Drone Detection Using Convolutional Neural Networks with Acoustic STFT Features

Yoojeong Seo, Beomhui Jang and Sungbin Im
Soongsil University
Seoul, Korea
yjseo@soongsil.ac.kr, sbi@ssu.ac.kr

Abstract

As the use of drones has raised in the city, the regulation of malicious usage of drones is an important issue. However, it is difficult to detect a drone due to its miniaturization and modification. In this paper, we investigate the performance of using the convolutional neural networks (CNN) for detecting drones with the acoustic signals received by a microphone. Since the harmonic characteristics of drones are different from those of the objects that produce similar noise including scooters and motorcycles, the two-dimensional feature employed in the study is made of the normalized short time Fourier transform (STFT) magnitude. The performance of the proposed approach is evaluated in terms of detection rate and false alarm rate under various environments. The dataset used in this study consists of the measurements through experiments. The experiments are carried out in the open space with a hovering drone, which is DJI Phantom 3 or Phantom 4. The dataset contains 68,931 frames of drone sound and 41,958 frames of non-drone sound. For the 100-epoch model, the detection rate is 98.97 % and the false alarm rate is 1.28, while the 10-epoch model demonstrates the detection rate of 98.77 % and the false alarm rate of 1.62 %.

1. Introduction

As the use of drones in the city becomes more common, they cause various problems ranging from invasion of privacy to terrorism [2]. To solve these problems, many researchers are working on the anti-drones systems [12]. The anti-drone system is a system that detects and identifies drones, and disrupts and/or neutralizes them if they are identified as having malicious purposes. The initial purpose of the anti-drone system was to counteract the approach of a drone which wants to invade the sovereign airspace or seize military secrets in the battlefield. Therefore, the anti-drone systems developed for the military purpose did not need to consider a congested environment such as downtown. Recently, however, there is a need to develop anti-drone systems operating in the complex environments due to the problems such as leakage of industrial confidentiality and invasion of privacy.

The conventional sensor-based detection has been conducted using radar, vision sensors, acoustic sensors and RF sensor [5]. However, the conventional approach cannot take into consideration the complex environment and the false alarm rate is relatively high. Recently, the detection approach with machine learning suitable for a complex environment is under study. For example, images received from a vision sensor are utilized to detect a drone with a machine learning technique [10]. The performances of Gaussian mixture model (GMM), convolutional neural networks (CNN), and recurrent neural network (RNN) models learned with Mel-frequency cepstrum coefficients (MFCC) and mel-spectrogram of acoustic signals are investigated in terms of F-score [6].

In this study, acoustic sensors are employed for detection. Since they are very sensitive, the detection performance is depending on environments. Machine learning is utilized to compensate for this. The proposed approach relies on short time Fourier transform (STFT) of the received signal to distinguish a drone from motor-based devices with similar harmonic characteristics to a drone in an urban environment. This is based on the NASA study [4] indicating that humans can sufficiently distinguish the acoustic properties of drones even though they have similar harmonic characteristics. The machine learning model used in the study is a CNN model, which demonstrates high performance for the two-dimensional features in many applications. In the sonar signal processing, the received signal can be considered as images such as acoustic color and coherence-based synthetic aperture sonar (SAS). From this viewpoint, the CNN model can be applied to the acoustic signal processing like the image processing.

This paper is organized as follows. In the following section, the machine learning model and the feature extraction

978-1-5386-9294-3/18/$31.00 ©2018 IEEE

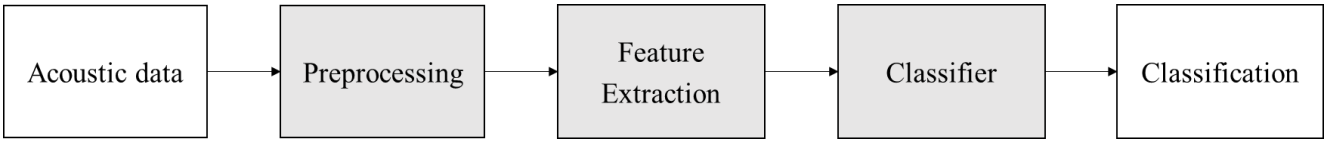
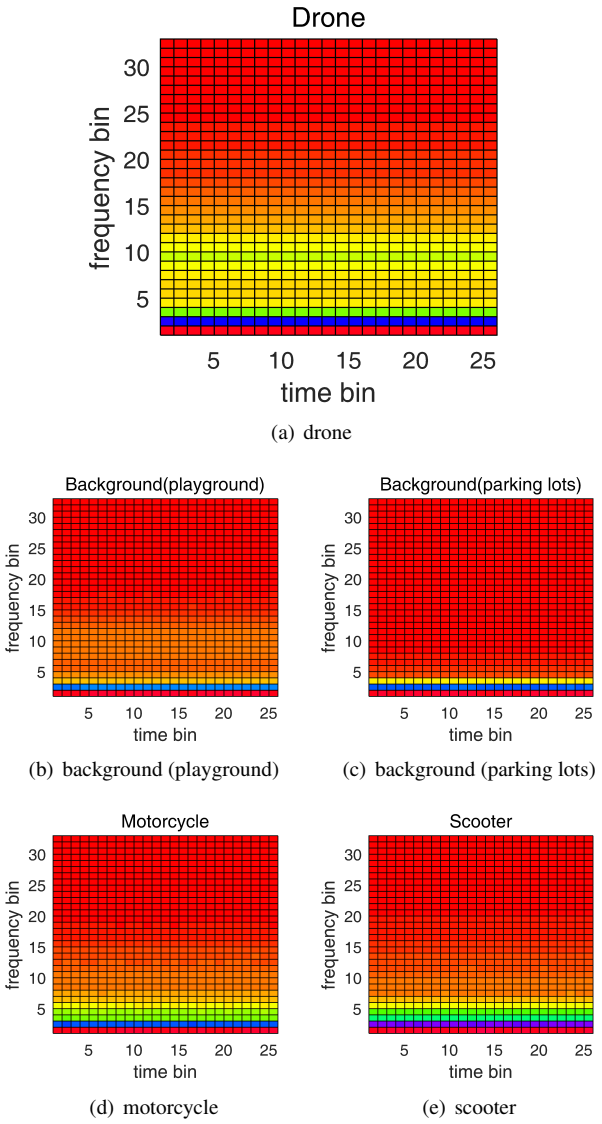


Figure 1. Flowchart of the proposed approach



(a) drone

(b) background (playground)

(c) background (parking lots)

(d) motorcycle

(e) scooter

Figure 2. Averages of features for the drone, background sounds at playground, parking lot, motorcycle and scooter

method are presented and the learning scheme employed in this study is described. In Section 3, the experiment setup and the evaluation results of the proposed model are presented. Finally, concluding remarks and future research direction are given in Section 4.

2. Methods

In this section, the architecture of the CNN model used in this study is introduced. How to extract the STFT-based feature from the acoustic data and the way of training the CNN model that determines the presence of drones from the received acoustic data are presented. Figure 1 shows the flowchart of the proposed approach. It can be divided into three stages: preprocessing on the received acoustic data, feature extraction consisting of magnitude of STFT and classifier based on the CNN model.

2.1. Preprocessing

In the preprocessing stage, the data are segmented into the frames of 20 ms, and each frame is 50 % overlapped with an adjacent frame. The reasoning for selecting the signal duration of 20 ms is based on the discussion presented in [1], which demonstrates that the minimum duration is about 10 to 30 ms so that an acoustic signal has valid information.

2.2. Feature Extraction

Short time Fourier transform (STFT) is an approach to reveal temporal characteristics of the frequency components. That is, by applying STFT to the signal to be analyzed, it is possible to grasp the temporal variation of the frequency characteristics of the signal. The STFT approach is suitable for analyzing instantaneous characteristics since it can reflect information that may be lost when analyzing the frequency characteristics as a whole. The STFT can be expressed as:

$$STFT\{x(k)\}(n,\omega) = \sum_{k=-\infty}^{\infty} x(k)w(k-n)e^{-j\omega k}, \quad (1)$$

where $x(k)$ represents the acoustic signal, while the window function $w(k)$ is the Hanning window. The constant j represents imaginary, and the parameters n and ω denote the time bin number and the frequency bin number, respectively.

The feature used in the study is obtained by normalizing the magnitude of STFT by dividing by its maximum value, which is given by

$$F(n,\omega) = STFT\{x(k)\}(n,\omega) / \max_{l,\kappa} STFT\{x(k)\}(l,\kappa). \quad (2)$$

Thus, the components of the feature $F(n, \omega)$ belong to $[0, 1]$.

Figure 2 shows the average features of the following signals, respectively.

- Drone (Phantom 3, Phantom 4)
- Background sound (playground)
- Background sound (underground parking lot)
- Motorcycle
- Scooter

The averaged features in Fig. 2, show the different characteristics from each other. According to Fig. 2(a), the energy of the drone signal is distributed evenly in the high frequency band as compared with other signals.

2.3. Classifier

Detection using acoustic signals is studied generally in two ways. One is based on the one-dimensional information such as magnitude or phase of the acoustic signal [9]. The other relies on the two-dimensional analysis such as STFT or wavelet. In this study, the CNN is employed as a machine learning method since it is suitable for the feature of 2D format. The CNN is one of major detection approaches for the image processing [8]. This approach, which is composed of multiple layered neural networks and classified as deep learning, can reflect the spatial relationship of the input data, which cannot be exploited in the conventional neural networks. By considering two dimensions, it is possible to improve the identification performance of 2D features because of preserving information, which may be lost when it is transformed into an one-dimensional vector. This is often used in face recognition, music analysis, image classification, etc. [7, 11].

The CNN model considered in the study consists of a convolution layer, a pooling layer, an active layer, and a full connection layer. In the convolution layer, the input data are convolved with the convolution kernel, and a feature extractor is generated by learning various weighted kernels. In the pooling layer, the dimension reduction of the data is carried out by taking prominent features in the previous layer through averaging or selection, thus the computational load of learning can be greatly reduced. The active layer applies an activation function to impart nonlinearity to information received from the previous layer. Representative activation functions include sigmoid and rectified linear unit (ReLU). Finally, the full connection layer is the output layer for returning the responses of the corresponding input to a particular class.

In the model developed in this study, the pooling layer uses the maxpooling method, which returns the maximum value in the $K \times K$ kernel. The ReLU function is used as the active function in the convolution layer while the softmax function is employed in the output layer. The ReLU function and the softmax function are expressed by the following equations, respectively.

Layer	Type	Dimension	Kernel	Connection Perceptage
1	Convolution	$6@22 \times 29$	5X5	156
2	Subsampling	$6@11 \times 14$	2X2	-
3	Convolution	$16@7 \times 10$	5X5	2416
4	Subsampling	$16@3 \times 5$	2X2	-
5	Flatten	$240@1 \times 1$	-	-
6	Fully connected	$20@1 \times 1$	-	4820
7	Fully connected	$2@1 \times 1$	-	42

Table 1. Dimensions for the convolutional neural networks. The convolutional layer dimensions are denoted as ⟨number of feature maps⟩@⟨feature map size⟩ (e.g. $6@5 \times 5$).

$$ReLU(x) = \max(0, x) \tag{3}$$

$$softmax(\mathbf{x})_k = \frac{e^{x_k}}{\sum_i e^{x_i}} \tag{4}$$

Note that x is a scalar in Eq. (3), while $\mathbf{x}$ in Eq. (4) is a vector and x_i represents the i-th component of the vector $\mathbf{x}$. The ReLU function solves the problem of gradient vanishing with a small amount of computation by making all negative values to be zero. The computational load of the ReLU is less than that of the conventional sigmoid function. The softmax function is used mostly in a multi-class environment and returns the value of the output layer stochastically. This makes it possible to identify the class having the greatest probability regardless of the magnitude of the value.

The operations of the proposed approach are as follows.

1. For the input learning acoustic data, 26×33 input feature is convolved with six kernels of the form 5×5.
2. The output of each kernel is maxpooled by a window of the form 2×2.
3. The maxpooled output is convolved with 16 kernels of the form 5×5.
4. The output of each kernel is maxpooled by a window of the form 2×2.
5. The results of the previous layer are transformed into a one-dimensional form to be represented by 240 nodes.
6. In the dense layer, 240 nodes are fully connected to 20 nodes.
7. Finally, 20 nodes are fully connected to output nodes of the same amount as the output class. The result is returned to softmax, the active function.

Type	Time (s)	Description
Drone	690	Two kinds of drones hovering in 5, 10, 15 and 20m
Playground	270	Background noise collection in a playground
Parking lots	50	Background noise collection in an underground parking lot
Motorcycle	50	Two-wheeled motorcycle sound at 5m outside
Scooter	50	Two-wheeled scooter sound at 5m outside

Table 2. Type, time and description of the data collections used in the experiment.

		Prediction class		
		Drone	No drone	Total
True class	Drone	34,169	426	34,595
	No drone	337	20,513	20,850
	Total	34,506	20,939	55,445

Table 3. Confusion matrix for 10 epochs.

		Prediction class		
		Drone	No drone	Total
True class	Drone	34,237	358	34,595
	No drone	267	20,583	20,850
	Total	34,504	20,941	55,445

Table 4. Confusion matrix for 100 epochs.

Table 1 shows the kernel and output dimensions used in each operation. The number of nodes used in the proposed model is 7,434, and all nodes are affected by learning.

3. Experiment and Result

In this section, the experimental environment and the data used in the experiment are described in details and the experiment results are presented.

Table 2 displays types, lengths and descriptions of the collected datasets used in the experiment. The sampling rate is $44,100$ Hz and the data are collected at a 10-second interval. The microphone used in this experiment is Yeti Pro, the product of the Blue microphone, which has a sampling rate range from 22 kHz to 192 kHz. The number of the reconstructed data is 882 samples of $20ms$ frame with 50% overlapping. As a result, the drone sound dataset consists of 68,931 frames while the non-drone sound dataset has 41,958 frames. Hence, the dataset ratio is approximately 1: 0.61. For feature extraction, STFT with 64-FFT is applied to generate a 26×33 feature matrix for each frame. Each feature matrix is normalized with respect to its maximum value.

Epoch	Train data		Test data	
	Accuracy	Loss	Accuracy	Loss
1	0.9216	0.2178	0.9648	0.1237
10	0.9850	0.0543	0.9862	0.0502
100	0.9938	0.0199	0.9887	0.0522

Table 5. Model accuracy and loss for 1,10, 100 epochs, respectively.

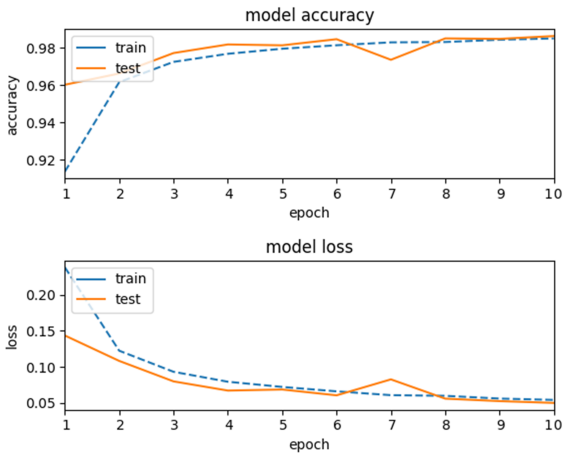


Figure 3. Model accuracy and loss during 10 epochs.

Learning the proposed model is carried out under the condition that the ratio of training to test data is 1:1. For randomly selected data with a batch size of 32, the proposed model is verified under two cases of 10 and 100 epochs. The model is implemented in Python 3.6 with Keras 2.1.5 [3], SciPy 0.19.1 and Tensorflow 1.2.1 on the system with 8-core 3.4-GHz CPU, 16GB RAM, and GTX 960.

Figure 3 shows the accuracy and loss of the proposed model during 10 epochs. The proposed model has an accuracy of 0.9850 and a loss of 0.0543 for the training data, while it has an accuracy of 0.9862 and a loss of 0.0502 for the test data. According to the confusion matrix shown in Table 3, the detection rate is 0.9877 and the false alarm rate is 0.0162.

Figure 4 shows the accuracy and loss of the proposed model during 100 epochs. The proposed model has an accuracy of 0.9938 and a loss of 0.0199 for the training data, while it has an accuracy of 0.9887 and a loss of 0.0522 for the test data. According to the confusion matrix in Table 4, the detection rate is 0.9897 and the false alarm rate is 0.0128. As the epoch repeats, the accuracy increases and the loss decreases. In Table 5, the accuracy and loss of the models with 1, 10, 100 epochs are summarized.

Table 6 shows the performance comparison of the pro-

SNR [dB]	**10-epoch model**				**100-epoch model**			
	Acc.	Loss	P_D	P_{FA}	Acc.	Loss	P_D	P_{FA}
5	0.5886	3.4833	0.3794	0.0642	0.7587	2.9891	0.8540	0.3993
10	0.7718	1.3140	0.7094	0.1247	0.8129	2.2980	0.8595	0.2644
15	0.8630	0.5475	0.8771	0.1603	0.8294	1.9857	0.8617	0.2243
20	0.8998	0.4312	0.9638	0.2062	0.8614	1.3155	0.8786	0.1671

Table 6. The performance comparison of 10- and 100-epoch models for each signal to noise ratio (SNR) [dB]. The model's performance is evaluated in terms of accuracy, loss, detection rate (P_D) and false alarm rate (P_{FA}).

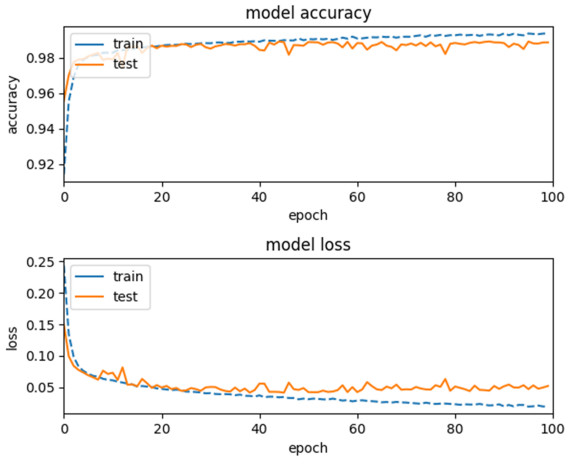


Figure 4. Model accuracy and loss during 100 epochs.

posed models according to signal to noise ratio (SNR). In this evaluation, SNR varies from 5 dB to 20 dB with increment of 5 dB. The models are evaluted using the test dataset corrupted by additive white Gaussian noise (AWGN). 10 and 100 learning epochs are used in the evaluation The criteria for performance evaluation is accuracy, loss, detection rate, and false alarm rate for the test dataset.

For higher SNR's, both models demonstrate better performance in terms of accuracy, detection rate, loss and false alarm rate. Note that the 10 epoch model shows a different trend of false alarm rate about SNR, which implies that the model is not sufficiently trained. For the lower SNR's, a test signal has a strong noise signal, whose spectral feature is different from that of a drone. This enables the model make correct decision to reduce the false alarm rate while for the higher SNR cases, a weak noise signal confuses the model to increase the false alarm rate.

The 100-epoch model also shows excellent performance for low SNR. According to these observations, the more epochs the model learns, the stronger it is about noise. For 15 and 20 dB SNR's, however, the performance of the 100-epoch is slightly inferior to that of the 10-epoch model due to overfitting, which is observed in Fig. 4. Therefore, it is possible to create a model, which is robust to noise and achieves high detection rate, with several epochs between 10 and 100.

4. Conclusion

In this study, a CNN model is proposed to detect drones using STFT characteristics of drone's acoustic signal. In this study, the acoustic signals of motorcycles and scooter are also employed in the experiments, since they have similar harmonic characteristics to drone's signal. Furthermore, the performance of the model is evaluated according to the number of training epochs. According to the experiment results, fairly lower false positive rates are observed for signals having a harmonic characteristic similar to a drone.

Since 50 % overlapped STFT used in the proposed approach is randomly selected, it is necessary to select a more suitable overlap size through further studies. In addition, since the collection environment is limited in this study, it is necessary to try to include various acoustic data sets composed of other classes.

For further study, outdoor electric motors equipments such as electric scooters and electric bicycles will be investigated for expanding the dataset. This research can be extended to develop drone identification and the direction and position estimation of drone using acoustic sensor.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A1B03032101).

References

[1] A. Bernardini, F. Mangiatordi, E. Pallotti, and L. Capodiferro. Drone detection by acoustic signature identification. *Electronic Imaging*, 2017(10):60–64, 2017.

[2] G. C. Birch, J. C. Griffin, and M. K. Erdman. Uas detection classification and neutralization: Market survey 2015. Technical report, Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States), 2015.

[3] F. Chollet et al. Keras. `https://keras.io`, 2015.

[4] A. W. Christian and R. Cabell. Initial investigation into the psychoacoustic properties of small unmanned aerial system noise. In *23rd AIAA/CEAS Aeroacoustics Conference*, page 4051, 2017.

[5] S. R. Ganti and Y. Kim. Implementation of detection and tracking mechanism for small uas. In *Unmanned Aircraft Systems (ICUAS), 2016 International Conference on*, pages 1254–1260. IEEE, 2016.

[6] 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. In *Signal Processing Conference (EUSIPCO), 2017 25th European*, pages 1858–1862. IEEE, 2017.

[7] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113, 1997.

[8] K. O'Shea and R. Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[9] L. R. Rabiner and R. W. Schafer. *Digital processing of speech signals*. Prentice Hall, 1978.

[10] A. Schumann, L. Sommer, J. Klatte, T. Schuchert, and J. Beyerer. Deep cross-domain flying object classification for robust uav detection. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–6. IEEE, 2017.

[11] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[12] R. L. Sturdivant and E. K. Chong. Systems engineering baseline concept of a multispectral drone detection solution for airports. *IEEE Access*, 5:7123–7138, 2017.