https://doi.org/10.2352/ISSN.2470-1173.2017.10.IMAWM-168

Drone detection by acoustic signature identification

Andrea Bernardini, Federica Mangiatordi, Emiliano Pallotti, Licia Capodiferro; Fondazione Ugo Bordoni; Rome, Italy

Abstract

In the last years, the ductility and easiness of usage of unmanned aerial vehicles (UAV) and their affordable cost have increased the drones use by industry and private users. However, drones carry the potential of many illegal activities from smuggling illicit material, unauthorized reconnaissance and surveillance of targets and individuals, to electronic and kinetic attacks in the worse threatening scenarios. As a consequence, it has become important to develop effective and affordable countermeasures to report of a drone flying over critical areas. In this context, our research chooses different short term parametrization in time and frequency domain of environmental audio data to develop a machine learning based UAV warning system which employs the support vector machines to understand and recognize the drone audio fingerprint. Preliminary experimental results have shown the effectiveness of the proposed approach.

Introduction

Nowadays commercial-grade and consumer-grade drones are a market product since the technology to control and operate unmanned aircraft is cheap, widely available and fast developing. The drones applications range from hobby and amusement to aerial surveillance and lastly for illegal and criminal activities[1]. The increasing drone efficiency, the installation of GPS control systems and autopilot allow drones to fly programmed routes without a pilot for several miles[2]. This multiplies the risk of airspace exposition of sensitive buildings, facilities and personals[3]. So in various contexts the necessity of detecting and restricting the illegal use of drones emerges. However, the speed and varying shape of drones make the discovery of flying drones a complex and difficult task especially when the unmanned aerial vehicles (UAVs) have small size and a single detection method, using RF or optical sensors, is employed.
On the other end the typical acoustic signature of most commercial drones suggest the opportunity to detect or boost the existing monitoring system by using a UAV sound recognition system [4, 5, 6].

Specifically, Mezei proposed a drone sound detection based on the correlation technique of audio fingerprinting [7].
Besides, Souli presented an environmental sound spectrogram SVM classification approach built on the reassignment of spectral patches [8]. This research proposes a cheap and portable drone detection system, that extracts and classifies the temporal and spectral features of the recorded environmental sounds to figure out if there is or not a drone near. The proposed system is independent but integrable with other common detection approaches as multi-sensor drone trackers that can use optical, thermal, infrared and RF array of sensors.
This work generalizes the application context of Mezei approach developing an acoustic signature identification framework that applies the bag of frames [9] with machine learning techniques. In addition, it exploited audio analysis both in temporal and frequency domain [10, 11] to obtain an higher accuracy.
The paper is organized as it follows. The first section describes the drone acoustic detection framework. The second section presents the selected audio descriptors.The third section discusses the automatic identification of drone sounds by using SVM. The forth session presents the result. Finally, concluding remarks are given in section six.

Proposed framework

In this paper, a drone detector engine is modelled as an intelligent system able to perceive the context in which it is deployed by monitoring the environmental sound. When a drone sound is recognized an alert is triggered by the system.
The proposed framework for the identification of the drone acoustic signature is shown in Fig. (1). Its schema is constituted by five main modules that perform the tasks detailed in the following.

- *Audio acquisition*
  The sounds produced by the surrounding environment are picked up by an audio sensor and converted in a digital format by a sound card. To maintain a good time resolution and a wide frequency bandwidth, it is assumed a sampling rate of 48 kHz and a linear encoding with 16 bits for sample.

- *Preprocessing*
  Each digital sound segment, recorded in a buffer memory, is broken into consecutive frames of predetermined duration (5 seconds); then the frames are normalized in the range [-1,1].

- *Short term analysis*
  To reduce the amount of data and identify discriminative meaningful information, each input frame is further segmented into sub-frames of 20ms using a moving Hamming window with overlap of 10ms. The sub-frames are processed by a bank of filters to compute the so called short term feature in both temporal and frequency domains.

- *Mid term analysis and frame modelling*
  Considering the sequence of the extracted local audio features, their statistics are computed on a mid-term window of 200ms. The made assumption is that the audio signal presents homogeneous physical characteristics during this temporal segment. Subsequently, the audio frame of 5s is represented by a signature vector, that is obtained by the concatenation of the set of feature statistics, associated with each segment of 200ms in the given frame.

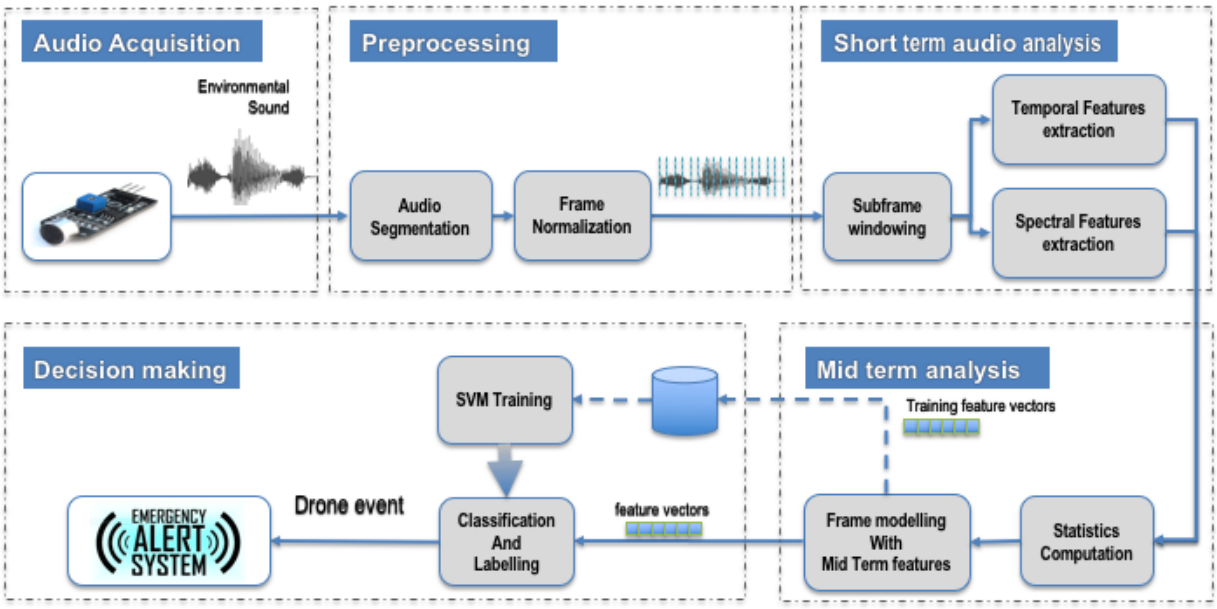

Figure 1: Drone acoustic signature identification framework

- *Decision making*
  The content of the environmental digital audio segment is ranked by a set of SVM classifiers, operating on frame basis. Specifically, there is a set of binary classifiers that process the audio signature vectors of the input frames. Each SVM classifier is trained according to the paradigm one against one to recognize the drone acoustic signature. If the majority of the frames is labelled with the tag "drone" into the given audio segment, it is assumed to recognize a flying drone in the surrounding environment.

Audio data description

A critical issue in the audio pattern recognition problem is the choice of features for constructing an accurate identification system. The audio features should be efficient, robust and physically interpretable so as to obtain a machine processable data representation containing the key properties of the audio signal.
In general, the environmental sounds can be generated by variety of sources under multiple contexts and no assumptions can be made about their spectral and temporal structure [12]. Besides, the corresponding audio signal is non-stationary in time, so that the signal can be assumed locally stationary only on short time range of 10-30 ms. This implies that the time and spectral behaviour of the audio signal can be considered practically homogeneous in a time range of few milliseconds.
Hence, to capture the heterogeneous audio information, two different temporal scales are considered during the analysis phase of environmental sound. Specifically, the audio segments are processed on a short time basis of 20ms to discriminate which set of temporal and frequency features is effective in the drone identification problem. Subsequently, to give prominence to salient audio features discarding the local detail, a mid term time analysis is performed on 200ms.

Short term analysis

Denoting with $x(n)$ the discrete time representation of normalized audio frame, the audio data $s(n)$ of each subframe are given by eq.(1).

$$\begin{aligned} s(n) &= x(n) \cdot w(m-n) \\ w(n) &= 0.54 - cos\Big(\frac{2\pi(m-n)}{L-1}\Big) n \in [0, L) \end{aligned} \tag{1}$$

whew $w(n)$ is an Hamming window of length L and m its time shift. To compute the raw features and locally characterize the corresponding audio waveform and spectrum shape, each sub-frame $s(n)$ is processed by a bank of specific filters on the basis of the feature algorithms detailed below.
In particular, the computed local features are: the Short-Time Energy (STE), the Zero Crossing Rate (ZCR) and the Temporal Centroid for time domain; the Spectral Centroid (SC), the Spectral Roll-Off (SRO), the Mel Frequency Cepstrum Coefficients for frequency domain.

- *Short Time Energy*
  It is computed according to the expression (2) and provides a measure of the energy variations of the environmental sound over time.

$$STE = \frac{1}{L} \sum_{i=0}^{L-1} | s(i) |^2 \tag{2}$$

- *Temporal centroid*
  It is defined as the temporal balancing point of the amplitude distribution of audio signal. It is expressed as:

$$C = \frac{\sum_{h=1}^{L} h \cdot s(i)}{\sum_{h=1}^{L} s(i)} \tag{3}$$

- *Zero crossing rate*
  It counts the average number of times where the audio

signal changes its sign within the short-time window. This features is particular useful to identify voiced subframe.

$$ZCR = \frac{1}{2(L-1)} \cdot \sum_{i=0}^{L-1} \mid sgn(s(i)) - sgn(s(i-1)) \mid \quad (4)$$

- *Spectral Centroid*
  It represents the balancing point of audio spectrum $p(f)$ specifying if lower or higher frequencies are contained in the spectrum.

$$\begin{aligned} SC &= \frac{\sum_f f \cdot p(f)}{\sum_f p(f)} \\ p(f) &= \mid \sum_{i=0}^{L-1} s(i) \cdot e^{-j2\pi f/L} \mid^2 \end{aligned} \quad (5)$$

- *Spectral roll-off*
  It defines the frequency below which a certain amout β of the spectral energy is concentrated. In this work, it is assumed $\beta = 0.9$

$$SRO = arg \max_m \sum_{f=1}^{m} p(f) \leq \beta \cdot \sum_{f=1}^{F} p(f) \quad (6)$$

- *Mel frequency coefficients*
  Mel frequency cepstral coefficients are the discrete cosine transform of the mel-scaled log-power spectrum $p(f)$. The main steps to compute these M cepstral coefficients are described below.
  - The M banks of Mel filters are used to map the power spectrum $p(f)$ onto the mel scale defined by the eq.(7). The frequency responses of these filter banks are triangular and equally spaced along the mel-scale.

$$f_{Mel} = 2.595 \log_{10}\Big(1 + \frac{f}{700}\Big) \quad (7)$$

  - The energies E_m at the output of mth filter is computed and then compressed into a logarithmic scale. Let be C_m the relative log value.
  - Given the M log energies C_m, the corresponding DCT coefficients are computed and constitute the Mel cepstral coefficients of audio signal $s(n)$.

$$c_i = \sum_{k=1}^{M} C_k \cdot cos\Big(\frac{\pi(2i+1)k}{2M}\Big) \quad (8)$$

  To address the drone sound identification problem, 13 MFCCs are extracted because it is found that they lead discriminative information.

Mid term analysis and features aggregation

After analyzing the subframes into the environmental audio frame, the relative sequences of low level features are processed statistically on a mid term time window, as mentioned in the second section. The goal is to obtain new salient mid term features with low sensitivity to the small variations of underlying audio signal. Then, a set of mid term robust features are aggregated in a global vector that is able to completely describe the perceptual physical property of the environmental audio frame.

The maximization of the expressive power of the audio descriptor can be obtained by selecting mid term window with a time length, that allows to decorrelate the various components of the vector, i.e. the selected set of mid term features.
Let denote $\{\phi_i\}$ the generic sequence of low level feature into an audio frame, and let be N the number of sub-frame contained into the generic mid term window w_j. Then the N local features $\{\phi_{j+k}, \quad k = 0,..N-1\}$ relative to w_j are processed to compute the first and second order statistics, written in eq(9).

$$\begin{aligned} \mu_j &= \frac{1}{N} \sum_{k=0}^{N-1} \phi_{j+k} \\ \sigma_j &= \frac{1}{N} \sum_{k=0}^{N-1} (\phi_{j+k} - \mu_j)^2 \end{aligned} \quad (9)$$

Given all mid term windows w_j in the environmental audio frame, it is considered a selection of these statistics for the various raw features to generate a vector of mid term features by concatenation. This vector is the global audio descriptor representing the audio signature vector of the frame. Besides, it is the basic unit processed by the classifier for the identification of drone sound.

Classifier for drone sound identification

The drone audio identification problem is addressed at frame level formulating a multiclass environmental audio recognition problem. Since the dependency of the mid term feature values from the properties of environmental sounds properties is not known a priori, a multiclass SVM based classifier is trained for estimating the multidimensional audio descriptors.
This projectual choice is based on two main motivations. First, SVMs are powerful machine learning techniques that are applied in various pattern recognition problems with excellent results [14, 15]. Secondly, a labelled training data set is the only information required to implement an SVM.
In this work, it is adopted the *one-against-one* strategy [16] to implement the SVM based classifier with k classes for drone sound identification.
This leads to $H = k(k-1)/2$ binary SVM classifiers, each one trained on data from two classes. The output class is selected among those chosen by most classifiers according to the *max wins* strategy.
If two classes obtained identical number of votes it is selected randomly one. Denoting with $\mathbf{v}$ the multidimensional audio descriptor for the input audio frame and with $f_j(\mathbf{v})$ the decision function of the generic classifier SVM_j, $j = 1,2,...H$, the output label can be expressed by eq(10) with $H = k(k-1)/2$.

$$label(\mathbf{v}) = arg \max_{1 \leq j \leq H} \{f_j(v)\} \quad (10)$$

Given a dataset of elements $\{\mathbf{v_n}, n = 1,2,..M\}$ belonging to two different audio classes $\mathscr{A}$ and $\mathscr{B}$, the model of the relative binary SVM_j classifier is computed exploiting the principle of the structural risk minimization. This implies minimizing a bound on the generalization error, rather than minimizing the mean square error [13]. Hence, the learning problem for the SVM_j can be formulated as the determination of the optimal separating surface that maximize the margin between the two classes $\mathscr{A}$ and $\mathscr{B}$. This

margin is given by the perpendicular distance of the closest data point $\mathbf{v_n}$ to the optimal separating surface. In the mapped space $\mathbf{z} = \phi(\mathbf{v})$, the optimal separating surface becomes the optimal decision hyperplane $\mathbf{w}^T \cdot \mathbf{z} + b = 0$ and the margin between the two classes is represented by the minimum distance $\mathbf{d_{min}}$ of the closest point $\mathbf{v_n}$ from the bounding plane $\mathbf{w}^T \cdot \mathbf{z} + b = 0$.

To model the SVM_j classifier, we search for the optimal values of the parameters $\mathbf{w}$ and b that maximize the minimum distance $\mathbf{d_{min}}$. These values identify the optimal decision hyperplane and define the SVM_j decision function $f_j(v)$ as:

$$\begin{aligned} f_j(v) &= sgn\Big(\sum_{n=1}^{M} a_n t_n K(\mathbf{v}, \mathbf{v_n}) + b\Big) \\ K(\mathbf{v}, \mathbf{v_n}) &= \phi(\mathbf{v})^{\mathbf{T}} \cdot \phi(\mathbf{v_n}) \\ \mathbf{w} &= \sum_{n=1}^{M} a_n t_n \phi(\mathbf{v_n}) \end{aligned} \tag{11}$$

where $\{t_n, n = 1, 2, ..M\}$ are the target label for the two classes, assuming with $t_n \in [-1, 1]$.
In general the audio descriptors $\mathbf{v_n}$ represent points linearly non-separable in the audio descriptor space but which may be separated in a non-linear way. This corresponds to the adoption of SVM_j with non-linear kernel function $K(\mathbf{v}, \mathbf{v_n})$. In particular this work consideres all SVM_j having a Gaussian Radial Basis kernel function (RBF) [13] expressed as $K(\mathbf{v}, \mathbf{v_n}) = e^{-\frac{|\mathbf{v}-\mathbf{v_n}|^2}{2\sigma^2}}$.

Experimental results

To evaluate the performance of the proposed system for drone sound recognition, we used a dataset containing five different typologies of environmental sounds corresponding to: drone flying, nature daytime, street with traffic, train passing, crowd. The dataset is created starting from background sounds collected from the web using a specific scraper for audio files. The selection of a balanced number of elements for the five main classes are the criteria for the construction of the dataset.

Table 1: Classes of the environmental audio frames

Classes	Total
Drone	868
Nature daytime	844
Crowd	840
Train passing	856
Street with traffic	864

Following these criteria, the scraper surfs the web discarding all digital audio file with sampling rate less than 48 KHz. The search process ends when a balanced amount of audio files is downloaded for each class specified in the queering set. Subsequently, each found audio file is manually validated to verify if it is correctly associated with the given label. A specialized software is used to divide each audio segment in frame of 5s and annotate them with the class label relative to the audio segment. In table 1 it is reported the class of environmental sounds and the relative number of audio frames in the dataset used for the train and test process. Namely, it corresponds to six hours of environmental sounds. As suggested by the *one against one* strategy, ten SVM binary classifiers are trained and tested, one for each possible couple of class. For each SVM classifier, several experiments are performed to determine the optimal discriminative subset of aggregated features in combination with the optimal SVM kernel among those linear, polynomial, RBF and sigmoid.
The different choices are compared in terms of accuracy and precision using the k-fold cross validation procedure (k=5) to prevent the overfitting problem.
The experiments has revealed that the RBF kernel performed better than the others with audio descriptor vectors constituted by the first order statistics of raw features.
After defining the typology of SVM kernel, we have trained the different binary classifier using specific subset of constructed dataset. For example in modelling the SVM that classifies drone frames versus crowd frames, we considered 868 segments labeled as drone and 840 segments labeled as crowd. Then the 50% of files of these two classes is used for training while the remaining for the test.
The obtained performance for the adopted ten SVMs are shown in Table 2 in terms of accuracy and precision.

Table 2: SVMs Performance

SVM-Classifier	Accuracy	Precision
Drone/Crowd	0,964	0,984
Drone/Nature_daytime	0,992	0,983
Drone/Train_passing	0,978	0.983
Drone/Street_with_traffic	0,964	0,987
Crowd/Nature_daytime	0,959	0,919
Crowd/Train_passing	1	1
Crowd/Street_with_traffic	0,8911	0,782
Nature_daytime/Street_with_traffic	0,991	1
Nature_daytime/Train_passing	0,996	0,991
Street_with_traffic/Train_passing	1	1

To conclude we have tested the overall SVMs based classifier, giving in input the same test frame to all ten SVM. By using the max wins strategy, the output label was identified. The resulting precision of the drone recognition is 98,3 % .

Conclusions

This work investigates the efficacy of machine learning techniques to face the problem of drone detection in the context of critical areas protection. To this end an empirical analysis of the environmental sounds, recorded by the audio sensors in the critical areas, is performed. The extracted time-frequency fingerprint is adopted by the warning system to recognize drone sounds.

References

[1] E. Vattapparamban, . Gven, A. . Yurekli, K. Akkaya and S. Uluaa, "Drones for smart cities: Issues in cybersecurity, privacy, and public safety," 2016 International Wireless Communications and Mobile Computing Conference (IWCMC), Paphos, 2016, pp. 216-221.

[2] A. Harrington,"Who controls the drones? [Regulation Unmanned Aircraft]," in Engineering & Technology, vol. 10, no. 2, pp. 80-83, March 2015.

[3] K. Hartmann and K. Giles, "UAV exploitation: A new domain for cyber power," 2016 8th International Conference on Cyber Conflict (CyCon), Tallinn, 2016, pp. 205-221.

[4] S. Chu, S. Narayanan, C. c. J. Kuo and M. J. Mataric, "Where am I? Scene Recognition for Mobile Robots using Audio Features," 2006 IEEE International Conference on Multimedia and Expo, Toronto, Ont., 2006, pp. 885-888.

[5] A. Rabaoui, M. Davy, S. Rossignol and N. Ellouze, "Using One-Class SVMs and Wavelets for Audio Surveillance," in IEEE Transactions on Information Forensics and Security, vol. 3, no. 4, pp. 763-775, Dec. 2008.

[6] R. Serizel, V. Bisot, S. Essid and G. Richard, "Machine listening techniques as a complement to video image analysis in forensics," 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 2016, pp. 948-952.

[7] J. Mezei, V. Fiaska and A. Molnr, "Drone sound detection," 2015 16th IEEE International Symposium on Computational Intelligence and Informatics (CINTI), Budapest, 2015, pp. 333-338.

[8] Sameh Souli, Zied Lachiri, and Alexander Kuznietsov,"Using Three Reassigned Spectrogram Patches and Log-Gabor Filter for Audio Surveillance Application". In Proceedings, Part I, of the 18th Iberoamerican Congress on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - Volume 8258 (CIARP 2013), Jos Ruiz-Shulcloper and Gabriella Sanniti Di Baja (Eds.), Vol. 8258. Springer-Verlag New York, Inc., New York, NY, USA, 527-534.

[9] J.-J. Aucouturier, B. Defreville, and F. Pachet, The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music, J. Acoust. Soc. Amer., vol. 122, no. 2, pp. 881891, Aug. 2007.

[10] Xuan Guo, Yoshiyuki Toyoda, Huankang Li, Jie Huang, Shuxue Ding, and Yong Liu. 2012. "Environmental sound recognition using time-frequency intersection patterns". Appl. Comp. Intell. Soft Comput. 2012, Article 2 (January 2012), 6 pages

[11] Mitrovi, Dalibor, Matthias Zeppelzauer, and Christian Breiteneder. "Features for content-based audio retrieval." Advances in computers 78 (2010): 71-150.

[12] S. Chu, S. Narayanan and C. C. J. Kuo, "Environmental Sound Recognition With TimeFrequency Audio Features," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1142-1158, Aug. 2009.

[13] C. Bishop, "Pattern Recognition and Machine Learning", Information Science and Statistics ed.Springer, 2007

[14] S. Theodoridis and K. Koutroumbas, Pattern Recognition, Academic Press. ed., 2008.

[15] P. Dhanalakshmi, S. Palanivel and V. Ramalingam, "Classification of audio signals using SVM and RBFNN," Expert Systems with Applications: An International Journal, vol. 36, pp. 6069-6075, 2009.

[16] Hsu,C.-W.,Lin,C.-C,"A comparison of methods for multi-class support vector machines".J.IEEE Transactions on Neural Networks 13,pp.415-425, 2002

Author Biography

Andrea Bernardini received his Dr. Ing. degree in Computer Engineering at the University of Rome "Roma Tre". In 2010, he was Visiting Researcher at the Institute for Computing, Information and Cognitive Systems (ICICS) of the University of British Columbia (UBC). In 2002 he joined The Fondazione Bordoni, where he works as researcher in Information processing and management Department. His research interests include User Experience, Data Mining and User Modeling.

Licia Capodiferro received her Dr. Ing. degree in Electronic Engineering from the University of Rome La Sapienza, Italy. In 1987 she joined the Fondazione Ugo Bordoni where she currently works as head of the Department of Information Processing and Management. Her main research interests are in the field of multimedia processing, with a focus on algorithms that allow the use of images and videos on the different types of terminals.

Federica Mangiatordi received the M.Sc. Degree in Electronic Engineering at University of Rome La Sapienza and the PhD in Electronic Materials, Optoelectronics and Microsystems from the University of Roma TRE. She works at Fondazione Ugo Bordoni from 2007. Her research interest concern multimedia retrieval, image restoration algorithms, novel metrics for full reference and no-reference image objective quality assessment.

Emiliano Pallotti received the Laurea Degree in Telecommunications Engineering at the University of Rome La Sapienza, Italy, and PhD in Electronic Materials, Optoelectronics and Microsystems from the University of Roma TRE. In 2007 he joined the Fondazione Ugo Bordoni where his research activities are in the field of on computational algorithms and video process- ing techniques based on multiresolution image representation in wavelet domain.