

How Far Can a Drone be Detected?

A Drone-to-Drone Detection System Using Sensor Fusion

Juann Kim¹^a, Youngseo Kim²^b, Heeyeon Shin³^c, Yaqin Wang⁴^d and Eric T. Matson⁴

¹Dept. Software, Sangmyung University, Cheonan, Republic of Korea

²Dept. Human Centered AI, Sangmyung University, Seoul, Republic of Korea

³Computer Science and Engineering, Kyung Hee University, Yongin, Republic of Korea

⁴Computer and Information Technology, Purdue University, West Lafayette, U.S.A.

Keywords: Drone Detection, Audio Classification, Object Detection, Machine Learning, Deep Learning, Decision Fusion, Sensor Fusion.

Abstract: The recent successes of drone detection models show that leveraging the decision fusion of audio-based and vision-based features can achieve high accuracy, instead of only using unitary features. In this paper, we propose to estimate how far can a drone be detected in different distances. Drone-to-drone dataset for were collected separately using a camera and a microphone. The data are evaluated using deep learning and machine learning techniques to show how far can a drone be detected. Two different types of sensors were used for collecting acoustic-based features and vision-based features. Convolutional Neural Network (CNN) and Support Vector Machine (SVM) are utilized with audio-based features, which are Mel-Frequency Cepstral Coefficients (MFCC) and Log-Mel Spectrogram. YOLOV5 is adopted for visual feature extraction and drone detection. Ultimately, by using the sensor fusion of both domains of audio and computer vision, our proposed model achieves high performances in different distances.


1 INTRODUCTION


The application of Unmanned Aerial Vehicles (UAVs), or drones, is increasing rapidly in diverse fields including agriculture, construction, technical service, health care, and delivery systems. The benefits of drones are enormous: operating without a pilot, applying diverse fields, low-cost infrastructure, and etc. Especially, Countering Unmanned Aerial System (CUAS) is required to detect and track malicious drones that approach protected or secure areas. Drone flights in the Air Exclusion Zone have repeatedly occurred. For instance, a man was detained since he flew his UAV 100 feet above near the White House in 2015. (H. Abdullah, 2015) Due to this, the importance of drone detecting and further drone localization comes to the fore. Various domains, including Radar and Lidar, many types of camera and microphone were applied to drone detection and localization.


In this paper, the low-cost sensors, camera and


microphone, are used for drone-to-drone detection. By the sensor fusion, two different sensors can compensate each other. The experiment is conducted in the scenario that two drones are facing each other in the air. The distance between the target drone and the moving plane of the detecting drone are set from 20m to 60m to experiment with how far can the target drone be detected. The collected dataset is used in developing Machine Learning and Deep Learning models to detect a drone using various sensors. This paper focuses on UAV detection by certain range using audio-based and vision-based approaches.

In previous research, various feature extraction methods were proposed such as MFCC, Log Mel-spectrogram, Short Time Fourier Transform (STFT), (S. Al-Emadi, A. Al-Ali, A. Mohammad, and A. Al-Ali, 2019), (Y. Wang, F. E. Fagian, K. E. Ho, and E. T. Matson, 2021). From these extraction methods, various studies have succeeded in detecting drones using MFCC, (S. Jeon, J. -W. Shin, Y. -J. Lee, W. -H. Kim, Y. Kwon, and H. -Y. Yang, 2017). In this work, SVM and a Convolutional Neural Network are used for drone detection with audio data. Log Mel-spectrogram and MFCC are evaluated for feature extraction. For Computer Vision, a state-of-the-art structure, you only look once - YOLOV5 is applied

^a <https://orcid.org/0000-0002-0923-0115>

^b <https://orcid.org/0000-0002-9019-2135>

^c <https://orcid.org/0000-0002-3423-3780>

^d <https://orcid.org/0000-0003-2954-0855>

to detect the drone. After comparing different models and feature extraction methods, we choose the CNN model and MFCC feature for the sensor fusion. In our proposed fusion system, the YOLOV5 model first detects the detecting drone using visual data; then, the falsely detected data are reclassified with a pre-trained CNN model based on audio data. Overall, the main contributions of this work can be summarized as follows:

- We gather the drone-to-drone audio and video data that is collected at a distance of 20 to 60 meters manually.
- We propose a novel drone detection scheme that reduces the error rate using the proposed sensor fusion.

The rest of the work contains five sections. Section 2 reviews several related works to organize the problems. In Section 3, our methodology is introduced, which includes data collection and data processing. In Section 4, the experiments of proposed system are conducted to evaluate the optimal performances for each domain and also for the sensor fusion of both domains. Lastly, Section 5 suggests the conclusions and future works.

2 RELATED WORK

2.1 Radio Frequency and Radar-Based Approach

Currently, various methods have been used for drone detection and drone localization. In (Choi B, Oh D, Kim S, Chong J-W, Li Y-C., 2018), distance estimation or drone localization was done in two ways. Firstly, the implemented FMCW radar system result with only one drone showed the maximum distance between the drone and the radar system was greater than about 1005 to 1010 m. Meanwhile, when the two drones were flying at the same time, one frame of the detection results in a range of around 339 m. The distance of drone-to-drone detection using radar is shorter than using only one drone flying. However, radar-based detection is not optimized for plastic material drone detection and small drone at widely varying ranges (Liu, Hao, et al., 2017). Also, radar is a high-cost sensor.

2.2 Audio-Based Deep Learning Approach

A radar system has a small cross-section, and radio frequency (RF) based systems do not operate well

when GPS communication signals are small; therefore their performances are limited. However, the microphone array overcomes the shortcomings of the sensors and shows excellent performance in drone localization and drone tracking. In (Christnacher, F., Hengy, S., Laurenzis, M., Matwyschuk, A., Naz, P., Schertzer, S., & Schmitt, G., 2016), four microphone sensors were used to predict the direction of drone arrival (DOA), and localization is performed by obtaining azimuth and elevation angles by a multi-signal classification algorithm (MUSIC). In fact, it showed a very low performance. Meanwhile, in (Sedunov, A., Sutin, A., Sedunov, N., Salloum, H., Yakubovskiy, A., & Masters, D., 2016), (H. Salloum, A. Sedunov, N. Sedunov, A. Sutin and D. Masters, 2015), Acoustic Aircraft Detection (AAD) systems were developed and built. This system can detect and track small airplanes and helicopters, whereas it does not consider a situation with multiple noises.

2.3 Vision-Based Deep Learning Approach

Research on drone detection systems using Computer Vision is one of the traditional methods that is widely used in the past. Furthermore, research on drone detection systems using computer vision-based technology has been shown to be sufficiently accurate and commercially available by experiments conducted by (Deng, S., Li, S., Xie, K., Song, W., Liao, X., Hao, A., & Qin, H, 2020). Among CNN-based deep learning models, YOLO, a one-stage model, is easy to detect drone objects in real-time and can respond in a very short time. Thus, it can be applied to systems such as CUAS.

2.4 Depth Estimation and Distance Prediction

Depth estimation and distance prediction have made enormous progress in recent years and achieved significant results with the advance of deep learning (Al-malioglu, Yasin, et al., 2019), (Wu, Zhenyao, et al., 2019), (Feng, Tuo, and Dongbing Gu., 2019). In the early stage, (Aswini, N., S. V. Uma, and V. Akhilesh., 2022) applied object detection model-YOLOV3 and mathematical principles for obstacle distance estimation. They established the maximum distance to the obstacle is a 30m. Beside, we set the distance to the drone to a maximum of 60m. On the other hand, (Yip, Daniel A., et al., 2020) designed a sound level measurements from audio recordings that provides objective distance estimation. However, our paper utilizes

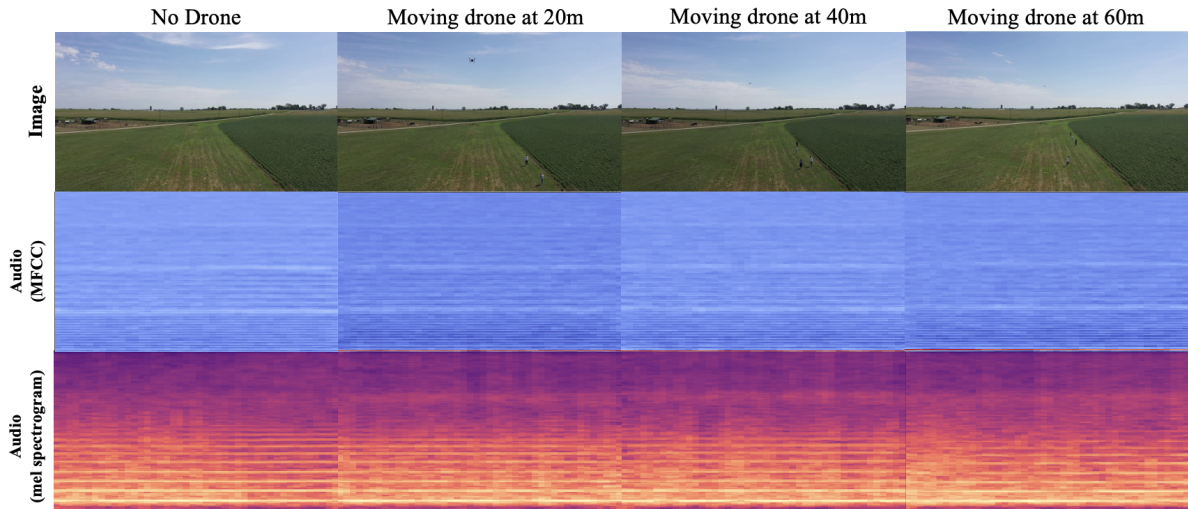


Figure 1: **Visual of data samples.** We use the manually collected audio and image input data. From the first left column, there are no drone data, drone data at a 20m distance, drone data at 40m, and lastly the drone data at 60m. Then, from the first top row, there are image samples, mel spectrogram feature map from audio data and MFCC feature maps from audio data.

both sound and visual source for drone distance prediction.

3 METHODOLOGY

3.1 Data Collection

The data collection method of this paper is the similar as that of previous work (Kim, J, Lee, D, Kim, Y, Shin, H, Heo, Y, Wang, Y, & Matson, E. T, 2022). We use DJI Matrice200 as the target drone and DJI Mavic2 Pro as the detecting drone to collect data in this research. Similar to (Alaparthi, V., Mandal, S., & Cummings, M., 2021), the negative dataset, i.e., no drone data, is also collected for the drone detection, which includes environment noises while no drone is flying in the air such as wind or bird sound. Various previous studies (Liu, Hao, et al., 2017), (Hu, Yuanyuan, et al., 2019), (Al-Emadi, Sara, et al., 2019) have implemented with the camera and microphone placing on the ground to collect data containing the target drone. In contrast, our dataset was collected from the camera and the microphone of detecting drone while two drones were flying in the air at the same time, facing each other.

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 distance with plane of detecting drones moving horizontally and vertically by 20m, 40m, and 60m. With the fixed distance, the tar-

get drone was moving randomly at the camera range.

The weather condition varies in the days of data collection. The different weather conditions include windy, sunny, and foggy days with different humidity levels and wind speed. So, the background images and noise are included in the data, while other environmental factors being the same that of the drone data.

Audio data consists four classes in .wav format. The data are collected in the environment with other various noises such as wind, bird, cow, insect, traffic, airplane, etc.

When collecting vision data, the raw mp4 video files are split into images per 30 frames. Each image has 640 x 640 resolution, and the image format is jpg.

For each domain of audio and image data, 1029 data samples are collected for each class. Thus, total 4116 data samples are collected for each domain, as shown in Table 1.

3.2 Audio Data Augmentation

The raw audio data is split into one second which can sufficiently represent acoustic-based features in training and testing (S. Seo, S. Yeo, H. Han, Y. Ko, K. E. Ho, and E. T. Matson, 2020), (Casabianca, Pietro, and Yu Zhang, 2021), (S. Al-Emadi, A. Al-Ali, A. Mohammad, and A. Al-Ali, 2019).

Table 1: The Number of Dataset.

Split \ Class	No drone		20m		40m		60m	
	Audio	Image	Audio	Image	Audio	Image	Audio	Image
Train	720	720	719	719	719	719	719	719
Validation	204	204	205	205	205	205	205	205
Test	105	105	105	105	105	105	105	105

Before conducting feature extraction, pitch shifting is used for audio data augmentation in order to improve performance in generalization. Pitch shifting is a methodology to raise or lower the pitch of the audio data without affecting the speed of the sound. In (J. Salamon and J. P. Bello, 2017), pitch shifting augmentation shows the best positive impact on performance and is the only method that does not have negative impacts on any types of environmental sound classification. Therefore, the total number of data samples doubled as 8232 from the original dataset.

3.3 Audio Feature Extraction

The audio features are extracted using two feature extraction methods: MFCC and Log Mel-Spectrogram. Also, MFCC provides useful features to capture periodicity from the fundamental frequencies brought on drone's rotor blades (Jeon, S., Shin, J. W., Lee, Y. J., Kim, W. H., Kwon, Y., & Yang, H. Y., 2017). Meanwhile, the Log Mel-Spectrogram has a low false alarm rate but a weak drone detection ability. However, the MFCC has a strong drone detection ability while having a high False Alarm Rate compared to Log Mel-Spectrogram (Dong, Qiushi, Yu Liu, and Xiaolin Liu, 2022). For the hyper-parameter, the number of mels is used as 128 which is the default value, and the number of MFCC is also unified as 128. The examples of extracted feature map of four classes are shown in Figure 1.

3.4 Vision Data Processing

To the purpose, train the model for drone detection, all of the ground truth objects in the picture require to be labeled first. This dataset is labeled using the "LabelImg" (heartexlabs, 2014). The coordination of the bounding box including the location information of drones is generated as text files.

4 EXPERIMENT

4.1 Overview

In this paper, the low-cost sensor fusion system for detecting the target drone by three intervals is proposed. The camera and the microphone used for this system are attached to the Drone (A. Patle and D. S. Chouhan, 2013). Generally, drone detection results using visual-based features show a high performance (Madasamy, K., Shanmuganathan, V., Kandasamy, V., Lee, M. Y., & Thangadurai, M., 2021). However,

the camera cannot perform its role properly in situations where vision is obstructed. In the dataset we collected, weather conditions are the main factors for the obstruction including cloudy and foggy conditions. This can be compensated by using additional sensor for drone detection, which is audio-based features. Therefore, drone detection is done based on vision-based features using the YOLOV5 model. Then, the falsely detected vision data is reclassified using audio data with a CNN model. The falsely detected data is specifically the ones that are classified as False Negative (FN) and False Positive (FP). The proposed system is described in Figure 2.

The sound of the detecting drone with the microphone attached is considered the background noise when detecting another drone. Although two drone sounds are simultaneously recorded, drone detection is successfully presented in this paper. Practically, in (Kim, J, Lee, D, Kim, Y, Shin, H, Heo, Y, Wang, Y, & Matson, E. T, 2022), while the microphone is attached to the detecting drone, another drone is detected through an audio signal up to 20m with an accuracy of 88.96%.

4.2 Audio Classification

4.2.1 Background

Machine Learning and Deep Learning approaches are well-known for achieving high performances for the drone detection system using audio data. In (Seo, Y., Jang, B., & Im, S., 2018), comparing to SVM, CNN showed a decrease in false positives and an increase in the correct detection rate. Similarly, (Seo, Y., Jang, B., & Im, S., 2018) also obtained the result of the Deep learning model which shows a higher performance than that of SVM, with 8.31% improvement. In this experiment (Seo, Y., Jang, B., & Im, S., 2018), the performance of SVM and CNN are both evaluated with two different features, which are Mel-Spectrogram and Mel Frequency Cepstral Coefficient. Various kernels of SVM are applied to classify the features.

SVM acquires an optimal hyperplane containing positive and negative samples with the principle of structural risk minimization (Winters-Hilt, S., Yelundur, A., McChesney, C., & Landry, M., 2006). Meanwhile, the CNN model also demonstrates high performance for the two-dimensional features in many applications such as audio-based features (Seo, Y., Jang, B., & Im, S., 2018). CNN model is composed of multiple layered neural networks including a convolution layer, a pooling layer, an active layer, and a full connection layer.

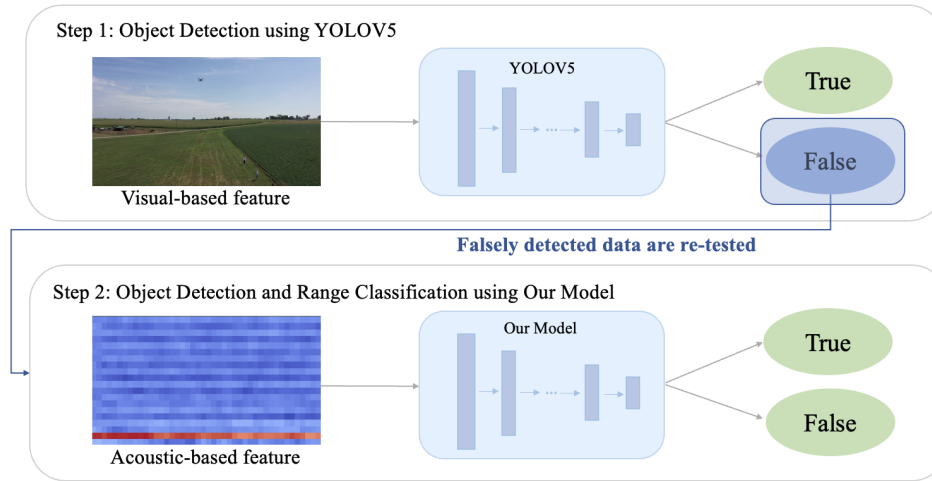


Figure 2: Overview of the drone detection system.

4.2.2 Machine Learning Training

The input dimension of the Support Vector Machine should be 1-dimension (1 x N) for each data sample. The Principal Component Analysis (PCA) is applied for dimension reduction from (2 x N) to (1 x N). Also, the N, the hyper-parameter of PCA named n_components, is set as 128.

Grid search refers to the process of training a certain model with all possible combinations of different hyper-parameters within the range specified by the user and eventually obtaining the optimal hyper-parameter that shows the highest performance.

Three different kernels, Gaussian Radial Basis Function (RBF), sigmoid, and polynomial kernels, are used. One of the most commonly used kernel functions is the radial basis function. Each data point has a "bump" added to it.

$$K(x, x_i) = e^{-\gamma \|x - x_i\|^2} \quad (1)$$

Here γ , r and d are kernel parameters.

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \quad (2)$$

The polynomial kernel function is directional. In other words, the direction of the two vectors in low-dimensional space determines the output. This is due to the dot product in the kernel. The magnitude of the vector x influences both the output and the vector's magnitude.

$$K(x, x_i) = (1 + x \cdot x_i^T)^d \quad (3)$$

d is the degree of kernel function.

In the kernel functions, a gamma hyper-parameter defines how far the influence of a single training point

reaches and the range set as $1e-3$ to $1e-6$ multiplied by 0.1 intervals. Also, C is another hyper-parameter that controls the trade-off between smooth decision boundary and classifying training points correctly and the range set from $1e-3$ to $1e3$ (A. Patle and D. S. Chouhan, 2013). For the model's stability, 5-fold cross validation is applied.

4.2.3 Deep Learning Training

Early Stopping technique is used for training in order to prevent overfitting. The monitor and the patient hyper-parameters are set as a validation loss and 15 respectively. Softmax is used for the last activation layer as an activation function. The two different optimizers, Stochastic Gradient Descent (SGD) and Adam, are used for evaluation. Also, 5-fold cross-validation is applied for obtaining a more accurate estimate of model prediction performance.

4.2.4 Result and Analysis

As a result of using 5-fold cross validation procedure with MFCC feature extraction for the SVM model, the accuracy is obtained as 65.5%. Among different combinations of hyper-parameters, the highest performance is shown when the kernel function is set to RBF, $1e-5$ for gamma, and 10 for C . On the other hand, the result of the 5-fold cross validation of SVM using Log Mel features is 34.5%. Thus, it can be seen that better results are obtained when MFCC is used for the feature extraction for the SVM model.

Overall, from Table 2 to Table 5, CNN based on MFCC features using Stochastic Gradient Descent as an optimizer shows the highest and the most stable performance in multi-class classification. The model can detect up to 40 meters with more than 70% accu-

racy.

In comparison of two models, the CNN model shows better performances than the SVM model for drone detection based on the audio-based features. Therefore, the CNN model based on MFCC features is employed for the second step of the proposed drone detection system as shown in Figure 2.

Table 2: CNN-MFCC Adam.

Class	Precision	Recall	F1	Accuracy
no drone	74.0%	66.0%	69.7%	66.0%
20m	75.0%	75.0%	74.7%	74.9%
40m	54.3%	57.3%	56.0%	57.5%
60m	59.0%	61.3%	60.3%	62.2%

Table 3: CNN-Mel Adam.

Class	Precision	Recall	F1	Accuracy
no drone	75.7%	70.7%	73.0%	70.8%
20m	74.5%	68.0%	70.7%	67.9%
40m	55.3%	54.7%	54.7%	54.6%
60m	61.0%	69.0%	64.0%	69.8%

Table 4: CNN-MFCC SGD.

Class	Precision	Recall	F1	Accuracy
no drone	75.3%	75.0%	75.0%	75.2%
20m	78.0%	78.3%	78.0%	78.4%
40m	60.7%	70.7%	65.0%	70.5%
60m	73.3%	60.3%	66.0%	61.0%

Table 5: CNN-Mel SGD.

Class	Precision	Recall	F1	Accuracy
no drone	48.3%	45.7%	44.4%	45.4%
20m	60.3%	50.3%	54.0%	50.1%
40m	42.3%	63.0%	50.7%	62.9%
60m	45.7%	33.0%	38.3%	33.3%

4.3 Vision Object Detection

To detect the drone in images, 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 (Ultralytics, "YOLOV5"), which introduces a type of powerful object detecting model. The backbone network obtains feature maps of different sizes from input images via the pooling layer and convolution layer. The total structure is shown in Figure 3.

First, CBL is a block that is fundamentally used to extract features containing of leaky ReLU, batch normalization and the convolution layer. SPP enhances performance by pooling different 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 depress the massive inference computations caused by duplicate gradient information. Then, they are combined again in the cross-stage hierarchy method proposed in the paper (Wang, Chien-Yao, et al.,). This way, the spread out gradient information can have a huge correlation difference by transition the transformation and concatenation steps. Furthermore, CSP can considerably impair computational effort and improve inference cost and accuracy.

5 Backbone networks - YOLOv5-n,s,m,l,x are used. Each model is distinguished by depth multiple and width multiple, and can be organized. The larger the depth multiple value, the more BottleneckCSP() is repeated to become a deeper model. Moreover, the larger the width multiple, the higher Convolution filter number of the corresponding layer.

The training is performed through the SGD optimizer with a momentum of 0.937 and weight decay of $1e-5$. Also, for other hyper-parameters, the model initializes the learning rate as 0.01 and the batch size of 16. The iteration is set to 30. Our model architecture has 270 layers and 17K parameters. The evaluation performances for drone detection tasks are measured by precision, recall, and accuracy.

4.4 Sensor Fusion

As previously mentioned in Figure 2, the fusion method includes two steps. The first step is drone detection using the YOLOV5 model based on vision-based features. Then, from the first step, the falsely detected data by the YOLOV5 model is re-classified by the pre-trained CNN model as shown in the second step. This proposed system including two steps shows the highest performance among three methodologies: using only audio-based features, only vision-based features, and the decision fusion of both features. As shown in Figure 4, for the distance of 40m, the accuracy of the fusion method reached 88% which is about 10% to 20% higher than the accuracies of the two individual methods of using only one sensor. Furthermore, for the drone detection in different distances, it can be clearly noticed that the performance decreases as the distance from the detecting drone increases.

5 CONCLUSION AND FUTURE WORKS

The proposed system combines a camera and a microphone to perform drone detection and distance interval estimation. First, YOLOV5 is trained with im-

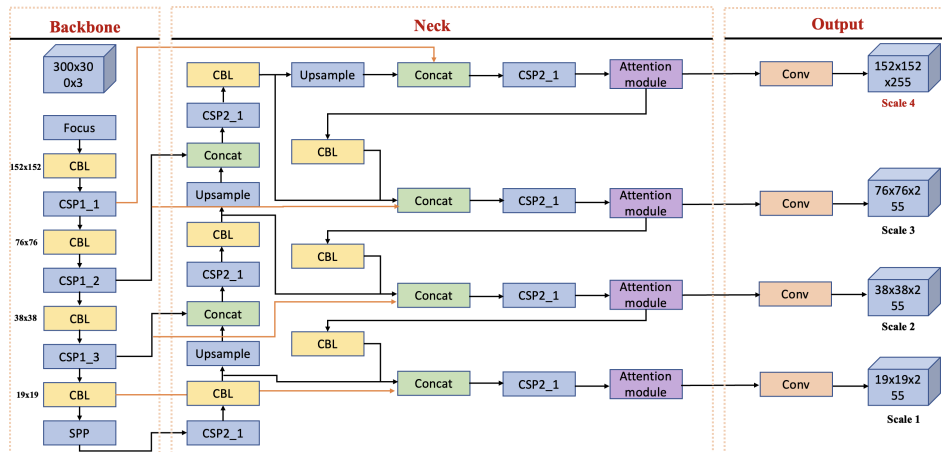


Figure 3: YOLOV5 Architecture.

age data with different ranges of distances. From the tested result, certain data which are classified as False Negative (FN) and False Positive (FP) are again re-classified CNN models using MFCC features that are pre-trained with audio data.

Although the microphone is attached directly to the detecting drone, the model is able to detect the target drone flying in 20m with the accuracy of 78%.

When the distance becomes far off from 20m to 40m, the performance of our proposed system is 10% higher than when using only vision and 17% higher than when using only audio. Even if the distance of the plane of the target drone is 60m away from the detecting drone, it is possible to detect the drone with 80% high performance as shown in Figure 4.

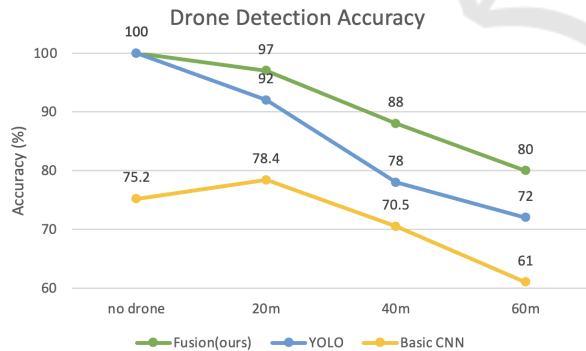


Figure 4: The detection accuracy decline as the distance between the detecting drone and the moving plane of target drone increases.

Figure 4 shows that the detection performance decreases as the distance between the detecting drone and the moving plane of target drone increases. This research has a limitation of only one type of target drone being used. In the future work, various types of drones will be used. Also, we plan to apply other deep

learning methods, such as LSTM and RCNN, will be used to compare the performances and find the best model for drone-to-drone detection using audio and computer vision sensor.

REFERENCES

- AA. Patle and D. S. Chouhan, "SVM kernel functions for classification," 2013 International Conference on Advances in Technology and Engineering (ICATE), 2013, pp. 1-9, doi: 10.1109/ICATE.2013.6524743
- Al-Emadi, S., Al-Ali, A., & Al-Ali, A. (2021). Audio-based drone detection and identification using deep learning techniques with dataset enhancement through generative adversarial networks. *Sensors*, 21(15), 4953.
- Alaparthi, V., Mandal, S., & Cummings, M. (2021, March). Machine Learning vs. Human Performance in the Real-time Acoustic Detection of Drones. In 2021 IEEE Aerospace Conference (50100) (pp. 1-7). IEEE.
- Almalioglu, Y., Saputra, M. R. U., De Gusmao, P. P., Markham, A., & Trigoni, N. (2019, May). GANVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In 2019 International conference on robotics and automation (ICRA) (pp. 5474-5480). IEEE.
- Aswini, N., S. V. Uma, and V. Akhilesh. "Drone to Obstacle Distance Estimation Using YOLO V3 Network and Mathematical Principles." *Journal of Physics: Conference Series*. Vol. 2161. No. 1. IOP Publishing, 2022.
- Casabianca, Pietro, and Yu Zhang. "Acoustic-based UAV detection using late fusion of deep neural networks." *Drones* 5.3 (2021): 54.
- Choi B, Oh D, Kim S, Chong J-W, Li Y-C. Long-Range Drone Detection of 24 G FMCW Radar with E-plane Sectoral Horn Array. *Sensors*. 2018; 18(12):4171. <https://doi.org/10.3390/s18124171>
- Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., & Dahl, G. E. (2019). On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*.

- Christnacher, F., Hengy, S., Laurenzis, M., Matwyschuk, A., Naz, P., Schertzer, S., & Schmitt, G. (2016, October). Optical and acoustical UAV detection. In *Electro-Optical Remote Sensing X* (Vol. 9988, pp. 83-95). SPIE.
- Deng, S., Li, S., Xie, K., Song, W., Liao, X., Hao, A., & Qin, H. (2020). A global-local self-adaptive network for drone-view object detection. *IEEE Transactions on Image Processing*, 30, 1556-1569.
- Dong, Qiushi, Yu Liu, and Xiaolin Liu. "Drone sound detection system based on feature result-level fusion using deep learning." *Multimedia Tools and Applications* (2022): 1-23.
- Feng, Tuo, and Dongbing Gu. "SGANVO: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks." *IEEE Robotics and Automation Letters* 4.4 (2019): 4431-4437.
- 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>
- H. Salloum, A. Sedunov, N. Sedunov, A. Sutin and D. Masters, J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," in *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279-283, March 2017, doi: 10.1109/LSP.2017.2657381.
- H. Liu, Z. Wei, Y. Chen, J. Pan, L. Lin and Y. Ren, "Drone Detection Based on an Audio-Assisted Camera Array," 2017 IEEE Third International Conference on Multimedia Big Data (BigMM), 2017, pp. 402-406, doi: 10.1109/BigMM.2017.57.
- Hu, Y., Wu, X., Zheng, G., & Liu, X. (2019, July). Object detection of UAV for anti-UAV based on improved YOLO v3. In *2019 Chinese Control Conference (CCC)* (pp. 8386-8390). IEEE.
- Heartexlabs, "labelImg", [github.com](https://github.com/Heartexlabs/labelImg) (2014)
- Jeon, S., Shin, J. W., Lee, Y. J., Kim, W. H., Kwon, Y., & Yang, H. Y. (2017, August). Empirical study of drone sound detection in real-life environment with deep neural networks. In *2017 25th European Signal Processing Conference (EUSIPCO)* (pp. 1858-1862). IEEE.
- Kim, J., Lee, D., Kim, Y., Shin, H., Heo, Y., Wang, Y., & Matson, E. T. (2022). Deep Learning Based Malicious Drone Detection Using Acoustic and Image Data (No. 9335). *EasyChair*.
- Madasamy, K., Shanmuganathan, V., Kandasamy, V., Lee, M. Y., & Thangadurai, M. (2021). OSDDY: embedded system-based object surveillance detection system with small drone using deep YOLO. *EURASIP Journal on Image and Video Processing*, 2021(1), 1-14.
- S. Al-Emadi, A. Al-Ali, A. Mohammad and A. Al-Ali, "Audio Based Drone Detection and Identification using Deep Learning," 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), 2019, pp. 459-464, doi: 10.1109/IWCMC.2019.8766732.
- Sedunov, A., Sutin, A., Sedunov, N., Salloum, H., Yakubovskiy, A., & Masters, D. (2016). Passive acoustic system for tracking low-flying aircraft. *IET Radar, Sonar & Navigation*, 10(9), 1561-1568.
- S. Jeon, J. -W. Shin, Y. -J. Lee, W. -H. Kim, Y. Kwon and H. -Y. Yang, "Empirical study of drone sound detection in real-life environment with deep neural networks," 2017 25th European Signal Processing Conference (EUSIPCO), 2017, pp. 1858-1862, doi: 10.23919/EUSIPCO.2017.8081531.
- S. Seo, S. Yeo, H. Han, Y. Ko, K. E. Ho and E. T. Matson, "Single Node Detection on Direction of Approach," 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), 2020, pp. 1-6, doi: 10.1109/I2MTC43012.2020.9129016.
- Seo, Y., Jang, B., & Im, S. (2018, November). Drone detection using convolutional neural networks with acoustic STFT features. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-6). IEEE.
- Ultralytics, "YOLOV5", [github.com https://github.com/ultralytics/YOLOV5](https://github.com/ultralytics/YOLOV5)
- Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 390-391).
- Winters-Hilt, S., Yelundur, A., McChesney, C., & Landry, M. (2006, September). Support vector machine implementations for classification & clustering. In *BMC bioinformatics* (Vol. 7, No. 2, pp. 1-18). BioMed Central.
- Wu, Z., Wu, X., Zhang, X., Wang, S., & Ju, L. (2019). Spatial correspondence with generative adversarial network: Learning depth from monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7494-7504).
- Y. Wang, F. E. Fagian, K. E. Ho and E. T. Matson, "A Feature Engineering Focused System for Acoustic UAV Detection," 2021 Fifth IEEE International Conference on Robotic Computing (IRC), 2021, pp. 125-130, doi: 10.1109/IRC52146.2021.00031.
- Yip, D. A., Knight, E. C., Haave-Audet, E., Wilson, S. J., Charchuk, C., Scott, C. D., ... & Bayne, E. M. (2020). Sound level measurements from audio recordings provide objective distance estimates for distance sampling wildlife populations. *Remote Sensing in Ecology and Conservation*, 6(3), 301-315.