electronics

Article

A Gunshot Recognition Method Based on Multi-Scale Spectrum Shift Module

Jian Li [1], Jinming Guo [1], Mingxing Ma [1], Yuan Zeng [1], Chuankun Li [1,*] and Jibin Xu [2]

[1] National Key Laboratory of Electronic Testing Technology, North University of China, Taiyuan 030051, China
[2] Hunan Vanguard Group Co., Ltd., Changsha 410100, China
* Correspondence: chuankun@nuc.edu.cn

Abstract: In view of the issues such as the larger network model and lower recognition accuracy of the current gunshot recognition networks, a neural network based on a multi-scale spectrum shift module is proposed in this paper to fully mine the relevant information among the gunshot spectrums. This network employs the architecture of a densely connected convolutional network and uses a multi-scale spectrum shift module on the branch to realize the interaction among spectrum information. This spectrum shift replaces the under-sampling operation among the spectrums, realizes the globalized feature extraction of the spectrum, avoids the loss of information during the under-sampling process, and further improves the quality of the spectrum feature map. Experiments were conducted based on the NIJ Grant 2016-DN-BX-0183 gunshot dataset and YouTube dataset on gunshots that have been open to the public, both of whose classification accuracy reached 83.2% and 95.1%, respectively, with the size of the network model being controlled at around 16 MB. The experimental results indicate that, compared with other existing methods for convolutional neural network, the proposed network can mine globalized time-frequency information better and effectively, and has a higher accuracy of gunshot recognition.

Keywords: gunshots; log-mel spectrum; spectrum shift; convolutional neural network

Citation: Li, J.; Guo, J.; Ma, M.; Zeng, Y.; Li, C.; Xu, J. A Gunshot Recognition Method Based on Multi-Scale Spectrum Shift Module. *Electronics* **2022**, *11*, 3859. https://doi.org/10.3390/electronics11233859

Academic Editor: Stefanos Kollias

Received: 31 October 2022
Accepted: 21 November 2022
Published: 23 November 2022

1. Introduction

Millions of people have suffered from the severe and long-term psychological effects that arise out of gun violence or its threat to individuals, families, and their broader communities. In the United States, nearly 134,000 people had gunshot injuries in 2017. However, quick responses to gun violence can effectively reduce casualties, which means most gun violence can be recognized by video surveillance due to the rapid development of human gesture recognition technology; that is, whenever detecting instant supervisory pictures containing acts of gun-related violence such as running with guns and aiming the gun, the supervisory personnel can be notified immediately and police can be mobilized. However, this method of capturing human gestures to judge gun violence through video surveillance has many shortcomings. First, video surveillance devices cannot achieve omnidirectional detection due to video blind zones. In cases of gun violence, the gunman deliberately keeps out of the areas where the cameras are present when committing a crime, resulting in many acts of gun violence that cannot be recorded by surveillance devices, making it difficult to realize accurate and timely early warning. Second, on cloudy days or at night, the recognition accuracy is greatly reduced by blurred video caused by dim light. Such problems existing in video surveillance will create enormous challenges to the decidability of the system. Compared with video surveillance, audio surveillance does not have the above-mentioned problems, and the sound signal is not affected by light and weather, and the acoustic sensing probe is smaller and more concealed, available for surveillance of public areas in an all-weather, wide-range, and invisible way. The gunshot signals collected by audio surveillance usually contain vital information, including the type

of weapon, the size of the firepower, etc., which provides the police not only with clues to solve the case, but also the firepower intelligence of the criminals, and can help prevent follow-up criminal behaviors.

Researchers have carried out a great deal of work on gunshot recognition with audio data. Alain et al. [1] designed a sound recognition system in 2000 by extracting the sub-band energy features of the sound to detect and recognize the impact sounds such as glass sounds, screams, gunshots, explosions, and slamming doors by means of the recognition methods of Gaussian mixture model (GMM) and hidden Markov model (HMM). However, this system performed poorly under strong noise conditions. In 2005, Clavel et al. [2] explored the problem of gunshot detection in a surveillance environment, and comprehensively considered the impact of different background noises on the detection results. They extracted mel-frequency cepstrum coefficients, energy features, and spectrum moment from each frame of audio, from which their respective first-order and second-order differences were extracted, and the 13-dimensional features were selected as the input features of the subsequent GMM model by principal component analysis methodology. In 2007, Valenzise et al. [3] detected and localized gunshots and screams mainly in the surveillance environment, and developed a system for square surveillance in Milan, which employed two parallel GMMs to distinguish the gunshots as well as screams from background sounds. As seen from the final results, the system obtained 93% accuracy and a 5% missed detection rate under the signal-to-noise ratio of 10 dB. Busse et al. [4] proposed a shooting classification model to determine whether there is a gunshot in the audio signal through a support vector machine (SVM) classifier, resulting in an accuracy of 70.39%.

The purpose of the above-mentioned gunshot recognition system is merely to detect whether there are any gunshots in the environment but fails to provide more effective information on firearms. Many researchers have proposed schemes to identify the information such as the types and calibers of firearms. Kiktov et al. [5] extended the intelligent acoustic event detection (AED) system to capture the types of firearms from the noise monitoring stations in which gunshots were collected. An 80% detection success rate was obtained through the hidden Markov model (HMM) classification and a Viterbi-based decoding algorithm. Ahmed et al. [6] utilized mel-frequency cepstrum coefficient (MFCC) and support vector machine (SVM) classifiers for the efficient classification of gun types. Djeddou et al. [7] extracted gunshot audio signals in a noisy environment, from which the MFCC features of the signals and the GMM classifiers to be used were extracted, to classify five kinds of gun audio signals, with a classification rate of 96.29%. In recent years, data-driven deep learning methods have developed rapidly and have been widely used in the aspects of image fusion [8] and object detection [9]. Deep learning realizes complex function approximation through nonlinear mapping and demonstrates the powerful capability to extract essential features of datasets from a few sample sets. Therefore, the sound recognition method by way of deep learning has gradually become the mainstream research direction. Raponi et al. [10] designed a multi-layer convolutional neural network that could effectively identify the category, caliber, and model of guns from a dataset of 59 different types of firearms extracted from YouTube videos with an accuracy rate of over 90%. Ryan Lilien [11] collected 6000 individual gunfire audio signals from 18 kinds of guns in the field, for which the method of transfer learning was used to improve the recognition accuracy of gun categories and calibers, and the 14-layer CNN network was first pre-trained with the AudioSet Dataset, a large dataset, and then gunshot audio was trained, and finally achieved an accuracy of 78.2%.

All the above gunshot recognition methods are of some limitations. The gunshot recognition methods based on HMM and GMM have low robustness. When the actual test environment is greatly different from the simulation environment, the gunshot cannot be accurately identified. The method of gunshot recognition based on SVM has high robustness but requires many experiments to adjust parameters and select kernel functions, and the recognition ability is poor under strong noise conditions. The method of gunshot recognition based on deep learning can realize a large amount of model calculation and

high hardware requirements and is difficult to achieve the purpose of a real-time alarm. In order to meet the requirements of high robustness and portable device deployment, this paper proposes a small neural network based on multi-scale spectrum shift, which ensures that the model has a certain generalization ability, reduces the network parameters, and realizes the requirements of real-time discrimination of gun types. Ensuring that the model has a certain generalization ability reduces the network parameters and realizes the requirements of real-time discrimination of gun types.

2. Principles Relating to Sniper Gunshot Recognition

Common gunshots include two kinds of transient sound signals: the first is the muzzle blast formed when the bullet is pushed out of the muzzle by the high temperature, high pressure, and high-speed airflow generated from the explosion of the gunpowder in the bullet during the firing process [12]; the second is the shock wave produced by the projectile flying at supersonic speed in the atmosphere, also called Mach wave [13]. Figure 1 shows the gunshot waveform signals that are collected actually.

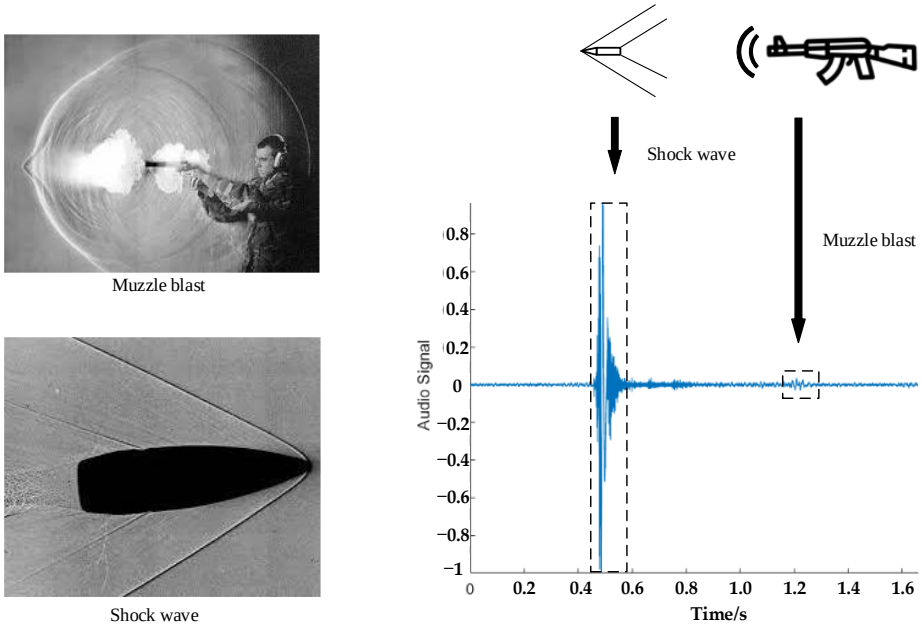


Figure 1. Schematic diagram of a muzzle sound wave.

Not all firearms may generate Mach waves, since most of the common pistol projectiles are subsonic without Mach waves when flying in the air. Mach waves, as the characteristics of the bullet itself, are an important basis for judging the caliber of guns [14]. Therefore, simultaneously considering the two kinds of audio components of muzzle wave and Mach wave in gunshot recognition can effectively recognize the gun type and caliber.

3. Design of the Gunshot Recognition Method

3.1. Feature Extraction

It is very difficult to capture useful information related to the research from the original one-dimensional audio data that are used as features due to a large amount of noise inside gunshot data collected under realistic conditions. In order to solve this problem, a relatively popular method at present is to map the audio data to the Mel spectrogram or MFCC spectrogram; that is, the input gunshot information is filtered by the method of Mel spectrogram feature extraction to improve the accuracy of gunshot recognition. In this paper, the log-mel spectrogram is used as the input of the network, whose principle is to design a filter by imitating the auditory characteristics of the human ears and then recognize the original audio data when processed by the filter.

First, the short-time Fourier transform is performed on the input gunshot signal with the specific steps as follows:

(1) Framing: the input audio signal x (n) is divided into multiple short time frames, the length of each frame is 43 ms, and the frameshift is 21 ms.
(2) Windowing: each short-time frame is multiplied by Hamming window to increase the continuity of the left and right ends of the frame.
(3) FFT: A 1024-point fast Fourier transform is performed on each frame signal to convert.
(4) Taking power spectrum: the modulus square is taken for the frequency and spectrogram of the sound signal to obtain the spectrum line energy of the signals.

Secondly, the obtained power spectrum is filtered through the equal-height Mel filter banks: the Mel filter banks consist of M triangular filters (M = 40 in this paper), and the filters are concentrated in the low-frequency region and sparsely distributed in the high-frequency region. The frequency response of the triangular filter is:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \le k \le f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \quad (1)$$

where $0 \le m \le M$, $f(m)$ is the center frequency. For the power spectrum obtained when processed by Formula (1), the operation of frequency multiplication and accumulation is performed with M triangular filters, respectively, to obtain M energy values of the corresponding frequency band of the data in that frame.

Finally, the amplitude value (color) of the obtained Mel spectrogram is converted into decibels by taking the logarithm to obtain the result approximate to the homomorphic transformation, as shown in Figure 2.

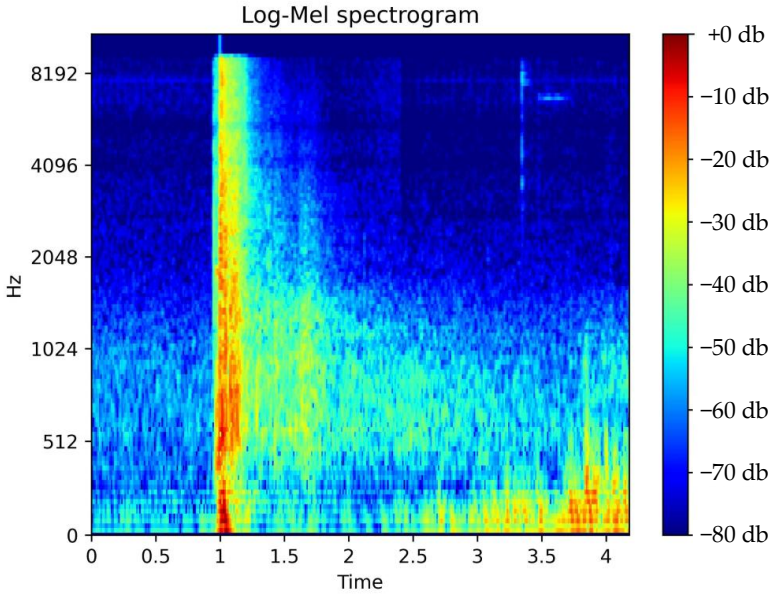


Figure 2. Log-Mel spectrogram.

3.2. Multi-Scale Spectrum Shift Dense Neural Network

In order to be able to make greater use of the convolutional neural network to mine effective time-frequency information from the spectrogram, a multi-scale spectrum shift densely connected convolutional network is proposed in this paper to explore the correlated features among the frequency spectrums, which not only solves the problem that convolution can merely extract fixed adjacent spectrum information but also extracts the globalized information among different spectrums without inter-spectrum under-

sampling, so as to mine more effective time-frequency features and improve the accuracy of gunshot recognition.

Dense neural network (DenseNet) has been widely used in various tasks in terms of computer vision. Compared with the traditional convolutional neural networks, the DenseNet realizes the information flow through each convolutional layer in a densely connected way, which can effectively alleviate the problem of vanishing gradients in deep networks, and the densely connected method can effectively use the information from different layers to further mine new useful features. In this paper, a densely connected method is used to construct a multi-scale spectrum shift densely connected convolutional network. As shown in Figure 3, this network consists of two convolutional layers, three spectrum shift-dense modules, and a global pooling layer (global pooling), of which each convolutional layer contains a 2D convolution operation, a batch normalization layer (BN layer), and activation functions (ReLU). An end-to-end learning approach is applied to the network framework proposed in this paper to realize a direct mapping from the log-mel spectrogram to the corresponding gunshot recognition. First, the log-mel spectrogram is entered into a convolutional layer to extract the shallow feature map, where the time dimension step of the convolution kernel is set to 2, and the obtained feature map is input into the three spectrum shift-dense modules, and then the output feature map is input to the convolution layer to extract features, and finally, the probability value of sound classification is output through the global pooling layer.

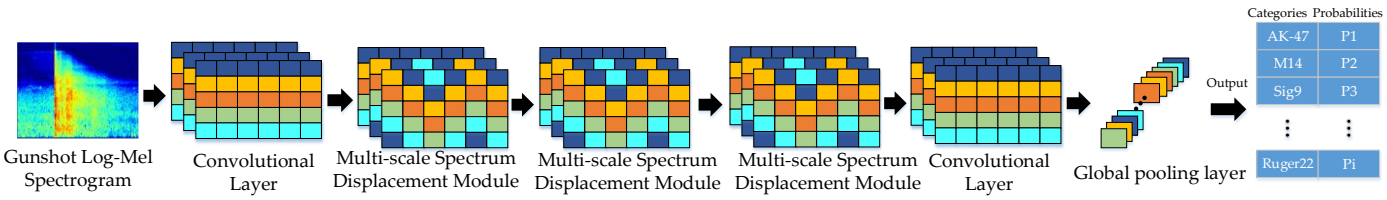


Figure 3. Multi-scale spectrum shift dense neural network.

As for the traditional convolutional neural network, the effective time-frequency features can be extracted by the convolutional layer under the local receptive field through a large number of convolution kernels, and then the mining of global features can be realized through nonlinear activation functions and under-sampling, but under-sampling will cause the loss of useful information. To reduce the loss of information, the spectrum width of the spectrogram has to be relatively small, resulting in the under-sampling being performed just on the time axis during the under-sampling process to retain all the information on the frequency axis. However, when using the receptive field of the convolution kernel, just the local information of the adjacent spectrums can be extracted. To solve this problem, a multi-scale spectrum shift dense module is proposed in this paper, as shown in Figure 4. The shallow feature map is input to two paths, where two convolution layers are in the main path above to extract local features among adjacent spectrograms, and the time step of the first convolution layer is set to 2 for under-sampling in terms of time; two convolution layers and a multi-scale spectrum shift module are in the branch underneath to realize the position transform among the spectrograms by first passing through the first convolution layer (with the time step being set to 2), and then passing through the multi-scale spectrum shift module to extract the information among other adjacent spectrograms through convolution; and finally, the information from the upper and lower paths is densely connected to realize the fusion of different spectrum information.

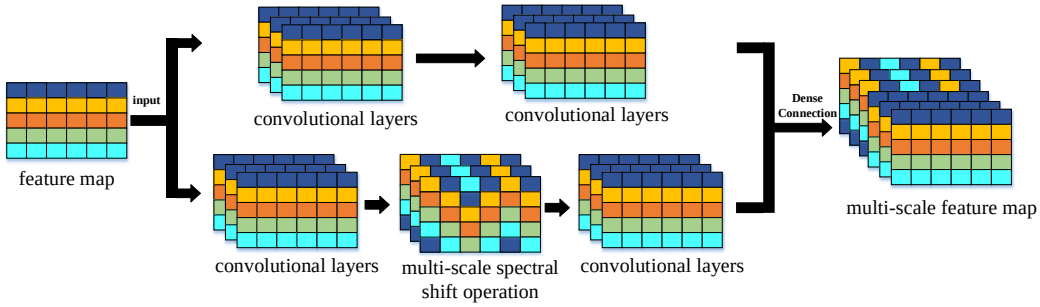


Figure 4. Multi-scale spectrum shift dense neural module.

A multi-scale spectrum shift module is designed in this paper to mine useful information among different spectra of gunshots, as shown in Figure 5. For simplicity, only the spectrum and feature channels are displayed, and different spectrum features are represented by different colors in each row, traditional 2D convolution operations are carried out between different channels, i.e., it is operated individually along each row so that only the information among adjacent spectra can be extracted. In order to mine more information among spectrums, the feature transform is divided into two paths in this paper, in one of which one-quarter of the feature channel is moved down for one bit, and the downmost feature channel is moved topmost; meanwhile, one-quarter of the feature channel is moved up for one bit, and the topmost feature is moved downmost; the remaining feature channel positions remain unchanged. The other channel that remains unchanged is moved down for two bits, and the downmost feature channel is moved topmost; at the same time, the other feature channel is moved up for two bits, and finally, the two feature maps are connected in parallel so that when the convolution operation for each row is performed, the information among different spectrums can be extracted to generate more effective time-frequency features.

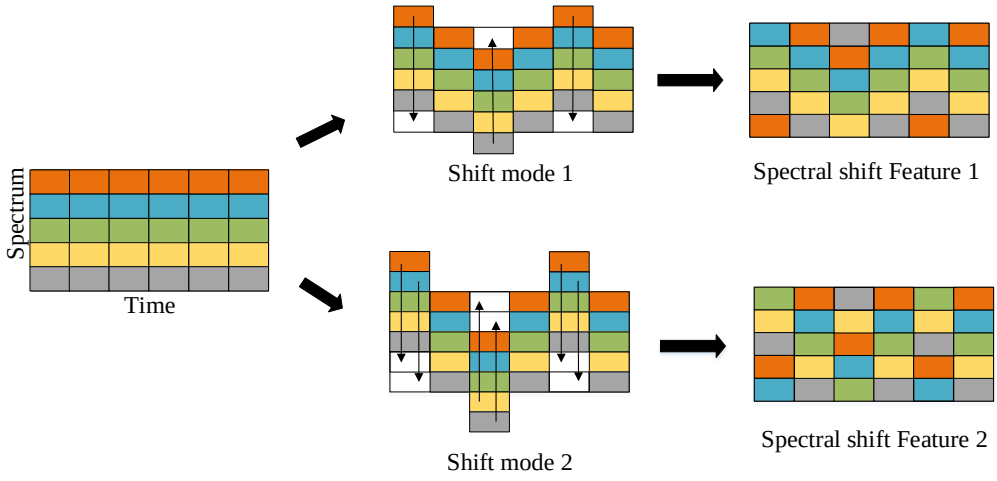


Figure 5. Multi-scale spectrum shift operation.

Similar to DenseNet, in the spectrum shift dense module, a fixed number of convolution kernels are adopted in the lower branch, that is, the convolution kernel in the first layer is set to $3 \times 3 \times 96$, and that in the second layer is set to $3 \times 3 \times 24$. For the main path above, different numbers of convolution kernels are adopted according to different spectrum shift dense modules. The convolution kernels of the four spectrum shift dense modules are set as $3 \times 3 \times 48$, $3 \times 3 \times 96$, and $3 \times 3 \times 192$. The overall network parameters are shown in Table 1.

Table 1. Overall network parameters.

Operation	Kernel Size	Output Size
Convolution Layer 1	$3 \times 3 \times 48$	$40 \times 256 \times 48$
Multi-scale spectrum shift dense module 1	$3 \times 3 \times 48$	$40 \times 128 \times 92$
Multi-scale spectrum shift dense module 2	$3 \times 3 \times 96$	$40 \times 64 \times 120$
Multi-scale spectrum shift dense module 3	$3 \times 3 \times 192$	$40 \times 32 \times 216$
Convolution Layer 2	$3 \times 3 \times$ Number of categories	$40 \times 32 \times$ Number of categories
Global Pooling Layer	-	Number of categories

4. Experimental Results and Data Comparison

The methods for gunshot recognition proposed by researchers at present have their own advantages and disadvantages, but their accuracy for gunshot recognition varies greatly. The accuracy of the early methods for gunshot recognition may exceed 90%, whereas the accuracy of the method proposed by Ryan Lilien [11] is only 78%; when similarly using the convolution neural network, the recognition accuracy of the multi-layer convolution neural network designed by Raponi et al. [10] can be up to 90%. This is because there is no standard gunshot database for public places in the industry currently, and the researchers need to sort out the data and build their own database. However, constrained by the environment, it is impossible to directly record gunshots in public places. Many researchers did not record the audio of gunshots on the spot but downloaded related gunshot videos and audio over the Internet, which have been artificially screened (since the gunshot signals with poor sound quality are not uploaded), so this kind of dataset has a high signal-to-noise ratio and fine signal quality, and higher recognition accuracy. However, the gunshot signals collected by some researchers actually have not been artificially screened for high-quality signals, resulting in poor recognition accuracy.

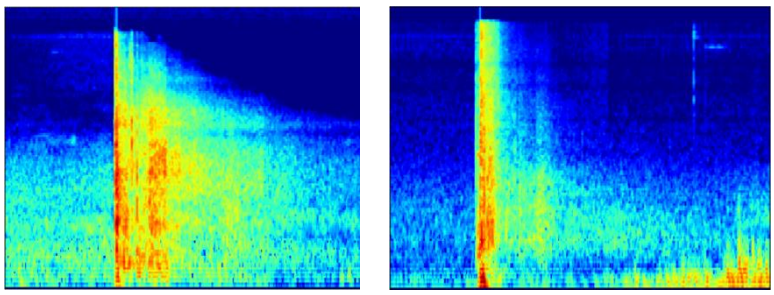
For the purpose of result comparability with those of other researchers, the NIJ Grant 2016-DN-BX-0183 Project gunshots dataset [11] with the gunshot signals being collected on the spot and the YouTube gunshots dataset [15] with gunshot signals captured from YouTube is employed in this paper for verification. First, the audio files in the training set from the two gunshot data sets are input into the network proposed in this paper for network model training, and then the audio files in the validation set are input into the trained network model, after which the obtained classification results are compared with the real results to acquire the corresponding classification accuracy. Simultaneously, to better reflect the advantages of the network proposed in this paper, an ablation experiment is proposed when comparing the accuracy between the network in this paper and the existing network, to further verify the improvement of the sound recognition accuracy of the method proposed in this paper. The specific details are described as follows:

4.1. Dataset

The YouTube gunshots dataset consists of the gunshot recordings collected from public videos from YouTube with a total of 851 samples from eight kinds of gun models including AK-47, AK-12, and M249. Each audio segment lasts 2 s with the sampling frequency at 44.1 KHz. A large portion of these gunshots is obtained from game sound effects or shot by firearm vloggers at close range, whose sound effects are post-processed when uploading to the website, whereas some of the unique signals of gunshots are missing. At the same time, each audio sample is artificially screened by the author to ensure that it is rarely mixed with irrelevant audio signals or noise, which makes some gaps between the gunshots in this dataset and the real gunshots.

Unlike the collecting method for the YouTube gunshots dataset, the NIJ Grant 2016-DN-BX-0183 Project gunshots dataset was recorded in empty rural areas in Arizona, USA, in the summer, where the audio information of 18 kinds of guns in 20 different locations was recorded on three different recording devices (Zoom H4N, iPhone 7, Samsung Galaxy S7), respectively, and the whole dataset contains about 6000 individual gunshot audio files,

including 18 kinds of firearms, such as SportKing22, HKUSP, Remington700, M16, and other firearms. Compared with the YouTube gunshots dataset, it was mixed with a large amount of noise, and the gunshots were recorded from multiple angles, which is more inclined to the gunshots under actual conditions. The recording method was more professional and more in line with the real conditions of shooting incidents. For the classification network, this is a huge challenge. Figure 6 shows the log-mel spectrogram of AK-47 gunfire on the YouTube gunshots dataset and NIJ Grant gunshots dataset.



(**a**) (**b**)

Figure 6. Log-mel spectrogram of AK-47 gun sound in (**a**) YouTube gunshots dataset (**b**) NIJ Grant gunshots dataset.

4.2. Experimental Details

Taking Keras as the environment and TensorFlow as the backend, this experiment completed the training and verification of the datasets on the Quadro RTX 6000 GPU, of which the batch size was set to 128 and 64 on the NIJ Grant 2016-DN-BX-0183 Project gunshots dataset and YouTube gunshots dataset, respectively, SGD was used as the optimization function, the initial learning rate was set to 0.1, and data enhancement was performed by mixing data samples. Both data sets are evaluated using 5-fold cross-validation. During the training process, the optimal model was saved according to the results of the validation set, the learning rate was reduced at 150, 180, and 210 rounds of iterations, and the training was stopped at 240 rounds.

4.3. Experiment Results

After 240 rounds of iterations, the multi-scale spectrum shift dense neural network produced a 16.8 MB weight model. Table 2 shows the results of 5-fold cross-validation on the NIJ Grant 2016-DN-BX-0183 project dataset and the YouTube gunshots dataset. Figure 7 shows the accuracy curve of one of the fold data under the NIJ Grant 2016-DN-BX-0183 project dataset.

Table 2. 5-fold cross-validation results.

Cross-Validation	NIJ Grant 2016-DN-BX-0183 Project Gunshots Dataset	YouTube Gunshots Dataset
1 fold	81.42%	96.25%
2 fold	80.50%	91.35%
3 fold	84.21%	97.52%
4 fold	86.12%	93.10%
5 fold	83.75%	97.28%
average	83.20%	95.1%

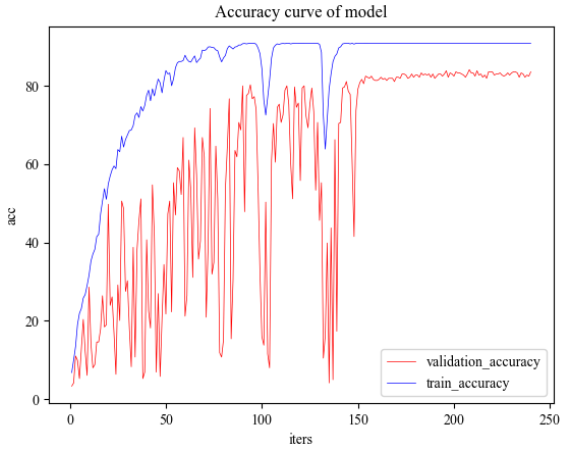


Figure 7. Network accuracy curve.

Table 3 describes the accuracy of network recognition of 18 kinds of gunshots on the NIJ Grant 2016-DN-BX-0183 Project gunshots dataset.

Table 3. Classified recognition of 18 kinds of gunshots.

Gun Model	Accuracy Rate
BoltAction22	100%
Colt1911	90%
Glock9	55.56%
Glock45	88.89%
HKUSP	100%
Kimber45	50%
Lorcin380	87.5%
M16	83.3%
MP40	100%
Remington700	87.5%
Ruger22	100%
Ruger357	100%
Sig9	53.85%
Smith&Wesson22	100%
Smith&Wesson38special	100%
SportKing22	71.43%
WASR-10	22.22%
WinchesterM14	66.67%

Meanwhile, to verify the effectiveness of the multi-scale spectrum shit module, two contrast models were selected in this paper, in one of which the spectrum shift module in each multi-scale spectrum shift dense module was turned into a single-scale spectrum shift (only one path of feature transform method), and the spectrum shift module in each multi-scale spectrum shift dense module was removed in another, whereas the other network structures and parameters remained unchanged to build a basic network, which was compared with the network proposed in this paper. The results are shown in Table 4. Through the verification by experiments, the results of the proposed network in this paper are higher than that of the basic network, indicating that the spectrum shift module can better mine time-frequency information.

Table 4. Validation of spectrum displacement module.

Model	NIJ Grant 2016-DN-BX-0183 Project Gunshots Dataset	YouTube Gunshots Dataset
Baseline	79.6%	90.1%
Single-scale spectrum shift dense neural network	81.5%	92.8%
Multi-scale spectrum shift dense neural network	83.2%	95.1%

However, without a unified and authoritative dataset in the task of gunshot recognition, the datasets adopted in different articles differ from one another, and the recording methods of gunshots in these datasets are quite different: different recording methods and places of the gunshot audios captured from video websites, recorded in the shooting range, and collected in the open countryside can have a huge influence on the final result of the model. Therefore, there is no comparability among different data sets in terms of the accuracy of gunshot recognition. The effects of some gunshot recognition methods are listed in Table 5, which are compared with their corresponding papers under the same dataset.

Table 5. Comparison of effects of multiple gunshot recognition networks.

Method	Dataset	Accuracy Rate	Ours Accuracy Rate
Gaussian mixture model (GMM) [16]	10 gun models and 100 gunshots	90%	94.1%
Hierarchical GMM classification [7]	5 gun models and 230 gunshots	96.29%	98.2%
HMM [5]	4 gun models and 372 gunshots	80%	90.1%
LS-LDA and maximum likelihood decision fusion [15]	14 gun models and 840 gunshots	94.1%	95.4%
Hidden Markov model (HMM) [17]	5 gun models and 46 gunshots	95.65%	96.5%
Convolutional neural networks [10]	59 gun models and 3655 gunshots	90%	93.2%
Relief feature selector [18]	8 gun models with 851 gunshots	94.48%	95.1%
Transfer learning [11]	18 gun models with 6000 gunshots	78.2%	83.2%

The confusion matrix of the validation set is shown in Figure 8. The recognition accuracy of the network for the WASR-10 gun type is relatively lower, making it difficult to distinguish the gunshot of the Remington700 from that of the WASR-10, probably because the caliber of the two guns is the same, and the Remington700 has a smaller sample size than other gun types.

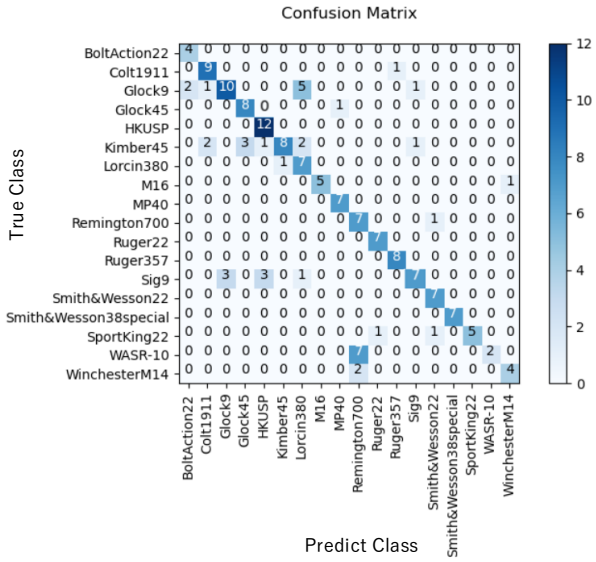


Figure 8. Validation set obfuscation matrix.

5. Conclusions

In this paper, we propose an efficient method of gunshot recognition, which uses multi-layer neural networks to extract the representation features in the log-mel gunshot spectrogram, and realizes the interaction between gunshot spectrum information through the internal spectrum displacement module of the network, so as to fully tap the relevant information between gunshot spectrum. This method avoids the loss of information in the down-sampling process and further improves the quality of the gun sound spectrum feature map. On the gunshot data set of the NIJ Grant 2016-DN-BX-0183 project, the identification accuracy of this method for the gun model can reach 80%, and the model size does not exceed 17 MB. We have proven that the network proposed in this paper has certain advantages in accuracy and model size, and has strong application values. In future research, we will further improve the gunshot recognition method, initialize the weight of the network model by means of large dataset pre-training, and reduce the number of network parameters by means of knowledge distillation, which may result in higher recognition accuracy and a smaller network model.

Author Contributions: Conceptualization, J.L., J.G., M.M., C.L. and Y.Z.; methodology, J.L., J.G., M.M. and C.L.; software, J.L. and J.G.; investigation, J.L., J.G., J.X., C.L. and Y.Z.; writing, J.L. and J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by the National Science Foundation of China (No. 61901419) and Fundamental Research Program of Shanxi Province (No. 20210302124031).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dufaux, A.; Besacier, L.; Ansorge, M.; Pellandini, F. Automatic sound detection and recognition for noisy environment. In Proceedings of the 2000 10th European Signal Processing Conference, Tampere, Finland, 4–8 September 2000.
2. Clavel, C.; Ehrette, T.; Richard, G. Events Detection for an Audio-Based Surveillance System. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005.
3. Valenzise, G.; Valenzise, G.; Gerosa, L.; Gerosa, L.; Tagliasacchi, M.; Antonacci, F.; Sarti, A. Scream and gunshot detection and localization for audio-surveillance systems. In Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2007, London, UK, 5–7 September 2007; pp. 21–26.
4. Busse, C.; Krause, T.; Ostermann, J.; Bitzer, J. Improved Gunshot Classification by Using Artificial Data. In Proceedings of the Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics, Porto, Portugal, 18–20 June 2019.
5. Kiktova, E.; Lojka, M.; Pleva, M.; Juhar, J.; Cizmar, A. Gun type recognition from gunshot audio recordings. In Proceedings of the 3rd International Workshop on Biometrics and Forensics (IWBF 2015), Rome, Italy, 6–7 May 2005; pp. 1–6.
6. Ahmed, T.; Uppal, M.; Muhammad, A. Improving efficiency and reliability of gunshot detection systems. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, UK, 26–31 May 2013.
7. Djeddou, M.; Touhami, T. Classification and modeling of acoustic gunshot signatures. *Arab. J. Sci. Eng.* **2013**, *38*, 3399–3406. [CrossRef]
8. Tokozume, Y.; Harada, T. Learning environmental sounds with end-to-end convolutional neural network. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017.
9. Erhan, D.; Szegedy, C.; Toshev, A.; Anguelov, D. Scalable Object Detection using Deep Neural Networks. In Proceedings of the 2013 IEEE Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
10. Raponi, S.; Ali, I.; Oligeri, G. Sound of Guns: Digital Forensics of Gun Audio Samples meets Artificial Intelligence. *arXiv* **2020**, arXiv:2004.07948. [CrossRef]
11. Lilien, R. *Development of Computational Methods for the Audio Analysis of Gunshots*; NCJRS; 2016-DN-BX-0183; National Institute of Justice: Washington, DC, USA, 2018.
12. Arslan, Y. Impulsive Sound Detection by a Novel Energy Formula and Its Usage for Gunshot Recognition. *arXiv* **2017**, arXiv:1706.08759.

13. Hawthorne, D.L.; Horn, W.; Reinke, D.C. A system for acoustic detection, classification, and localization of terrestrial animals in remote locations. *J. Acoust. Soc. Am.* **2016**, *140*, 3182. [CrossRef]
14. Aguilar, J. Gunshot Detection Systems in Civilian Law Enforcement. *J. Audio Eng. Soc.* **2015**, *63*, 280–291. [CrossRef]
15. Sánchez-Hevia, H.A.; Ayllón, D.; Gil-Pita, R.; Rosa-Zurera, M. Maximum likelihood decision fusion for weapon classification in wireless acoustic sensor networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1172–1182. [CrossRef]
16. Khan, S.; Divakaran, A.; Sawhney, H.S. Weapon identification using hierarchical classification of acoustic signatures. In *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense VIII*; Society of Photo-Optical Instrumentation Engineers(SPIE): Orlando, FL, USA, 15–17 April 2009.
17. Morton, K.D., Jr.; Torrione, P.A.; Collins, L. Classification of acoustic gunshot signatures using a nonparametric Bayesian signal model. In *Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense X*; Society of Photo-Optical Instrumentation Engineers(SPIE): Orlando, FL, USA, 25–28 April 2011.
18. Tuncer, T.; Dogan, S.; Akbal, E.; Aydemir, E. An Automated Gunshot Audio Classification Method Based on Finger Pattern Feature Generator and Iterative Relieff Feature Selector. *Adıyaman Üniversitesi Mühendislik Bilim. Derg.* **2021**, *8*, 225–243.