Research Article

CNN-Based Multiterrain Moving Target Recognition Model for Unattended Ground Sensor Systems

Kunsheng Xing, Nan Wang, Wei Wang, Yaoyao Miao, and Yaping Chen

College of Intelligent Science, National University of Defense Technology, Changsha 410073, China

Correspondence should be addressed to Nan Wang; acnwang@aliyun.com

Received 10 May 2022; Revised 30 July 2022; Accepted 2 August 2022; Published 31 August 2022

1. Introduction

The use of unattended ground sensor systems to achieve detection of moving targets in intrusion alert areas is an important research direction in the field of military and civilian security protection. The detected targets usually include humans, various types of vehicles, and ultra-low-altitude vehicles. Effective identification of moving targets is important for border control [1], protection of key facilities [2, 3], and reduction of illegal hunting. Among many detection means, seismic sensors are considered to be the most effective means of reconnaissance at present because of their long detection distance, low cost, light weight, and low power consumption. However, there are some shortcomings in using seismic signals as the only means of detection. The same target will exhibit different seismic signal characteristics under different ground geological conditions [4], and this characteristic poses a great challenge for identifying moving vehicles or humans in different ground geological conditions. Compared to seismic signals propagated by the ground as a medium, sound signals propagate in the air with little distortion. In addition, because of the damping effect of grass, the seismic signal triggered by a human walking on grass is very weak but can trigger a larger sound (see Figure 1). Therefore, adding sound signals as supplementary information to seismic signals can effectively improve the accuracy of recognition.

Among the ground target recognition techniques based on seismic or acoustic signals, machine learning is considered as the best solution at present. According to the development process of machine learning, machine learning can be divided into traditional machine learning (TML) and deep learning (DL). TML methods used for moving target recognition based on ground seismic signals include but
are not limited to the following: support vector machine (SVM) [5, 6], k-nearest neighbour (K-NN) [7], decision tree (DT) [8], naive Bayesian algorithm (NB) [9], artificial neural networks (ANN) [10], boost [11], and Gaussian mixture model (GMM) [12]. Although TML has high computational efficiency, its classification accuracy is highly dependent on the selection of features. And it is difficult for the manually selected features to portray the nonlinear relationship between the seismic signal and the target. DL is a new method emerging in recent years, among which convolutional neural networks are most widely used in image recognition, crack diagnosis [13, 14], synthetic aperture radar automatic target recognition [15–17], etc. Wang et al. proposed a 24-layer end-to-end convolutional neural network (CNN) model called Vib-CNN [18] to achieve the recognition of three types of moving targets (humans, wheeled vehicles, and tracked vehicles) for classification. Xu et al. proposed a parallel recurrent neural network (PRNN) to detect moving humans [19]. Experiments show that the algorithm achieves 98.4% accuracy for mobile person detection. Jin et al. used LFCC feature extraction and the CNN to validate the publicly available dataset SITEX02 with an accuracy of 92.28% [20]. Bin et al. proposed an edge-oriented intelligence approach named the compressed-aware edge convolutional neural network (CS-ECNN) [21]. The algorithm was successfully implemented for deployment on FPGAs.

Despite the excellent performance of these deep networks, there are still two problems that remain unsolved. First, these networks have poor generalization performance for moving target recognition over multiple terrains because the seismic waves generated by moving targets are susceptible to ground patterns. Second, with the deepening and widening of neural networks, the computational complexity of deep neural network models increases dramatically and unattended ground sensors have to adopt more costly hardware solutions. In response, this paper proposes a moving target recognition method called MFC-TinyNet that balances recognition accuracy, multiterrain generalization capability, and low complexity. The method consists of two parts, the Mel-frequency spectrum feature extraction and CNN. The Mel-frequency spectrum improves the generalization capability of the model for multiple terrains through multimodal information fusion of sound and seismic signals and achieves sensor recognition of moving human and vehicles on multiple terrains of concrete roads, asphalt roads, dirt roads, and grass. The CNN includes both traditional convolutional layers and depthwise separable convolutional layers. The convolutional neural network with the addition of the depthwise separable convolutional layer can effectively reduce the size of the network while maintaining the same target recognition accuracy, solving the problem that the model size is too large to be deployed on low-power hardware devices.

The rest of this paper is structured as follows. Section 2 will present a new dataset and describe the signal acquisition device for the dataset, the signal acquisition environment, and the way that the dataset is constructed. Section 3 focuses
on the identification method of this paper. Section 4 shows the experimental results of the proposed method in this paper with other traditional machine learning and deep learning methods on this dataset. Section 5 will conclude this study and provide an outlook on future research.

2. Multitopography Dataset

The dataset used in this paper is collected by an unattended ground sensor (see Figure 2) developed by our team. The hardware equipment of this sensor and the dataset acquisition environment will be described in this section.

2.1. Unattended Ground Sensor Hardware. The unattended ground sensor consists of three parts: sensor module, processor module, and communication module. In the process of acquiring sound and seismic signals, the unattended ground sensors communicate with the PC via USB connection. The sensor module consists of a MEMS accelerometer (seismic sensor), an audio sensor, and an AD converter. The processor module contains the storage unit, the processing unit, and the microprocessor. The microprocessor is a 32-bit high-performance STM32H743 based on the ARM core. In the moving target identification stage, the identification results will be sent to the command center through wireless communication. The communication module uses Zigbee self-organizing network communication, and the on-board communication chip model is CC2630. The composition of the sensor nodes is shown in Figure 3.

The sound sensor type is a MEMS microphone, and the seismic sensor type is an MEMS accelerometer. Since the frequency of the seismic and sound signals triggered by the movement of vehicles and human is mainly concentrated below 500 Hz [4], the sampling frequency of both sound and vibration signals is 1015 Hz. The signal collected by the seismic sensor is the acceleration signal of vibration, the sensitivity is 900 mV/g, the full scale acceleration is ±3 g, and the maximum noise is 8 μg. The sensitivity of the sound sensor is ~38 dB, the maximum sound pressure level is 130 dB, and the signal-to-noise ratio is 65 dB.

2.2. Recording Environments. The terrain in the dataset will include four different road environments: concrete road, tarmac road, dirt road, and grassland (see Figure 4). The types of moving targets collected include humans, small tracked vehicles, and background noise (no targets). The location of the dataset collection is located in Changsha city, Hunan province.

The unattended ground sensor will record the sound and seismic signals as the target passes through the alert zone, which is a circular area centered on the unattended ground sensor. The diameter of the warning area for humans is 10 m, and the diameter of the warning area for small tracked vehicles is 15 m.

2.3. Data Preprocessing. Data preprocessing consists of two steps: DC bias removal and dataset construction. To improve the accuracy of model classification, the DC bias generated by the acquisition instrument needs to be removed from the original signal. First, a fast Fourier transform is applied to the signal to remove the zero frequency signal. Then, the sound and seismic signals after removing the DC component are obtained using the fast Fourier inverse transform. Next is the construction of the dataset, which contains two parts: the training set and the test set. The training set is the data sample used for model fitting, and the test set is used to evaluate the generalization ability of the final model. To ensure a more effective evaluation of the generalization ability of the model, the training set and the test set are composed of sound and seismic signals collected at different times. Table 1 shows the size of the processed dataset.

The left side of the bracket in the table is the dataset of the training datasets of this kind of data, and the right side of the bracket is the size of the test datasets of this kind of data.

3. MFC-TinyNet Algorithm Analysis

In this section, a ground target recognition method is presented. In order to overcome the poor generalization performance of different terrains and achieve model lightweight, MFC-TinyNet is proposed by referring to feature extraction methods in the field of speech recognition and image classification methods in computer vision. As seen in Figure 5, MFC-TinyNet takes the original sound and vibration signals as the input of the model, extracts the two-dimensional time-frequency domain feature matrices of the sound and seismic signals of the same time period using the Meier inverse spectrum feature extraction method, and then fuses the two two-dimensional feature matrices to form a three-dimensional feature matrix. The feature matrix is then fed into a convolutional neural network. After the computation

![Figure 2: Unattended ground sensors.](image)

![Figure 3: Unattended ground sensor hardware structure.](image)
of the convolutional neural network, the recognition result of the moving target is output.

The Mel-frequency spectrum is a sound feature extraction method based on the physical feature of the human ear. It is widely used in the fields of speech recognition, sound classification, music genre classification, and voiceprint recognition. As shown in Figure 6, the extraction of the Mel-frequency spectrum is divided into the following steps: pre-emphasis, short-time Fourier transform (STFT), Mel filter bank, and logarithm [17]:

(1) Pre-emphasis

The transmission of the seismic signal and sound signal in the medium is accompanied by the loss of energy. When the seismic and sound signals have a certain scale, the higher the signal frequency, the more serious the energy loss during the propagation process. Therefore, the high-frequency part of the input signals is compensated to offset the loss of the high-frequency part in the propagation process. This signal processing method is called pre-emphasis. The transfer function of the pre-emphasis system is shown in equation (1) as follows:

$$H(z) = 1 - az^{-1}. \quad (1)$$

(2) Short-time Fourier transform

Windowing and framing the pre-emphasis signals, then, fast Fourier transforms each frame signal to extract its amplitude spectrum. Generally, the window used is the Hamming window. The window length is 20 ms to 40 ms, and the window is moved to about half of the window length.

(3) Mel filter bank

The Mel filter bank refers to the filter bank following Mel scale distribution. The Mel scale is a nonlinear scale used to describe the different sensitivity of the human ear to sound in different frequency domains. Because the traditional Mel-frequency spectrum feature is aimed at speech signal
**Figure 6**: Mel-frequency spectrum extraction process.

**Figure 7**: Depthwise separable convolutional layer.
recognition, it is different from the moving target recognition based on sound and seismic signals in this subject, so it is necessary to modify the frequency parameters of the Mel scale distribution formula. The distribution formula of the Mel scale is modified as follows:

\[
m = 2595 \times \ln \left(1 + \frac{f}{500}\right),
\]

where \(f\) is the frequency expressed in Hertz and \(m\) is the corresponding Mel frequency. In practice, we use 50 Mel frequencies.

(4) Take the logarithm

Take the logarithm of the data output by the Mel filter bank to obtain the Mel-frequency spectrum feature.

The seismic and sound signals are extracted with Mel-frequency spectrum features and then will be input into a CNN to obtain classification results. Based on the performance characteristics of seismic and sound signals on the Mel-frequency spectrum, a convolutional neural network named TinyNet is proposed in this paper. The network has two convolutional blocks and two fully connected layers in total, where each convolutional block includes one traditional convolutional layer, a depth-separable convolutional layer [22], and a

![Network structure](image)
maximum pooling layer. Deep convolution consists of depth convolution and point convolution. Each convolution kernel convolves only the current input channel, and the number of output feature map channels is exactly the same as the input. The size of the convolution kernel is \(1 \times 1 \times M\), and \(M\) is the number of channels in the previous layer. Therefore, the output of pointwise convolution will combine depth convolution in a weighted manner and the output image channels are the same as the pointwise convolutional filter channels (see Figure 7). Adding the depthwise convolution and pointwise convolutional layer to the convolutional block can significantly reduce the computational effort and network layer parameters with essentially no loss of recognition accuracy. All hidden layers use the ReLu function as the nonlinear activation function, and the output layer uses the Sigmoid activation function. To solve the overfitting problem of the network, dropout functions are added after each convolutional block and fully connected layer. The MFC-TinyNet network structure is shown in Figure 8.

The network uses the cross-entropy loss as the objective function. Therefore, the loss function of the classifier is as follows:

\[
L(\Theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} t_{ij} \log \left( p(y_{ij}|x_i; \Theta) \right),
\]

where \(N\) is the number of samples, \(C\) is the number of target types, \(t_{ij}\) represents the probability that the \(i\) sample belongs to category \(j\), and \(\Theta\) represents a collection of all parameters.

4. Experimental

4.1. Evaluation Metrics. In the unattended ground sensor system, the accuracy rate, false alarm rate, and underreporting rate are usually used to evaluate the advantages and disadvantages of the model. Accuracy [23] refers to the proportion of all correctly predicted samples in the total number of samples, and its calculation formula is \(\text{ACC} = (TP + TN)/(TP + TN + FP + FN)\). Among them, correctly classifying intrusion events is called true positive (TP), correctly classifying noise samples is called true negative (TN), missing intrusion events are called false negatives (FN), and misjudging noise as an intrusion event is called false positive (FP). False alarm rate refers to the proportion of samples that misjudge noise as intrusion events in the total number of intrusion events. The calculation formula is \(\text{FAR} = FP/(TP + TN)\). The underreporting rate refers to the proportion of samples that misjudge the intrusion event as noise in the total number of samples. The calculation formula is \(\text{UR} = FN/(TP + TN)\).

4.2. Multitopography Dataset Experimental Results

(1) The result of MFC-TinyNet

Using sound signal and seismic signal as the input of the MFC-TinyNet model, the model is trained with 200 epochs. The model basically reached a stable state after 50 epochs. Then, the classification ability of the trained model is evaluated with the test datasets. The median results of 11 repeated
Only the seismic signal is used as the input of the MFC-TinyNet model, and the model is trained with 200 epochs. In order to avoid random factors, the median results of 11 repeated trials were selected for discussion. The performance of the model on the test datasets is shown in Figures 11 and 12. Its overall accuracy is 94.93%, the false alarm rate is 3.87%, and the underreporting rate is 1.20%.

Comparison With the Benchmark Methods: To verify the effectiveness of MFC-TinyNet, the classification performance of the state-of-the-art machine learning algorithms proposed by other scholars, in the field of ground-based moving target recognition based on seismic signals, is used as a comparison with the classification performance of MFC-TinyNet. These machine learning algorithms include genetic algorithm-optimized SVM (GA-SVM) [6], Vib-CNN [18], LFCC-CNN [20], and CS-ECNN [21]. Among them, GA-SVM is a traditional machine learning method that uses the genetic algorithm to optimize the penalty parameters and Gaussian kernel parameters of the Gaussian kernel support vector machine. CS-ECNN and LFCC-CNN are non-end-to-end deep neural networks that use compressed sensing and LFCC as feature extraction methods, respectively. Vib-CNN is an end-to-end deep learning method. Each model will use multitopography datasets as the training set and test set of the model. Each model will undergo 11 experiments, and each experiment will be trained for 200 epochs. The median classification accuracy of the test set of the 11 experiments is discussed as the classification effect of the model, and the computation time of the algorithm on the test dataset is recorded. All methods are run on a desktop computer with an Intel Core i710th CPU and a GeForce RTX 2070S GPU.

The classification results of each model are shown in Table 2. The classification accuracies of GA-SVM and CS-ECNN are 74.09% and 72.09%, respectively, which is much lower than the accuracies of other deep learning algorithms, and the generalization ability for the multiterrain is poor. LFCC-CNN classification accuracy is 84.86%, which basically achieves the effect of identifying moving targets on...
Table 2: Performance comparison between the proposed method and the benchmark methods.

| Model                                      | Accuracy | False alarm rate | Underreporting rate | Training times | Testing times |
|--------------------------------------------|----------|------------------|---------------------|----------------|---------------|
| GA-SVM                                     | 74.09%   | 13.80%           | 12.11%              | 52107 s        | 1614 ms       |
| CS-ECNN                                     | 72.09%   | 1.51%            | 27.21%              | 513.96         | 2014 ms       |
| LFCC-CNN                                    | 87.77%   | 0.15%            | 12.08%              | 411.73 s       | 8612 ms       |
| Vib-CNN                                     | 98.16%   | 0.43%            | 1.41%               | 43298 s        | 11131 ms      |
| MFC-TinyNet (without sound signal)         | 94.93%   | 3.87%            | 1.20%               | 480.93 s       | 1626 ms       |
| MFC-TinyNet (the method of this paper)     | 98.49%   | 0.43%            | 1.08%               | 493.58 s       | 2613 ms       |

5. Conclusion and Future Work

In this paper, we propose a ground moving target recognition method based on the Mel-frequency spectrum and CNN for distinguishing environmental noise (no target), human intrusion, and tracked vehicle intrusion. And the sound and seismic signals of a set of moving targets are recorded by a self-researched UGS. The accuracy of the method for target classification reached 98.49% after experimental validation. Compared with other deep learning methods, this method has two advantages of low hardware requirements for implementation and applicability to a variety of terrain conditions. However, the proposed method in this paper is only limited to identifying the type of intrusion target and cannot determine the relative position of the intrusion target and the sensor. Since the dataset collection site in this paper is located in an urban area, there are a large number of humans moving around the periphery of the unattended ground sensor alert area, resulting in a low signal-to-noise ratio in the dataset and a small alert radius for the unattended ground sensors. If the ambient noise and moving target signals are implemented, the alert area will be greatly enhanced. In future work, based on the difference between environmental noise and target signal in the frequency domain, the separation of environmental noise and target signal using deep clustering method will be a promising research direction. In addition, the combination of sensor arrays, generalized mutual correlation time delay estimation techniques, and deep neural networks will likely enable the localization of multiple intrusion targets.

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study is supported by the National Defense Science and Technology Key Laboratory Fund Project (WDZC20215250302).

References

[1] A. A. Yahoo and B. B. G. C. ComNicolae, "Unattended ground sensor borders—the forgotten solution for Afghanistan," International Conference Knowledge-Based Organization, vol. 27, no. 1, pp. 8–13, 2021.
[2] R. Levy, J. Moras, and B. Pannetier, "Vibrating beam MEMS seismometer for footstep and vehicle detection," IEEE Sensors Journal, vol. 17, no. 22, pp. 7306–7310, 2017.
[3] F. Li, J. Clemente, M. Valero, Z. Tse, and W. Z. Song, "Smart home monitoring system via footstep induced vibrations," IEEE Systems Journal, vol. 14, pp. 3383–3389, 2019.
[4] Y. Tian and H. Qi, "In target detection and classification using seismic signal processing in unattended ground sensor systems," International Conference on Acoustics, vol. 4, pp. 4172–4172, 2002.
[5] K. Bin, Y. Long, X. Tong, and J. Lin, "Ground moving target detection with seismic fractal features," Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2021.
[6] Z. Zhong and H. Li, "Recognition and prediction of ground vibration signal based on machine learning algorithm," Neural Computing and Applications, vol. 32, pp. 1937–1947, 2020.
[7] A. R. S. G. Jin, "Target detection and classification using seismic and PIR sensors," IEEE Sensors Journal, vol. 12, pp. 1709–1718, 2012.
[8] R. Narayanaswami, A. Gandhe, A. Tyurina, and R. K. Mehra, "Sensor fusion and feature-based human/animal classification for unattended ground sensors," in 2010 IEEE International Conference on Technologies for Homeland Security (HST), pp. 344–350, Waltham, MA, USA, 2010.
[9] T. Damarla, A. Mehmood, and J. Sabatier, "Detection of people and animals using non-imaging sensors," in 14th International Conference on Information Fusion, pp. 1–8, Chicago, IL, USA, 2011.
[10] V. W. Porto, L. J. Fogel, and D. B. Fogel, "Using evolutionary computation for seismic signal detection: a homeland security application," in IEEE International Conference on Computational Intelligence for Homeland Security & Personal Safety, pp. 62–66, Venice, Italy, 2004.

[11] H. O. Park, A. A. Dibazar, and T. W. Berger, "Cadence analysis of temporal gait patterns for seismic discrimination between human and quadruped footsteps," in 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1749–1752, Taipei, Taiwan, 2009.

[12] A. A. Dibazar, A. Yousefi, H. O. Park, L. Bing, S. George, and T. W. Berger, "Intelligent acoustic and vibration recognition/alert systems for security breaching detection, close proximity danger identification, and perimeter protection," in 2010 IEEE International Conference on Technologies for Homeland Security (HST), pp. 351–356, Waltham, MA, USA, 2010.

[13] Y. Yu, C. Wang, X. Gu, and J. Li, "A novel deep learning-based method for damage identification of smart building structures," Structural Health Monitoring, vol. 18, no. 1, pp. 143–163, 2019.

[14] Y. Yu, M. Rashidi, B. Samali, M. Mohammadi, T. N. Nguyen, and X. Zhou, "Crack detection of concrete structures using deep convolutional neural networks optimized by enhanced chicken swarm algorithm," Structural Health Monitoring, vol. 1, no. article 147592172110535, 2022.

[15] M. Amrani, A. Bey, and A. Amamra, "New SAR target recognition based on YOLO and very deep multi-canonical correlation analysis," International Journal of Remote Sensing, vol. 42, pp. 1–20, 2021.

[16] M. Amrani, F. Jiang, Y. Xu, S. Liu, and S. Zhang, "SAR-oriented visual saliency model and directed acyclic graph support vector metric based target classification," Ieee J-Stars, vol. 11, pp. 3794–3810, 2018.

[17] M. Amrani and F. Jiang, "Deep feature extraction and combination for synthetic aperture radar target classification," Journal of Applied Remote Sensing, vol. 11, no. 4, p. 42616, 2017.

[18] Y. Wang, X. Cheng, P. Zhou, B. Li, and X. Yuan, "Convolutional neural network-based moving ground target classification using raw seismic waveforms as input," IEEE Sensors Journal, vol. 19, no. 4, pp. 5751–5759, 2019.

[19] T. Xu, N. Wang, and X. Xu, "Seismic target recognition based on parallel recurrent neural network for unattended ground sensor systems," IEEE Access, vol. 7, pp. 137823–137834, 2019.

[20] G. Jin, B. Ye, Y. Wu, and F. Qu, "Vehicle classification based on seismic signatures using convolutional neural network," IEEE Geoscience and Remote Sensing Letters, vol. 16, pp. 628–632, 2019.

[21] K. Bin, J. Lin, and X. Tong, "Edge intelligence-based moving target classification using compressed seismic measurements and convolutional," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2022.

[22] A. G. Howard, M. Zhu, B. Chen et al., "MobileNets: efficient convolutional neural networks for mobile vision applications," 2017, http://arXiv.1704.04861.

[23] S. Liu, W. Jiang, L. Wu, H. Wen, and Y. Wang, "Real-time classification of rubber wood boards using an SSR-based CNN," IEEE Transactions on Instrumentation and Measurement, vol. 69, pp. 8725–8734, 2020.