DRONE CLASSIFICATION WITH MOTION AND APPEARANCE FEATURE USING CONVOLUTIONAL NEURAL NETWORKS

by

Eunsuh Lee

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

Department of Computer and Information Technology

West Lafayette, Indiana

August 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Prof. Eric T. Matson, Chair

Department of Computer and Information Technology

Prof. Anthony H. Smith,

Department of Computer and Information Technology

Prof. John A. Springer,

Department of Computer and Information Technology

Approved by:

Dr. Eric T. Matson

Head of the Graduate Program

ACKNOWLEDGMENTS

Special thanks to my advisor, Prof. Eric T. Matson, for guidance and support in the last two years. I feel grateful to have him as my academic advisor. I would also like to thank my parents for encouragement, the committee members, Prof. Anthony H. Smith and Prof. John A. Springer for constructive comments, my lab mates and friends, help and friendship.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

UAV	Unmanned Aerial Vehicles
FAA	Federal Aviation Administration
EM	expectation maximization
GAN	Generative Adversarial Networks
YOLO	You Only Look Once
GFD	Generic Fourier descriptor
CNN	Convolutional Neural Network
kNN	k Nearest Neighborhood
SNR	Signal-to-noise Ratio
RCS	Radar Cross Section
RF	Radio Frequency
AE	Autoencoder
ROC	Receiver Operating Characteristics
AUC	Area under curve
PRC	Precision-Recall
SVM	Support Vector Machine
VGG	Visual Geometry Group
ResNet	Residual Neural Network
HOG	Histogram of Oriented Gradients
MBH	Motion Boundary Histogram
BoVM	bag of visual words

ABSTRACT

With the advancement in Unmanned Aerial Vehicles (UAV) technology, UAVs have become accessible to the public. However, recent world events have highlighted that the rapid increase of UAVs is bringing with it a threat to public privacy and security. Thus, it is important to think about how to prevent the threats of UAVs to protect our privacy and safety. This study aims to provide an alternative way to substitute an expensive system by using 2D optical sensors that can be easily utilized by people. One of the main challenges for aerial object recognition with computer vision is discriminating other flying objects from the targets, in the far distance. There are limitation to classify the flying object when it appears as a set of small black pixels on the frame. The movement feature can help the system to extract the discriminative feature, so that the classifier can classify the UAV and other objects, such as a bird. Thus, this study proposes a drone detection system using two elements of information, which are appearance information and motion information to overcome the limitation of a vision based system.

CHAPTER 1. INTRODUCTION

This thesis focuses on improving the performance of detecting UAVs using optical sensors. The proposed method is analyzing the motion and appearance captured by the vision sensor. This will help any individual with a limited budget for applying the system. In this chapter, the reason why we need this system and how the system will be conducted will be described.

1.1 Background

Enabled by recent technological advances, the use of Unmanned Aerial Vehicles (UAV) is rapidly increasing in a wide range of applications and various situations, so that futuristic ideas are quickly transformed into reality. For better agricultural production and management, according to (Murugan, Garg, & Singh, 2017), drones can be used to monitor a farm field remotely and also manage the health of crops. The drone also can be used for package delivery (Scott & Scott, 2018) and ad hoc access point network connection (Deruyck, Wyckmans, Martens, & Joseph, 2016). Even though there are many beneficial applications of UAVs, there is an associated risk to public safety. The usage of UAVs raises several social and security problems and concerns that the drones can be used to conduct attacks or invade someone's privacy (Wilson, 2014) (Bonetto, Korshunov, Ramponi, & Ebrahimi, 2015) (Clarke & Bennett Moses, 2014) (Philpott, Kwasa, & Bloebaum, 2018). A couple of problems occurred by drones have been reported. For instance, recently, the world's largest oil processing factory in Saudi Arabia and major oil fields are attacked by a drone (Hubbard, Karasz, & Reed, 2019). In addition to these cases, more and more issues related to drones threats have been reported. Small drones can be dangerous to aircraft, and there have been reports of accidents around airports involving UAVs (Wild, Murray, & Baxter, 2016). Thus, it is critical to restrict small UAVs from entering protected airspace and regulate the usage of civilian UAVs to prevent the potential threat.

Even though drone threats are an emerging issue, it is hard to regulate the operation of UAVs because, as the cost of commercial small UAVs is decreasing, the number of hobbyists using them is increasing rapidly. The drones are developed and modified in various shapes. It is hard to detect whether the drones carry improvised explosives or drugs. Moreover, with various appearance because of its cheap cost, it is becoming difficult to identify drones that could occur potential incidents and make those threatening drones stop the operation when necessary. To control the usage of drone, the Federal Aviation Administration (FAA) has made the law of all drones mandatory starting from 2015 (Morris & Thurston, 2015). However, the regulation only can be applied to the number of registered drones that are monitored by the FAA.

There are several detection technologies using various sensors based on RADARs(Rahman & Robertson, 2018) (Shin, Jung, Kim, Ham, & Park, 2017), vision(Prates et al., 2018), or sound (Case, Zelnio, & Rigling, 2008) to find the UAVs in the sky. However, the drones cannot be detected efficiently with conventional methods, such as RADARs. RADARS are widely deployed for observing and identifying aerial objects like aircraft because it is hard to sense the small size object or small electromagnetic signatures (*A Drone, Too Small for Radar to Detect, Rattles the White House - The New York Times*, n.d.). The sound-based detection is only suitable for scenarios that the drone is in the small boundary from the sensor because of ambient noise(Bisio, Garibotto, Lavagetto, Sciarrone, & Zappatore, 2018). The alternative technology, the RF fingerprint classification of drones is widely used on time-domain techniques and it is not effective for real-time systems (Ezuma, Erden, Anjinappa, Ozdemir, & Guvenc, 2019). Also, those solutions with multiple sensors network systems are costly so that it is hard for the individual to use. The method with the optical sensor is also one of the traditional methods for object detection, but there are some limitations that the performance highly depends on the resolution of the camera (Gökçe, Üçoluk, Şahin, & Kalkan, 2015). However, the optical sensor is the most approachable sensor to the public and state-of-the-art technology helps the system to overcome the constraints related to computational power.

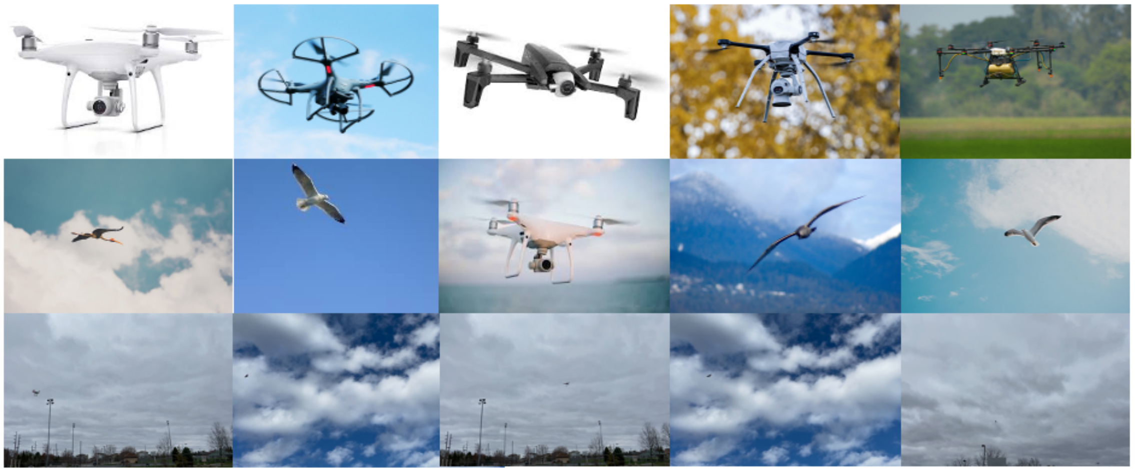


Figure 1.1. Examples of drone and bird have different size and features with the real captured image.

This study proposed to use a feature of movement and appearance together to solve the limitation of extracting the features from the extremely small object. Humans use not only appearance information but also movement pattern to recognize objects. Similar to the human vision capability, the convolutional neural network model is also available to do the classification task based on extracting the flight motion pattern and appearance information. The movement of a drone has differences compared to the movement of bird. The movement of birds is randomly attributed. However, the movement of drone is more clear, as it can also move up and down. These differences can be captured by convolutional neural network models. The state-of-the-art deep learning models studied on the motion recognition show that using both information increased the accuracy of recognition. This study therefore tries to show the differences of using convolutional neural network model applied to the drone and bird domain with temporal and spatial features.

First, in order to apply the models, it is important to gather the large enough trainable dataset. There are no publicly available datasets for the evaluation. Therefore, I have collected a couple of datasets that reflect the problem of detecting drones in outdoor environments. To extract the optical flow clearly, the video dataset filed by stilled camera is used. The flying objects are captured as small part of the frame. The average portion of object in the frame occupies less than 50 pixels and its color or shape is hard to be detected. The datasets are manually labeled as drone and nondrone data.

Second, this study presents a method for classifying flying objects such as drones and birds that possibly occupy in front of fixed sky background, and are filmed by a fixed camera. In order to classify the flying objects, it requires combining motion and appearance information, as neither of the two alone is not enough to determine its class. Therefore a deep learning technique and fusion methods which operate on the spatial field and temporal field are proposed in this study. These approaches, which are applied on different domains in the previous studies, are used in this research.

With this gathered dataset and approach, this study provides the drone classification system by vision sensor and shows how the result is different between using individual features and using both features.

1.2 Statement of Problem

The approach of using high-accuracy detection devices require excessive costs and difficulty of acquiring such equipment to the public. The relatively cheap and familiar optical sensor, the 2D camera has limitations regards to the resolution. The target object in the long-distance can be shown as a dark spot in the frame, so it could be hard to return the valuable appearance features to classify whether it is UAV or not.

1.3 Scope

The scope of this research is providing the probability of development of UAV detection based on the vision system using the appearance information and movement information together. The autonomous system is not a fully integrated, end-to-end solution, but rather alarms the people that there could be an unauthorized UAV breaking into the private area.

1.4 Significance

UAVs are easily accessible in the market, and some people build and rebuild their own drones in various shapes. Due to the maneuverability and small size of the UAVs, the traditional object detection method cannot meet the requirements of detection accuracy. The approach of classification with analyzing the motion and shape feature together to give the possibility to enhance the accuracy of solution. With applying this approach to the specific domain containing the UAVs, the affordable solution with cheap cost and high performance is expected.

1.5 Research Question

How is the accuracy different between using appearance information, motion information and using appearance information and motion information together?

1.6 Assumptions

- The flying objects in front of the prototype system are assumed to be controlled by an operator with the intent of committing a crime. The objects would not be overlapped while they are flying. There is only one UAV in flight at one time. The UAV does not have any payloads. The aerial objects have constant weight, height, and other physical conditions while they are flying. The detection and classification system are delivered in a stable environment with enough power.

1.7 Limitations

- to the limitation of the vision system, this study will be based on the vision system, so the experiment would require the environment which can show the presence of an object at least.

- The hardware performance of the machine used for training the machine learning model can limit the accuracy of the model.

1.8 Delimitations

- There are various types of commercial UAVs, but it is difficult to gather all UAVs. Thus, only one type of UAV is used in the experiment. The one UAV is flying in the dataset. Thus, only one or two birds are captured in the dataset, to make simliar to drone dataset. If there is differences in the number, the number of object can be one of feature.
- To extract the clear features, the video is filmed by the fixed camera. This study aims showing the possibility of imporving the performance of vison system. Thus, he dataset for the study is very small. It is hard to extract the features from the moving camera.
- To minimize the confusing factor for the prototype system and focus on detecting drones, the experiment will occur on the background with only drone or bird.

1.9 Summary

This paper begins by introducing why the drone detection and classification system is needed. The vision detection system is needed because of the accessibility and price of sensor. In the vision system, using motion features and spatial feature simultaneously gives the possibility to improve the accuracy of the model.

In Chapter 2, the previous studies are discussed. There are a lot of methods with different sensors. First, many methodologies with using vision sensor, acoustic sensor, RF sensor and Radar sensors are reviewed. Second, the deep learning approaches using spatio-temporal information are discussed.

In Chapter 3, the methodology is described. The architecture of models and methods of how the video dataset are described.

In Chpater 4, the result and anaylsis of expreiment are discussed.

Finally, the study closes this thesis in Chapter 5 with concluding limitations and the future work discussion.

CHAPTER 2. REVIEW OF LITERATURE

Several methods have been studied to detect drones. Sensors such as RADAR, vision, and sound are used. Furthermore, methods of using multiple sensors together has been studied. This section will explain which methods were previously used to detect drones. Second, we will review how the previous studies used appearance features and motion features when using the vision method. This chapter will give a summary of the research that has been proposed to classify the UAVs with the various sensors and how the method using the motion and appearance information has been conducted and applied to the different domains.

2.1 RADAR detection

RADAR is an object detection method that transmits the electromagnetic signals which interact with the target and verifies whether the object is a drone or not by analyzing the reflection from the object. This interaction causes a shift in the carrier frequency due to the Doppler effect. In addition to the main Doppler shift (Chen, Li, Ho, & Wechsler, 2006), if the object has vibrating or rotating structures such as propellers or engines of drones, the micro-motion will produce various frequency modulation on the signal. The dynamic features of the object can be gathered by Doppler processing of the radar, and hence it enables the detection of the small moving objects with a low RADAR Cross Section.

According to Samiur Rahman and Duncan A. Roberton (Rahman & Robertson, 2018), the micro-Doppler signature of a drone is compared with that of birds. Birds and UAVs have different flying motions according to features illustrated by the RADAR. UAVs mostly fly by using a propeller blade rotation while birds flap their wings. These flapping micro-Doppler frequencies at a much lower frequency-band are compared to the micro-Doppler frequencies induced by the rotating propellers of the micro-UAVs. The study shows that the unique micro-Doppler K-band and W-band RADAR signatures of the micro-UAVs and birds give a classification of drones and birds.

In (De Quevedo, Urzaiz, Menoyo, & López, 2019), a small and low powered RADAR system based on the ubiquitous concept is introduced. The RADAR system detects a micro-UAV by Doppler processing and digital beamforming. In (Ezuma, Ozdemir, Anjinappa, Gulzar, & Guvenc, 2019), multiple transmit/receive antenna configurations for UAVs detection is proposed. The study concludes that the distinctive micro-Doppler signature of the UAVs can be used for the automatic recognition system to distinguish the aerial objects.

Although RADAR-based detection using the mini Doppler effect has been a the mainstream method, they generally fail to classify another flying object. Also, it has a limitation when the objects account for small radar cross-section. There are some modified UAVs to avoid a RADAR detection. For example, the stealth UAVs which are designed to avoid RADAR signal, have a low RADAR cross-section (RCS). It has a too low RCS to be detected by conventional RADARs. The United States White House's surveillance radar could not catch the entering micro-UAV flying across the fence (*A Drone, Too Small for Radar to Detect, Rattles the White House - The New York Times*, n.d.) because of low RCS value. Moreover, the RADAR detection method is still questionable for its cost-effectiveness and feasibility because of a vast amount of data output and sensitivity to the background, such as clouds.

2.2 Acoustic detection

The acoustic-based method could be considered as an effective detection system because the sound sensors are easily deployable at a reasonable price. Audio signal processing is considered to be more provident compared to image processing because of its small computing cost. The acoustic signatures, which are the unique acoustic data of UAVs extracted by arrays of microphones, are compared with the recorded signals that are collected in the market to find a match. This unique acoustic signature is produced because of the motor and propeller, as well as the vibration of the structure. There are previous studies to differentiate the sound from a drone from the background noise.

Ellen E.Case, Anne M. Zelnio, and Brian D. Riglingtime proposed the economical and mobile acoustic array for tracking and localization of UAV (Case et al., 2008). The signal processing software consisted of the calibration procedure and the beamforming algorithm. The calibration is used to determine the 2D coordinate location of the microphones. The target can be detected and tracked by the beamforming algorithm.

Acoustic-based detection has the advantage that it has a reasonable price. However, noisy ambient environments with a complex jumble of music, traffic, and human sounds are challenges for using the sound dataset for drone detection. Moreover, the new drone models that are not included in the dataset are hard to analyze. The fundamental challenge of the audio-based systems is the practical range of commercial microphones. The commercial microphones mostly operate well and give good results in a range of 25-30 ft (Bisio et al., 2018).

2.3 Radio Frequency fingerprinting

There are UAV detection methods using the Radio Frequency (RF) fingerprint. RF fingerprint-based detection is analyzing features of the RF signals emitted from UAV controllers (Ezuma, Erden, et al., 2019).

In (Zhao, Chen, He, & Wu, 2018), the authors use the Generative Adversarial Networks (GAN) to generate a large dataset and the classification model of GAN to distinguish the types of signal. The accuracy of recognition in the indoor environment is about 95% and this method could be applied to the outdoor environment. Through the combination of Auxiliary Classification GAN and Wasserstein GAN, it provided the possibility to classify the signal of UAV in real-time. However, the computation cost can be increased because of the expectation-maximization (EM) algorithm, which gives the estimated threshold. Moreover, a large amount of Gaussian calculation needs high computational cost.

According to Ezuma et al. (Ezuma, Erden, et al., 2019), a developed system converts the raw RF signal into frames in the wavelet domain to reduce the size of the data and remove any bias and noise in the signal. After preprocessing the data, the k Nearest Neighborhood (kNN) classification is applied to the dataset. The result shows that the average accuracy of detection is 96.3. In this paper, the authors also mentioned about different signal-to-noise ratio (SNR) levels. When SNR is higher than 12 dB, the accuracy is 100% while the performance recoded not good when SNR is below 10dB.

The RF signal is an important and unique feature of drones so that it is useful to detect the drones. However, if the drone is autonomously operated, it is hard to use RF Fingerprint-based system because there is no signal from the controller. Moreover, there is a lack of a large enough comprehensive dataset to analyze RF signals. Furthermore, the existing methods have limitations for low signal and noise so that most previous experiments are conducted in the indoor environment.

2.4 Vision detection

Vision-based detection methods can be applied with one or more optical sensors to detect a drone. The frame of the video is used for analysis to detect a drone. One method presented in the previous research using deep learning (Aker & Kalkan, 2017) uses a single-shot object detection model, You Only Look Once (YOLO) v2, which is the second version of YOLO (Ralphs, 2018). This model is applied to the artificial dataset the authors created. This convolutional neural network identifies multiple objects and classifying them over different frames. The YOLO deep learning algorithm learns the optimal features from the captured object frames by using a linear regression-based. The YOLO v2 is very fast but removing the dropout. The authors applied this detection system to the artificial dataset, which is the combination of the images of UAVs and bird images with different background videos. The vast variance and the scale of the dataset return the high performance of drone detection with real-time processing. However, training CNN networks requires a massive amount of data and expensive computational cost.

According to Eren Unulu, Emmanuel Zenou, and Nicolas Riviere, a computer-vision approach based on a Generic Fourier descriptor (GFD) can be used for drone detection and classification between drone and bird (Unlu, Zenou, & Riviere, 2018b) (Unlu, Zenou, & Riviere, 2018a). The author proposed using speeded-up robust features (SURF) to detect key feqatures of micro-UAVs. The Region Growing algorithm is introduced to classify the pixels of drones or birds.In comparison to CNN, the speed of this method is much faster detection for micro-UAVs.

The vision application is widely used for general drone monitoring systems. However, most of vision-based techniques have a common significant problem. The performance of the optical sensors depends on the ambient condition, such as lighting or background. Besides, the detection of UAVs may not perform well when it is applied to the large surveyed area. However, the recent approach with the optic sensor not only using static features but also motion features solves the problem the vision has. Thus, even with the low resolution, the model could classify the object with the combined features. 145

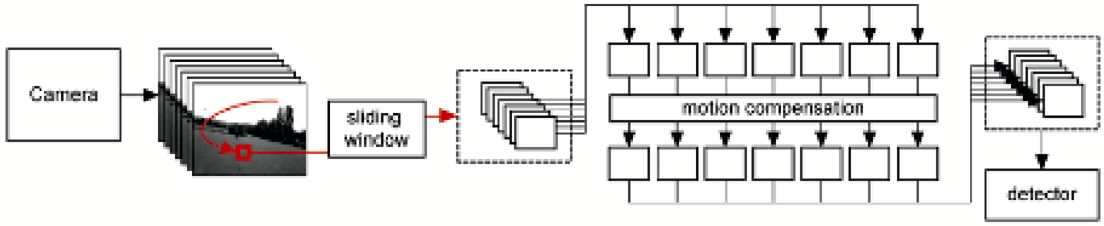


Figure 2.1. Detecting flying objects using appearance-based and motion-based methods simultaneously in *Vision-based detection of aircrafts and UAVs Artem ROZANTSEV* (n.d.)

There is a research named *Vision-based detection of aircrafts and UAVs Artem ROZANTSEV* (n.d.) that uses motion and appearance information on the drone and aircraft domain. Rozantsev presents a combination of the two methods to detect aircraft and UAVs using a camera on board a UAV to solve the problem of collision between UAVs and aircrafts. The motion features are extracted to detect flying objects by motion compensation. The still frame information is processed to decide if the object is a target object using convolutional neural networks. The authors generated the synthetic dataset to train on large dataset. This approach has similar concept with this study. Both studies are using motion and appearance information to classify the flying object. However, the applied domain and architecture are different. In this study, the domain contains birds and UAVs, not aircraft. Also, the approach studied by Rozantsez is divided into two steps. It detects the flying object first by using motion information. However, in this study, the movement feature will be used for classification and both features which are movement information and appearance feature are used parallel. The sequence of operation is not separated.

2.5 Two-stream Models

Several sensors are used to detect and classify drones and other flying objects. Among them, the camera sensor has advantages in terms of price and accessibility. The method using the vision sensor focuses on the extroverted part of the object, that is, the appearance feature. It is essential to gain higher detection accuracy with reduced risk of threats from a drone. In order to further increase the accuracy score, this study tries to compare the action recognition methods that use appearance features and motion features simultaneously. Determining whether a drone is in a video scene can also be said to be similar to the action recognition domain because it involves the process of analyzing movement and identifying the drone’s outward and spatial information. This section describes how various action recognition methods have been studied.

2.5.1 Spatio-Temporal stream

Video recognition research has come a long way with the advances in image recognition methods. These methods are often adapted to sequential data and video data. Many researchs use spatial feature and motion features simultaneously.

In the previous study Laptev, Marszałek, Schmid, and Rozenfeld (n.d.), many video action recognition approaches are based on shallow high-dimensional encodings of local spatio-temporal features, which are the Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF). The datasets are films of humans from some parts of the movie. Local spatio-temporal features provide a short video representation. These features are encoded to Bag of Features and pooled over several spatial-temporal grids. Then, features are combined with multi-channel non-linear Supporting Vector Machines (SVMs). There can be problems with getting useful information because of occlusions, scale changes, and complex backgrounds. The occlusions are the problem that there can be some parts are hidden. In a later work, it was shown that using dense sampling of local features performs better than using sparse interest points.

Instead of using computing local video features, there are other methods to represent the shallow video. In Wang, Kläser, Schmid, and Liu (2011), they first introduced the methods that adjust local descriptors to support the local regions. Feature trajectories are sampled in dense points from each frame and track the points based on the information that is achieved from the dense optical flow field. The best result from the trajectory-based method was achieved by the Motion Boundary Histogram (MBH), which is a gradient-based feature. It is separately computed on the horizontal and vertical components of optical flow. It is similar to the method that this study applied to the domains in the next section. A combination of several features was shown to increase accuracy.

Temporal data means that the data that represents a state in time, such as a motion pattern. Many research works have been devoted to modeling the temporal structure for action recognition. Gaidon et al. Gaidon, Harchaoui, and Schmid (2011) annotated a sequence of each small movement for each video in the large dataset and proposed Actom Sequence Model for action detection. Temporal visual features are represented in this paper. The actoms, which are sequence of atomic action unit, are detected automatically by a non-parametric model. The proposed method in this paper focuses on the small sequence of movements. However, when we try to find the pattern of flight, we should analyze sufficiently long videos to see the trajectory of the flying object. The atomic movement is not proper to achieve the flying pattern of an object. Savarese, DelPozo, Niebles, and Fei-Fei (2008) proposed a model for human action recognition by using a fixed single temporal ordering of fundamental poses. Fernando modeled the temporal evolution of bag of visual words (BoVW) representation. However, these methods have some limitations that they are hard to assemble end-to-end learning. It is crucial to provide a full package of solutions for users to use the system more comfortable.

In the paper Zhu, Lan, Newsam, and Hauptmann (2017), instead of obtaining the optical flow, the authors proposed a motion vector to recognize the action with spatial data together. Motion vectors represent movement patterns of image blocks that resemble optical flows in terms of describing local motions. Motion vectors have the advantage of not having to perform additional calculations. However, the lack of fine and accurate motion information is not suitable for flying objects domain, because the flying object can be too far away from the sensor. In addition, the MotionNet has difficulties in regions that are highly saturated or have dynamic textures, for example, water, sky, and mirrors. This difficulty is because the constant brightness assumption does not hold. In this paper, the goal is to distinguish flying objects. Therefore, since the background is mostly sky, it is not suitable for using MotionNet in this study.

CHAPTER 3. RESEARCH METHODOLOGY

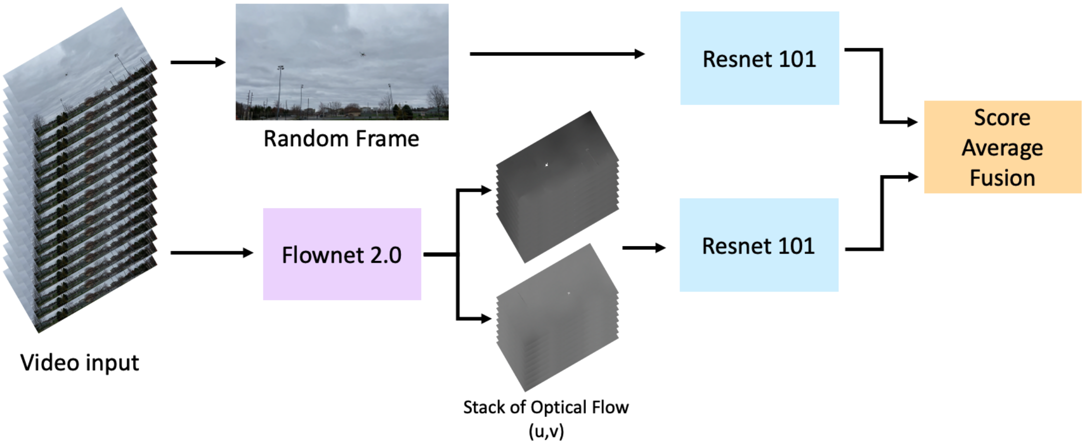


Figure 3.1. Architecture of the detection system.

The method for classification is based on the paper name on ”Two-Stream Convolutional Networks for Action Recognition in Videos” by (Simonyan & Zisserman, 2014a) published in the conference on Neural Information Processing System 2014. The video can be basically divided into two elements, which are spatial and temporal factors. This paper proposes the approach of applying spatial stream Convolutional Networks and temporal stream Convolutional networks together with some fusion methods. Traditionally, fusion methods are divided into early fusion and late fusion. The early fusion builds descriptors by fusion descriptors of different models. On the other hand, late fusion makes the decisions obtained by each classifier, separately. The method of this study is based on these approaches.

3.1 Spatial Training

Spatial training is applying Convolutional Network on stilled images. This stream is finding the patterns from the image set and classify them into several classes. The patterns are extracted from the image and map to the latent vector space. The ResNet (He, Zhang, Ren, & Sun, 2015) is used to find the features of classes, and the SVM is used for the classification. For making the diverse on the dataset, the input frame is randomly flipped.

3.2 Optical Flow

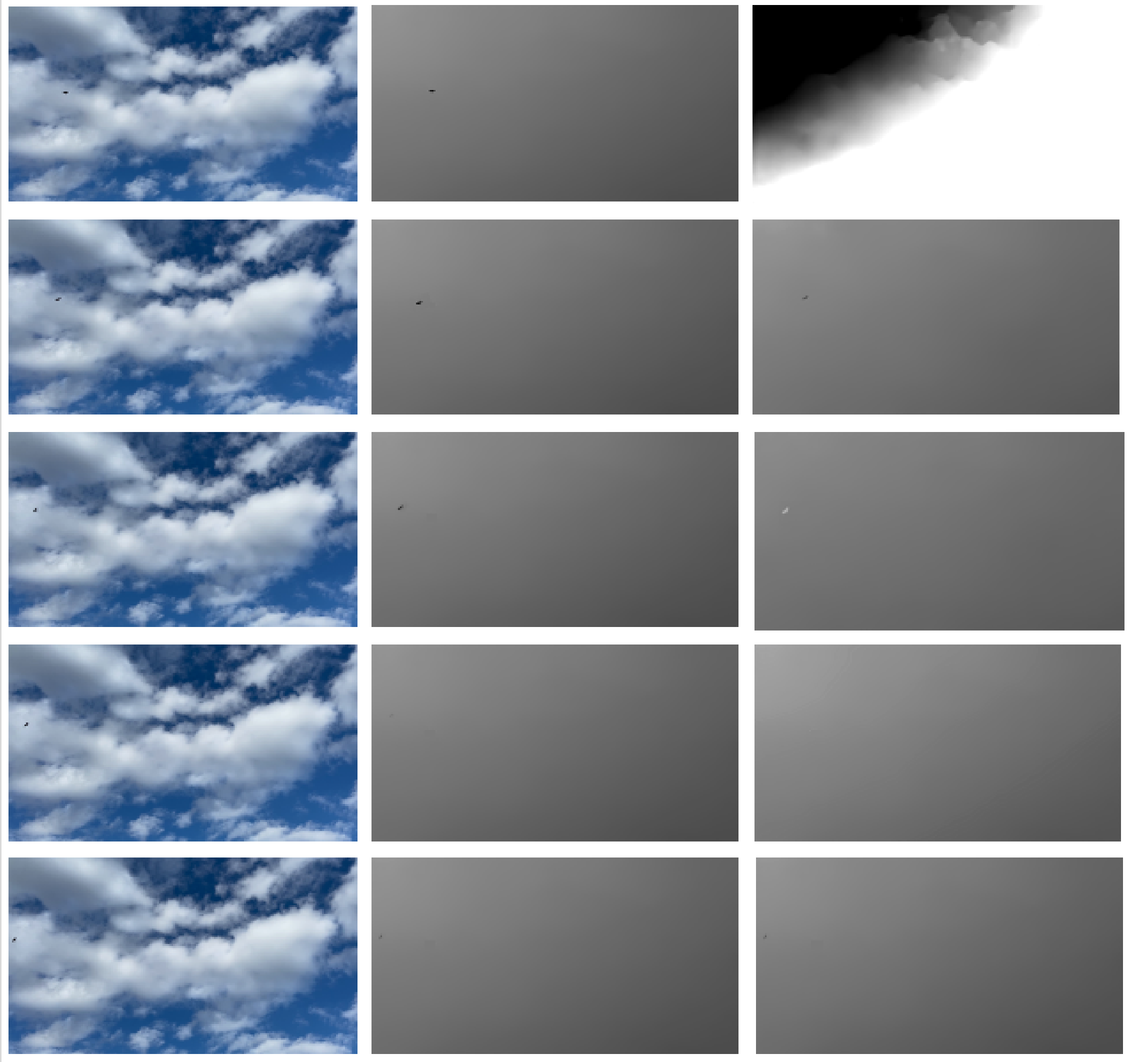


Figure 3.2. The divided optical flow into horizontal and vertical components from the frame. These are sequence of five frames from bird dataset.

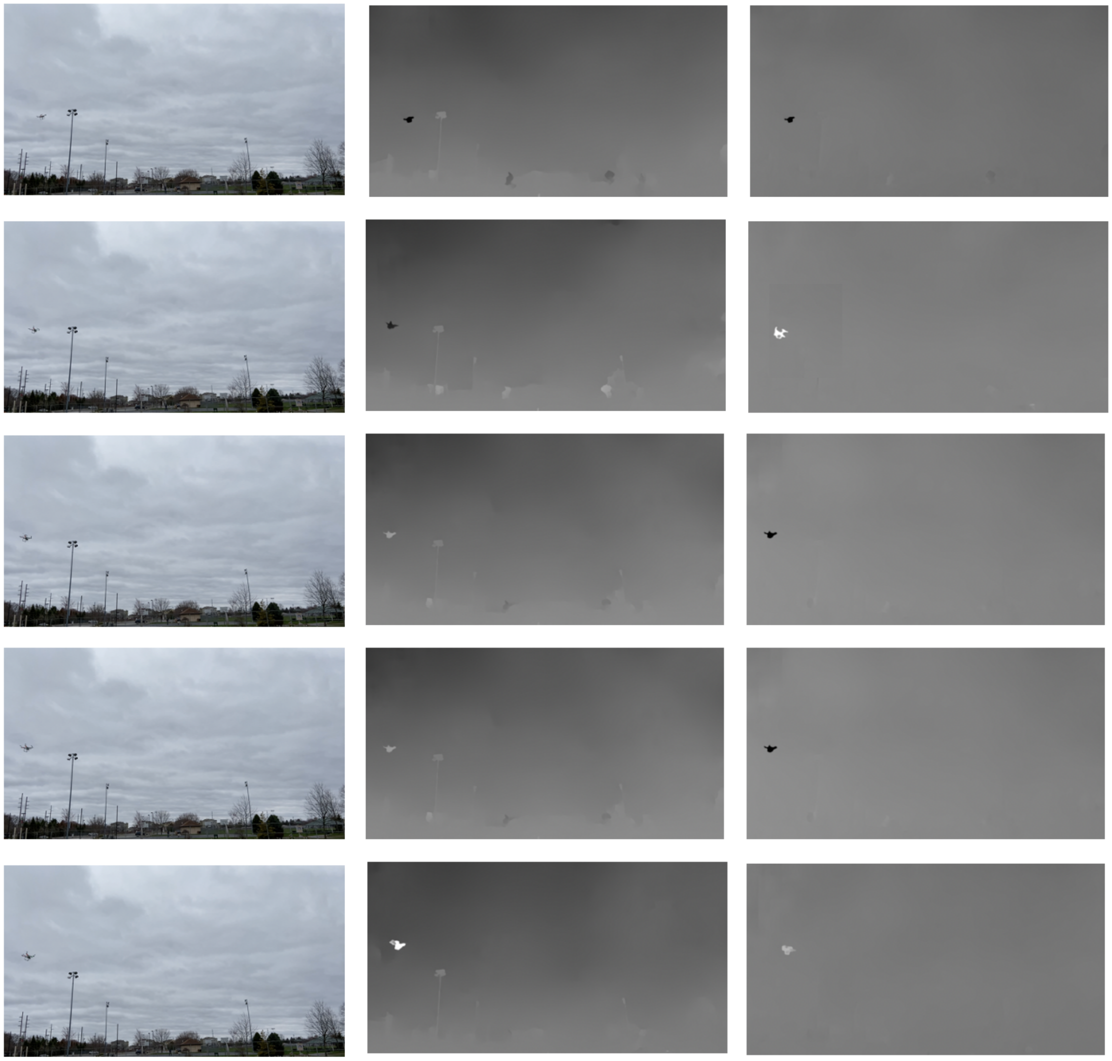


Figure 3.3. The divided optical flow into horizontal and vertical components from the frame. These are sequence of five frames from drone dataset.

Opitcal Flow is mainly used for observing the motion of the object. There is a traditional method to compute the optical flow from the OpenCV toolbox. However, it is hard to find the proper threshold number by using handcrafted approaches, and there is an aperture problem. The aperture problem is that the optical flow is not directly measurable. There are a few applications of optical flow in Deep Learning. One of the state-of-the-art deep learning methods to calculate an optical flow is FlowNet 2.0. Flownet2.0 is significantly faster than traditional methods and generates very accurate results. However, computing optical flow by using FlowNet would introduce a bottleneck, so the optical flow is pre-computed before training. The tide is stored in the horizontal and vertical components and compressed using JPG format.

3.3 Motion Training

In this section, the temporal stream architecture is described. The input of this model is formed by stacking optical flow. A dense optical flow is a displacement vector field between consecutive frames.

There is a traditional method to compute the optical flow from the OpenCV toolbox. However, it is hard to find the proper parameters by handcrafted methods, and there are some problems such as aperture problem. Thus, for the efficiency and accuracy, it is better to use Deep Learning for calculating the optifal flow. There are a few applications of optical flow in Deep Learning. Flownet 2.0(Ilg et al., 2016) is one of the state-of-the-art Convolutional Neural Networks (CNN) for optical flow estimation. Flownet2.0 is much faster than the traditional methods and it do not need to do hand-craft stage to find the parameters for training. However, Flownet2.0 has a big model size and needs high computational power. Computing optical flow can be a bottleneck in time aspect, so the optical flow should be pre-computed before training. The horizontal and vertical optical flows can be calculated by using FlowNet.

The horizontal and vertical components of the optical flow vector field are seen as two channels of an images. For representing the movement of drone across the sequence of frames, the optical flow data is stacked from N consecutive frames and generates 2N input channels. The vectors are concatenated. Let w and h be the width and height of a video; a Convolutional Networks input shape is [1, w, h, 2N]. For making the diverse on the dataset, the input frame is horizontally flipped. The horizontal and vertical optical flows are extracted and stacked, then the stacked features are used for the classificationl.

3.4 Transfer Learning

The ImageNet (Krizhevsky, Sutskever, & Hinton, 2017) is a project which tries to provide an extensive image database for research purposes. It contains more than 14 million images and 20,000 classes. The models in this study use the pre-trained the models on the ImageNet dataset. There are some benefits to use a pre-trained model. The pre-trained model can provide a useful starting point because deep networks have a large number of unknown parameters. Transfer learning is the idea that model can learn new tasks based on previously learned tasks. For example, the knowledge gained while learning to recognize cars could apply when trying to recognize trucks. The learning process can be faster, more accurate, and needs less training data. Transfer learning is useful when there are insufficient data for a new domain applied by a neural network, and there is a big pre-existing data pool that can be transferred to the problem. Thus, getting a piece of knowledge from the pre-trained model is an essential part of the training.

However, the channel that ImageNet, used for the research, has some differences compared to the models and input data used in this study. Thus, it needs some modification when pre-trained weights are applied to different domains. Pre-training has turned out to be an effective method to initialize deep ConvNets when the domain dataset is not big enough to be trained. When we train the RGB images as input, many of the previous approaches used ImageNet as initialization for training spatial networks. In this study, other modalities, such as optical flow field are used. The optical flow field has different visual aspects of video data. The distributions of it are different for that of RGB images. Thus, a cross-modality pre-training technique is used in this study to utilize RGB models. First, optical flow fields are divided into the interval from 0 to 255 by a linear transformation, as same as the RGB range. Then, the weights of the first convolution layer of the RGB model are modified to apply to the optical flow domain. Here, the average of weights across RGB channels is replicated. It makes the number of copies the same as the input channel number of optical flow fields. This initialization methods show a good performance for temporal networks and reduce the over-fitting problem.

3.5 Early feature level fusion

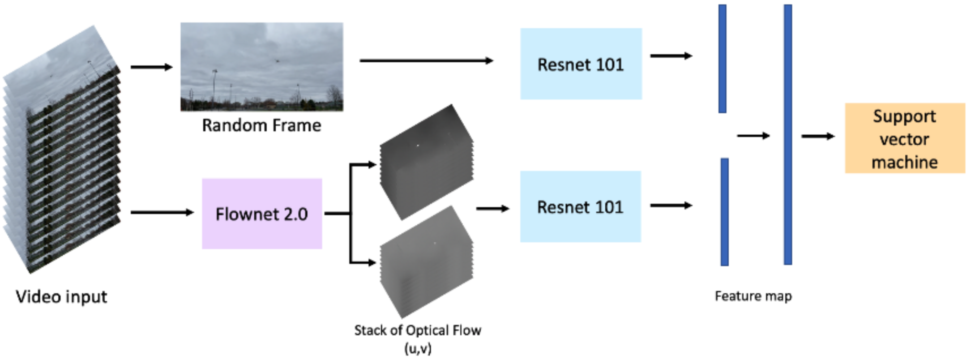


Figure 3.4. Early fusion architecture for the detection system. The features are extracted from the appearance stream and motion stream individually.

Early fusion is applied to pre-processed dataset. Data features are extracted from the each model and combined to the synchronized features. It is challenging when dastaset has different characteristics such as continuous features and descrete features. It is hard to transform each extracted vector to the one single vector. Once the dataset are combined together, they are classified by using a decision metrics such as SVM.

The training pipeline of the early fusion methods is processed as one model. Each model extracts the features. Then, a single model is trained to learn the combined one feature.

$$p = h([v_1, v_2])$$

Two features containing different information and different dimension are merged into one feature by concatenation. The size of the merged data is twice the size of each feature from the spatial stream and the motion stream. Then merged feature is classified by a SVM. The SVM is mostly used for two-class discrimination problems because SVM performs for classifying binary class and high dimension dataset by using kernel trick. The detection model is binary classifiers with two classes representing the UAV flying and the birds flying. The SVM also has good performance with a small set of training data.

3.6 Late feature level fusion

Late fusion uses the same dataset independently, and the extracted features are predicted in each prediction level, such as the softmax layer. Then the final decision is decided using the prediction scores from the result of each model. There are several calculation methods for using the late fusion. One of the late fusion methods is average score fusion. Average score fusion is averaging the score from the softmax layer. In this study, the spatial network and temporal network are trained individually on the same sequential dataset.

$$p = F(h_1(v_1), h_2(v_1, v_2))$$

The model makes the prediction by selecting the highest score, which is obtained by averaging the scores from each softmax layer of the model.

3.7 Summary

The method from ”Two-Stream Convolutional Networks for Action Recognition in Videos” (Simonyan & Zisserman, 2014a) published in the conference on Neural Information Processing System 2014 is used as a method. The three randomly selected frames, and ten sequential optical flows are used as a dataset. The two divided components of optical flows are used as input for the temporal stream, and the three still images are used as input for the spatial stream. The two streams are trained individually.

For the fusion, the early fusion and the late fusion are used. Early fusion has combined the features that are extracted from two separate streams, while the late fusion used the output score of two separate streams.

These methods are applied to the dataset containing a drone and a bird. In the next section, it explains how the dataset is gathered to be applied and the result of experiments using different frameworks. Then it shows how the spatial stream, temporal stream, and Spatio-temporal stream are different.

CHAPTER 4. EXPERIMENTS

This section explains how the dataset is gathered. For the experiments, the different ResNet backbones are used. The results of using different backbones are compared. Those backbones are used in each spatial stream, temporal stream, and spatio-temporal stream. The experiments show the difference accuracy from the spatial stream, temporal stream, and spatio-temporal streams with different fusion methods.

4.1 Hardware setup

An Intel i5-6600 is used as CPU for the experiment. For the GPU, NVIDIA GeForce GTX 1080 and NVIDIA TITAN Xp is used for training and testing. Linux OS is usded.

<u>4.2 Dataset</u>

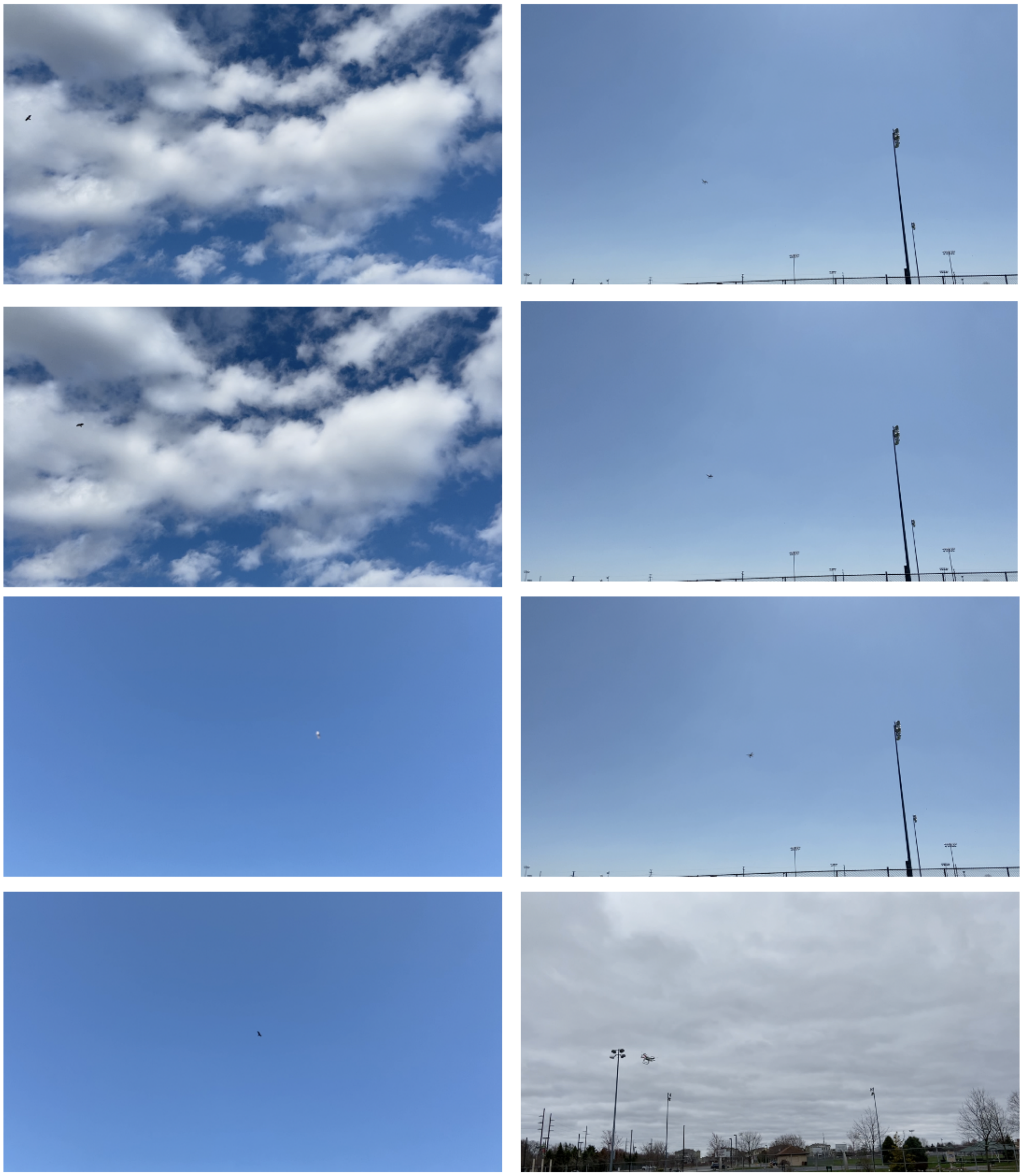


Figure 4.1. Examples of drone dastaset and bird dataset in the landscape

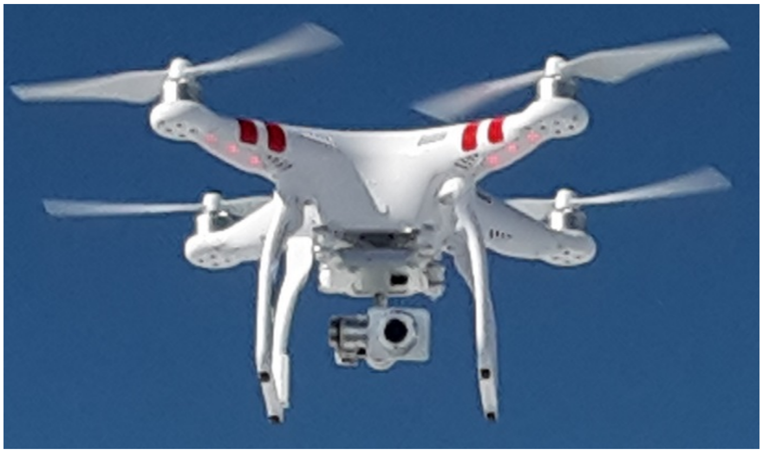


Figure 4.2. DJI Phantom 2 drone

A dataset contains two classes, which are the video of the flying drone and the video of the other flying object. It was difficult to hard to gather open data with a consistent. Besides, the background should not move as much as possible to obtain more accurate optical flow information of the object. This is because the optical flow is obtained as information about pixels that differ between frames. If it was taken with a moving camera, it is difficult to extract the movement of the drone and bird. Therefore, the data for the experiment was taken with a camera that does not move. The drone used in the video containing the drone is the DJI Phantom 2 model. The drone has four arms and is 14.6 × 8.3 × 13.2 inches. The drone was filmed from a mobile phone camera with a resolution of 1792 × 828 at a distance of 30-50m. Each video data is 5 to 6 seconds, and frames are captured in 4 fps. The bird video data only contains less than five birds flying because only one drone appears in the drone video data. Bird flying video data, like drone video data, is 5 to 6 seconds, and frames are captured in 4 fps. The frames are extracted by using FFmpeg. The number of drone flying video data are 63, and the number of bird flying video data are 54. The unbalanced number of data in each class can generate a biased result, so the number of data in each class is almost the same.

Table 4.1. *Dataset*

Case	Training	Testing	Total
Drone flying video	51	12	63
Bird flying video	42	12	54
Total	93	24	117

The input feature for the early fusion model is extracted from the spatial stream model and motion stream model. For the training, 80 percent of data is used. For the testing, 20 percent of data is used. The class scores for the whole video are then obtained by averaging the scores across the sampled frames and crops therein.

4.3 Training and Testing

Two models are trained from scratch, independently. The spatial stream model is trained from scratch. The model trained with 300 epoch and one batch size. The network weights are learned using the mini-batch stochastic gradient descent with momentum as 0.9. For the motion model, a number of epochs are set to 100, and batch size is also 1. An initial learning rate decreases according to a fixed schedule, which is kept the same for all training sets. Learning rate schedules seek to adjust the learning rate during training by reducing the learning rate according to a pre-defined schedule. Both learning rate schedules include time-based decay, step decay, and exponential decay. For the optimizer, the stochastic gradient descent method is used.

4.4 Evaluation

In this study, we measure the performance of the stream convolutional neural networks with several different backbones. Moreover, to make an experiment for transfer learning, we compare the ConvNet with using pre-trained model and the ConvNet without using pre-trained model. Both spatial stream convolutional neural network are trained with ResNet as backbone. The motion stream neural network in this study uses the 10 stacked optical flow to get the information of movement. There can be differences between the length of stacked optical flows.

The early fusion method and late fusion method are used in this study. The early fusion is evaluated by F1-score. F1-score is used as metric for evaluation of classification models. While accuracy score represents the ratio of correct predictions, it cannot reflect the bias of the dataset. F1-score shows the precision and recall with true-positive, false-positive, and false-negative. True-positives mean the cases when the predictions are correct. False-positive means that there is a drone while there is a bird. False-negative means the model cannot predict the scene correctly.

The late fusion methods are evaluated by calculating the number of correct prediction and cross entropy loss. The cross entropy loss is a one of the methods to optimize classification model. It minimizes the distance between two probability distribution.

4.5 Analysis

This section is analysis of the results of the experiment. There are four results from the models. These result is the results of spatial training model, temporal training model, early fusion, and the late fusion.

Table 4.2. *Results*

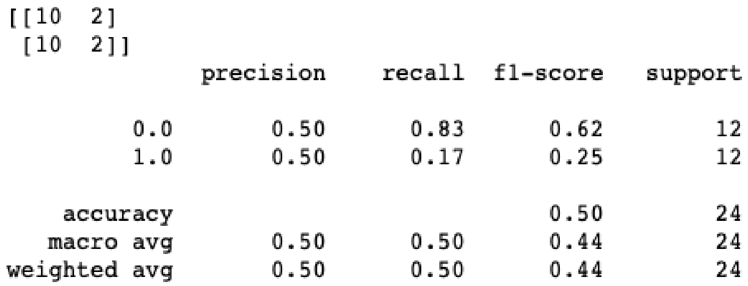
Case	ResNet34	ResNet52	ResNet101	ResNet152
Spatial Stream	50.0000	59.0909	63.6363	66.6666
Motion Stream	50.0000	63.6363	64.1616	66.6966
Late Fusion	50.0000	72.7272	81.3636	83.3636

The datasets experiment with several different backbones. ResNet-34,Resnet-50, ResNet-101 and ResNet-152 as backbone are used. The ResNet is deeper when the number followed by the ResNet is increased.

After appearance of the AlexNet (Krizhevsky et al., 2017) at the LSVRC2012 classification contest, deep Residual Network was arguably the most groundbreaking work in the computer vision area in the last few years. ResNet (He et al., 2015) makes it possible to train up to hundreds or even thousands of layers and still achieves compelling performance. For taking advantage of its powerful representational ability, the performance of many computer vision applications other than image classification has been boosted, such as object detection and face recognition. The ResNet-50 model consists of 5 stages, each with a convolution and Identity block. Each convolution block has three convolution layers, and each identity block also has three convolution layers. The deep learning models are considered to have better performance as the deeper they are. The number followed by ResNet is the number of layers the ResNet model has.

The ResNet-34 has the minimum number of layer among the ResNets used in the experiment. The results of each models which are spatial stream and motion stream are 50 percent ac curacies. This results mean that the system can classify the drone and bird 50 percent. The test classes have same number of drone and bird images, so this results mean that the system cannot classify the bird and drone well. The result means that the model randomly classify the images. The results of the other ResNets show that the late fusion has better performance compared to each stream. The result of running spatial stream which only uses the appearance features is 59.0909 and the result of running the motion stream which uses the movement features is 63.6363. The result of the late fusion is 72.7272 which is higher than running the spatial stream or motion stream. The late fusion result of running ResNet-101 is 81.3636 while the result of running the spatial stream is 63.6363 and the result of running the motion stream is 64.1616. The accuracy score of running the late fusion is 83.3636 which is higher than the scores of the spatial stream and the motion stream. From these result, the late fusion generally have higher performance than using the appearance information and the motion information separately.

Figure 4.3 shows that the results of Resnet have generally better performance as the number of layers increase. The results of ResNet-52, ResNet-101 and ResNet-152 are 59.0909, 63.6363 and 66.666. This increasing pattern depends on the depth of ResNet is also showed in the motion stream and the late fusion. The results of the motion streams slightly increase, and also the late fusion. The deeper model is, the higher accuracy is.

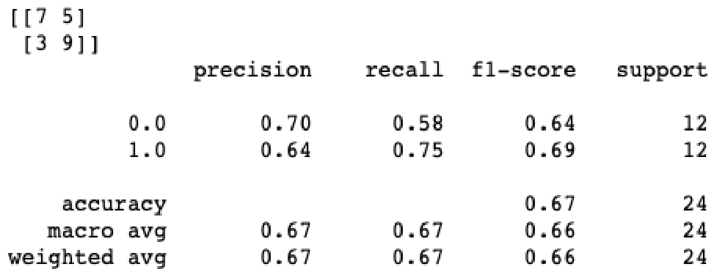


```
[[10  2]
 [10  2]]
              precision    recall  f1-score   support

         0.0       0.50      0.83      0.62        12
         1.0       0.50      0.17      0.25        12

    accuracy                           0.50        24
   macro avg       0.50      0.50      0.44        24
weighted avg       0.50      0.50      0.44        24
```

Figure 4.3. The confusion matrix and classification report of Early Fusion ResNet-50 result. The left two by two matrix is the confusion matrix.

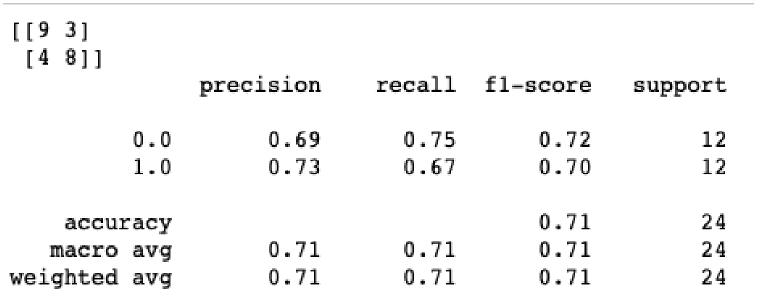


```
[[7 5]
 [3 9]]
              precision    recall  f1-score   support

         0.0       0.70      0.58      0.64        12
         1.0       0.64      0.75      0.69        12

    accuracy                           0.67        24
   macro avg       0.67      0.67      0.66        24
weighted avg       0.67      0.67      0.66        24
```

Figure 4.4. The confusion matrix and classification report of Early Fusion ResNet-101. The left two by two matrix is the confusion matrix.



```
[[9 3]
 [4 8]]
              precision    recall  f1-score   support

         0.0       0.69      0.75      0.72        12
         1.0       0.73      0.67      0.70        12

    accuracy                           0.71        24
   macro avg       0.71      0.71      0.71        24
weighted avg       0.71      0.71      0.71        24
```

Figure 4.5. The confusion matrix and classification report of Early Fusion ResNet-152 result. The left two by two matrix is the confusion matrix.

Early Fusion obtained F1-score for accuracy. Early Fusion is a method for classifying using the application feature from three frames using ResNet and the motion feature using ten stacked optical flows. SVM is used for classification. From the result of SVM, the F1-score, which shows the score according to the precision and recall are achieved. Precision is the fraction of action experiences among, while recall is the fraction of the total number of actual costs that were restored. From the result with training the spatial stream and motion stream with ResNet-34 backbone, the results are 50 percent accuracy. This means that it is hard to believe the result, because it seems the moldel makes a decision randomly. Thus, ResNet-50, ResNet-101, and ResNet-152 are used for the early fusion.

The f1-score is the relationship between the precision and recall. It is important value to analyze how much the system detect the target when there is a target or not in detail. There are two classes in the dataset, which are drone and bird. The 0.0 category means a drone class and 1.0 category means a bird class. In the drone class, the precision means how much the decision is right when the model decides that a drone exists in the data while the recall means how much the systems decides that a drone is exists in a data when there is a drone in a data. The recall and precision are both important, but there is a trade-off. Thus, it is important to analyze both values, and f1-score to analyze the relation between the recall and the precision. In the confusion matrix, if the numbers of False-Positive and False-Negative are small, the accuracy is high.

Figure 4.4 shows the confusion matrix of using the ResNet-50 and the early fusion method. In the drone class, the precision is 0.50 and the recall is 0.83. In the bird class, the precision is 0.50 and the recall is 0.17. This result means that the system classifies almost every data as drone. Thus, the recall of drone class and the precision of bird class are high, but the recall is small. The precision of both classes are all 0.50 which means that the decisions made by the system has only 50 percent accuracy. ResNet-50 can detect the drone in high possibility, but it also classify a bird as a drone, so the user can get the alarm that there is a drone frequently even thought it is a bird.

Figure 4.5 shows the confusion matrix of using the ResNet-101 and the early fusion method. In the drone class, the precision is 0.70and the recall is 0.58. In the bird class, the precision is 0.64 and the recall is 0.75. This result is better than using the ResNet-50 as a backbone. However, the recall score is small, so that it is difficult to detect a drone even though a drone is in front of the system.

Figure 4.6 shows the confusion matrix of using the ResNet-152 and the early fusion method. In the drone class, the precision is 0.69 and the recall is 0.75. In the bird class, the precision is 0.73 and the recall is 0.67. The ResNet-152 performs better than the ResNet-101 and the ResNet-50. In the drone class, the recall is higher than the recall value of using the ResNet-101 and lower than the recall value of using the ResNet-50. It means that the system with ResNet-152 backbone detects the drone when there is a drone in the data. The recall value is smaller than the recall value of ResNet-50, but the accuracy of the decision is hgiher than ResNet-50, because of higher precision value.

The accuracy results of ResNet-50, ResNet-101, and ResNet-152 are lower than the late and ResNet-50 has lower accuracy result than motion stream. The accuracy scores of the early fusion increase when the models become deeper. However, the results of using the early fusion is lower than the result of using the late fusion, and some of results even performs worse than using individual stream.

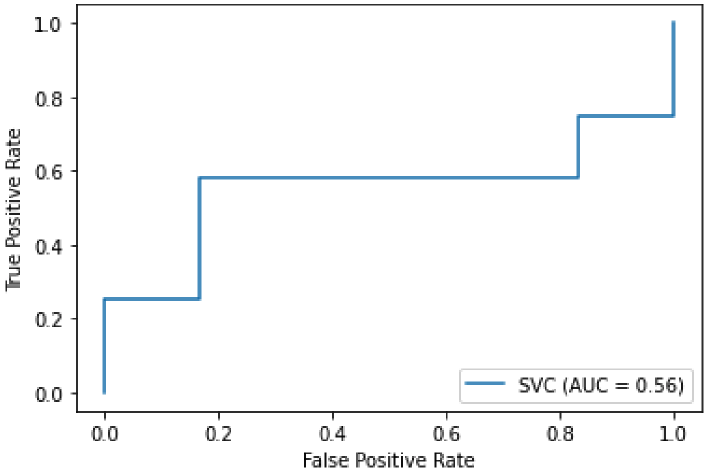


Figure 4.6. The ROC curve of using ResNet-50

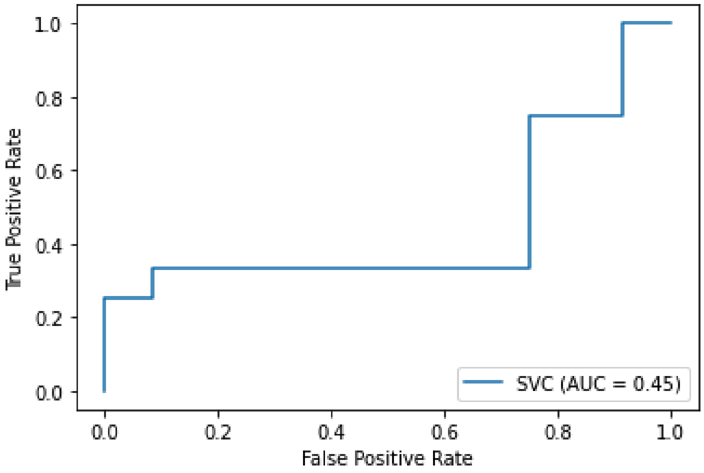


Figure 4.7. The ROC curve of using ResNet-101

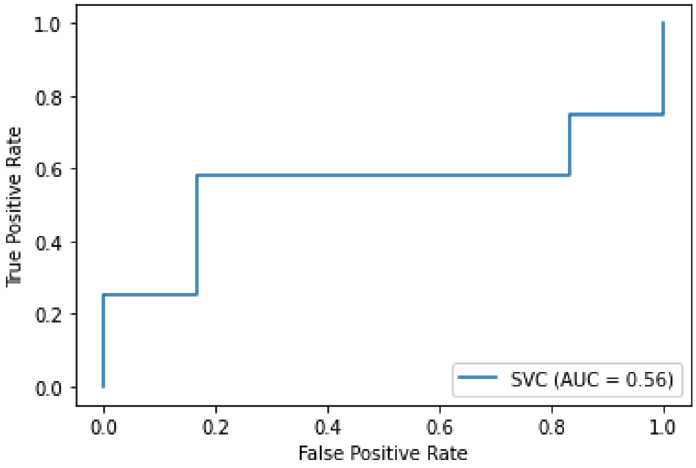


Figure 4.8. The ROC curve of using ResNet-152

The ROC curve is a plot the shows the ratio between the true positive rate and the false positive rate. The AUC is the area under the curve. If AUC is high and closes to one, then it means the the performance of the model is good. From the Figure 4.7 to Figure 4.9, The AUC of the ResNet-152 and the ResNet-50 are higher than the ResNet-101. However, it is hard to say that the ResNet-50 has better performance than ResNet-101, because the true positive rate of the ResNet-50 is smaller than the true positive rate of ResNet-101. Thus, it is difficult to say that the system with ResNet-50 backbone is better than the system with ResNet-101 baackbone.

4.6 Summary

This section explains how the dataset is gathered. The videos contain the flying drone and flying bird in the sky. The videos are filmed by a fixed camera to gain more clear features. Thus, the background is fixed in the film. The goal of this study is the comparison between using appearance and motion information separate, and together, so the size of flying objects is small. The input of the spatial stream is the still frames, and the input of the temporal stream is stacked optical flows. The frames are randomly extracted from the video. The optical flows are essential to analyze the motion feature. To see the pattern, ten optical flow frames that are divided into two components are stacked.

For the experiments, the different ResNet backbones are used. The results of using different backbones are compared. Those backbones are used in each spatial stream, temporal stream, and spatio-temporal stream. The experiments show the difference accuracy from the spatial stream, temporal stream, and spatio-temporal streams with different fusion methods.

The experiments show that the deeper ResNet, which is ResNet-152, shows a better result than ResNet-50 and ResNet-101. Also, it shows that the spatio-temporal model using the late fusion method returns better results than using the models separately. However, the early fusions are not returned to a better result than using the models independently.

From the result, it is better to use deeper ResNet with late fusion. However, as a network is deeper, there is a trade-off that the complexity is higher. Thus, it is important to consider the computing power while using ResNet-152 with late fusion.

CHAPTER 5. CONCLUSION

Recently, several works have explored various methods to exploit spatio-temporal information. This study introduced the comparison between three methods of detecting drones visually, with affordability and accessibility. The data contains two different classes which are a drone and bird. The average time of video dataset is five seconds to six seconds and the flying objects are filmed on the fixed background. After the data collection phase, two features are extracted by CNN model. Two features are frames for appearance features and optical flow for motion features.

Through the comparison between the late fusion method and the early fusion methods, using both appearance features and motion features with the late fusion performs better than using the information individually. The late fusion is using the average of the scores from individual trained result while the early fusion is feature fusion. Moreover, the depth of ResNet impacts the result of the classification. The deeper ResNet performance better.

5.1 Limitation and Future Work

There are limitations regarding modeling and methodology.

5.1.1 Methodology

There are various backbones, such as the Visual Geometry Group (VGG) Simonyan and Zisserman (2014b). The different backbone can return the better result. The parameters that are used in the experiment are not adjusted. If the hyper-tunning paramters are used for the training, the better result can be expected.

The size of bird or drone contains small pixels in the frames, less than 50 pixels. The CNN models that are used in the experiment are not actually targeting the small object classification. They are hard to extract the features from the small objects in the frame. Using the models aiming small object classification can help to increase the performance of system.

This system focused on the moving object on the fixed background. In the future work, adding tracking step to capture the only flying object with moving background can be considered. With tracking step, it can alarm the users more precisely. Furthermore, if the flying object can be cropped and zoomed, the accuracy of classification can be higher.

5.1.2 Dataset

The two-stream convolutional networks are applied to the drone and bird domain using spatial feature and motion features. The CNN model was used, but the CNN model requires many data. However, the published drone video data was minimal, and it was not easy to collect data that met specific criteria. The dataset should be collected more to sufficient train and test. More dataset for training means the CNN can extract more precise. With more dataset, and enough training, the neural network is expected to detect in much higher accuracy. Furthermore, only one type of drones is used in the experiment. There are various shape and types of drones. The drone can be various. The non-drone dataset also only includes a bird class. Thus, the non-drone dataset can be various through including the other types of flying objects.

REFERENCES

Aker, C., & Kalkan, S. (2017). Using deep networks for drone detection. *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017*(August). doi: 10.1109/AVSS.2017.8078539

Bisio, I., Garibotto, C., Lavagetto, F., Sciarrone, A., & Zappatore, S. (2018). Unauthorized Amateur UAV Detection Based on WiFi Statistical Fingerprint Analysis. *IEEE Communications Magazine*, *56*(4), 106–111. doi: 10.1109/MCOM.2018.1700340

Bonetto, M., Korshunov, P., Ramponi, G., & Ebrahimi, T. (2015). Privacy in mini-drone based video surveillance. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*, *2015-Janua*, 1–6. doi: 10.1109/FG.2015.7285023

Case, E. E., Zelnio, A. M., & Rigling, B. D. (2008). Low-cost acoustic array for small UAV detection and tracking. *National Aerospace and Electronics Conference, Proceedings of the IEEE*(August 2008), 110–113. doi: 10.1109/NAECON.2008.4806528

Chen, V. C., Li, F., Ho, S. S., & Wechsler, H. (2006). Micro-doppler effect in radar: Phenomenon, model, and simulation study. *IEEE Transactions on Aerospace and Electronic Systems*, *42*(1), 2–21. doi: 10.1109/TAES.2006.1603402

Clarke, R., & Bennett Moses, L. (2014). The regulation of civilian drones' impacts on public safety. *Computer Law and Security Review*, *30*(3), 263–285. Retrieved from `http://dx.doi.org/10.1016/j.clsr.2014.03.007` doi: 10.1016/j.clsr.2014.03.007

De Quevedo, D., Urzaiz, F. I., Menoyo, J. G., & López, A. A. (2019, 9). Drone detection and radar-cross-section measurements by RAD-DAR. In *Iet radar, sonar and navigation* (Vol. 13, pp. 1437–1447). Institution of Engineering and Technology. doi: 10.1049/iet-rsn.2018.5646

Deruyck, M., Wyckmans, J., Martens, L., & Joseph, W. (2016). Emergency ad-hoc networks by using drone mounted base stations for a disaster scenario. *International Conference on Wireless and Mobile Computing, Networking and Communications*. doi: 10.1109/WiMOB.2016.7763173

A Drone, Too Small for Radar to Detect, Rattles the White House - The New York Times. (n.d.). Retrieved from `https://www.nytimes.com/2015/01/27/us/white-house-drone.html`

Ezuma, M., Erden, F., Anjinappa, C. K., Ozdemir, O., & Guvenc, I. (2019). Micro-UAV Detection and Classification from RF Fingerprints Using Machine Learning Techniques. *IEEE Aerospace Conference Proceedings*, *2019-March*. doi: 10.1109/AERO.2019.8741970

Ezuma, M., Ozdemir, O., Anjinappa, C. K., Gulzar, W. A., & Guvenc, I. (2019). Micro-UAV detection with a low-grazing angle millimeter wave radar. *IEEE Radio and Wireless Symposium, RWS*. doi: 10.1109/RWS.2019.8714203

Gaidon, A., Harchaoui, Z., & Schmid, C. (2011). Actom Sequence Models for Efficient Action Detection. , 3201–3208. Retrieved from `http://lear.inrialpes.fr/people/last-name` doi: 10.1109/CVPR.2011.5995646{\"{i}}

Gökçe, F., Üçoluk, G., Şahin, E., & Kalkan, S. (2015). Vision-Based Detection and Distance Estimation of Micro Unmanned Aerial Vehicles. *Sensors*, *15*(9), 23805–23846. doi: 10.3390/s150923805

He, K., Zhang, X., Ren, S., & Sun, J. (2015, 12). Deep Residual Learning for Image Recognition. Retrieved from `http://arxiv.org/abs/1512.03385`

Hubbard, B., Karasz, P., & Reed, S. (2019). Two Major Saudi Oil Installations Hit by Drone Strike, and U.S. Blames Iran. *The New York Times*, 1–4. Retrieved from `https://www.nytimes.com/2019/09/14/world/middleeast/saudi-arabia-refineries-drone-attack.html`

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2016, 12). FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. Retrieved from `http://arxiv.org/abs/1612.01925`

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6). doi: 10.1145/3065386

Laptev, I., Marszałek, M., Schmid, C., & Rozenfeld, B. (n.d.). *Learning realistic human actions from movies* (Tech. Rep.). Retrieved from `www.weeklyscript.com.`

Morris, R., & Thurston, G. (2015). INTERIM FINAL RULE REGULATORY EVALUATION: Registration and Marking Requirements for Small Unmanned Aircraft. (December). Retrieved from `https://www.faa.gov/news/updates/media/2015-12-13_2120-AK82_RIA.pdf`

Murugan, D., Garg, A., & Singh, D. (2017). Development of an Adaptive Approach for Precision Agriculture Monitoring with Drone and Satellite Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(12), 5322–5328. doi: 10.1109/JSTARS.2017.2746185

Philpott, R., Kwasa, B., & Bloebaum, C. (2018). Use of a Value Model to Ethically Govern Various Applications of Small UAS. *Drones*, *2*(3), 24. doi: 10.3390/drones2030024

Prates, P. A., Mendonca, R., Lourenco, A., Marques, F., Matos-Carvalho, J. P., & Barata, J. (2018). Vision-based UAV detection and tracking using motion signatures. *Proceedings - 2018 IEEE Industrial Cyber-Physical Systems, ICPS 2018*, 482–487. doi: 10.1109/ICPHYS.2018.8390752

Rahman, S., & Robertson, D. A. (2018). Radar micro-Doppler signatures of drones and birds at K-band and W-band. *Scientific Reports*, *8*(1), 1–11. Retrieved from `http://dx.doi.org/10.1038/s41598-018-35880-9` doi: 10.1038/s41598-018-35880-9

Ralphs, C. (2018). Better, faster, stronger. *TLS - The Times Literary Supplement*, *2018-June*(6009), 28. doi: 10.5860/lrts.53n4.261

Savarese, S., DelPozo, A., Niebles, J. C., & Fei-Fei, L. (2008). Spatial-temporal correlatons for unsupervised action classification. In *2008 ieee workshop on motion and video computing, wmvc.* doi: 10.1109/WMVC.2008.4544068

Scott, J. E., & Scott, C. H. (2018). Models for Drone Delivery of Medications and Other Healthcare Items. *International Journal of Healthcare Information Systems and Informatics*, *13*(3), 20–34. doi: 10.4018/IJHISI.2018070102

Shin, D. H., Jung, D. H., Kim, D. C., Ham, J. W., & Park, S. O. (2017). A Distributed FMCW Radar System Based on Fiber-Optic Links for Small Drone Detection. *IEEE Transactions on Instrumentation and Measurement*, *66*(2), 340–347. doi: 10.1109/TIM.2016.2626038

Simonyan, K., & Zisserman, A. (2014a). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems* (Vol. 1, pp. 568–576). Neural information processing systems foundation.

Simonyan, K., & Zisserman, A. (2014b, 9). Very Deep Convolutional Networks for Large-Scale Image Recognition. Retrieved from `http://arxiv.org/abs/1409.1556`

Unlu, E., Zenou, E., & Riviere, N. (2018a). Generic Fourier Descriptors for Autonomous UAV Detection. *ICPRAM 2018 - Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods*, *2018-Janua*(January), 550–554. doi: 10.5220/0006680105500554

Unlu, E., Zenou, E., & Riviere, N. (2018b). Using shape descriptors for uav detection. *IS and T International Symposium on Electronic Imaging Science and Technology*, 2761–2766. doi: 10.2352/ISSN.2470-1173.2018.09.SRV-128

Vision-based detection of aircrafts and UAVs Artem ROZANTSEV (Tech. Rep.). (n.d.).

Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2011). Action recognition by dense trajectories. In *Proceedings of the ieee computer society conference on computer vision and pattern recognition.* doi: 10.1109/CVPR.2011.5995407

Wild, G., Murray, J., & Baxter, G. (2016). Exploring Civil Drone accidents and incidents to help prevent potential air disasters. *Aerospace*, *3*(3). doi: 10.3390/aerospace3030022

Wilson, R. L. (2014). Ethical issues with use of Drone aircraft. *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, ETHICS 2014*. doi: 10.1109/ETHICS.2014.6893424

Zhao, C., Chen, C., He, Z., & Wu, Z. (2018). Application of auxiliary classifier wasserstein generative adversarial networks in wireless signal classification of illegal unmanned aerial vehicles. *Applied Sciences (Switzerland)*, *8*(12). doi: 10.3390/app8122664

Zhu, Y., Lan, Z., Newsam, S., & Hauptmann, A. G. (2017, 4). Hidden Two-Stream Convolutional Networks for Action Recognition. Retrieved from `http://arxiv.org/abs/1704.00389`