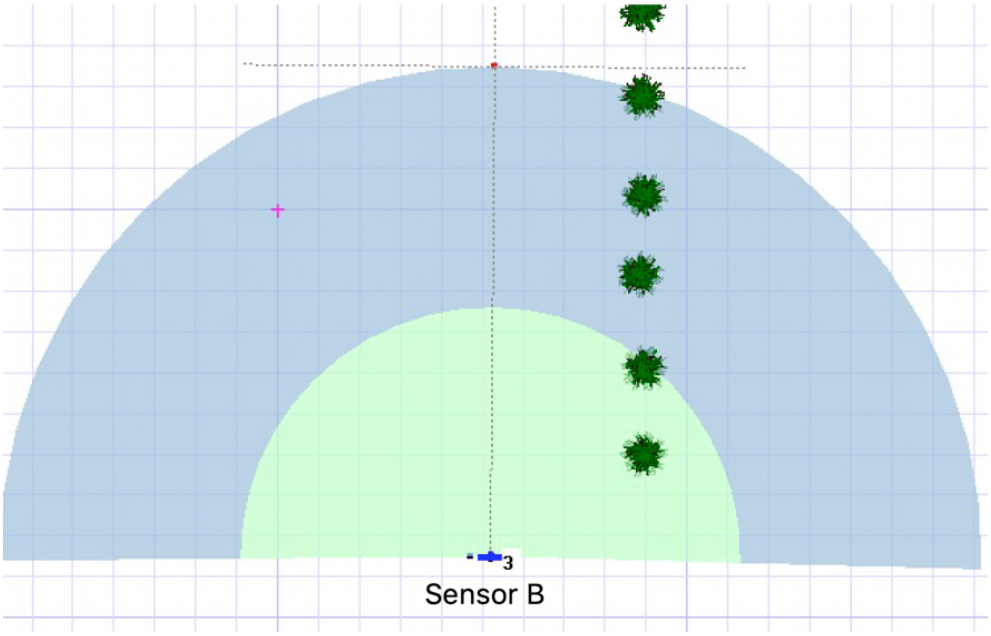


Figure 61 – The angle of camera B

Note – Camera resolution 1920x1080, focal length 0.84 meters, camera height 9 meters, camera pixel density 25x14 pixels/m

The view from camera B is depicted in (figure 62). The camera's resolution is 1920x1080. The camera is installed in 9 meters.

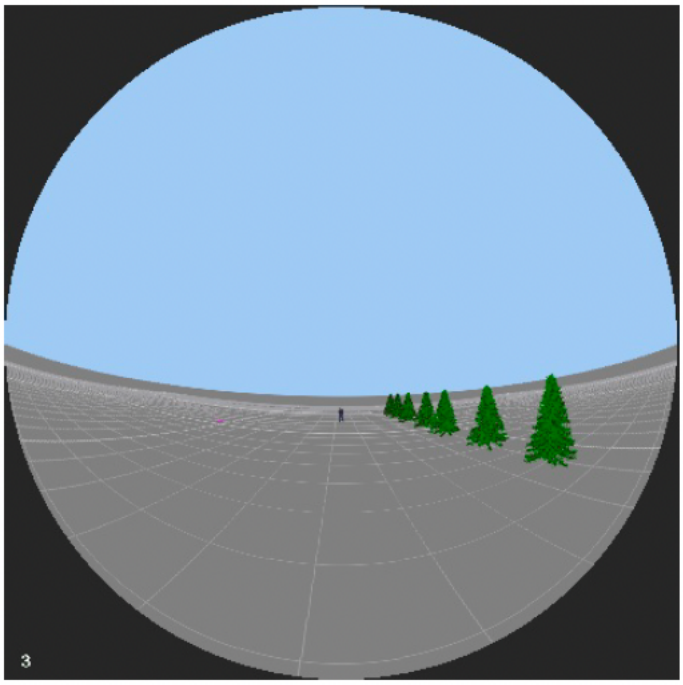


Figure 62 – View from camera B

The main condition in fusion of three camera sensors as below:
If IoU<0.5, then output of each camera sensor is 0.
If IoU≥0.5, then output of each camera sensor is 1.

Based on these conditions following final result can output different class labels (table 10).

Table 10 – Sensor fusion results based different input values

Sensor inputs			Output
A	B	C	
0	0	0	0, class label “No drone”
0	1	1	1, class label “Drone”, Alert
1	0	1	1, class label “Drone”, Alert
1	1	1	1, class label “Drone”, Alert

Below these results were obtained on the platform UnityPro (SchneiderElectric), (figures 61, 62, 63).

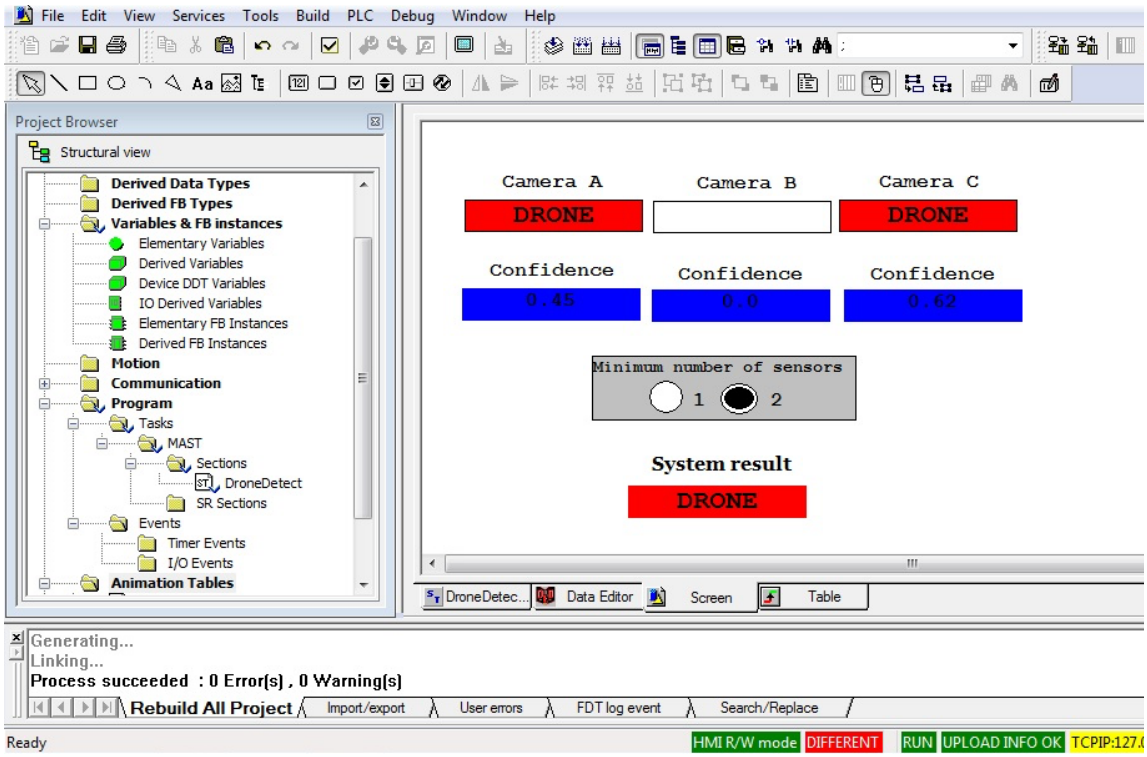


Figure 61 – The case when the drone is flying in the area of camera A and C

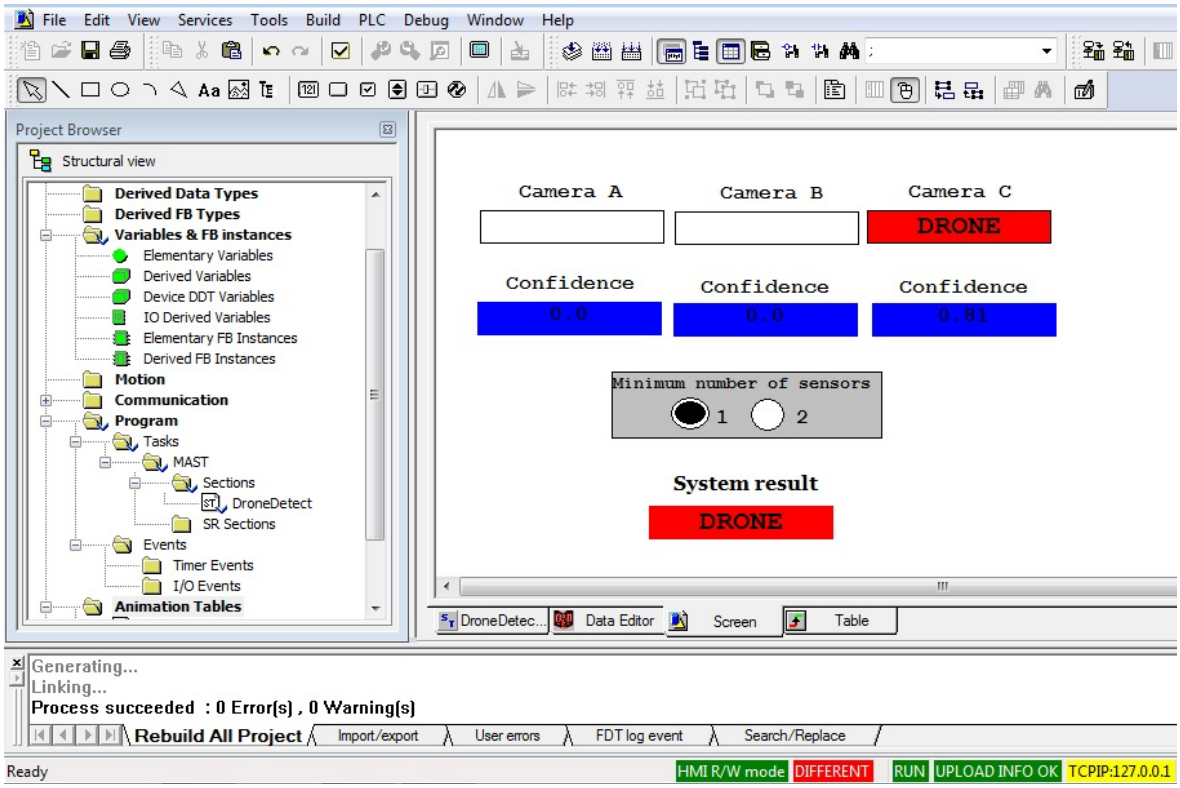


Figure 62 – The case when the drone is flying in the area of camera C

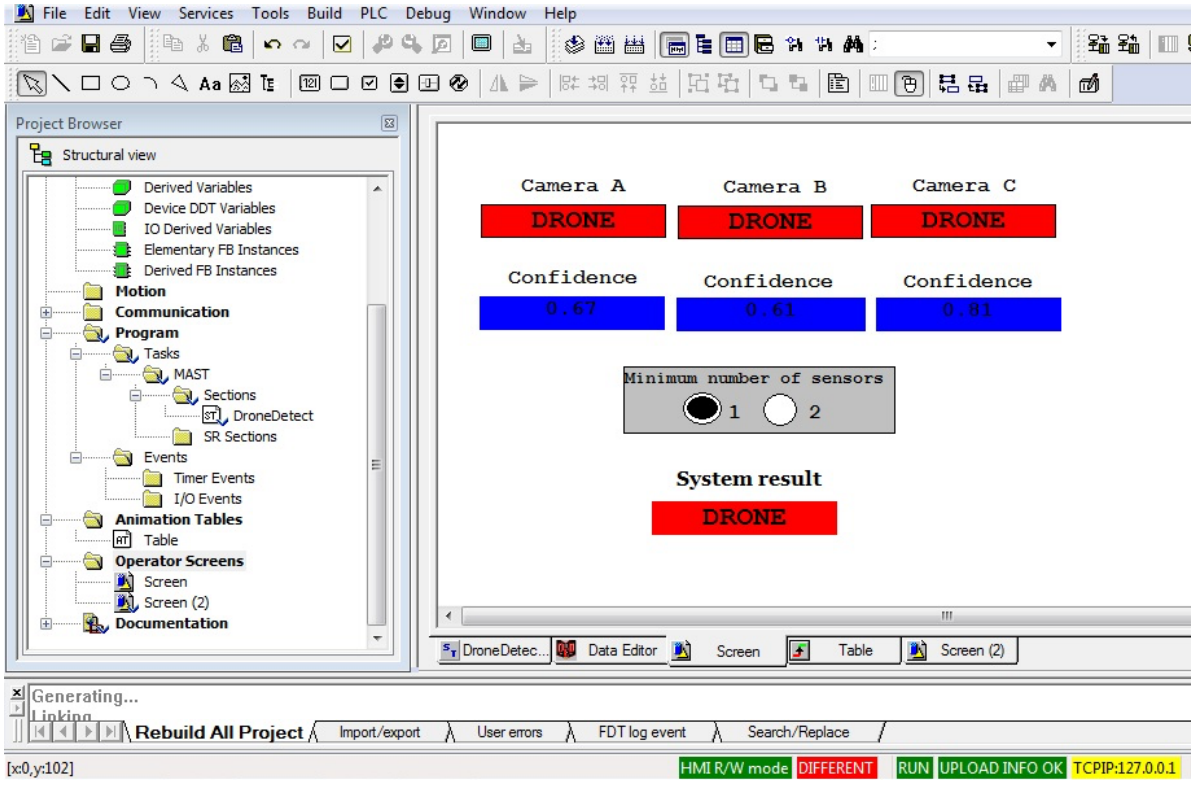


Figure 63 – The case when in drone is flying in the center, and all cameras can capture it

Chapter's conclusion. This chapter presented a proposed a decision-level based sensor fusion system based on voting method. Different configurations of sensor

integration were analyzed and a common decision result was considered by voting the output results of several camera sensors. Multi-sensor detection modes for a three-sensor system were determined. Three camera system was chosen in order to escape the blind spots.

CONCLUSION

This dissertation work has focused on research of effective UAV detection using optical sensors. The five chapters were considered as below:

The first chapter is dedicated on detailed theoretical anaclasis started from security UAV threats to literature review of related works on UAV detection and classification methods based on different drone detection technologies. Particular attention is paid to methods for detecting and classifying UAVs based on visual data, since this work is devoted to the study of effective UAV detection using image sensors.

The second chapter focused on image acquisition, video signal processing methods, Image processing techniques, as well as data preparation steps.

In the third chapter, moving object detection based two-points background subtraction and moving object classification methods with the help of deep CNN neural networks were analyzed. The scientific results of the research direction of recognizing and classifying the visual data of the unmanned aerial vehicle were systematically compared through literature reviews.

The fourth chapter considers a real-time approach for drone detection that is comparable to current algorithms. It was clarified that dividing the drone detection task into the detection and classification steps can solve the given task effectively. The results of the experiment demonstrated both the benefits and drawbacks of the proposed approach. The main limitation of the proposed detector lies on its heavily influenced performance on the moving background existence. In order to solve this limitation and reach higher detection accuracy the next chapter is dedicated to a sensor fusion approach combining multi-angle visual information from several camera sensors.

In the fifth chapter, a decision-level based sensor fusion system was proposed based on voting method. Different configurations of sensor integration were analyzed and a common decision result was considered by voting the output results of several camera sensors.

As a result of the work, all the set tasks have been fulfilled, the relevance, scientific novelty and practical importance of the research work have been fully revealed and proven in the dissertation work. And as an application of the work, a scientific project for "National Security and Defense" was won on the basis of the competition, and currently the research continues as a future work based on the bimodal method combining LiDAR and camera sensors.

REFERENCES

1 Counter drone tactics: Which drones are a real threat, and which aren't? // https://www.ifsecglobal.com/drones/counter-drone-tactics-which-drones. 19.02.2021.

2 Военные перехватили дрон над зданием Минобороны Казахстана // https://ru.sputnik.kz/20190403/nur-sultan-dron-zapusk-minoborony. 03.04.2019.

3 Дрон с казахстанским трамадолом перехватили пограничники Узбекистана // https://www.lada.kz/another_news/73444-dron-s. 16.09.2019.

4 Впервые в казахстанскую колонию пытались передать запрещённые предметы при помощи дрона // https://www.caravan.kz/news/vpervye. 4.09.2020.

5 Владелец дрона в Актобе задержан за полеты над воинской частью // https://elitar.kz/ru/materialy/news/vladelec-drona-v-aktobe-zaderzhan. 14.06.2019.

6 Siddiqi M.A., Iwendi C., Jaroslava K. et al. Analysis on security-related concerns of unmanned aerial vehicle: attacks, limitations, and recommendations // Math Biosci Eng. – 2022. – Vol. 19, Issue 3. – P. 2641-2670.

7 Counter drone tactics: Which drones are a real threat, and which aren't? // https://www.ifsecglobal.com/drones/counter-drone-tactics-which-drones. 19.02.2021.

8 Дроны в руках террористов: что дальше? // https://russiancouncil.ru/analytics-and-comments/interview/drony-v-rukakh-terroristov-chto. 02.08.2021.

9 Radar // https://en.wikipedia.org/wiki/Radar. 17.03.2023.

10 Famili A. et al. Securing your Airspace: Detection of Drones Trespassing Protected Areas // https://arxiv.org/abs/2111.03760. 05.11.2021.

11 Rudys S., Laučys A. et al. Hostile UAV Detection and Neutralization Using a UAV System // Drones. – 2022. – Vol. 6, Issue 9. – P. 250-1-250-18.

12 Park S., Kim H.T. et al. Survey on Anti-Drone Systems: Components, Designs, and Challenges // IEEE Access. – 2021. – Vol. 9. – P. 42635-42659.

13 Khan M.A., Menouar H., Eldeeb A. et al. On the Detection of Unauthorized Drones - Techniques and Future Perspectives: A Review // IEEE Sensors Journal. – 2022. – Vol. 22, Issue 12. – P. 11439-11455.

14 Nassi B., Shabtai A., Masuoka R. et al. SoK - Security and Privacy in the Age of Drones: Threats, Challenges, Solution Mechanisms, and Scientific Gaps // https://arxiv.org/abs/1903.05155. 12.03.2019.

15 Kim J. et al. Real-time uav sound detection and analysis system // IEEE Sensors Applications symposium (SAS). – Glassboro, NJ, 2017. – P. 1-5.

16 Siriphun N. et al. Distinguishing drone types based on acoustic wave by IOT device // Procced. 22nd internat. computer science and engineering conf. (ICSEC). – Chiang Mai, 2018. – P. 1-4.

17 Park S. et al. Combination of radar and audio sensors for identification of rotor-type Unmanned Aerial Vehicles (UAVs) // Procced. IEEE Sensors. – Busan, 2015. – P. 1-4.

18 Anwar M.Z. et al. Machine learning inspired sound-based amateur drone detection for public safety applications // Transactions on Vehicular Technology. – 2019. – Vol. 68, Issue 3. – P. 2526-2534.

19 Liu H.et al. Drone detection based on an audio-assisted camera array // Procced. 3rd internat. conf. on Multimedia Big Data (BigMM). – Laguna Hills, CA, 2017. – P. 402-406.

20 S. Jeon et al. Empirical study of drone sound detection in real-life environment with deep neural networks // Procced. 25th European Signal Processing conf.e (EUSIPCO) – Kos, 2017. – P. 1858-1862.

21 Bernardini A. et al. Drone detection by acoustic signature identification // Electronic Imaging. – 2017. – Vol. 2017, Issue 10. – P. 60-64.

22 Flak P. Drone detection sensor with continuous 2.4 GHz ISM band coverage based on cost-effective SDR platform // IEEE Access. – 2021. – Vol. 9. – P. 114574-114586.

23 Drone Detection: Everything you Need to Know: Can Drones be Detected? // https://www.911security.com/en-us/knowledge-hub/drone-detection. 13.03.2023.

24 Yang S., Qin H., Liang X. et al. An Improved Unauthorized Unmanned Aerial Vehicle Detection Algorithm Using Radiofrequency-Based Statistical Fingerprint Analysis // Sensors. – 2019. – Vol. 19, Issue 2. – P. 274-1-274-22.

25 Taha B., Shoufan A. Machine Learning-Based Drone Detection and Classification: State-of-the-Art in Research // IEEE Access. – 2019. – Vol. 7. – P. 138669-138682.

26 Medaiyese O.O., Ezuma M. er al. Wavelet transform analytics for RF-based UAV detection and identification system using machine learning // Pervasive and Mobile Computing. – 2022. – Vol. 82. – P. 101569-1-101569-32.

27 Andrasi P., Radisic T., Mustra M. et al. Night-time detection of UAVs using thermal infrared camera // Transp. Res. Procedia. – 2017. – Vol. 28. – P. 183-190.

28 Wu H., Li W., Li W. et al. A Real-time Robust Approach for Tracking UAVs in Infrared Videos // Procced. 2020 IEEE/cvf conf. on Computer Vision and Pattern Recognition Workshops (CVPRW). – Seattle, WA, 2020. – P. 4448-4455.

29 Vanek B., Peni T., Bauer P. et al. Vision only sense and avoid: A probabilistic approach // Procced. 2014 American Control conf. – Portland, OR, 2014. – P. 1204-1209.

30 Seidaliyeva U.O., Utebayeva D.Zh., Ilipbayeva L.B. et al. Survey on different drone detection methods in restricted flight areas // Vestnik KazNRTU. – 2019. – Vol. 136. – P. 483-488.

31 Jiang X., Hadid A., Pang Y. et al. Deep Learning in Object Detection and Recognition. – Berlin: Springer, 2019. – 224 p.

32 Unlu E., Zenou E., Rivière N. Using shape descriptors for UAV detection // Electronic Imaging 2017. – Burlingam, 2018. – P. 1-6.

33 Dong Q., Zou Q. Visual UAV detection method with online feature classification // Procced. 2nd Information Technology, Networking, Electronic and Automation Control conf. (ITNEC). – Chengdu, 2017. – P. 429-432.

34 Boddhu S.K., McCartney M. et al. A Collaborative Smartphone Sensing Platform for Detecting & Tracking Hostile Drones // Procced. conf. Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent. – Baltimore, 2013. – P. 874211-1-874211-11.

35 Wang Z., Qi L., Tie Y. et al. Drone detection based on FD-HOG descriptor // Procced. internat. conf. on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). – Zhengzhou, 2018. – P. 433-4333.

36 Baghaei M.R., Tabatabaee P.H., Hashemi M. et al. A new framework for detecting and tracking drones // Indian Journal of Fundamental and Applied Life Sciences. – 2014. – Vol. 14, Issue S4. – P. 612-621.

37 Wagoner A.R., Schrader D.K., Matson E.T. Survey on Detection and Tracking of UAVs Using Computer Vision // Procced. 1st IEEE internat. conf. on Robotic Computing (IRC). – Taichung, 2017. – P. 320-325.

38 Toma A., Cecchinato N., Drioli C. et al. Onboard Audio and Video Processing for Secure Detection, Localization, and Tracking in Counter-UAV Applications // Procedia Computer Science. – 2022. – Vol. 205. – P. 20-27.

39 Schumann A., Sommer L., Klatte J. et al. Deep cross-domain flying object classification for robust UAV detection // Procced. 14th IEEE internat. conf. on Advanced Video and Signal Based Surveillance (AVSS). – Lecce, 2017. – P. 8078558-1-8078558-6.

40 Saqib M. et al. A study on detecting drones using deep convolutional neural networks // Procced. 14th IEEE internat. conf. on Advanced Video and Signal Based Surveillance (AVSS). – Lecce, 2017. – P. 8078541-1-8078541-5.

41 Aker C., Kalkan S. Using deep networks for drone detection // Procced. 14th IEEE internat. conf. on Advanced Video and Signal Based Surveillance (AVSS). – Lecce, 2017. – P. 8078539-1- 80785391-6.

42 Craye C., Ardjoune S. Spatio-temporal Semantic Segmentation for Drone Detection // Procced. 16th IEEE internat. conf. on Advanced Video and Signal Based Surveillance (AVSS). – Taipei, 2019. – P. 8909854-1-8909854-5.

43 Magoulianitis V., Ataloglou D., Dimou A. et al. Does Deep Super-Resolution Enhance UAV Detection? // Procced. 16th IEEE internat. conf. on Advanced Video and Signal Based Surveillance (AVSS). – Taipei, 2019. – P. 8909865-1-8909865-5.

44 Nalamati M., Kapoor A., Saqib M. et al. Drone Detection in Long-range Surveillance Videos // Procced. 16th IEEE internat. conf. on Advanced Video and Signal Based Surveillance (AVSS). – Taipei, 2019. – P. 8909830-1-8909830-5.

45 David de Iglesia, Miguel Mendez, Raquel Dosil, Iago Gonzalez. Drone detection CNN for close- and long-range surveillance in mobile applications // Procced. 16th IEEE internat. conf. on Advanced Video and Signal Based Surveillance (AVSS). – Taipei, 2019. – – P. 8909856-1-8909856-5.

46 Wu M., Xie W., Shi X. et al. Real-time drone detection using deep learning approach // Machine Learning and Intelligent Communications: proceed. internat. conf. – Hangzhou, 2018. – P. 22-32.

47 Park J., Kim D.H. et al. A Comparison of Convolutional Object Detectors for Real-time Drone Tracking Using a PTZ Camera // Procced. 17th internat. conf. on Control, Automation and Systems (ICCAS). – Jeju, 2017. – P. 696-699.

48 Peng J., Zheng Ch., Lv P. et al. Using Images Rendered by PBRT to Train Faster R-CNN for UAV Detection // Procced. Internat. conf. in Central Europe on

Computer Graphics, Visualization and Computer (Vision'2017). – Plzen, Czech Republic, 2018. – P 13-18.

49 Yoshihashi R. et al. Differentiating Objects by Motion: Joint Detection and Tracking of Small Flying Objects // https://arxiv.org/pdf/1709.04666.pdf. 17.09.2018.

50 Rozantsev A., Lepetit V., Fua P. Detecting Flying Objects Using a Single Moving Camera // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017. – Vol. 39, Issue 5. – P. 879-892.

51 Sobue H., Fukushima Y., Kashiyama T. et al. Flying Object Detection and Classification by Monitoring Using Video Images // Proceed. of the 25th international conf. on Advances in Geographic Information Systems (ACM SIGSPATIAL). – NY., 2017. – P. 1-4.

52 Hu Y., Wu X., Zheng G. et al. Object Detection of UAV for Anti-UAV Based on Improved YOLO v3 // Proceed. of the 38th Chinese Control conf. – Guangzhou, 2019. – P. 8386-8390.

53 Thai V.-P., Zhong W., Pham T. et al. Detection, tracking and classification of aircraft and drones in digital towers using machine learning on motion patterns // Procced. Integrated Communications, Navigation and Surveillance conf. (ICNS). – Herndon, 2019. – P. 8735240-1-8735240-8.

54 Lee D.R., La W.G., Kim H. Drone detection and identification system using artificial intelligence // Procced. 9th internat. conf. on Information and Communication Technology Convergence (ICTC). – Jeju, 2018. – P. 1131-1133.

55 Aledhari M., Razzak R., Parizi R.M. et al. Sensor Fusion for Drone Detection // Procced. IEEE 93rd Vehicular Technology conf. (VTC2021-Spring). – Helsinki, 2021. – P. 9448699-1-9448699-7.

56 Dudczyk J. wt al. Multi-Sensory Data Fusion in Terms of UAV Detection in 3D Space // Sensors. – 2022. – Vol. 22, Issue 12. – P. 4323-1-4323-23.

57 Svanström F., Alonso-Fernandez F., Englund C. Drone Detection and Tracking in Real-Time by Fusion of Different Sensing Modalities // Drones. – 2022. – Vol. 6, Issue 11. – P. 317-1-317-38.

58 Jovanoska S., Brötje M., Koch W. Multisensor Data Fusion for UAV Detection and Tracking // Procced. 19th internat. Radar sympos. (IRS). – Bonn, 2018. – P. 8447971-1-8447971-10.

59 Dumitrescu C., Minea M., Ciotirnae P. UAV Detection Employing Sensor Data Fusion and Artificial Intelligence // Information Systems Architecture and Technology: proceed. of 40th Anniversary internat. conf. on Information Systems Architecture and Technology – ISAT 2019. – Cham: Springer, 2019. – P. 129-139.

60 Klein L.A. A Boolean algebra approach to multiple sensor voting fusion // In IEEE Transactions on Aerospace and Electronic Systemsю – 1993. – Vol. 29, Issue 2. – P. 317-327.

61 Image Acquisition (Introduction to Video and Image Processing) Part 1 // http://what-when-how.com/introduction-to-video-and-image-processing. 17.03.2023.

62 Various Color Models used in Digital Image Processing // https://levelup.gitconnected.com/various-color-models-used-in-digital. 12.04.2022.

63 A Complete Guide to Data Augmentation // https://www.datacamp.com/tutorial/complete-guide-data-augmentation 15.11.2022.

64 Comprehensive Guide to Different Pooling Layers in Deep Learning // https://analyticsindiamag.com/comprehensive-guide-to-different. 26.08.2021.

65 Alzubaidi L., Zhang J., Humaidi A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions // J Big Data. – 2021. – Vol. 8, Issue 1. – P. 53-1-53-74.

66 Seidaliyeva U., Akhmetov D., Ilipbayeva L. et al. Real-Time and Accurate Drone Detection in a Video with a Static Background // Sensors. – 2020. – Vol. 20, Issue 14. – P. 3856-1-3856-19.

67 Andrewssobral/Bgslibrary // https://github.com/andrewssobral. 10.06.2019.

68 Dertat A. Review: MobileNetV2-Light Weight Model (Image Classification) // https://towardsdatascience.com/review-mobilenetv2. 19.05.2019.

69 Huyvnphan/PyTorch_CIFAR10 // https://github.com/huyvnphan/PyTorch-CIFAR10. 01.06.2020.

70 Sandler M., Howard A., Zhu M. et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks // In proceed. IEEE/CVF conf. on Computer Vision and Pattern Recognition. – Salt Lake City, 2018. – P. 4510-4520.

71 Wild Birds in a Wind Farm: Image Dataset for Bird Detection // http://bird.nae-lab.org/dataset/ 1.05.2020.

72 Flach P.A. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics // In proceed. of the ICML. – Washington, 2003. – P. 194-201.

73 Classification Evaluation Metrics: Accuracy, Precision, Recall, and F1 Visually Explained // https://txt.cohere.com/classification-eval-metrics. 07.06.2022.

74 Coluccia A., Saqib M., Sharma N. et al. Drone-vs-Bird Detection Challenge at IEEE AVSS2019 // In Proceed. of the 2019 16th IEEE internat. conf. on Advanced Video and Signal Based Surveillance (AVSS). – Taipei, 2019. – P. 1-7.

75 Brena R.F., Aguileta A.A., Trejo L.A. et al. Choosing the Best Sensor Fusion Method: A Machine-Learning Approach // Sensors. – 2020. – Vol. 20, Issue 8. – P. 2350-1-2350-22.

APPENDIX A

Выписка 2 из Протокола № 5 заседания
Национального научного совета по приоритетному направлению
"Национальная безопасность и оборона" от 11-12 августа 2022 года

Дата проведения заседания : 11-12 августа 2022 года

Дата принятия решения : 12 августа 2022 года

Председательствовал : Бердибеков Айдар Токтамысович

Характер вопроса : Рассмотрение заявок по конкурсу на грантовое финансирование молодых ученых по проекту ""Жас Ғалым""на 2022-2024 годы КН МОН РК

№	ИРН объекта	Наименование	Заявитель	Научный руководитель	Балл ГНТЭ	Балл ННС	Дополнительный балл ННС (наличие софинансирования)	Общий средний балл	Запрашиваемая сумма на 2022 год	Запрашиваемая сумма на 2023 год	Запрашиваемая сумма на 2024 год	Общая запрашиваемая сумма на 2022-2024 годы	Одобренная сумма на 2022 год	Одобренная сумма на 2023 год	Одобренная сумма на 2024 год	Общая одобренная сумма 2022-2024 годы	Решение Совета	Обоснование
1	AP14971031	Исследование и внедрение бимодальной системы обнаружения беспилотных летательных аппаратов в режиме реального времени.	Некоммерческое акционерное общество "Казахский Национальный Исследовательский технический университет имени К.И. Сатпаева"	Сейдалиева Ұлжалгас Омиртаевна	27,66	10,93	0	38,59	2995315	7993743	7972269	18961327	2995315	7993743	7972269	18961327	Одобрено	Согласно п.39, п.40 постановления Правительства Республики Казахстан от 16 мая 2011 года №519 «О национальных научных советах», Советом было принято решение одобрить проект
2	AP14972866	Терроризм и экстремизм в религиозной сфере: механизмы реабилитации и деradикализации в Казахстане и зарубежом.	Некоммерческое акционерное общество "Евразийский Национальный университет имени Л.Н. Гумилева"	Темирбаев Тагтат Тулобаевич	24,66	11,14	0	35,8	2480670	7658073	7769061	17907804	2480670	7658073	7769061	17907804	Одобрено	Согласно п.39, п.40 постановления Правительства Республики Казахстан от 16 мая 2011 года №519 «О национальных научных советах», Советом было принято решение одобрить проект
3	AP14971907	Разработка надежной системы обнаружения подозрительных БПЛА на частотной основе с использованием SDR и акустических сигнатур	Некоммерческое акционерное общество "Казахский Национальный Исследовательский технический университет имени К.И. Сатпаева"	Утебаева Дана Жолдыбайкызы	23,66	12	0	35,66	2977847,92	7995738,31	7993325	18966911,23	2977847,92	7995738,31	7993325	18966911,23	Одобрено	Согласно п.39, п.40 постановления Правительства Республики Казахстан от 16 мая 2011 года №519 «О национальных научных советах», Советом было принято решение одобрить проект
4	AP14971555	Проектирование и внедрение системы обеспечения безопасности в режиме реального времени в закрытых помещениях с применением методов машинного обучения	Республиканское государственное предприятие на праве хозяйственного ведения "Институт механики и машиноведения имени академика У.А. Джолдасбекова"	Досбаев Жандос Махсутулы	23,33	11,73	0	35,06	2980890	7995322	7966966	18943178	2980890	7995322	7966966	18943178	Одобрено	Согласно п.39, п.40 постановления Правительства Республики Казахстан от 16 мая 2011 года №519 «О национальных научных советах», Советом было принято решение одобрить проект
5	AP14972687	Независимая антикоррупционная экспертиза нормативных правовых актов как средство обеспечения деятельности правоохранительных органов и защиты национальной безопасности	Учреждение образования "Алматы Менеджмент университет"	Арын Айжан Арыновна	23	11,39	0	34,39	2856630	7342067	7742812	17935509	2856630	7342067	7742812	17935509	Одобрено	Согласно п.39, п.40 постановления Правительства Республики Казахстан от 16 мая 2011 года №519 «О национальных научных советах», Советом было принято решение одобрить проект

Председатель Совета Бердибеков Айдар Токтамысович

Figure A.1 – Minute on the acceptation of a scientific project by the "Zhas Galym 2022-2024"

APPENDIX B

Conducting experimental studies at international research institutions

PURDUE UNIVERSITY | Office of Globalization
POLYTECHNIC INSTITUTE

February 12, 2020

Ulzhalgas O Seidaliyeva
KazNRTU named after K.I.Satpayev
22,a, Satpayev Street
Almaty
Kazakhstan

Re: Offer to Extend Appointment as Visiting Scholar at Purdue University

Dear Ms. Seidaliyeva:

On behalf of Dean Bertoline and the Purdue Polytechnic Institute, it is my sincere pleasure to extend your appointment as a Visiting Scholar in the Computer and Information Technology Department at Purdue University April 30, 2020 through August 31, 2020. This offer is contingent upon the satisfaction of various conditions as described in this letter.

Visiting Scholars are invited to the University to engage in scholarly activities for their own academic enrichment and that of the department in which they have an appointment. Professor Eric Matson will serve as your principal point of contact while you are at Purdue University. Although you will have no formal departmental duties, we hope that you will become an active member of our scholarly community and will participate in University events. It is expected that you will work on CUAV research for image analysis of drones while at Purdue.

Applicable Terms & Conditions affecting Visiting Scholars

Your Visiting Scholar appointment does not carry any salary or benefits. Purdue University will continue to provide you a $1400 per month living allowance for the additional 4 months (May 2020 – August 2020). You will be eligible to purchase a parking permit during the length of your appointment, but prior to leaving the University, we ask that you return your permit to Parking Facilities. The permit is non-transferable. In addition, you will be issued a Purdue identification card, be able to use library facilities, and your name will be listed in the University directory and on appropriate mailing lists.

As a Visiting Scholar at Purdue University, your appointment is subject to all applicable Purdue University policies, as they may be amended from time to time. It is your responsibility to become acquainted with the following policies, which are specifically incorporated into this letter:

1. C-12 "Classes of Purdue University Appointments for Personnel Not on the University Payroll" http://www.purdue.edu/policies/human-resources/c-12.html

2. I.A.1 "Intellectual Property"
   www.purdue.edu/policies/academic-research-affairs/ia1.html.

Please note that policy I.A.1 referenced above requires Visiting Scholars who create intellectual property ("IP") in the course of their appointment with Purdue University to execute a general assignment of such IP in favor of Purdue, subject to certain exceptions, including one for certain scholarly and instructional copyrightable works. By accepting this offer letter, you will be making a prospective assignment of Purdue Intellectual Property (as defined in policy I.A.1) that you create in the course of your appointment by the University.

Conditional Offer

This offer is also contingent upon your obtaining and maintaining appropriate immigration status to permit you to work as a Visiting Scholar.

This letter and the policies referenced above contain the entire agreement concerning your appointment with the University. If these terms are acceptable and if you assent to the assignment of Purdue Intellectual Property, as described above and defined in Policy I.A.1, please sign where indicated below and return a signed copy to Misty Clugh, mclugh@purdue.edu at your earliest convenience.

The faculty and staff join me in welcoming you to Purdue University and look forward to continue working with you. We trust that it will be mutually rewarding.

Sincerely,

Robert F. Cox, Ph.D.
Senior Associate Dean for Globalization
Purdue Global Fellow, Office of Corporate and Global Partnerships
Professor of Construction Management Technology

EMBRY-RIDDLE
Aeronautical University
PRESCOTT, ARIZONA

Ronald A. Madler, Ph.D.
madler@erau.edu
Dean
College of Engineering
T: 928-777-3896

January 18, 2019

Dear Ulzhalgas Seidaliyeva,

It is my pleasure to invite you to the Prescott Arizona Campus of Embry-Riddle Aeronautical University to work with Dr. Akhan Almagambetov of the Computer, Electrical and Software Engineering Department. The university is processing a research/scholar J-visa for you. Prof. Almagambetov will be working with you to engage in cultural activities on and off campus. You will participate in the International Student Festival on February 23, 2019, which will showcase international food, activities, and performances by the international student body at Embry-Riddle. This festival attracts over 600 visitors from campus and the local community. In addition, several cultural trips to Arizona landmarks are planned while the students are on campus.

You will work under Dr. Almagambetov's direction during the period of 9 January 2019 through 31 July 2019 to complete your research topics. Research on automatic control for preservation of laminar flow via computer vision methods. Current flow visualization methods, for the most part, use manual methods of fluid dynamics analysis. This research will attempt to automate some aspects of this process, in effect providing users with a real-time computational fluid dynamics (CFD) capability during testing. Instantaneous feedback will be used to correct any flow anomalies, leading to laminar flow.

Your dissertation topic, "Computer-vision based computational flow control for aerospace applications," closely aligns with the research currently being performed at Embry-Riddle. As previously stated, this research will attempt to automate aspects of flow visualization during wind-tunnel testing, with the goal of providing real-time feedback for the correction of turbulent flow. Embry-Riddle currently has state-of-the-art wind tunnel testing facilities, which will be instrumental in completing this research.

In addition to the generous financial support of your university, Dr. Almagambetov has additional financial resources to cover any shortfalls that may occur in your housing and living expenses. We thank you for your understanding as our university learns how to process the visiting researcher J-visa.

Sincerely,

Ronald A. Madler
Dean, College of Engineering

3700 Willow Creek Road
Prescott, AZ 86301-3720