AN INTELLIGENT ROLLING BEARING FAULT DIAGNOSIS METHOD OF CNN BASED ON MNIST DATABASE OF HANDWRITTEN DIGITS

Kai Wen1, Maohua Xiao[1,2,],Cunyi Zhang[1], Dan Wu[1], Nong Gao[2], Jing Zhang[1]*
[1] College of Engineering, Nanjing Agriculture University, Nanjing 210031, China
[2] Faculty of Engineering and the Environment, University of Southampton, SO17 1BJ, UK
**Correspondence: xiaomaohua@njau.edu.cn; Tel.: +86-13951756153*

Abstract - An intelligent rolling bearing fault diagnosis method of convolutional neural network (CNN) based on MNIST database of handwritten digits was proposed. The equal length interception of the vibration signal of the rolling bearing was performed, and the intercepted signal was reconstructed into two-dimensional pictures in sequence. In order to ensure the authentic and effective restoration of the fault characteristics of the vibration signal, no preprocessing was performed on the vibration signal, and direct interception was performed to reconstruct the two-dimensional pictures. To realize intelligent fault diagnosis of rolling bearings, the failure pictures were input as feature maps, and a CNN classifier model based on MNIST database of handwritten digits was established. In order to verify the validity of the rolling bearing fault diagnosis method proposed in this paper, the open-end rolling bearing data from Case Western Reserve University and the rolling bearing data collected under laboratory conditions were experimentally verified respectively. The experimental results were then compared with the traditional BP neural network, BP neural network optimized by particle swarm optimization, wavelet energy entropy support vector machine optimized by particle swarm optimization, support vector machine based on variational mode decomposition mutual approximation entropy and BP neural network optimized by genetic algorithm on the effectiveness of fault pattern recognition. The results show that this method can identify effectively the fault type of the rolling bearing, and the diagnosis efficiency is higher. The method has strong feature extraction and recognition capabilities.

Keywords: Rolling Bearing; Fault Diagnosis; MNIST; Convolutional Neural Network.

1. Introduction

Rolling bearing, an important rotating machinery in the petroleum, chemical, metallurgical and other machinery industries has the advantages of high speed, low power consumption and low noise [1]. Because of the development of modern heavy industry, high speed and heavy load are still regarded as the tendency of mechanical equipment. The impact of rolling bearing performance on the entire manufacturing and engineering applications is also increasing. Therefore, it is very important to develop rolling bearing fault diagnosis technology [2]. It is of important value in engineering application, improving equipment operation efficiency, equipment safety performance, and reducing equipment maintenance cost [3]. Therefore, this paper takes the rolling bearing as the research object and studies its fault diagnosis method.

Entropy can realize the quantitative measurement of information and effectively represent the complexity of time series. In recent years, the concepts of entropy of information entropy, approximate entropy and sample entropy have been applied to the field of mechanical fault diagnosis. Sample entropy was first proposed as an improved method proposed by the literature [4] for the modal problem of approximate entropy. The literature [5] applies Sample Entropy to the fault diagnosis of rolling bearings. The literature [6] applies Permutation Entropy (PE) to vibration signal mutation detection, which has achieved good diagnostic results. Nevertheless, the above method has its own drawbacks. For example, Sample Entropy is slow in calculation, poor in real-time performance, and the similarity measure is prone to mutation; although PE is simple in concept calculation [7], it does not consider the difference between the mean and amplitude values of amplitude. Moreover, the above methods are all based on a single-scale analysis method of time series.

Convolutional neural network (CNN) is a Multi-Layer Perceptron (MLP) designed to identify two-dimensional feature maps. It is a kind of deep learning network model with multiple hidden layers. It can transform low-level features into higher-level features through layer-by-layer feature transfer to realize the learning and expression of features [8]. Compared with shallow networks such as BP neural

network, SVM and others, CNN has stronger learning and expression capabilities for complex features, faster computing speed. At present, CNN has been widely used in speech recognition [9], handwriting recognition, face recognition [10], behavior detection, and text classification [11].

As a network model with excellent recognition performance, CNN has been preliminarily used in the fault diagnosis and has raised the level and increased the efficiency of fault diagnosis. For example: Chen Zhi-qiang et al. [12]. Considered the feature maps in combination with the statistical characteristics of gearbox vibration signals as the CNN input, and realized the diagnosis and identification of gearbox faults through CNN. The method showed better recognition rate and efficiency than other methods.

In view of the above problems, this paper proposes an intelligent rolling bearing fault diagnosis method of CNN based on MNIST database of handwritten digits, and rolling bearing faults could be diagnosed.

2. Construction of Convolutional Neural Network

2.1. Convolutional neural network

The CNN is mainly a feature extractor composed of input layer, convolution layer, pooling layer, full connection layer and output layer [13]. Generally, feature maps showing a number of neurons arranged in a matrix are found in the convolution layer. Neurons in each feature map share weights, and the shared weight is a filter, also known as a convolution kernel. The convolution kernel is usually numbers arranged in a matrix and it will learn the appropriate weights in the network model training. The sharing of the weights of the convolution kernels can reduce the connections between the various layers of the network and the risk of overfitting as well [14]. The pooling layer mainly includes two forms, mean pooling and max pooling. The pooling layer selects the feature maps of the convolution layer, which reduces the parameters of the model.

2.2. Neural network of MNIST database of handwritten digits

2.2.1. MNIST database of handwritten digits

MNIST is a database of handwritten numbers created by the National Institute of Standards and Technology (NIST) [15]. MNIST is an entry-level computer vision database that contains a total of ten samples from number 0 to number 9 which are stored in the sample matrix as gray values. The database contains 60,000 training samples and 10,000 test samples [16]. The pixels of all sample images are 28×28 pixels. Some digital samples are taken from the database, as shown in Figure 1.

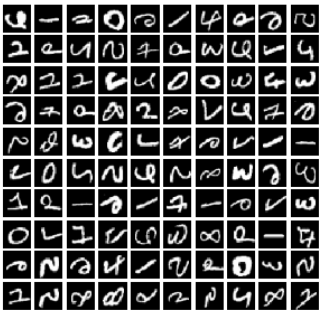

Figure 1: Some sample pictures in MNIST.

Each picture contains 28x28 pixels, and a digital array can be used to represent the picture, as shown in Figure 2.

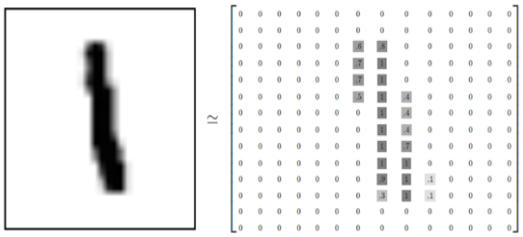

Figure 2: Digital array diagram.

First, the array is expanded into a vector of length 28x28=784, and so is the picture. Then, the images in the MNIST database are expanded into points in the 784-dimensional vector space and show a more complex structure. The entire training database is a tensor of [60000, 784], as shown in Figure 3. The first dimension is used to index the picture and the second dimension is used to index the pixels in each picture. Each element of the tensor here represents the intensity value between 0 and 1 of a pixel in a picture.

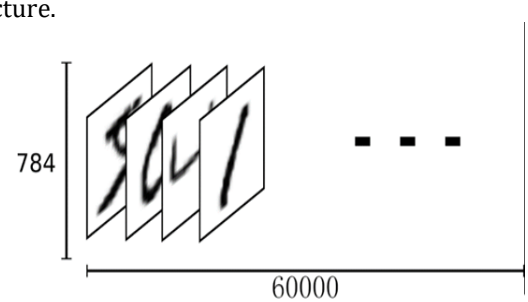

Figure 3: Digital tensor diagram.

2.2.2. One-hot vectors

Since a given label is required to identify number 0 to number 9 in the MNIST database of handwritten digits, one-hot vectors are introduced to label MNIST. A one-hot vector except for one digit is 1 and all other digits are 0. Therefore, in the MNIST database of handwritten digits introduced in this paper, the number n indicates a 10-dimensional vector whose number is only 1 in the nth dimension (starting from 0).

For example, the number 1 will be represented as [0,1,0,0,0,0,0,0,0,0]. Therefore, the training database is a [60000, 10] label number matrix, as shown in Figure 4.

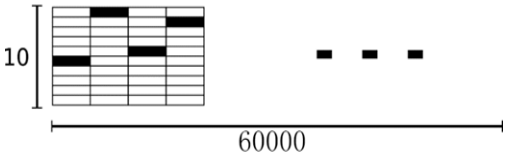

Figure 4: Digital matrix diagram.

2.2.3. Convolutional neural network structure of MNIST database of handwritten digits

Each picture in the MNIST database of handwritten digits represents a number from 0 to 9, and it is hoped by CNN that the probability that each picture represents each number can be identified.

For example, our model may speculate that the probability of a picture containing the number 9 is 80% but the probability of judging it as 8 is 5% (because both 8 and 9 have a small circle in the top half).

In this paper, the CNN network model of MNIST database of handwritten digits is constructed by two convolutional layers, two pooling layers and two full connection layers.

The last layer of the Softmax classifier classifies the 10 labels after the full connection layer in terms of probability. The main parameters are shown in Table 1.

Table 1: The parameters in CNN of MNIST.

Network layer	Input map	Output map	Convolution kernel	Stride
Convolution layer 1	28 x 28(1)	28 x 28(32)	5 x 5	1
Pooling layer 1	28 x 28(32)	14 x 14(32)	2 x 2	2
Convolution layer 2	14 x 14(32)	14 x 14(64)	5 x 5	1
Pooling layer 2	14 x 14(64)	7 x 7(1024)	2 x 2	2
Full connection layer 1	7 x 7(1024)	1 x 1(10)	7 x 7	1
Full connection layer 2	1 x 1(10)	1 x 1(10)	7 x 7	1

Figure 5 is the CNN network model of MNIST database of handwritten digits. It can be clearly seen in the figure that after a series of convolution and pooling operations, the image will eventually be classified by the Softmax classifier according to probability. The RELU layer in the figure mainly exists as an activation function, which is designed for the nonlinearity of the network and enhancement of robustness.

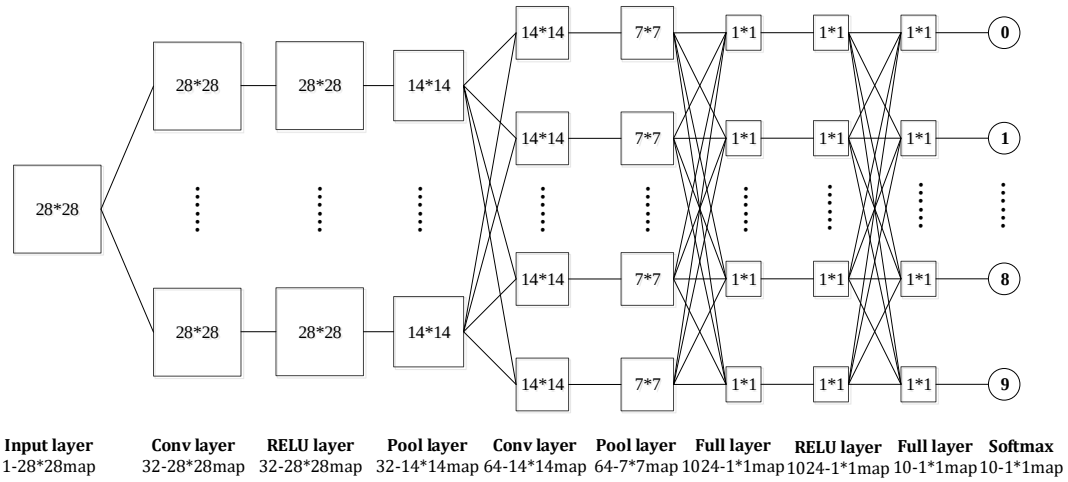

Fig.5 Model of CNN of MNIST

From number 0 to number 9, 10,000 pictures are selected for each category to train. We select 100 pictures to train for a batch, and final test shows the recognition accuracy rate of the network model of handwritten digits after 20 times iterative loop is 99.19%.

2.3. Network training

The CNN training method adopts the batch sample input method. The training process is shown in Figure 6. It includes two parts, the forward propagation of the data and the backward propagation of the error [17]: First, the training parameters of the network is set and the weight of the network is initialized. The input feature map is processed by the convolution layer, pooling layer, and full connection layer and then transmitted to the output layer. The output of the former layer is the input of the next layer. Then, the error between the actual output and the expected output is back-propagated through the algorithm layer by layer [18]. The error is distributed to each layer, and then the weights and bias of the network are adjusted until it meets the convergence conditions so as to achieve supervised training of the network.

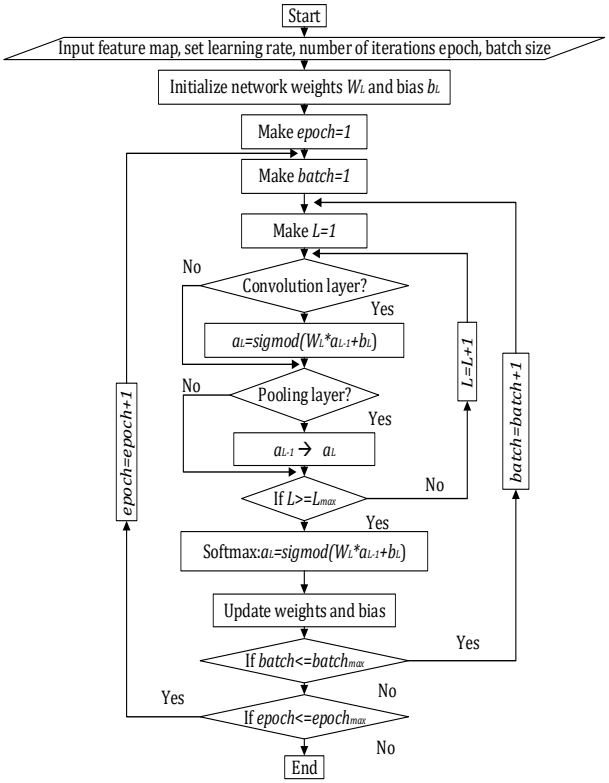

Figure 6: CNN training process.

3. Instance of Rolling Bearing Fault Diagnosis

3.1. Two-dimensional reconstruction of vibration signal

The one-dimensional time-domain signal acquired by the sensor represents different types of faults. Since the input of the convolutional neural network is preferably two-dimensional pictures, the collected source signals are clipped in equal length, and the intercepted signals are reconstructed into two-dimensional pictures in sequence.

After the two dimensional reconstruction, the signal of different fault types formed different groups with specific fault characteristics.

In order to ensure that the fault pictures can restore the fault characteristics of the vibration signal, the vibration signal collected by the sensor is not pre-processed, and the interception is directly reconstructed into a two-dimensional picture, as shown in Figure 7. The main formula of the two-dimensional reconstruction signal is as follows:

$$P = \begin{bmatrix} x(t) & \cdots & x(t+n-1) \\ \vdots & \ddots & \vdots \\ x(t+(m-1)n) & \cdots & x(t+mn-1) \end{bmatrix}$$

where P represents the reconstructed picture with the fault features and *x(t)* represents the collected vibration signals.

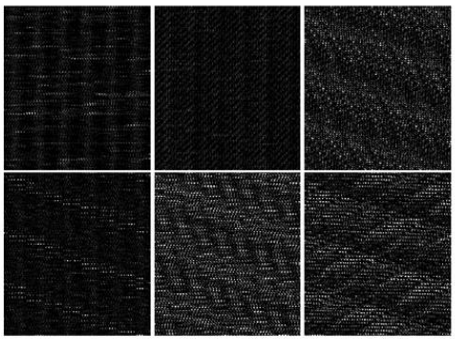

Figure 7: Images of the vibration signals for different bear faults

3.2. Data description

The structure of the CNN used for the fault pattern recognition of the rolling bearing is similar to that of the MNIST database of handwritten digits. The bearing fault vibration signal needs to be intercepted by equal length to reconstruct the fault picture, and the number of the fault signal pictures generated is limited. In this paper, the CNN training model based on MNIST database of handwritten digits recognition is used in the rolling bearing fault experiment. Based on this network model, the fault signal pictures are diagnosed. In order to verify the validity of rolling bearing fault diagnosis method of CNN based on MNIST database of handwritten digits proposed in this paper, the open source rolling bearing fault data of Case Western Reserve University and the rolling bearing fault data collected under laboratory conditions were respectively verified by experiments.

Table 2 and Table 3 show the relevant descriptions of bearing fault data collected by the Western Reserve University and laboratory conditions.

Table 2: Rolling bearing fault data of Case Western Reserve University

Fault location	No	Rolling element			Inner ring			Outer ring		
Fault diameter	\	0.007	0.014	0.021	0.007	0.014	0.021	0.007	0.014	0.021
Test data	380	380	380	380	380	380	380	380	380	380

Table 3: Rolling bearing fault data under laboratory conditions.

Bearing number	1	2	3	4	5	6
Fault type	Inner ring corrosion	Outer ring corrosion	Inner and outer ring compound corrosion	Inner and outer ring compound spalling	Inner ring and rolling element compound spalling	Outer ring and rolling element compound spalling
Test data	380	380	380	380	380	380

The one-dimensional vibration signals collected by the sensor are cut into one segment for each 900 data, reconstructed into a 30 x 30 two-dimensional picture, and used as an input of a convolutional neural network to perform feature extraction.

3.3. Experiment setup

The MNIST database of handwritten digits model is called. Based on the model, the fault pictures after the reconfiguration of the fault signal of the rolling bearing are input, and the input fault pictures are subjected to corresponding convolution and pooling operations. The specific parameters are as follows in Table 4. Meanwhile, the traditional BP neural network, the BP neural network optimized by particle swarm optimization and the BP neural network optimized by the genetic algorithm are compared with the rolling bearing fault diagnosis method of CNN based on MNIST database of handwritten digits set.

Table 4: Model of CNN based on fault images.

Network layer	Kernel height	Kernel width	Kernel depth	Stride
Convolution layer 1	3	3	6	1
Pooling layer 1	2	2	1	2
Convolution layer 2	3	3	12	1
Pooling layer 2	2	2	1	2
Convolution layer 3	3	3	12	1

3.4. Analysis of diagnosis results

Accuracy rates for fault diagnosis of bearing fault data by BP neural network, BP neural network optimized by particle swarm optimization, wavelet energy entropy support vector machine optimized by particle swarm optimization, support vector machine based on variational mode decomposition mutual approximation entropy, BP neural network optimized by genetic algorithm and the CNN based on MNIST database of handwritten digits proposed in the paper are shown in Table 5 and Table 6.

Table 5: Fault diagnosis results of rolling bearing of Case Western Reserve University.

Method	Normal state	Inner ring fault	Outer ring fault	Rolling element fault	Average diagnostic rate
BP	0.8556	0.8487	0.86	0.7404	0.8262
PSO-BP	0.9326	0.9746	0.9684	0.8465	0.9305
PSO-WSVM	0.9070	0.9390	0.9270	0.9401	0.9283
VMD-CAE-SVM	0.9333	0.9667	0.9333	0.9667	0.9500
GA-BP	0.9306	0.9742	0.9702	0.8386	0.9284
CNN	0.9912	0.9934	0.9978	0.9042	0.9717

Table 6: Fault diagnosis results of rolling bearing under laboratory conditions.

Method	Fault 1	Fault 2	Fault 3	Fault 4	Fault 5	Fault 6	Average diagnostic rate
BP	0.8678	0.8786	0.8682	0.8812	0.8764	0.8708	0.8738
PSO-BP	0.9425	0.9408	0.9426	0.9388	0.9512	0.9448	0.9435
PSO-WAVM	0.8801	0.8612	0.8421	0.9270	0.8911	0.8735	0.8792
VMD-CAE-SVM	0.9367	0.9583	0.9324	0.9101	0.9023	0.9093	0.9249
GA-BP	0.9312	0.9514	0.9402	0.9468	0.9502	0.9426	0.9437
CNN	0.9896	0.9912	0.9932	0.9944	0.9968	0.9942	0.9932

Table 5 is an analysis result of the diagnosis of the open-source rolling bearing fault data by Western Reserve University. It can be seen from the table that the accuracy of the diagnosis of the rolling bearing in the normal state, the inner ring fault and the outer ring fault is significantly greater than that of the rolling element fault.

The error here is mainly due to the influence of the position deviation when the sensor acquires the signal. At the same time, it can be found through comparison that the BP neural network optimized by the particle swarm algorithm and genetic algorithm has greatly improved the diagnostic rate compared with the traditional BP neural network.

This is mainly because the increase of the algorithm optimizes the input eigenvalues and greatly improves the convergence ability of the BP network.

However, no matter it is a neural network optimized by the particle swarm optimization or a neural network optimized by genetic algorithm, the input values are all calculated manually and the calculation efficiency is low.

When a large sample is used, the feature value calculation will take more than half of the time.

Moreover, since the eigenvalues are calculated manually, errors cannot be avoided, and even errors will surely affect the diagnostic rate of the method. The fault diagnosis method of CNN based on MNIST database of handwritten digits directly reconstructs the original vibration signal for two-dimensional pictures instead of calculating manually.

The reconstructed fault signal pictures are used as the input layer of the CNN network to make a diagnosis. The fault diagnosis method of CNN based on the MNIST database of handwritten digits not only has high diagnostic efficiency, but also has a much higher diagnostic rate than traditional neural network methods.

The accuracy of the fault signals collected under the laboratory condition shown in Table 6 also reflects that the rolling bearing fault diagnosis method of CNN based on MNIST database of handwritten digits is obviously higher than other kinds of fault diagnosis methods based on traditional neural networks. Because convolutional neural networks generally have large data sets, a normally trained convolutional neural network will require tens of thousands of pictures as a training set.

When the continuous vibration signal data is not sufficient to create a training set, other models that have been trained for migration learning shall be used, but not all models can be used for migration learning. Generally speaking it is necessary to fine-tune the model so that the model can adapt to the objects that need to be identified. If you can collect enough continuous fault vibration signals to make a data set, you can use the fault signal pictures to do the training for fault diagnosis. The pictures can be rotated, stretched, cut, and other treatments to enhance the robustness of the CNN model.

4. Conclusions

In this paper, the rolling bearing fault diagnosis method of CNN based on MNIST database of handwritten digits is used. The one-dimensional vibration signal collected by the sensor is intercepted and reconstructed into two-dimensional fault images and input into the convolutional neural network. Finally, the open source bearing data of Western Reserve University and the bearing failure data collected under the laboratory conditions are respectively tested, and the experimental results are compared with several fault diagnosis methods based on traditional BP neural networks. The following conclusions are obtained:

(1) The rolling bearing fault diagnosis method of CNN based on MNIST database of handwritten digits can effectively identify the failure mode of rolling bearing, and the accuracy of diagnosis is obviously higher than other methods based on traditional neural network.

(2) The rolling bearing fault diagnosis method of CNN based on MNIST database of handwritten digits does not need to calculate the eigenvalue manually, which can effectively improve the efficiency of diagnosis.

(3) By optimizing and improving the structure parameters and training parameters of CNN, the correct recognition rate of faults and the stability of recognition performance can be effectively improved.

Acknowledgments

The research is funded partially by the National Key Research and Development Program of China (2016YFD0701103), Key Research and Development Program of Jiangsu Province (BE2018127).

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

References

[1] SHI Kunju, LIU Shulin, JIANG Chao, ZHANG Hongli. Rolling Bearing Feature Frequency Extraction using Extreme Average Envelope Decomposition [J]. Chinese Journal of Mechanical Engineering, 2016, 29(05):1029-1036.

[2] Peng Wang. Fault diagnosis method of rolling bearing based on A(dB) value[A]. Wuhan Zhicheng Times Cultural Development Co., Ltd. Proceedings of 2017 International Conference on Advances in Materials, Machinery, Electrical Engineering(AMMEE 2017)[C].Wuhan Zhicheng Times Cultural Development Co., Ltd:,2017:5.

[3] LI Min, YANG Jianhong, WANG Xiaojing. Fault Feature Extraction of Rolling Bearing Based on an Improved Cyclical Spectrum Density Method [J]. Chinese Journal of Mechanical Engineering, 2015, 28(06): 1240-1247.

[4] Yang Yu, Yu Dejie, Cheng Junsheng. Fault Diagnosis Method of Rolling Bearing Based on HILBERT Marginal Spectrum [J]. Journal of Vibration and Shock, 2005, 24(1): 70-72

[5] Zhao Zhihong, Yang Shaopu. A bearing fault diagnosis method based on sample entropy [J]. Journal of Vibration and Shock, 2012, 31(6): 136-140

[6] Feng Fuzhou, Rao Guoqiang, Si Aiwei. Research on permutation entropy algorithm and its application in vibration signal mutation detection [J]. Journal of Vibration Engineering, 2012, 25(2): 221-224.

[7] RUQIANG YAN, YONGBIN LIU, ROBERT X. GAO. Permutation entropy: A nonlinear statistical measure for status characterization of rotary machines [J]. Mechanical Systems and Signal Processing, 2011, 29: 474- 484

[8] SUN J, CAO W F, XU Z B. Learning a convolutional neural network for non-uniform motion blur removal [J]. CVPR 2015, 2015 10(5): 48 — 56.

[9] ABDEL-HAMID O, MOHAMED A R, JIANG H. Convolutional neural networks for speech recognition [J]. IEEE /ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(10): 1533 — 1545.

[10] LIU M Y, LI S X, SHAN S G. AU-inspired deep networks for facial expression feature learning [J]. Neurocomputing, 2015, 159(8): 126-136.

[11] ZHU A N, WANG G Y, DONG Y. Detecting text in natural scene images with conditional clustering and convolution neural network [J]. Journal of Electronic Imaging, 2015, 24(5): 55-66.

[12] CHEN Z Q, LI C, SANCHEZ R V. Gearbox fault identification and classification with convolutional neural networks [J]. Shock and Vibration, 2015, 15(8): 18-28.

[13] Meziane Iftene. Very High Resolution Images Classification by Fusing Deep Convolutional Neural Networks [A]. Research Institute of Management Science and Industrial Engineering. Proceedings of 2017 5th International Conference on Advanced Computer Science Applications and Technologies (ACSAT 2017) [C].Research Institute of Management Science and Industrial Engineering:,2017:5.

[14] Du Guiming. Speech Recognition Based on Convolutional Neural Networks [A]. IEEE Beijing Section. Proceedings of 2016 IEEE International Conference on Signal and Image Processing（ICSIP）[C].IEEE Beijing Section:, 2016:4.

[15] Ernst Kussul, Tatiana Baidyk. Improved method of handwritten digit recognition tested on MNIST database [J]. Image and Vision Computing, 2004, 22(12).

[16] Dan Claudiu Cireşan, Ueli Meier,Luca Maria Gambardella, Jürgen Schmidhuber. Deep Big Multilayer Perceptrons for Digit Recognition [M]. Springer Berlin Heidelberg: 2012-06-15.

[17] LECUN Y, BOTTOU L, BENGIO Y. Gradient based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11):2278-2324

[18] P. Zhou, G. Zhou, Z. Zhu, C. Tang, Z. He, W. Li and F. Jiang. “Health Monitoring for Balancing Tail Ropes of a Hoisting System Using a Convolutional Neural Network”, Applied Science, 2018, x\Vol. 8, No. 8. Doi: 10.3390/app8081346.