Optical Engineering

OpticalEngineering.SPIEDigitalLibrary.org

Pattern recognition based on time-frequency analysis and convolutional neural networks for vibrational events in *φ*-OTDR

Chengjin Xu
Junjun Guan
Ming Bao
Jiangang Lu
Wei Ye

SPIE.

Chengjin Xu, Junjun Guan, Ming Bao, Jiangang Lu, Wei Ye, “Pattern recognition based on time-frequency analysis and convolutional neural networks for vibrational events in *φ*-OTDR,” *Opt. Eng*. **57**(1), 016103 (2018), doi: 10.1117/1.OE.57.1.016103.

Pattern recognition based on time-frequency analysis and convolutional neural networks for vibrational events in φ-OTDR

Chengjin Xu,[a] Junjun Guan,[b] Ming Bao,[b] Jiangang Lu,[a] and Wei Ye[a,*]
[a]Zhejiang University, College of Control Science and Engineering, Hangzhou, China
[b]Chinese Academy of Sciences, Institute of Acoustics, Beijing, China

Abstract. Based on vibration signals detected by a phase-sensitive optical time-domain reflectometer distributed optical fiber sensing system, this paper presents an implement of time-frequency analysis and convolutional neural network (CNN), used to classify different types of vibrational events. First, spectral subtraction and the short-time Fourier transform are used to enhance time-frequency features of vibration signals and transform different types of vibration signals into spectrograms, which are input to the CNN for automatic feature extraction and classification. Finally, by replacing the soft-max layer in the CNN with a multiclass support vector machine, the performance of the classifier is enhanced. Experiments show that after using this method to process 4000 vibration signal samples generated by four different vibration events, namely, digging, walking, vehicles passing, and damaging, the recognition rates of vibration events are over 90%. The experimental results prove that this method can automatically make an effective feature selection and greatly improve the classification accuracy of vibrational events in distributed optical fiber sensing systems. [DOI: 10.1117/1.OE.57.1.016103]

Keywords: optical fiber sensing system; φ-optical time-domain reflectometer; convolutional neural network; time-frequency analysis; pattern recognition.

Paper 171083 received Jul. 13, 2017; accepted for publication Dec. 7, 2017; published online Jan. 3, 2018.

1 Introduction

Distributed optical fiber sensing techniques have been widely used in many fields, such as security,[1] military, and pipeline transportation,[2] because of the optical fiber's small size, light weight, lack of power supply requirement, antielectromagnetic interference, high sensitivity,[3] and long detection distance. Among the various systems, the distributed optical fiber sensing system based on phase-sensitive optical time-domain reflectometer (φ-OTDR), with the advantages of a simple structure, multipoint vibration detection and positioning,[4–7] and high positioning accuracy, has become the new hotspot in the present distributed optical fiber sensing research field.

Vibration signals collected in distributed optical fiber sensing systems can reflect the characteristics of vibration sources. Thus, the pattern identification of vibrational events is an achievable function in distributed optical fiber sensing systems. The difficulty in achieving this function is figuring out the best methods of feature extraction. The existing feature extraction methods are based on some certain time-domain features or frequency-domain features.[8–9] Considering that vibration signals are nonstationary and manual feature extraction needs expert knowledge, using time-frequency analysis to obtain time-frequency diagram of vibration signals and automatically extracting deep features of time-frequency diagram by convolutional neural network (CNN) is a more efficient and more universal method.

The CNN is one of the most important models in the area of deep learning, especially in image processing. It is a type of feed-forward artificial neural network designed to recognize two-dimensional data. Its unique network structure can resist influences from image translation, scaling, or distortion in some degree. Krizhevsky et al.[10] use a deep CNN with rectified linear units (reLU) defined as an activation function for natural image classification and obtain the best result in ImageNet competition. Due to proved success of CNN in image recognition,[11] it is also successfully applied to speech recognition, face recognition, target detection, and some other fields.[12] Badshah et al.[13] use time-frequency representation of a speech signal as input and a CNN as a classifier to detect speech emotion. This method combining time-frequency analysis and CNN has a lower model complexity and a better recognition performance.

In this paper, we use time-frequency analysis to transform different types of vibration waveform data collected by an intensity-demodulated φ-OTDR into corresponding image data which express more abundant information about time-frequency characteristic and then classify them by a CNN. Spectrum subtraction and support vector machine (SVM) are implied to improve the algorithm performance. As experimental results show, this method can accurately classify four vibration event patterns, namely, knocking, shaking, crushing, and watering, under conditions of low signal-to-noise ratio (SNR) and can overcome the shortcoming of manual feature selection of the existing pattern recognition methods in distributed optical fiber sensing systems.

Moreover, considering that phase-demodulated φ-OTDRs have better sensitivity and higher SNRs than intensity-demodulated φ-OTDRs and can perfectly express the

*Address all correspondence to: Wei Ye, E-mail: wye@zju.edu.cn

0091-3286/2018/$25.00 © 2018 SPIE

frequency characteristics of the vibration sources, which are also the input characteristics of the recognition method proposed in this paper,[6,7] we strongly believe that this method can be applicable for distributed optical fiber sensing systems based on phase-demodulated φ-OTDRs.

2 Principle of Classification

2.1 *Time-Frequency Analysis*

Time-frequency diagrams obtained by time-frequency analysis show the combined information in time domain and frequency domain, which directly reflect change of the frequency components of the signals with time. Common time-frequency transform methods include short-time Fourier transform (STFT), continuous wavelet transform, and S-transform. Because spectrograms obtained by STFT have succeeded in practical application in the fields of music, sonar, radar,[12] and speech processing[13] and has a rich image color information that is helpful to image recognition, we choose STFT as the time-frequency analysis method in this paper.

2.1.1 *Short-time Fourier transform*

The STFT is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. In practice, the procedure for computing STFTs is to divide a longer time signal into shorter segments of equal length by a window function, which is nonzero for only a short period of time and then compute the Fourier transform separately on each shorter segment. Mathematically, this is written as

$$\mathrm{STFT}(t,f)=\int_{-\infty}^{\infty}x(\tau)w(t-\tau)e^{-j2\pi ft}\mathrm{d}\tau, \tag{1}$$

where $x(t)$ is a time-domain signal, $w(t)$ is the window function, commonly a Hamming window or a Gaussian window centered around zero, and τ is the center of the window function.

When the window function of STFT has been chosen, the resolution is fixed as long as the window length of the window function is fixed. Since the time resolution is inversely proportional to the frequency resolution, they are unable to be optimal at the same time. When the window length becomes longer, the time resolution becomes lower, and when the window length becomes shorter, the frequency resolution becomes higher. Thus, it is important to choose an appropriate type of window function and a reasonable window length.

2.1.2 *Spectrogram*

After the signal in time domain in each segment is transformed to the frequency domain by fast Fourier transform (FFT), the amplitude frequency is rotated by 90 deg, and the spectral amplitude mapped to a gray level (0 to 255) value. The higher the amplitude is, the darker the corresponding region. Time-frequency representation of a signal is referred to as spectrogram.

Due to the randomness of the vibration signal and the complexity of its frequency spectrum performance, we use the Hamming window function to segment the signal into the small frames. The Hamming window function can reduce the discontinuities of the signal at the edge of frame and eliminate high-frequency interference and energy leaks. It is formulated with the following equation:

$$w(n)=0.54-0.46\cos[2n\pi/(N-1)], \qquad 0\le n\le N-1, \tag{2}$$

where N is the length of the window.

By using high-resolution FFT to process the vibration signal, we found that the lowest frequency of its spectrum peak was around 25 Hz. Therefore, the frequency resolution should be lower than 25 Hz. With a system sampling rate of 5000 Hz, the maximum value of the number of FFT could be calculated as 200. To obtain a high-frequency resolution and facilitate the computing, the number of FFT is set as 128 and the time length of each frame is 64 ms.

Take a digging signal as an example. The total time length of the signal is 2.4 s. After using STFT to transform the time-domain signal to the frequency domain, the time-frequency representation of the digging signal is shown in Fig. 1.

2.1.3 *Spectral subtraction*

Spectral subtraction[14] is the most common method for dealing with wide-band noise. Its principle is obtaining the spectrum of the pure speech by subtracting the estimated noise

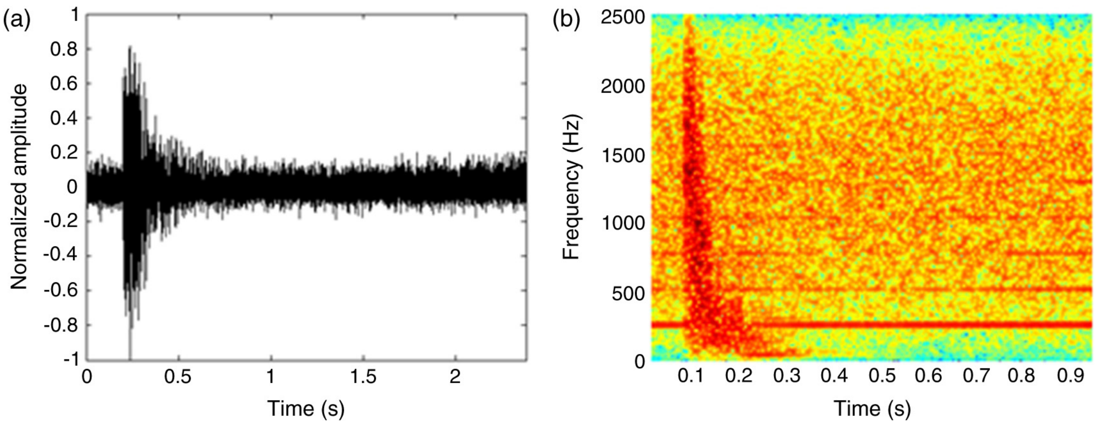

Fig. 1 An example of a knocking signal: (a) the original time-domain waveform of the knocking signal and (b) the spectrogram of the knocking signal.

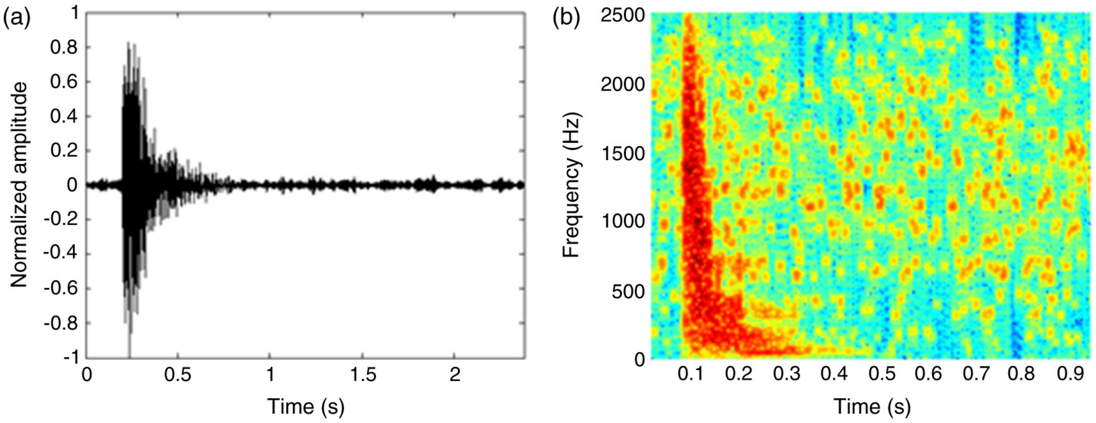

Fig. 2 The effect of spectral subtraction on the vibration signal: (a) the time-domain waveform of the knocking signal after noise reduction and (b) the spectrogram of the knocking signal after noise reduction.

amplitude spectrum from the amplitude spectrum of the noisy signal. The effect is equivalent to carrying out a type of equalization to the noisy signal in the transform domain. Compared with other noise reduction methods, spectral subtraction has the least constraint, the most direct physical meaning, the least amount of calculation, and the desired effect.

Assuming that the correlation between noise in the noisy signal and the vibration signal itself is zero, that is, the two are independent of each other and additive in the frequency domain, the additive model of the signal can be expressed as

$$X_w(\omega) = S_w(\omega) + N_w(\omega), \tag{3}$$

where $X_w(\omega)$ is the short-time Fourier spectra of the small frame of noisy signal, $S_w(\omega)$ and $N_w(\omega)$ are the short-time Fourier spectra of noise component and effective component of the signal frame, respectively.

The magnitude squared of the short-time Fourier spectrum gives the short-time power spectrum

$$|Y_w(\omega)|^2 = |S_w(\omega)|^2 + |N_w(\omega)|^2 + S_w(\omega)N_w^*(\omega) + S_w^*(\omega)N_w(\omega). \tag{4}$$

Since $s(m)$ and $n(m)$ are independent of each other, the cross-correlation statistical mean is 0. The short-time power spectrum of the pure signal is estimated as

$$|\hat{S}_\omega(\omega)|^2 = |Y_\omega(\omega)|^2 - E[|N_\omega(\omega)|^2], \tag{5}$$

where $S_\omega(\omega)$ is the short-time power spectrum estimation of the pure signal.

The phase is restored and then the inverse Fourier transform is used to obtain the denoised time-domain signal

$$\hat{s}(m) = \mathrm{IFFT}[|\hat{s}_w(\omega)|e^{i\varphi(\omega)}]. \tag{6}$$

Since the noise is locally stationary, to estimate the noise power spectrum, it is assumed that the noise at no vibration is close to the noise at the time of vibration. Thus, the noise power spectrum at the time of vibration can be estimated from the short-time power spectrum of the silent frame without vibration according to Welch's method.

After using spectral subtraction to denoise the original knocking signal in Fig. 1, the time-frequency presentation of the denoised signal is shown in Fig. 2.

It can be seen from Fig. 2(a) that the signal quality improves after noise reduction, and meanwhile the signal intensity and time-domain characteristics of the effective signals are not weakened. Figure 2(b) shows the broadband noise in noisy signals is significantly suppressed and other noise with multiple harmonics generated by the acquisition process is eliminated. By this method, adverse effects from noise in vibration signals on the spectrogram classification are reduced.

2.2 *Convolutional Neural Network*

CNN is a hierarchical neural network composed of a sequence of layers. A typical model usually consists of several convolutional layers where image contents are represented as a set of feature maps obtained after convolving the input with a variety of filters, which are learned during the training phase. Pooling layers are introduced after convolutional layers to reduce the image size and accumulate maximum activation features from convolutional feature maps. Furthermore, CNNs may also contain fully connected (FC) layers where each neuron of the input layer is connected with every neuron in the layer. A sequence of convolutional, pooling, and FC layers form a feature extraction pipeline that models the input data in abstract form. Finally, a soft-max layer performs the final classification task based on this representation.

The proposed CNN model, shown in Fig. 3, has a structure similar to that used in AlexNet. The network contains five convolutional layers, two FC layers, and a soft-max layer. Each convolutional layer is followed by an ReLU and a max-pooling layer. The ReLU is a nonlinear function expressed as $f(x) = \max(0, x)$. It is applied as an activation function after the convolutions. Max-pooling following the ReLU passes on the maximum value in each 2×2 block. The first convolutional layer takes the $224 \times 224 \times 3$ image and applies 1285×5 filters. Through an ReLU and a 2×2 max-pooling layer, the volume of the original image becomes $110 \times 110 \times 128$. The next two convolutional layers both have 1283×3 filters and are followed by an ReLU and 2×2 max-pooling, resulting in a $26 \times 26 \times 128$ image volume. The resulting volume is fed into the

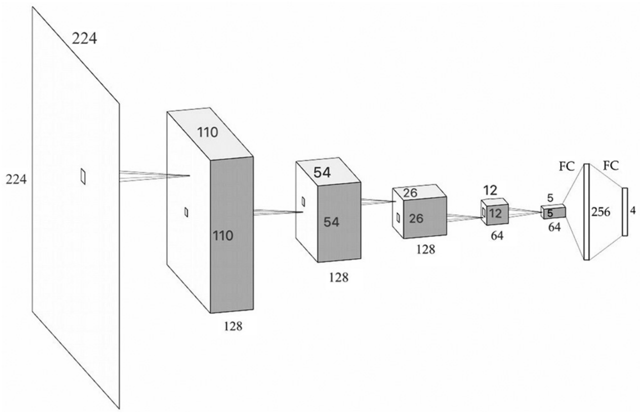

Fig. 3 Visual depiction of the Image Network architecture (the soft-max classifier is not pictured).

following two layers, each of which consists of 643×3 filters, an ReLU, and a 2×2 max-pooling layer. After all convolution and pooling, the $4 \times 4 \times 64$ image volume is flattened into a 1024×1 vector. The first and second FC layers reduce the size of the feature vector to 256 and 4 in turn, and then the soft-max classifier takes the 4×1 vector and outputs the final result.

2.3 *Support Vector Machine*

SVM[15] has advantages in the small sample, nonlinear, and high-dimensional expression and is widely used in pattern recognition. Its basic idea is that there is always an optimal hyperplane to separate a binary dataset through the support vector. Assuming that the given N-classes training sample set is inseparable in the low-dimensional space, it is expressed as (x_i, y_i), where $i = 1,2,\ldots,n$, n is the training sample number, x_i represents the input training samples, and y_i represents the class markers of data samples. The input samples are projected using nonlinear mapping onto a high-dimensional space R^n to construct an optimal hyperplane as

$$f(x) = sign\left[\sum_{s} w^T \cdot K(x_i, x) + b\right], \tag{7}$$

where *sign* means two-value functions, w is a normal vector to the hyperplane, $K(x_i, x)$ is the kernel function, b is the hyperplane position, and s is the support vector.

In the binary classification problem, classification results are determined by the symbol value after data samples to be measured are input to the classification function. To deal with N-classes classification problems (where $N > 2$), the combinatorial classifier is often used. This type of classifier is a combination of all possible binary subclassifiers, which are $N(N-1)/2$ in total. Subsequently, the voting method is used to determine the class with the most votes, which is used as the class of the input data sample.

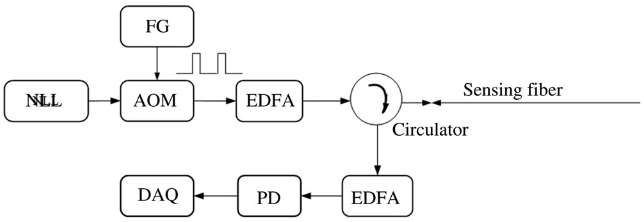

Fig. 4 The experimental setup of a φ-OTDR system: NLL, narrow linewidth laser; AOM, acousto-optic modulator; PD, photodetector; FG, function generator; EDFA, erbium-doped fiber amplifier; and DAQ, data acquisition card.

3 Experimental Setup

The experimental system shown in Fig. 4 uses a narrow linewidth erbium-doped laser with a wavelength of 1550 nm, a linewidth of 0.1 kHz, and an output power of 40 mW. The continuous light emitted by the laser is modulated by an acousto-optic modulator and a function generator into pulsed light with a pulse width of 200 ns and a frequency of 5 kHz. Then, the pulse light is amplified by an erbium-doped fiber amplifier (EDFA) and launched into a single-mode sensing fiber with a total length of 40 km by a circulator. The backscattered Rayleigh light generated from the sensing fiber is amplified by the EDFA again and finally converted into an electrical signal by a photodetector. The received signal traces are sampled by a data acquisition card with a 25-MHz sampling rate. Monitoring software was developed for data processing and alarm display.

4 Experimental Procedure

4.1 *Algorithm Execution Process*

The flowchart of vibration signal identification in distributed optical fiber sensing system is shown in Fig. 5.

First, a differential operation and a normalization operation are performed on each data sample as a preprocessing step to obtain the normalized vibration waveform without DC offset. Subsequently, we apply spectral subtraction to reduce the noise in the vibration signals. The denoised

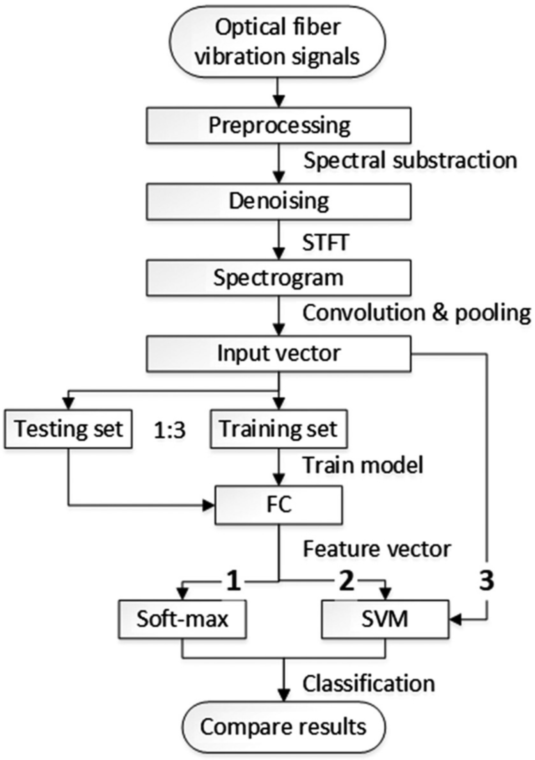

Fig. 5 The classification algorithm flowchart.

signals are transformed into spectrograms by STFT. And then all spectrograms are randomly divided into a training set and a testing set, the proportion of which are 75% and 25%, respectively. Next, the proposed CNN is trained on the training set. In the training process, spectrograms are flattened into feature vectors after all convolutional, pooling, and FC layers. There are three following methods used for classification:

1. A soft-max layer performs the classification task, the input of which is the output of the FC layer.
2. Replace the soft-max classifier with an SVM, the input of which is still the output of the FC layer.
3. Replace the soft-max classifier with an SVM, the input of which is the input of the FC layer.

4.2 *Vibration Event Simulation Experiments*

A segment of a 300-m-long optical fiber, led out in the vibration position of 20 km in the sensing fiber, was buried in the shallow soil of 10-cm depth. Striking the ground above the buried sensing fiber with shovels simulated digging events. Walking on the same position simulated walking events. Driving a car through this position simulated vehicle-passing events. Shaking the sensing fiber exposed on the ground simulated fiber-damage events.

Two hundred trials were carried out in each of the four vibration modes in the vibration position and around the vibration position, and data samples collected in the nearest five consecutive position nodes were stored. Through this method, 4000 data samples were obtained after all trials. Considering the fact that the system should respond to vibration events within 3 s, the duration and the total sampling point number of a single data sample were set as 2.4 s and 4800, respectively.

4.3 *Time-Frequency Analysis*

We used spectral subtraction to reduce the noise in the vibration signals with the estimated noise power spectrums that were estimated by the average power spectrums of the optical signals within the last 0.6 s before the occurrence of vibration events in the corresponding position nodes. Then the denoised vibration signal waveforms were transformed into spectrograms by STFT. During the STFT, both the length of window function and the number of FFT point were 128, and the overlapping rate was 50%. The resolution of spectrograms was 224 × 224, and the transformation process was encoded in Python and takes 0.3 s for each spectrogram. The typical spectrograms of all types of vibration signals after noise reduction are shown in Fig. 6.

4.4 *Network Training*

Two separate datasets comprised of 4000 spectrograms from original signals and 4000 spectrograms from denoised signals, respectively. Each dataset was assigned into a training set and a testing set. For each vibration mode, 750 spectrograms generated in 150 trials were selected as part of the training set and other 250 sets were taken as testing data.

We used training sets from two datasets to train two CNNs. The training processes were performed in Caffe, running on NVIDIA GTX TITAN X Pascal GPU with 12-GB onboard memory. Each training process was an iterative process. With every iteration, the loss and accuracy were obtained, and the hyperparameters were changed to optimize for the next training. We set the batch size, the dropout value, and the initial learning rate as 100, 0.5, and 0.001, respectively, and run the training process for 40 epochs. Each network was tested on the testing phase every 20 iterations. The training took around 40 min on each dataset.

Figure 7 shows the train loss and test accuracy curves of two CNNs trained by different datasets. The CNN1 was trained and tested on the dataset in which spectrogram samples were transformed from original vibration signals,

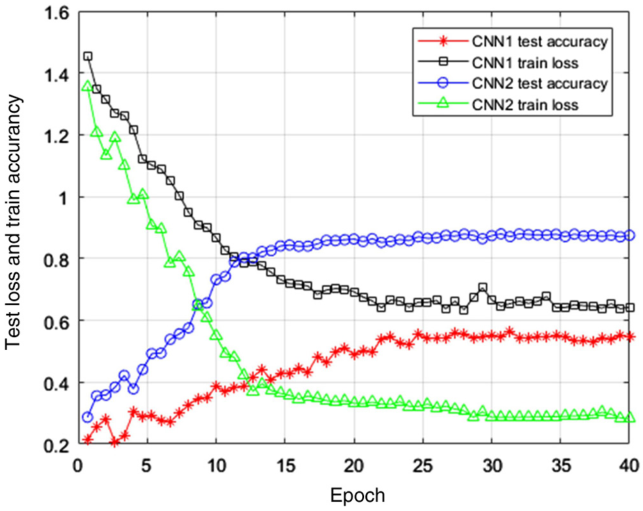

Fig. 7 The train loss and test accuracy curves for CNN1 and CNN2, which were trained by original spectrograms and enhanced spectrograms after noise reduction, respectively.

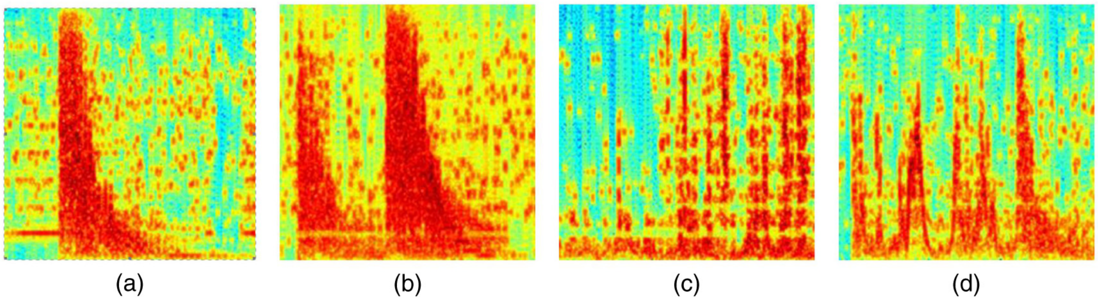

Fig. 6 The typical spectrograms of all types of denoised vibration signals: (a) the digging signal, (b) the walking signal, (c) the vehicle-passing signal, and (d) the damaging signal.

Table 1 Recognition accuracy of vibration signals using the CNN2.

	Digging (%)	Walking (%)	Vehicle (%)	Damaging (%)
Digging	97.2	0.8	2.0	0.0
Walking	1.6	96.0.	2.4	0.0
Vehicle	2.4	5.2	74.0	18.4
Damaging	0.4	0.8	14.0	84.8

while the CNN2 was trained and tested on spectrograms of denoised vibration signals. In CNN1 training process, the training loss, which represented the network error, decreased slowly until hovering around 0.65 after 25 epochs, and the best testing accuracy was only 55%. On the contrary, the test accuracy of the CCN2 grew up to 88%, and a training loss of 0.30 was achieved after 30 epochs. The figure indicates that the noise reduction process on vibration signals based on the spectral subtraction improved remarkably the classification performance of the CNN2.

5 Experimental Result

5.1 *Classification Based on Convolutional Neural Network*

The CNN2 trained on spectrograms generated from denoised vibration signals was proved to have a good performance on vibration event recognition for distributed optical fiber sensing systems. The results of the CNN2 on the test set are shown in Table 1.

Each row in Table 1 represents the recognition outcomes of each type of vibration signals on the testing set from the CNN2. It can be seen that the CNN2 classified four types of vibration signals with 97.2%, 96.0%, 74.0%, 84.8% accuracy on the testing set, respectively. These results indicate that this type of CNNs can be trained to classify very different types of vibration events, especially distinguish between instantaneous effects (digging and walking) and long-time effects (vehicle passing and damaging), but have a difficulty in accurately recognizing the vehicle signals. Though using a CNN is a more complicated way of carrying out this classification, these results demonstrate that this method successfully yields better results than other methods[8–9] and avoids manual feature selection that takes much more time in new application scenarios.

5.2 *Classification Based on Convolutional Neural Network and Support Vector Machine*

In the RCNN proposed by Girshick et al.,[16] an SVM classifier is applied to replace the soft-max classifier in the CNN. Girshick et al. extracted a 4096-dimensional feature vector extracted from each mean-subtracted 227 × 227 RGB image using a CNN containing five convolutional layers and two connected layers, and then performed classification with an SVM classifier comprised of a set of category-specific linear SVMs. This method improved mean average precision by more than 30% compared to the performance of traditional CNN approach on object recognition. It have been proved that an SVM has a better performance than a soft-max classifier, especially in solving problems of small sample size.

Table 2 Recognition accuracy of vibration signals using a nonlinear SVM classifier.

	Digging (%)	Walking (%)	Vehicle (%)	Damaging (%)
Digging	98.0	1.2	0.8	0.0
Walking	0.4	96.0	2.4	1.2
Vehicle	2.4	5.2	79.2	13.2
Damaging	0.8	3.2	12.8	83.2

Table 3 Recognition accuracy of vibration signals using a linear SVM classifier.

	Digging (%)	Walking (%)	Vehicle (%)	Damaging (%)
Digging	100.0	0.0	0.0	0.0
Walking	0.0	97.6.	1.6	0.8
Vehicle	1.2	5.6	85.6	7.6
Damaging	1.2	0.8	8.0	90.0

We first extracted the output vectors of the last FC layer in the CNN2 as feature vectors in the SVM classifier. Considering that each feature vector only had a size of 4×1, we used the radial basis function as the kernel function in the SVM classifier. The classification results are shown in Table 2. From Tables 1 and 2, we can see that classification results of the soft-max classifier and the nonlinear SVM classifier were approximately the same, except that the recognition rate of vehicle signals rose up to 79.2%. This was because the outputs of the last FC layer in a CNN were highly abstracted and reflected the classification results to a great degree. Thus, it was inefficient to replace the soft-max classifier with a complex classifier.

Then, we extracted the input vectors of the first FC layer in the CNN2 as feature vectors in the SVM classifier. The number of feature parameters was 1600, more than the sample number of each type of vibration events. Thus, we chose the linear kernel function as the SVM's kernel function. Table 3 shows the classification performance of the linear SVM classifier. It indicates that this classification method achieved a classification accuracy of 93.3%, much better than other methods. Meanwhile, the successful application of a linear SVM proved that the input vectors of the first FC layer had a good expression on spectrogram features. And we only needed to optimize the parameter C when using a linear SVM.

6 Conclusion

Aiming at the problem of vibration identification in distributed optical fiber sensing systems, a pattern recognition method based on time-frequency analysis and CNN is proposed in this paper.

Vibration signals are enhanced by spectral subtraction to weaken the influence of the noise signal on vibration signal features and then represented as spectrograms that act as the input to CNNs. The CNN model comprising of five convolutional and two FC layers extracts features from these spectrograms generated from different types of vibration events. The soft-max classifier and the SVM classifiers are used to classify these features.

To examine the algorithm performance, a set of experiments were carried out. Two CNNs were trained on the original spectrogram dataset and the enhanced spectrogram dataset, respectively, which were generated from four different types of vibration signals collected in a φ-OTDR distributed optical fiber sensing system. The CNN trained on enhanced spectrograms had a better classification accuracy of 88%. By replacing the soft-max classifier in the CNN with a nonlinear SVM classifier or a linear SVM classifier, the classification accuracy of the proposed algorithm rose up to 89.1% and 93.3%, respectively. The total time of time-frequency analysis and classification process was below 0.6 s. Experimental results show that this method greatly improved the recognition accuracy of φ-OTDR systems under a complex noise environment especially when a linear SVM was applied and would be able to work in real time.

Disclosures

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supporter in part by the National Natural Science Foundation of China (Nos. 61590925, U1609212, and U1509211), in part by the National Key Research and Development Plan of China (No. 2017YFC0210403), and in part by the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDA06020201).

References

1. J. P. F. Wooler and R. I. Crickmore, “Fiber optic sensors for seismic intruder detection,” *Proc. SPIE* **5855**, 278–281 (2005).
2. F. Tanimola and D. Hill, “Distributed fibre optic sensors for pipeline protection,” *J. Nat. Gas Sci. Eng.* **1**(4), 134–143 (2009).
3. X. Bao and C. Liang, “Recent progress in distributed fiber optic sensors,” *Sensors* **12**(7), 8601–8639 (2012).
4. J. C. Juarez et al., “Distributed fiber-optic intrusion sensor system,” *J. Lightwave Technol.* **23**(6), 2081–2087 (2005).
5. J. C. Juarez and H. F. Taylor, “Field test of a distributed fiber-optic intrusion sensor system for long perimeters,” *Appl. Opt.* **46**(11), 1968–1971 (2007).
6. Z. Wang et al., “Coherent Φ-OTDR based on I/Q demodulation and homodyne detection,” *Opt. Express* **24**(2), 853–858 (2016).
7. A. Masoudi, M. Belal, and T. P. Newson, “A distributed optical fibre dynamic strain sensor based on phase-OTDR,” *Meas. Sci. Technol.* **24**(8), 085204 (2013).
8. C. Xu et al., “Pattern recognition based on enhanced multifeature parameters for vibration events in φ-OTDR distributed optical fiber sensing system,” *Microwave Opt. Technol. Lett.* **59**(12), 3134–3141 (2017).
9. B. Wang et al., “Improved wavelet packet classification algorithm for vibrational intrusions in distributed fiber-optic monitoring systems,” *Opt. Eng.* **54**(5), 055104 (2015).
10. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Int. Conf. on Neural Information Processing Systems*, pp. 1097–1105, Curran Associates Inc. (2012).
11. O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” *Int. J. Comput. Vision* **115**(3), 211–252 (2015).
12. T. S. Jordan, “Using convolutional neural networks for human activity classification on micro-Doppler radar spectrograms,” *Proc. SPIE* **9825**, 982509 (2016).
13. A. M. Badshah et al., “Speech emotion recognition from spectrograms with deep convolutional neural network,” in *Int. Conf. on Platform Technology and Service*, IEEE (2017).
14. S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust. Speech Signal Process.* **27**(2), 113–120 (1979).
15. C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Min. Knowl. Discovery* **2**(2), 121–167 (1998).
16. R. Girshick et al., “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 580–587 (2013).

Chengjin Xu received his BS degree from the College of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2014, where he is currently working toward his MS degree at the Institute of Industrial Process Control. His research interests include fiber-optic sensors, signal processing, and pattern recognition. He is a student member of SPIE.

Junjun Guan is a senior engineer at the Institute of Acoustic, Chinese Academy of Sciences. He received his BS degree in physics from Hunan University of Technology in 2010 and his MS degree in physics from Jishou University in 2014. His current research interests include distributed optical fiber sensors and acoustic metamaterial.

Ming Bao is a professor at the Institute of Acoustic, Chinese Academy of Sciences. He received his BS degree in automatic control engineering from Wuhan University of Technology in 1996 and his PhD from the Institute of Acoustic, Chinese Academy of Sciences, in 2008. His research interests include signal processing, acoustic sensor network, and algorithm development.

Jiangang Lu is a professor at the College of Control Science and Engineering, Zhejiang University, Hangzhou, China. He received his BS and PhD degrees from the College of Chemical and Biological Engineering, Zhejiang University, in 1989 and 1995, respectively. His research interests include online analyzers, industrial process control, data-driven modeling, and optimization.

Wei Ye is an associate professor at the College of Control Science and Engineering, Zhejiang University, Hangzhou, China. He received his BS and PhD degrees from the College of Optical Science and Engineering, Zhejiang University, in 1992 and 1998, respectively. His research interests include fiber-optic sensors, optical measurement instruments, and wireless sensor networks.