# Drone Detection based on Deep Learning with Vision and Acoustic Methodologies


Dongwhan Lee
*Software Convergence*
*Kyunghee University*
Yongin, Republic of Korea
derick_lee@khu.ac.kr

Youngseo Kim
*Dept. Human Centered AI*
*Sangmyung University*
Seoul, Republic of Korea
201910787@sangmyung.kr

Juann Kim
*Dept. Software*
*Sangmyung University*
Seoul, Republic of Korea
201920951@sangmyung.kr

Heeyeon Shin
*Computer Engineering*
*Kyunghee University*
Yongin, Republic of Korea
567didi@khu.ac.kr

Yeeun Heo
*Software Engineering*
*Soongsil University*
Seoul, Republic of Korea
gjdpdms2005@soongsil.ac.kr

Yaqin Wang
*Computer and Information Technology*
*Purdue University*
West Lafayette, United States
wang4070@purdue.edu

Eric T. Matson
*Computer and Information Technology*
*Purdue University*
West Lafayette, United States
ematson@purdue.edu



***Abstract*—Autonomous drones have been studied in a variousety of industries in recent years including delivery services and disaster protection. As the supply of low-cost unmanned aerial vehicles(UAV) has been increasing, a robust Collision Avoidance System(CAS) and CUAS(Counter unmanned aerial system) is critical in order to manage autonomous drone traffic control and prevent drone flights in secured area. For these systems, drone detection is one of the most important step in the overall process. The goal of this paper is to detect the target drone using the microphone of iPhone 6 and the buil-in camera of the detecting drone by training deep learning models based on vision and acoustic features. For evaluation, three methods are used: visual-based, audio-based, and the decision fusion of both features. As the individual features are able to capture what cannot be seen or heard by another, visual and audio-based features are complementary. Therefore, the combination of audio and vision-based features is used to obtain higher performance on drone-to-drone detection. For the environment setting, by flying two drones in the sky at a fixed Euclidean distance of 20m, image and audio data are collected from the detecting drone. Deep learning methods are applied to investigate the optimal performance. The result shows that the decision fusion of audio and vision-based features showed the highest accuracy among three evulation methods.**



***Index Terms*—unmanned aerial vehicles, collision avoidance system, acoustic-based, vision-based, drone detection, deep learning**


## I. Introduction

Recently, the demand for Unmanned Aerial Vehicles (UAV), known as drones, started to increase. With the growing number of UAVs, the importance of small and low-cost UAVs has considerably been expanded.

The benefits of drones are enormous: operating without a pilot, fusing diverse fields, no high-cost infrastructure, etc. As reported by global market research publishing and management consulting firm *Grand View Research*, the size of the global commercial drone market is expected to flourish to a compound annual growth rate (CAGR) of 57.5% from 2021 to 2028 [1].

UAVs have been utilized in various fields including agriculture, construction, technical service, health care, and delivery systems. In the case of a UAV delivery system, it is mentioned that *United Parcel Service (UPS)* and *Consumer Value Stores (CVS)* are preparing to introduce a medication prescription delivery system using UAVs for people in the largest retirement community in Florida [2]. The use of UAVs has provided for personal, commercial, and even government means since UAV markets have been accelerated [3].

With the rapid spread of the diverse field of UAVs, several concerns about drones are also heightened. In the matter of UAV delivery systems, several technologies are needed in managing drone safety. As proof, British Airways Airbus A320 collided with a UAV at London Heathrow on 17 April 2016 [4]. Consequently, this event shed a new light on the risk of UAV collision, so resolving security problems of UAVs is pivotal to prevent threatening circumstances. For the safe flight of autonomous UAVs, Collision Avoidance System (CAS) is essential. CAS is a system for reducing collisions among drones. For CAS, various sensors are attached to drone, so avoiding obstacles and preventing collisions can be done without any support on the ground. On account of constructing CAS effectively, it is prior to detecting a target UAV successfully.

As many UAVs are commercialized, UAVs are also used for malicious reasons. Drones were used to attack and invade, such as an assassination attempt on the president of Venezuela in 2018 [5]. To prevent this situation, Counter Unmanned Aerial System (CUAS) has been further developed. In CUAS, it is critical to detect, track, and eventually destroy malicious UAVs [6]. Hence, it is needed to detect specific drones in order to solve problems of collision avoidance, removing malicious drones, etc.

Researchers have shown that cameras or audio devices,


Identify applicable funding agency here. If none, delete this.

placed on the ground, were handled to test CAS experiments [7]. However, a little is researched based on two features, vision and acoustic. Hence, an unprecedented UAV detection system using two features above is introduced. The proposed system detects a flying UAV in the mid-air via data of video and audio. It is to be noted that a CAS test is conducted in the moving expanding the CUAS invasion system between moving UAVs with a fusion of two methodologies.

## II. Related Work

### A. CUAS (Counter Unmanned Aerial System)

In recent years, drone flights in the Air Exclusion Zone (AEZ) have repeatedly occurred. In 2015, a man was detained, since he flew his UAV 100 feet above Lafayette Park near the White House [8]. Hence, CUAS has been conducted to prevent these occurrences. CUAS is a system to detect and track UAVs that approach protected or secure areas. Following [9], researchers present a study on a shooting system using Class 1 UAVs, defined as small and transportable, with a human-in-loop, an autonomous and vision-based system. If a target UAV is continuously captured in several frames, a UAV pilot changes the mode to autonomous, and the detecting UAV approaches the target UAV. A complete procedure is to move the detecting UAV towards the target UAV and put the target UAV into an inoperable status. This research demonstrates that the proposed system depends on a pilot to be practiced or skilled.

### B. CAS (Collision Avoidance System)

If UAVs were to collide in the air, collisions can cause bigger accidents than that of land-based vehicles. For this reason, it is requisite to possess an effective CAS to reduce collisions between UAVs. For a CAS, various methods exist to detect and avoid target UAVs. The authors of [10] defined passive sensors as detecting the energy emission from scenery and objects in observations time, including visual and infrared camera. Passive sensors have limitations including a high dependency on image processing that extracts informative data from raw data, and the weather condition. However, passive sensors is beneficial as a mean to recognize collisions: small, light-weight, lower consumption, and easy to install on UAVs. In [11], the purpose is to prevent collisions of animals and vehicles. Though this system detects animals and avoids collisions at the maximum vehicle speed of 35km/h, the accuracy of this paper is fairly high, 82.5%.

### C. UAV Detection using Sensors

*1) Rader:* These days, various methods have been used for drone detection, including Radar, LiDAR, Computer vision, and Acoustic sensor. Each method has its own strengths and limitations when detecting UAVs. For one's limitations, other methods may complement or overcome the method of another when they are used together to detect a UAV. Most commercial products that are widely utilized for drone detection are based on radar[12]. More specifically, radar is used for binary drone and multi-drone detection. However, radar-based detection has limitations. Radar is not optimized for plastic material drone detection and small drone detection at widely varying ranges [12]. Furthermore, a lot of false positives are recorded as it is hard to distinguish the difference between drones and other targets such as birds [13]. On the other hand, in this paper, audio and vision-based features, which are not significantly affected by the size or the materials of drones, are used for drone detection.

*2) Computer Vision:* In the past few decades, several researches about UAV detection using computer vision are already conducted.[***] Vision-based object detection method is accurate enough to classify the classes and can localize the actual location. Moreover, a single camera, a small, self-contained, and portable device, is even more accessible to perform detection tasks rather than LiDAR and Radar.

*Craye* [14] and *Ulzhalgas* [15] used a single camera to detect UAV on the ground using a computer vision method. Each author generated an accuracy of 73.5%, and 74.2% respectively on detecting objects from particular frames of video data based on CNN.

*3) Acoustic:* Using the iPhone, portable mobile devices, as a microphone is easy to utilize by anyone and affordable for drone detection than other sensors, such as radar, LiDAR, etc. Especially, as audio signal processing based on sound data is less expensive than radar, and can be more economical than other sensors aforementioned.

The author in [7] presented a UAV detection system using multiple acoustic nodes along with machine learning models. This system evidances that the models are able to recognize the acoustic signals in a wider range under 3-dimesional spaces. By ssing the low-priced multiple acoustic nodes, this technique can detect UAVs at most distance, 75m. Additionally, the paper [7] deduces that based on audio features, a method using deep learning has higher performance that of machine learning.

*4) Fusion:* As previously stated, diverse sensors can be employed to detect UAVs. As proposed by several papers, a fusion technique of more than two domains critically influenced academic dialogue on improving performances of UAV detection. The combination of radar and audio sensors, suggested method in [16], can detect and track rotor type of UAVs. An electro-optical and acoustic based fusion system was deployed to detect, localize, and track UAVs [17]. In [18], vision information and inertial measurement unit (IMU) were collected using a monocular camera. Collected data processed by end-to-end deep neural network architecture with feature fusion resulted in less than 3% error.

**(sensor fusion, feature fusion)**

## III. Methodology

### A. System Overview

A proposed system has a single camera and audio device. The data were collected using a built-in camera of a drone. iPhone 6 was also attached to a drone as an audio device to record sound data.

The conducted experiments were assigned to two drones to encounter each other in mid-air. Audio and video data were

collected by the detecting drone while hovering. The target drone moved at the range of a camera grid (16:9) of detecting drone, with fixed euclidean distance.

Several terminologies were set to indicate distances between two drones. As depicted in Fig 1, indicating a front view, projected distance ($d_p$) and Altitude ($a$) were demonstrated. $d_p$ is the euclidean distance between two drones. *A* is defined as an altitude from the ground. Fig 2 provides a top view. Vertical Range ($v$) shows the distance between two drones when the target drone moves horizontally.

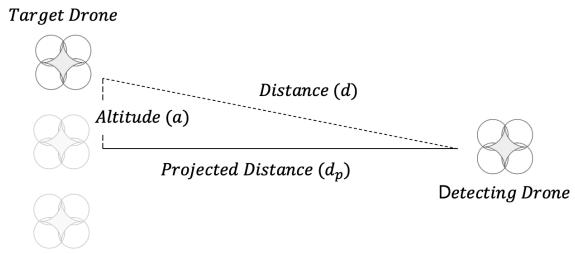



Fig. 1. Depiction of front view when flying drones encounter in mid-air

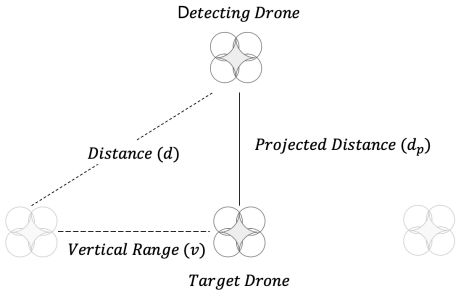



Fig. 2. Depiction of top view when flying drones encounter in mid-air

Drone data were sectioned into two classes: drone and no drone. Drone classification was performed to identify the existence of drones. A required processing flow can be segmented into five parts displayed in Fig 3: (1) capturing a surrounding data; (2) feature extraction; (3) training; (4) storing and sampling trained data; (5) feature classification.

### B. Data Collection

*1) Environment setting: DJI Matrice 200 V2* and *DJI Mavic 2 Pro* were used to collect acoustic signals and video files. Audio files were collected using *iPhone 6* attached to *Mavic 2 Pro* as shown in Figure 4, and video files were collected using the built-in camera, while both drones were flying in the air. *DJI Mavic 2 Pro*, which has a built-in camera and iPhone 6 attached, is defined as a *Detecting Drone*, and *DJI Matrice 200 V2* is defined as a *Target Drone*.

The data was collected in a wide field at Romney, Lafayette, IN, 47981. The weather conditions for the data collected varied by day. The different weather conditions include windy, sunny, and foggy days. Also, Table 1 shows the values of weather elements: temperature, wind speed, precipitation, and humidity.

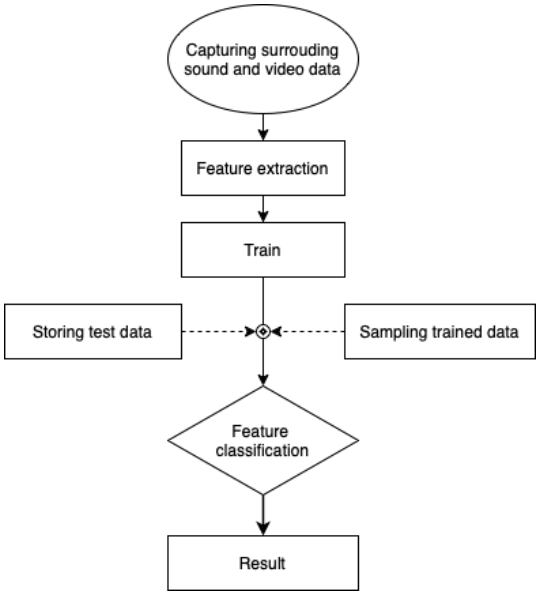



Fig. 3. Overview of drone detection system workflow

Fig. 4. iPhone 6 attached to DJI Mavic 2 Pro

TABLE I
THE STATISTICAL VALUES OF WEATHER ELEMENTS

| | Range | Median | Mean | Maximum | Minimum |
|---|---|---|---|---|---|
| **Temperature(°C)** | 17.2 ∼ 29 | 27 | 26.65 | 29 | 17.2 |
| **Wind(mi/h)** | 2 ∼ 12 | 7 | 7 | 12 | 12 |
| **Humidity(°C)** | 35 ∼ 87 | 64 | 61.9 | 87 | 35 |
| **Precipitation(" in last 24h)** | 0 ∼0.15 | 0 | 0.02 | 0.15 | 0 |

The detecting drone was hovering at the altitude of 10m. While hovering, the audio and video data were collected using the built-in camera and iPhone 6 attached to the detecting drone. Then, the target drone was maintaining the Euclidean distance with detecting drones by 20m. With the fixed Euclidean distance, the target drone was moving at a range of the camera range (16:9, 1920x1080), changing its vertical range and altitude within the screen. It can be shown in Figure 5. For no drone data, the detecting drone with attached iPhone6 was hovering at the altitude of 10m, without the target drone flying at the same time. So, the background image and noise data were collected, other environment factors being the same that of the drone data. For both audio and image files, 1509

TABLE II
NUMBER OF SAMPLES FOR EACH TWO CLASSES (NO DRONE DATA DRONE DATA)

| Type of data | Class | # Audio files | # Image files | # Label files | Total time(s) | Total time(s) |
|---|---|---|---|---|---|---|
| **Train** | drone | 1055 | 1055 | 906 | 2110 | 3026 |
| | no drone | 1055 | 1055 | - | | |
| **Validation** | drone | 300 | 300 | 600? | 600 | |
| | no drone | 300 | 300 | - | | |
| **Test** | drone | 158 | 158 | 134 | 316 | |
| | no drone | 158 | 158 | - | | |

samples are collected for every two classes, so, in total, 3018 samples are collected as shown in Table .

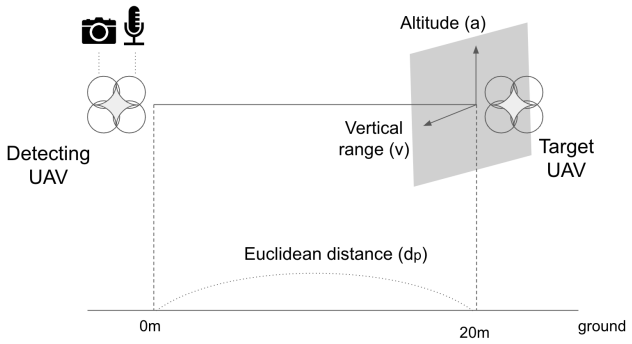



Fig. 5. A view of two drones while flying at the same time

While various previous drone detection-related papers collected the drone data when the target drone is hovering, this paper analyzes not only the hovering data but also the data with the target drone moving within the screen of detecting drone. Also, this experiment is done using the iPhone 6 with the basic performance. The strength (key contribution) of this paper is the data being collected in a more realistic situation with low price.

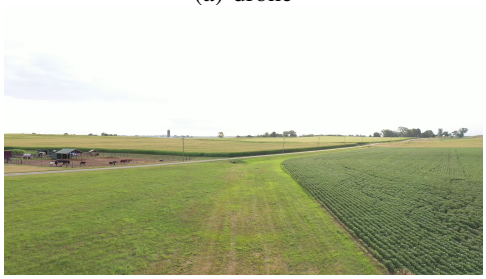
The acoustic signal was recorded by a "wav" format, and the sampling rate is 44.1kHz. Since the data are collected in the real world, a variety of background noises are also recorded for both drone data and no drone data. The noises include the sounds of insects, airplanes, human voices, animals including birds, and ground vehicles such as tractors and trucks. For vision, the video files are recorded in the same environment with noises.

In case of data collection for vision, *DJI Mavic 2 Pro* built-in camera, specification by CMOS (Complementary Metal Oxide Semiconductor) sensors with 20 million effective pixels was used to collect video files. As a loading task of the vision dataset, the raw "mp4" video files were split into images by 1 second. Each image has 1920 x 1080 resolution, and the image format is "jpg". Fig 6 shows images of drone data and no drone data.

### C. Feature extraction

*1) Audio:* Before the feature extraction, audio segmentation was done. In [19], splitting audio files into 1 seconds showed the highest performance than other time intervals when training the Deep Learning model.Therefore, for this paper, the audio files were split into 1 second for audio segmentation[20].

(a) drone



(b) no drone

Fig. 6. Visuals of sample images

Figure 8 shows the waveforms of no drone data and drone data recorded with the target drone, maintaining a 20m distance with the detecting drone. In order to classify the drone sound and various noises including birds and airplanes, Mel Frequency Cepstral Coefficients (MFCC) is used, which is a non-linear mapping of the original frequency according to the auditory mechanism of the human ear [21].

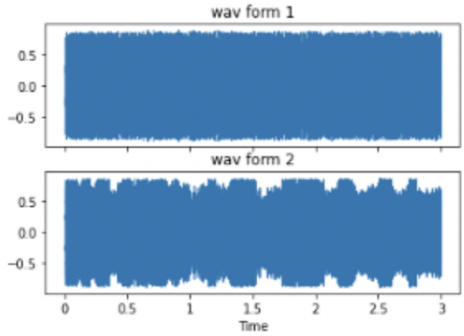



Fig2: wav form 1(none_Drone), wav form 2(20m)

Fig. 7. The waveform of no drone data and drone data

Furthermore, MFCC is widely used for audio classification and successfully used with Machine Learning approaches [22][23]. Also, it has valuable features to capture periodicity from the fundamental frequencies caused by the rotor blades of a UAV [24]. Therefore, the two different audio signals with different waveform are extracted using MFCC. In the audio analysis, the features are extracted using Librosa from the Python package. Each data for 1 second is composed of 44 x 20 size. The examples of extracted feature map of two classes are shown in Fig 9 [20].

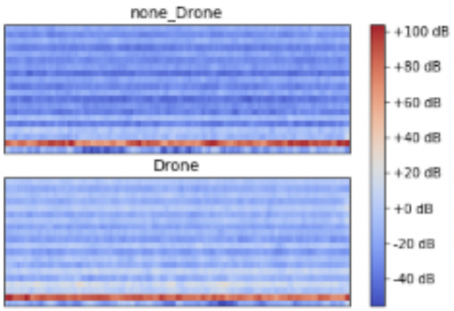



Fig.3: evaluation of MFCC feature between none_Drone and Drone sample

Fig. 8. MFCC Feature map of no drone data and drone data

*2) Vision:* In order to train the model for drone detection, all the ground truth objects in the images need to be labeled first. This dataset is labeled using the "LabelImg" [19], which is open source tools as shown in Fig 10. Moreover, the coordination of the bounding box including the location information of drones were generated as text files.

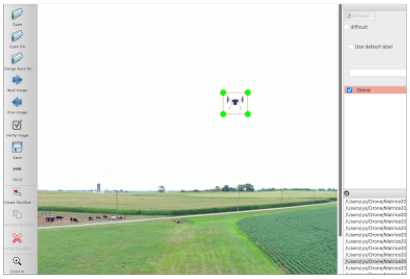
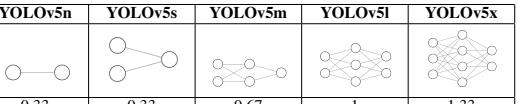


Fig. 9. Labeling software used to generate the ground truth bounding box for the drone detection task.

To extract image features, Convolution with Batch normalization and Leaky ReLU(CBL), Spatial Pyramid Pooling(SPP) and Cross Stage Partial(CSP) were used in the backbone layer of YOLOv5[25]. Backbone network extracts feature maps of various sizes from input images through the convolution layer and pooling layer. The overall structure is shown in Fig 12. First, CBL is a block that is essentially used to extract features consisting of the convolution layer, batch normalization and leaky ReLU. SPP improves performance by pooling various sizes of feature maps with filters and then merging them again. ]The CSP divides the feature map of the base layer into two parts to reduce the heavy inference computations caused by duplicate gradient information. Then, they are combined again in the cross-stage hierarchy method proposed in this paper[26]. In this way, the propagated gradient information can have a large correlation difference by switching concatenation and transition steps. In addition, CSP can greatly reduce the amount of computation and improve inference speed as well as accuracy.

5 Backbone networks - YOLOv5-n,s,m,l,x were used. Each

TABLE III
PARAMETERS & DIAGRAM OF YOLOV5 MODELS ACCORDING TO DEPTH_MULTIPLE AND WIDTH_MULTIPLE.

| Model | YOLOv5n | YOLOv5s | YOLOv5m | YOLOv5l | YOLOv5x |
|---|---|---|---|---|---|
| Diagram |  | | | | |
| depth_multiple | 0.33 | 0.33 | 0.67 | 1 | 1.33 |
| width_multiple | 0.25 | 0.5 | 0.75 | 1 | 1.25 |

model is distinguished by depth_multiple and width_multiple, and can be organized as shown in Table . The larger the *depth_multiple* value, the more BottleneckCSP() is repeated to become a deeper model. Also, the larger the *width_multiple*, the higher Convolution filter number of the corresponding layer.

### *D. Deep learning models*

*1) Audio:* Among various classifiers, CNN was used for this paper as it has shown high performance (outstanding results) in audio signal classification with spectral features such as MFCCs [30][31]. CNN was used for binary classification to distinguish the drone data and no drone data. The data for both no drone and drone data were split into training and testing data with percentages of 90% and 10%. 5 Fold Cross Validation was used for determining a more accurate estimate of model prediction on an unknown dataset.

For the parameters of CNN, four batch sizes, 16, 32, 64, and 128 were all tested with the learning rate of 0.0001 and TensorFlow Keras optimizer, Adam. Also, Early Stopping and model check point were used for saving the best model during training in order to limit the chance of the model being overfitted to the training dataset. From different parameters, 32 batch size outperformed compared to other batch sizes. When training, if there was no further reduction in the cross-entropy loss value after 10 consecutive epochs, training stopped [32][33].

*2) Vision:* In general, image classification refers to images in which only one object appears and is analyzed. In contrast, object detection involves both classification and localization tasks and is used to analyze more realistic cases in which multiple objects may exist in an image [34].

The object detection model can be divided into two main types: one-stage model and two-stage model. Compared to other two-stage object detection models including R-CNN[] and Faster R-CNN[], one stage model such as YOLO can calculate fast enough to conduct in real-time object detection task. Since object detection for CUAS and CAS has to be implemented in real-time, YOLOv5 is selected for appropriate model in drone detection in this research.

The YOLOv5 model can be represented by YOLOv5-n,s,m,l,x depending on the capacity of the model and the number of parameters.

Models with large capacities such as YOLOv5x can increase accuracy, but have the disadvantage of slow operation. Conversely, a lightweight model such as YOLOv5n is fast, but can not get outstanding performance in accuracy.

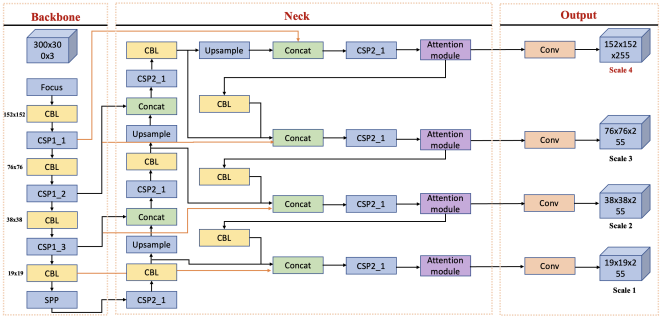



Fig. 10. YOLOv5 Architecture

## IV. Experiment

In this paper, the experiments are completed under the following working conditions: The processor is Intel(R) Xeon(R) CPU @ 2.20GHz, memory 8GB, GPU is Nvidia Tesla T4. The operating system is both Mac OS and Windows.

### A. Audio

The acoustic data is composed of 25 minutes of drone and no drone sounds. The training dataset was used for training the CNN model, and the test dataset was used to test the trained model. In the experiment, the trained CNN model using 5-fold validation was able to classify drone and no drone data with showing an accuracy of 83.44% and a consistent decrease in the loss. Figure 11 shows the accuracy and the loss of the CNN model with 200 epochs.

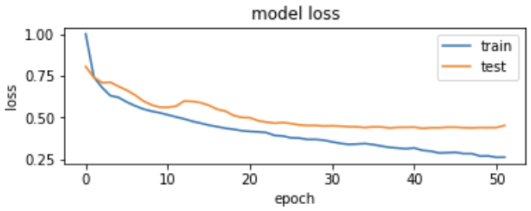



Fig. 11. fig 5.1 loss graph

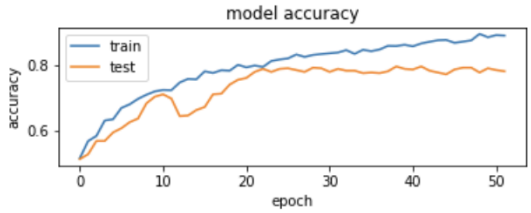



Fig. 12. fig 5.2 accuracy graph

Figure 13 shows the result of the binary classification of drone detection using the Confusion Matrix. For the evaluation of the model, three methods are obtained from the confusion matrix: precision, recall, and f1-score. Using these three methods of evaluation, Table  shows the results for the methods.

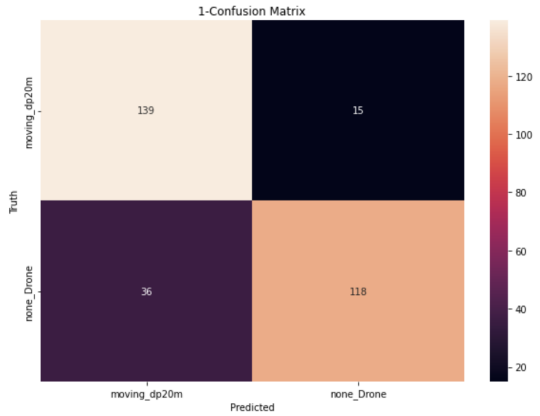



Fig. 13. confusion matrix

### B. Vision

*1) setup:* The training and validation dataset for the vision task include 2710 images, and the dataset for testing includes 308 images, which are collected in person. Input images are fixed with the size of 640 x 640 demanded by YOLOv5. For the hyperparameters, the batch size is 16 and epochs are 30(100). The Optimizer used for YOLOv5 is Stochastic Gradient Descent (SGD). In this paper, the comparison of experiments with 5 different YOLOv5 models is progressed to figure out which model is appropriate for drone-to-drone detection task. The evaluation performances for drone detection tasks are measured by precision, recall, F1-score, mean Average Precision (mAP), and accuracy.

*2) Training & Testing:* For the same conditions in the progress of training, the initial value for each parameter is an equal environment for all the different YOLOv5 models.

TABLE IV
Training results for YOLOv5 models in mAP, Precision, Recall, F1-score

| Models | mAP_0.5 | mAP_0.5:0.95 | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| YOLOv5n | 0.8093 | 0.5293 | 0.7084 | 0.8742 | 0.7826 |
| YOLOv5s | 0.8473 | 0.5682 | 0.718 | 0.8768 | 0.7895 |
| YOLOv5m | 0.8689 | 0.5842 | 0.7518 | 0.8831 | 0.8122 |
| YOLOv5l | 0.8101 | 0.5357 | 0.6896 | 0.9271 | 0.7909 |
| YOLOv5x | 0.8731 | 0.5961 | 0.7104 | 0.9073 | 0.7969 |

When testing was performed through the optimal weight obtained in the existing training process in the only vision task, the experiment comparing the five YOLOv5 models is the same as the results in Table . When comparing the performance of the model with the Model inference time, the YOLOv5n model showed the best performance.

### C. result

In the process of the drone-to-drone detection task, audio and vision tasks were conducted separately. During the experiment, acoustic signals and images were used. The method using vision can detect and locate objects. However, due to limited viewing angles, drone objects can only be detected if there is a drone object in the camera range. Furthermore,

TABLE V
TESTING RESULTS FOR YOLOV5 MODELS IN MAP, PRECISION, RECALL, F1-SCORE

| Models | mAP_0.5 | mAP_0.5:0.95 | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| YOLOv5n | 0.744 | 0.541 | 0.589 | 0.914 | 0.716362 |
| YOLOv5s | 0.824 | 0.6 | 0.646 | 0.914 | 0.756979 |
| YOLOv5m | 0.801 | 0.57 | 0.639 | 0.857 | 0.732116 |
| YOLOv5l | 0.76 | 0.552 | 0.619 | 0.886 | 0.728816 |
| YOLOv5x | 0.846 | 0.619 | 0.605 | 0.914 | 0.728071 |

there is also a limit to poor performance in foggy or noisy backgrounds. In the case of audio, it is difficult to clarify the exact location. However, for the strength, audio is not significantly affected by the weather or background, so it can robustly detect drones. Therefore, the method of using vision and audio together through decision fusion is considered a stronger drone detection method that can compensate for each other's shortcomings and showed higher accuracy compared to the results of each previous experiment.

**(table with audio, vision, a+v accuracy)**

## V. CONCLUSION AND FUTURE WORK

### A. Conclusion

In this paper, utilizing both the audio and vision-based features enabled the drone detection system that can be applied to CAS and CUAS. Different from previous related research, the Audio-based Drone-to-Drone detection task was done by using the built-in camera of the drone and an iPhone 6 that has basic performance. In addition, two drones were flying at the same time for drone detection, rather than recording from the ground. The result shows the possibility classifying the target drone sound regardless of the noise of detecting drone, as showing the accuracy of 00 and continuous decrease in loss. From the differentiation, the experiment opened the door to a new field of research. Therefore, this study can be an indicator of comparing the performance of drone-to-drone detection tasks using vision-based, audio-based, and the fusion of both features for future studies.

### B. Limitation

This research has a limitation of the detecting drone only hovering at the same location rather than moving in mid-air. In addition, the microphone of iPhone 6 for audio recording had a lot lower performance than that of lastest iPhones or other professional microphones. On the other hand, this is also a disproof that the drone-to-drone detection task could still show sufficient result only by using the acoustic device with low-performance.

Lastly, other limitations include only one type of target drone being used for the experiment, and the Euclidean distance between the target drone and detecting drone being fixed to 20m.

### C. Future work

In order to overcome these limitations, for future work, drone detection using various types and numbers of drones in different locations needs to be researched. By varying the number of drones to use or changing the state of drone movement, it is important to make the future research being applicable in a real environment or general situation.

Furthermore, this experiment using audio and image is not a real-time drone detection. So, since it is not a complete real-time method connected in one end-to-end process, it is necessary to develop a real-time drone detection system through further research.

Indeed, the range of availability of the drone-to-drone task needs to be specified in future work. By expanding the experiments based on this paper, a wider range of Euclidean distance, rather than 20m, should be conducted for collecting data of target drones. Eventually, finding the maximum Euclidean distance range available for drone detection can be researched.

## ACKNOWLEDGMENT

## REFERENCES


[1] Grand View Research, "Commercial Drone Market Size, Share Trends Analysis Report by Product, by Application, by End-use, by Region, and Segment Forecasts, 2021-2028", Apr. 26, 2021. [Online]. Available: https://www.grandviewresearch.com/Filters?search=Commercial+Drone+Marketsearch_submit=+

[2] M. Meisenzahl, "UPS will start delivering CVS prescriptions by drone in Florida — here's how it works." *INSIDER*, Apr. 29, 2020, [Online]. Available: https://www.businessinsider.com/ups-deliver-cvs-prescription-medications-by-drone-the-villages-2020-4.

[3] NH Motlagh et al., "UAV-based IoT platform: A crowd surveillance use case." *IEEE Communications Magazine*, vol. 55, no.2, 2017

[4] B. Stevenson, "A320 collision heightens UAV safety concerns. FlightGlobal", Apr. 21, 2016. [Online]. Available https://www.flightglobal.com/civil-uavs/a320-collision-heightens-uav-safety-concerns/120348.article.

[5] Science and Technology. "Counter-Unmanned Aircraft Systems (C-UAS)", Jul.22, 2022. [Online]. Avaliable: https://www.dhs.gov/science-and-technology/counter-unmanned-aircraft-systems-c-uas.

[6] X. GUAN et al., "A survey of safety separation management and collision avoidance approaches of civil UAS operating in integration national airspace system." *Chinese Journal of Aeronautics*, Apr.27, 2020

[7] Yang, Bowon, et al. "UAV detection system with multiple acoustic nodes using machine learning models." *2019 Third IEEE International Conference on Robotic Computing (IRC)*, 2019.

[8] H. Abdullah, "Man Detained for Flying Drone Near White House". NEWS, May. 15, 2015. [Online]. Available: https://www.nbcnews.com/news/us-20news/20man-20detained-20trying-20fly-20drone-20near-20white-20house-20n359011

[9] A. R. Wagoner, D. K. Schrader and E. T. Matson, "Towards a vision-based targeting system for counter unmanned aerial systems (CUAS)," *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 2017.

[10] J. N. Yasin, S. A. S. Mohamed, M. Haghbayan, J. Heikkonen, H. Tenhunen and J. Plosila, "Unmanned Aerial Vehicles (UAVs): Collision Avoidance Systems and Approaches," *IEEE Access* Access, vol. 8, pp. 105139-105155, 2020

[11] S. U. Sharma and D. J. Shah, "A Practical Animal Detection and Collision Avoidance System Using Computer Vision Technique," *IEEE Access*, vol. 5, pp. 347-358, 2017

[12] Liu, Hao, et al. "Drone detection based on an audio-assisted camera array." *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, 2017.

[13] B. Taha and A. Shoufan, "Machine Learning-Based Drone Detection and Classification: State-of-the-Art in Research," *IEEE Access*, vol. 7, pp. 138669-138682, 2019.

[14] Craye, Celine, and Salem Ardjoune. "Spatio-temporal semantic segmentation for drone detection." *2019 16th IEEE International conference on advanced video and signal based surveillance (AVSS)*, 2019.

[15] U. Seidaliyeva, D. Akhmetov, L. Ilipbayeva, and E. T. Matson, “Real-Time and Accurate Drone Detection in a Video with a Static Background,” *Sensors*, vol. 20, no. 14, p. 3856, Jul. 2020, doi: 10.3390/s20143856.
[16] S. Park et al., ”Combination of radar and audio sensors for identification of rotor-type Unmanned Aerial Vehicles (UAVs),” *2015 IEEE SENSORS*, 2015, pp. 1-4, doi: 10.1109/ICSENS.2015.7370533.
[17] Christnacher, Frank, et al., ”Optical and acoustical UAV detection.” *Electro-Optical Remote Sensing X*. vol. 9988, SPIE, 2016.
[18] Zhang, Xupei, et al., ”VIAE-Net: An End-to-End Altitude Estimation through Monocular Vision and Inertial Feature Fusion Neural Networks for UAV Autonomous Landing.” *Sensors*, 2021
[19] hy Audio Based Drone Detection and Identification using Deep Learning
[20] hy Acoustic-Based UAV Detection Using Late Fusion of Deep Neural Networks
[21] hy Drone Detection Based on an Audio-Assisted Camera Array
[22] hy Quad-rotor UAV Audio Recognition Based on Mel Spectrum with Binaural Representation and CNN
[23] hy Acoustic Features for Environmental Sound Analysis
[24] hy Single Node Detection on Direction of Approach
[25] Ultralytics, ”yolov5”, github.com https://github.com/ultralytics/yolov5 (accessed Aug. 1, 2022)
[26] Wang, Chien-Yao, et al., ”CSPNet: A new backbone that can enhance learning capability of CNN.” *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020.
[27] T. Wu, T. Wang and Y. Liu, ”Real-Time Vehicle and Distance Detection Based on Improved Yolo v5 Network,” *2021 3rd World Symposium on Artificial Intelligence (WSAI)*, 2021, pp. 24-28, doi: 10.1109/WSAI51899.2021.9486316.