Feature Extraction Method of Ultrasonic Signal Based on Wavelet Coefficients Cluster

Feng Zihong*,1 Miao Changyun2 and Bai Hua2

1Engineering Practice Teaching Training Center, Tianjin Polytechnic University, Tianjin 300387, P.R. China
2School of Electronics and Information Engineering, Tianjin Polytechnic University, Tianjin 300387, P.R. China

Abstract: Support Vector Machine is a very good solution to the classification problem of small sample, but when the input feature vector dimension is larger, the classifier has complex structure, long training time and degraded performance. In order to solve this problem, a feature extraction method based on wavelet coefficients cluster was put forward. All the wavelet coefficients was clustered, the energy value of wavelet coefficients in each cluster was calculated and used as the input feature vector of a classifier. The dimension of input data was reduced greatly, and at the same time the specific problem information was retained. Support Vector Machine was used to identify the defects in steel plate, the experiment results showed that the method has higher classification accuracy.

Keywords: Feature vector, feature extraction, support vector machine, wavelet transform.

1. INTRODUCTION

In industry, workpiece often needs to be tested by ultrasonic [1], during testing, a large number of ultrasonic echo signals need to be collected, and then processed and classified. In the process of classification, the classification accuracy depends mainly on extracted features and adopted classifier. The time-frequency localization characteristics of wavelet transform [2] makes it become a commonly used characteristics analysis method, but the dimension of discrete wavelet coefficients matrix of actual sampled signal is very big, in order to reduce its dimension, important features must be extracted further from the wavelet coefficients matrix. Previous feature extraction methods based on wavelet transform often used wavelet coefficients characteristics of a scale or some scales to classify [3-4], but obtained wavelet decomposition coefficients of different scales represent the size of signal component in different frequency range, if wavelet coefficients of some scales are discarded, signal information contained in these scales is bound to loss, so the classification accuracy will reduce.

In addition, the wavelet coefficients are variable for digital signal sequence position, in order to solve above problems a feature extraction algorithm based on wavelet coefficients cluster was proposed in this paper. First a wavelet coefficients matrix of sampled signal was obtained by using fast wavelet transform [5], and then for each row of the matrix, namely the wavelet coefficients cluster of each scale, coefficient characteristic containing large signal information was extracted, and coefficient characteristic containing small signal information was merged, finally the energy value of wavelet coefficients of each cluster was calculated and used as the input vector of a classifier. Proposed feature extraction method and support vector machine classifier [6-7] was combined to identify internal defects in medium plates, experiment results showed that proposed method has higher classification recognition accuracy.

2. FAST WAVELET TRANSFORM

The Daubechies wavelet basis has been widely used in signal analysis because of properties of ideal approximation precision and numerical stability [8]. The parameters of the corresponding discrete wavelet decomposition are given by real values sequence \((h_0, h_1, \ldots, h_N)\) and these values meet the following equation.

\[
\sum_{n=0}^{N} h_n = \sqrt{2}
\]

\[
\sum_{0}^{N} (-1)^n n^k h_n = 0, \quad k = 0, 1, \ldots, (N-1)/2
\]

\[
\sum_{0}^{N/2} h_{2k} h_{2k+2k} = 0, \quad k = 1, 2, \ldots, (N-1)/2
\]

where \(N\) is an odd natural number, the number of solutions is \(2^{(N-1)/2}\).

If the sampled signal of ultrasonic sensor is a certain finite data series \(a_0 = (a_{0n})(n = 0, 1, \ldots, L-1)\), \(L\) is length of the signal, \(N\) is order of the Daubechies wavelet basis, and then the expansion of wavelet basis meets the following relationship expression.

\[
a_{(m+1)n} = \sum_{k=2n}^{2n+N} h_{k-2n}a_{mk}
\]

where, \(n = \frac{1-N}{2}, \frac{1-N}{2}+1, \ldots, \frac{L}{2}-1\) (\(L\) is even number);

\(n = \frac{1-N}{2}, \frac{1-N}{2}+1, \ldots, \frac{L-1}{2}\) (\(L\) is odd number);

\[
d_{(m+1)n} = \sum_{k=2n+1-N}^{2n+1} (-1)^k h_{2n+1-k}a_{mk}
\]
where, \( n = 0,1,\ldots, \frac{L+N-3}{2} \) (\( L \) is even number); \( n = 0,1,\ldots, \frac{L+N-2}{2} \) (\( L \) is odd number). They are the decomposition of information contained in \( a_0 \), by computing equation (1), (2) and (3) the following equation can be obtained.

\[
a_{0n} = \sum_{k=-\infty}^{\infty} [h_{n-2k}a_{1k} + (-1)^nh_{2k+1-n}d_{1k}], n \in \mathbb{Z}
\]  

(6)

Namely the sequence \( a_0 \) can be reconstructed precisely by \( a_1 \) and \( d_1 \). The energy of sequence \( a_0 \) is assigned to \( a_1 \) and \( d_1 \) and the following definition is given.

Definition 1: The energy \( \|p\|^2 \) of a limited sequence \( p = (p_n) \) is defined as

\[
\|p\|^2 = \sum_{n=-\infty}^{\infty} p_n^2
\]  

(7)

If \( p \) is regarded as a vector, the square root of the energy \( \|p\| \) is the Euclidean norm \( \|p\|_2 \) of \( p \).

Theorem 1: The energies of the sequences \( a_0, a_1 \) and \( d_1 \) meets the following equation.

\[
\|a_0\|^2 = \|a_1\|^2 + \|d_1\|^2
\]  

(8)

With the recursion of equation (4) and (5), the signal \( a_0 \) can be decomposed into a sequence \( d_1, d_2, \ldots, d_M, a_M \). The sequence contains the same amount of information as \( a_0 \) and the following property is established.

\[
\|a_0\|^2 = \|a_M\|^2 + \sum_{m=1}^{M} \|d_m\|^2
\]  

(9)

The index \( M \) of sequences \( a_M = (a_{Mn}) \) and \( d_M = (d_{Mm}) \) is scale or decomposition level, coefficients \( a_{Mn} \) and \( d_{Mm} \) represent respectively approximation coefficients and detail coefficients, \( d_1 \) represents the coefficients of the highest frequency area, \( d_2 \) is the coefficients in the next frequency. For every measured signal the wavelet coefficients contained in \( d_1, d_2, \ldots, d_M \) and \( a_M \) are arranged in a matrix as shown in Fig. (1), where the part \( w \) is filled with zero value.

![Diagram](https://via.placeholder.com/150)

**Fig. (1).** Fast Wavelet Transform Coefficients Matrix of Sequence \( a_0 \).

By analyzing above derivation process we can know that sequences \( d_1, d_2, \ldots, d_M \) and \( a_M \) can be derived after wavelet of scale \( M \) is decomposed for a signal \( a_0 \) with fixed length \( L \), namely \( a_0 \) can be replaced completely with \( d_1, d_2, \ldots, d_M \) and \( a_M \), so the size of obtained wavelet coefficients matrix is \((M + 1) \times (L + N - 1/2)\) (for even number \( L \)) or \((M + 1) \times (L + N/2)\) (for odd number \( L \)), when \( L \) is larger, the obtained matrix will be very large, the structure of classifier is very complex. By extracting larger feature information and merging smaller feature information contained in \( a_0 \), the input feature dimension can be reduced, so the classifier structure will become simple, the classification process is relatively easy.

### 3. FEATURE EXTRACTION METHOD BASED ON WAVELET COEFFICIENT CLUSTER

#### 3.1. The Determination of Cluster

Fast wavelet transform was computed for \( K \) ultrasonic sampled signal, wavelet coefficients matrix of each signal \( s \) was obtained respectively.

\[
B_k = (b_{ij})_k \quad (k = 1, 2, \ldots, K)
\]  

(10)

where, \( i = 1, 2, \ldots, M + 1 \), \( j = 1, 2, \ldots, \frac{L+N-1}{2} \) (\( L \) is even number), \( j = 1, 2, \ldots, \frac{L+N}{2} \) (\( L \) is odd number), \( N \) is an odd nature number. The following two results need to be used for the cluster process.

Theorem 2 (central limit theorem) [9]: Let \( Y_k \) be an independent random variable sequence, for any \( k \in \mathbb{N} \) and a closed interval \([a, \beta]\), then \( p\{Y_k \in [a, \beta]\} = 1 \), and let:

\[
\begin{align*}
\rho_0 &= \sqrt{\frac{\sum_{k=1}^{n} \nu Y_k}{\sum_{k=1}^{n} \nu Y_k}} \\
Z_n &= \frac{\sum_{k=1}^{n} \nu Y_k - \sum_{k=1}^{n} \nu E Y_k}{\rho_0}
\end{align*}
\]  

(11)

Where \( n \in \mathbb{N}, EY_k \) and \( \nu Y_k \) represent respectively the expected value and variance of a random variable \( Y_k \), then if and only if \( n \to \infty, \rho_0 \to \infty \), then \( Z_n \) meets the standard normal distribution, that is \( Z_n \to N(0,1) \).

Theorem 3 [9]: Let \( Y_k \) be an independent \( N(0,1) \) distributed random variable sequence, \( Y \geq e^\varepsilon (\varepsilon \) is the Euler number) is a constant. Then, for any \( \varepsilon > 0 \), there is a natural number \( N(\varepsilon) \) so that the expected value \( EY_k \) of every random variable \( Z_N \):

\[
Z_N = \left[ Y_k, Y_k \geq \sqrt{\frac{2\ln N}{\gamma}}, k = 1, 2, \ldots, N \right]
\]  

(12)

meets the following inequality for every natural number \( N \geq N(\varepsilon) \):

\[
\frac{\gamma}{2\sqrt{\ln N}} < EY_k < (1 + e^{-\gamma/2\sqrt{\ln N}})
\]  

(13)

From the theorem 3 we know that when \( N \) is enough large, all independent random variables \( Z_k (k \in N) \) [9]:

\[
Z_k = \begin{cases} 
1, & Y_k \geq \sqrt{\frac{2\ln N}{\gamma}} \\
0, & Y_k < \sqrt{\frac{2\ln N}{\gamma}}
\end{cases}
\]  

(14)

follows the binomial distribution, then \( Z_N \):
\[ Z_N = \sum_{k=1}^{N} Z_k \]  
(15)

also follows the same binomial distribution.

Using the above results, clusters can be obtained from a group of \( K \) representative signals. In different wavelet coefficients matrix \( B_k \), the elements of the same location can be regarded as independent random variables, because they are signals sampled in different time, and different samples are independent of each other. A random variable sequences \((b_{ij1}), (b_{ij2}), \cdots, (b_{ijk})\) can be formed by the elements of the same position in the wavelet coefficients matrix \( B_k \), then, according to equation (11) after the random variables are transformed we can get a new random variable \( Z_{ij} \):

\[
Z_{ij} = \frac{\sum_{k=1}^{K} (b_{ij})_k - \sum_{k=1}^{K} E(b_{ij})_k}{P_k}
\]
(16)

where, \( E(b_{ij})_k \) is the even value of random variable \((b_{ij})_k\), and \( Z_{ij} \) follows the standard normal \( N(0,1) \) distribution, so we get the new matrix.

\[ Z = (Z_{ij}) \]
(17)

The matrix is statistical synthesis of \( K \) wavelet coefficients matrix, so it has statistical characteristics of wavelet coefficients matrix of \( K \) sampled signals. Because each row of the wavelet coefficients matrix represents a sample signal of different scales, namely represents the size of different frequency components, and that is, if the value of a row elements \( Z_{ij} \) is bigger in the matrix \( Z \), the information contained in the scale is larger, if the value of \( Z_{ij} \) is smaller, the information contained in the scale is smaller, so wavelet coefficients contained large information were used as the characteristic value, and wavelet coefficients contained small information were merged, that is the cluster process. In this paper, the mathematical statistics method was used for the cluster of wavelet coefficients. Due to the size of the wavelet coefficient indicates contained information content, and has nothing to do with the symbol of the wavelet coefficients, as a result, the wavelet coefficients cluster is expressed in absolute value, let matrix:

\[ B_k = (|b_{ij}|)_k \]
(18)

In fact, we can’t obtain the mathematical expectation \( E(b_{ij})_k \) of equation (16), but the average estimation algorithm can be adopted to replace the mathematical expectation, then the matrix in equation (16) and (17) can be expressed as:

\[
G = (g_{ij}) = \frac{1}{\sigma (R[\sum_{k=1}^{K} \hat{B}_k] \mu (R[\sum_{k=1}^{K} \hat{B}_k])] \cdot I \}
\]
(19)

where, \( R \) is an operator which when applied to any matrix \( A \), reduces the dimension of matrix \( A, \mu (A) \) and \( \sigma (A) \) denotes respectively the sample mean and standard variance of elements of the matrix \( A \), matrix \( I \) has the same matrix size as matrix \( \hat{B}_k \), but contains only element 1s. Then from the theorem 3 we can know that the element \( g_{ij} \) in \( G \) also follows \( N(0,1) \) distribution, so applying a threshold value \( T = \sqrt{2(\ln L/\gamma)} \) (\( \gamma \geq \varepsilon^2 \), \( L \) is the number of coefficients computed) to matrix \( G \), a binary matrix can be obtained.

\[
G_k = (\Theta(g_{ij} - T))
\]
(20)

For function \( \Theta(x) \), if \( x \geq 0, \Theta(x) = 1 \), and if \( x < 0, \Theta(x) = 0 \), so in matrix \( G_k \), wavelet coefficients value corresponding to element 1s is bigger and contains more sample information, in the same row it and its nearby wavelet coefficients of zero value were gathered together, and wavelet coefficients between two 1s were divided into two classes according to the center position, in different rows, that is to say wavelet coefficients cluster in different scales are not overlap. So there is a 1 in each cluster, if the whole row of matrix \( G_k \) does not contain 1, this row will be regarded as a cluster.

### 3.2. Feature Extraction

From the cluster process of wavelet coefficients described in the above subsection, cluster \( U_1, U_2, \cdots, U_c \) (the cluster number) can be determined from a set of typical signals. In this paper, the square root of energy of wavelet coefficients was used as signal characteristics, and then the number of signal characteristics is equal to the number of cluster, so the feature vector can be obtained with a three-step procedure. In the first step, the fast wavelet transform is applied to sampled discrete digital signal \( s \) to obtain the matrix \( B \) of wavelet coefficients; in the second step, matrix \( B \) was clustered according to the same matrix model in equation (20), \( U_1, U_2, \cdots, U_c \) represents the cluster, each row vector that is formed by elements of each cluster \( U_i \) in the matrix \( B \) is expressed by \( r_i \) \( (i = 1,2,\cdots,c) \); In the third step, the Euclidean norm of each vector \( r_i \) was determined as one feature \( u_i \), namely, each feature \( u_i \) was defined as the square root of energy of the wavelet coefficients corresponding cluster \( U_i \):

\[
\|r_i\|_2 = \frac{\sqrt{\sum_{v \in U_i} V^2}}{V^2}
\]
(21)

So the number \( c \) of features of a signal \( s \) is equal to the number of cluster determined with the foregoing method. From the above process we can see that each character represents a set of wavelet coefficients. That is, it expresses the time domain and frequency domain information of sampled signal \( s \), and in addition, the wavelet coefficients with different scales describe the characteristics of a certain frequency range. From the perspective of the process of feature extraction the following relation is established. It shows that the established eigenvector is robust to the noise in the corresponding signal \( s \).

\[
||a_0||^2 = \sum_{i=1}^{c} u_i^2
\]
(22)

### 4. CLASSIFICATION EXPERIMENT

Using neural network to train samples has many disadvantages such as low speed, easy to fall into local
Table 1. Cluster of coefficients computed with the fast wavelet transform.

(1-1)

<table>
<thead>
<tr>
<th>Scale m</th>
<th>Wavelet Coefficient of a Signal s</th>
<th>Cluster Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{d_{1(1)}, d_{1(3)}, \ldots, d_{1(15)}}</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>{d_{2(1)}, d_{2(3)}, \ldots, d_{2(15)}}</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>{d_{3(1)}, d_{3(3)}, d_{3(5)}}, {d_{3(10)}, d_{3(12)}, d_{3(15)}}</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>{d_{4(1)}, d_{4(3)}, d_{4(5)}}, {d_{4(6)}, d_{4(8)}, d_{4(10)}, d_{4(12)}, d_{4(14)}, d_{4(16)}, d_{4(18)}, \ldots, d_{4(15)}}</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>{d_{5(1)}, d_{5(3)}, d_{5(5)}, d_{5(7)}, \ldots, d_{5(15)}}</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>{a_{5(1)}, a_{5(3)}, a_{5(5)}, a_{5(7)}, \ldots, a_{5(15)}}</td>
<td>3</td>
</tr>
</tbody>
</table>

minimum value, and support vector machine (SVM) [10] can solve these problems very well. The SVM has been widely used in the field of pattern recognition and shows good performance. For the identification of workpiece internal defects, under the condition of lacking of prior knowledge and small sample SVM has more advantages than other methods. In this paper, previous proposed feature extraction method and SVM was combined to identify the internal defects of steel plates.

The research object of this identification experiment is three common internal defects of steel plates: hierarchical, osteoporosis and white spot [11]. Sampling frequency is 2.5MHz, 10 samples of each defect was collected respectively, and 1024 points near defect wave was intercepted, wavelet basis function used was Db7, the scale of wavelet transform M=5. Fast wavelet transform was used for each sample, and then decomposed wavelet coefficients were clustered according to proposed method, when binaryzation is applied to equation (20), if the threshold T is too large, characteristic value was more, the network structure was complex, but the classification accuracy is high, on the contrary, if the threshold T is too small, characteristic value is less, the network structure is simple, but the classification accuracy is reduced. Overall consideration, T=1.5 was selected, the cluster result was shown in Table 1, according to the cluster method in 3.1, corresponding to five levels of discrete wavelet decomposition of the original signal, matrix A was clustered into 14 classes U1, U2, \ldots, U14. In Table 1, scale 4 contains four clusters, and scale 3 contains two clusters, that means scale 4 carry more useful information than scale 3. So there sums to 3 * 10 = 30 feature vector of learning samples, that is, 30 groups of clustered wavelet coefficients, the dimension of the feature vector of each sample is 14.

Ten test samples of each type of defect were obtained, a total of thirty test samples, SVM was used as the identifier and radial basis function was used as the kernel function of SVM classifier. The one-to-one classification method was used to classify test samples, with 96% accuracy. In addition, the wavelet coefficients of scale 4 was used as extracted features, using the same identifier, the identify accuracy is 90%, and wavelet coefficients of scale 5 was used as extracted features, using the same identifier, the identify accuracy is 83%. The proposed feature extraction method kept all of the wavelet decomposition coefficients without losing any information, as well as the cluster of wavelet coefficients reduced data dimension, simplify the recognition process of recognizer. So proposed feature extraction method improved the accuracy of defect recognition and was effective and reliable for the defect type recognition of ultrasonic nondestructive testing.

CONCLUSION

Taking the internal defects of plates as identification examples, the fast wavelet transform was computed for the sampling signals, wavelet coefficients matrices were obtained, the cluster method of probability and statistics was used to extract the wavelet coefficients containing more sample information and merge the wavelet coefficients containing less sample information in matrices, the energy value of wavelet coefficients was calculated for each cluster and was used as the input feature vectors of SVM classifier. Each feature represented a set of wavelet coefficients, greatly reduced the dimension of input vector and facilitated the process of pattern recognition, at the same time retained specific problems information of measured signal, improved the classification accuracy.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was financially supported by project of the science and technology development fund of Tianjin City (20110831).

REFERENCES


