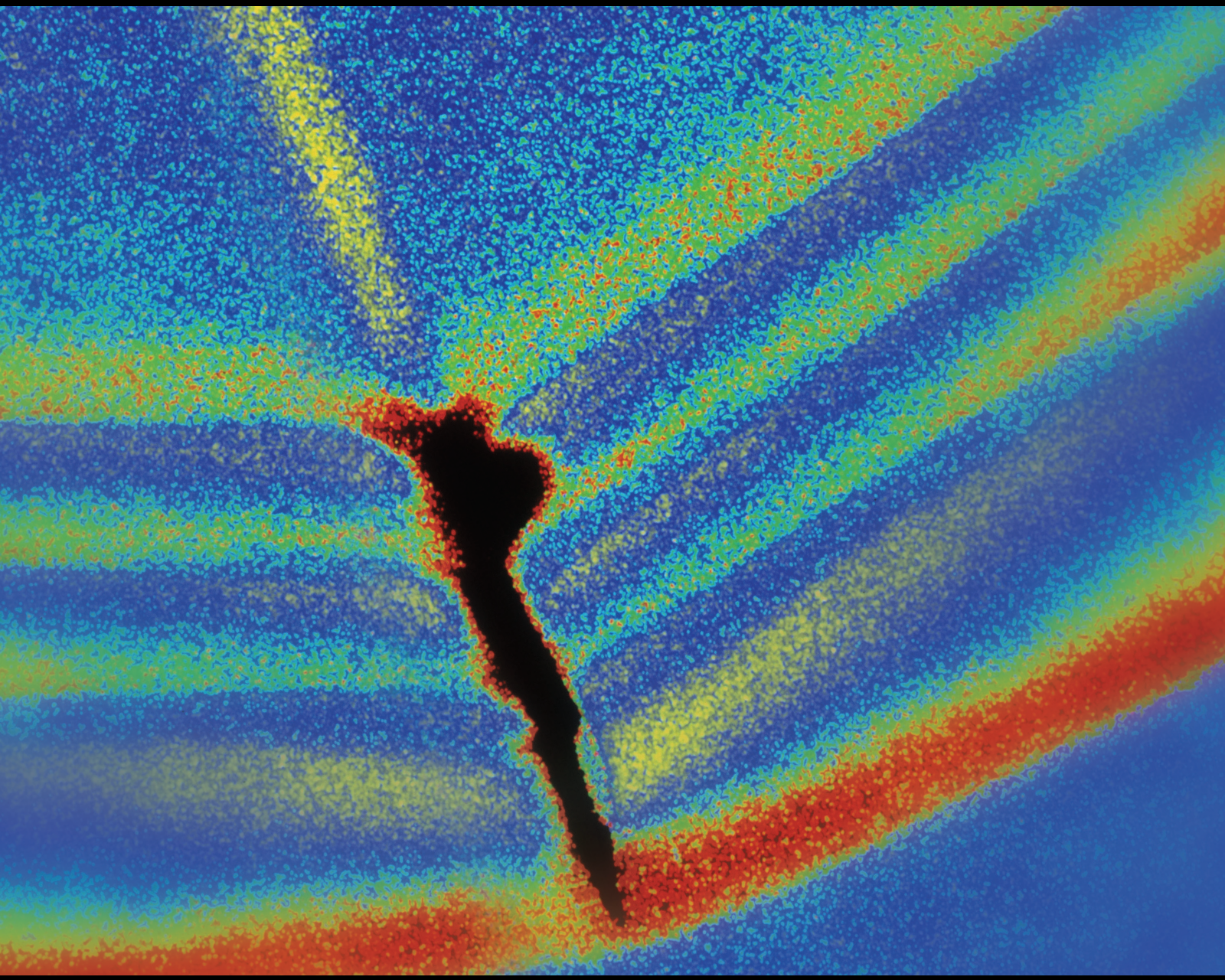
Shock and Vibration

Advances in Vibration Signal Time-Frequency Analysis for Defect Detection 2021

Lead Guest Editor: Shengwei Fei
Guest Editors: Chaoqun Duan and Abolfazl Gharaei



Advances in Vibration Signal Time-Frequency Analysis for Defect Detection 2021

Shock and Vibration

Advances in Vibration Signal Time-Frequency Analysis for Defect Detection 2021

Lead Guest Editor: Shengwei Fei
Guest Editors: Chaoqun Duan and Abolfazl Gharaei

Chief Editor

Associate Editors

Academic Editors

Contents

Hindawi
Shock and Vibration
Volume 2021, Article ID 2030128, 5 pages
https://doi.org/10.1155/2021/2030128

Research Article

An Adaptive Variational Mode Decomposition Technique with Differential Evolution Algorithm and Its Application Analysis

Yuanxin Wang

China Gezhouba Group Co., Ltd., Beijing 100022, China

Correspondence should be addressed to Yuanxin Wang; wangyuanxintdc@cggc.cn

Received 5 September 2021; Revised 23 September 2021; Accepted 27 October 2021; Published 11 November 2021

Academic Editor: Chaoqun Duan

Variational mode decomposition is an adaptive nonrecursive signal decomposition and time-frequency distribution estimation method. The improper selection of the decomposition number will cause under decomposition or over decomposition, and the improper selection of the penalty factor will affect the bandwidth of modal components, so it is very necessary to look for the optimal parameter combination of the decomposition number and the penalty factor of variational mode decomposition. Hence, differential evolution algorithm is used to look for the optimization combination of the decomposition number and the penalty factor of variational mode decomposition because differential evolution algorithm has a good ability at global searching. The method is called adaptive variational mode decomposition technique with differential evolution algorithm. Application analysis and discussion of the adaptive variational mode decomposition technique with differential evolution algorithm are employed by combining with the experiment. The conclusions of the experiment are that the decomposition performance of the adaptive variational mode decomposition technique with differential evolution algorithm is better than that of variational mode decomposition.

1. Introduction

Variational mode decomposition is an adaptive nonrecursive signal decomposition and time-frequency distribution estimation method [1]. It can decompose the input signal into several intrinsic mode functions with bandwidth constraints and make each group of intrinsic mode function focus near different central frequencies, which has good noise robustness [2]. In recent years, variational mode decomposition has been applied to prediction field [3, 4]. The improper selection of the decomposition number will cause under decomposition or over decomposition, and the improper selection of the penalty factor will affect the bandwidth of modal components; that is, the improper parameter combination of variational mode decomposition will influence the decomposition performance and the using effects in prediction field of variational mode decomposition. Hence, it is very necessary to look for the optimal parameter combination of the decomposition number and the penalty factor of variational mode decomposition.

Differential evolution algorithm is a population-based evolutionary optimization algorithm, which are motivated by the social behavior of animals [5–7]. Differential evolution algorithm is an efficient global optimization algorithm with simple structure, superior performance, and strong robustness, which can achieve a global optimal solution [8, 9]. Hence, differential evolution algorithm is used to look for the optimization combination of the decomposition number and the penalty factor of variational mode decomposition, which can obtain the variational mode decomposition model with superior performance. The method is called adaptive variational mode decomposition technique with differential evolution algorithm. As the prediction for birth population is helpful for decision-making of related department, China's birth data are used to testify the decomposition performance of the adaptive variational mode

decomposition technique with differential evolution algorithm, and the decomposition performance of the adaptive variational mode decomposition technique with differential evolution algorithm is judged by the prediction results for China's birth population.

2. Adaptive Variational Mode Decomposition Technique with Differential Evolution Algorithm

Variational mode decomposition is an adaptive non-recursive signal decomposition and time-frequency distribution estimation method. It can decompose the input signal into several intrinsic mode functions with bandwidth constraints and make each group of intrinsic mode function focus near different central frequencies, which has good noise robustness. The basic idea of variational mode decomposition technique is to transform the signal decomposition process into a constraint variational problem [10].

In the variational mode decomposition technique, the constrained optimization problem can be described by the squared L^2 norm form:

$$\begin{aligned}&\min_{(u_k,\omega_k)}\left\{\sum_{k=1}^{K}\left\|\partial_t\left[\left(\sigma(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-j\omega_k t}\right\|_2^2\right\}\\&\text{s.t.}\sum_{k=1}^{K}u_k(t)=s(t),\end{aligned}\tag{1}$$

where K is the decomposition number, $\sigma(t)$ is the Dirac distribution, $s(t)$ is a signal with a set of time series data, $u_k(t)$ is an intrinsic mode function of $s(t)$, ω_k is the center frequency of $u_k(t)$, ∂_t is the gradient operator, and $*$ is the convolution operator.

By introducing the penalty factor and Lagrangian multiplier, the constrained optimization problem can be converted into the mathematical formulation of an unconstrained optimization problem obtained with the induction of the augmented Lagrangian:

$$L_{(u_k,\omega_k,r)}=\alpha\sum_{k=1}^{K}\left\|\partial_t\left[\left(\sigma(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-j\omega_k t}\right\|_2^2+\left\|s(t)-\sum_{k=1}^{K}u_k(t)\right\|_2^2+\langle r(t),s(t)-\sum_{k=1}^{K}u_k(t)\rangle,\tag{2}$$

where α is the penalty factor and $r(t)$ is the Lagrangian multiplier.

The multiplication operator alternating direction method is used to search the optimal solution of the above variational problem; the updated formulas of u_k, ω_k, and r are shown below:

$$\begin{aligned}U_k^{n+1}(\omega)&=\frac{S(\omega)-\sum_{i\neq k}U_i(\omega)+0.5R(\omega)}{1+2\alpha(\omega-\omega_k)^2},\\\omega_k^{n+1}&=\frac{\int_0^\infty\omega|U_k(\omega)|^2\,d\omega}{\int_0^\infty|U_k(\omega)|^2\,d\omega},\\R^{n+1}(\omega)&=R^n(\omega)+\varepsilon\left(S(\omega)-\sum_{k=1}^{K}U_k^{n+1}(\omega)\right),\end{aligned}\tag{3}$$

where $S(\omega)$ is the Fourier transform signal of $s(t)$, $R(\omega)$ is the Fourier transform signal of $r(t)$, $U_k(\omega)$ is the Fourier transform signal of $u_k(t)$, n is the iterative times, and ε is the noise tolerance coefficient.

Decomposition number K and penalty factor α are two important parameters to determine whether variational mode decomposition can achieve good results. The improper selection of K will cause under decomposition or over decomposition, and the improper selection of α will affect the bandwidth of modal components, so it is very necessary to look for the optimal parameter combination of K and α.

Differential evolution algorithm is a population-based evolutionary optimization algorithm, which are motivated by the social behavior of animals. Differential evolution algorithm is an efficient global optimization algorithm with simple structure, superior performance, and strong robustness, which can achieve a global optimal solution. Hence, differential evolution algorithm is used to look for the optimization combination of the parameters K and α of variational mode decomposition. Differential evolution algorithm includes three operations: mutation operation, crossover operation, and selection operation [11, 12]. In mutation operation, the mutation vector is generated by the mutation operator. Generate the recombinant population by the crossover operation. In selection operation, the better individual is selected to produce the next offspring according to the selection operator. The process of obtaining the optimization combination of the parameters K and α of variational mode decomposition by differential evolution algorithm is as follows:

Step 1: as there are two parameters to be optimized, an initial population $\{X_i=(x_{i,1},x_{i,2})|i=1,2,\ldots,N\}$ is randomly generated in a given space according to equation (4), and the individual is composed of the parameters K and α of variational mode decomposition:

$$X_i=X_i^{\text{low}}+\text{rand}\cdot\left(X_i^{\text{up}}-X_i^{\text{low}}\right),\tag{4}$$

where X_i^{low} is the ith individual's lower bound, X_i^{up} is the ith individual's upper bound, and rand is the random value with uniform distribution from 0 to 1.

Step 2: average envelope entropy, shown in equation (5), is used as the objective function and computes the values of the objective function of the individuals:

$$G_{(K,\alpha)} = \frac{1}{K}\sum_{k=1}^{K}\left[-\sum_{m=1}^{M}\left(\frac{c_k(m)}{\sum_{m=1}^{M}c_k(m)}\right)\log_2\left(\frac{c_k(m)}{\sum_{m=1}^{M}c_k(m)}\right)\right], \quad (5)$$

where $c_k(m)$ is the kth intrinsic mode function's envelope entropy and M is the sampling numbers.

The optimization problem of the combination of the parameters K and α of variational mode decomposition is expressed as $\min\{G_{(K,\alpha)}\}$. The smaller the value of the objective function of the individual is, the better the combination of the parameters K and α of variational mode decomposition is.

Step 3: generate the mutated individuals with three different random individuals according to

$$y_{i,j}(t+1) = x_{z1,j}(t) + F_i(t)\left[x_{z2,j}(t) - x_{z3,j}(t)\right], \quad (6)$$

where $F_i(t)$ is the ith individual's scaling factor, $x_{z1,j}(t)$, $x_{z2,j}(t)$, and $x_{z3,j}(t)$ are three different individuals, and $i \neq z1 \neq z2 \neq z3$.

Step 4: generate the recombinant population by the crossover operation, if rand is less than or equal to CR or j is equal to j_{rand}, $u_{i,j}(t) = y_{i,j}(t)$, otherwise, $u_{i,j}(t) = x_{i,j}(t)$, where CR is the crossover rate and j_{rand} is the integer randomly chosen from 1 to 2.

Step 5: the individual with smaller value of the objective function is selected to produce the next offspring according to the selection operator.

Step 6: repeat the evolutionary cycle until the maximum number of iterations is reached. Otherwise, go to Step 2. Then, the optimal combination of the parameters K and α of variational mode decomposition is obtained.

3. Application Analysis and Discussion of the Adaptive Variational Mode Decomposition Technique with Differential Evolution Algorithm

The number of births has a great influence on society and economy of a country; the prediction for birth population is helpful for decision-making of related department. Hence, China's births data are used to testify the decomposition performance of the adaptive variational mode decomposition technique with differential evolution algorithm. The optimal combination of the parameters K and α is obtained by differential evolution algorithm, and the adaptive variational mode decomposition model with differential evolution algorithm is realized to obtain the variational mode decomposition model with superior performance. China's birth data are decomposed to form several intrinsic mode functions by using the adaptive variational mode decomposition technique with differential evolution algorithm.

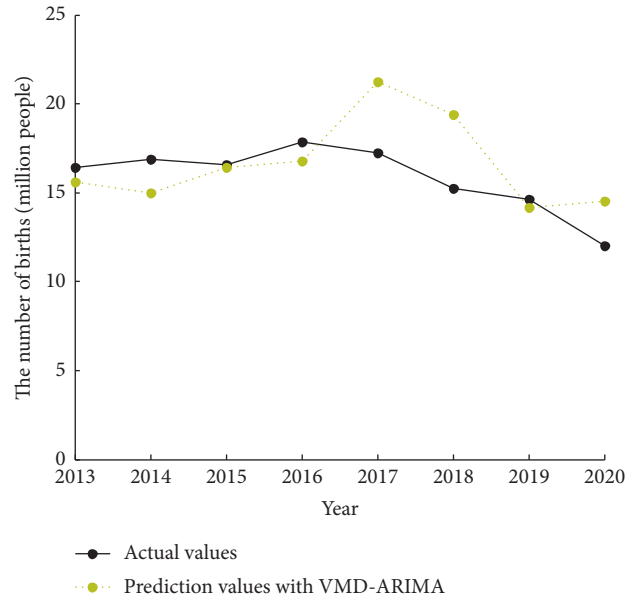


Figure 1: The actual China's birth population and the prediction results for China's birth population with VMD-ARIMA.

Autoregressive integrated moving average model is used to predict the several intrinsic mode functions of China's birth population; autoregressive integrated moving average is a simple and practical prediction method, which reduces the complexity of the model and is suitable for the prediction problem of small samples [13–15]. The intrinsic mode functions of China's birth population have an important influence on the prediction results for China's birth population, so the decomposition performance of the adaptive variational mode decomposition technique with differential evolution algorithm can be judged by the prediction results for China's birth population.

The actual China's birth population and the prediction results for China's birth population with variational mode decomposition-autoregressive integrated moving average (VMD-ARIMA) are given in Figure 1, and the actual China's birth population and the prediction results for China's birth population with adaptive variational mode decomposition technique with differential evolution algorithm-autoregressive integrated moving average (DAVMD-ARIMA) are given in Figure 2. By the trend analysis of the prediction results for China's birth population with DAVMD-ARIMA, China's birth population takes on decreasing trend in recent years, which is in accord with actual situation. The trend of the prediction results for China's birth population with DAVMD-ARIMA is near to the trend of the actual China's birth population than the trend of the prediction results for China's birth population with VMD-ARIMA.

According to the relative errors for China's birth population prediction of VMD-ARIMA and DAVMD-ARIMA given in Figure 3, the average relative errors for China's birth population prediction of VMD-ARIMA and DAVMD-ARIMA are calculated. The average relative error for China's

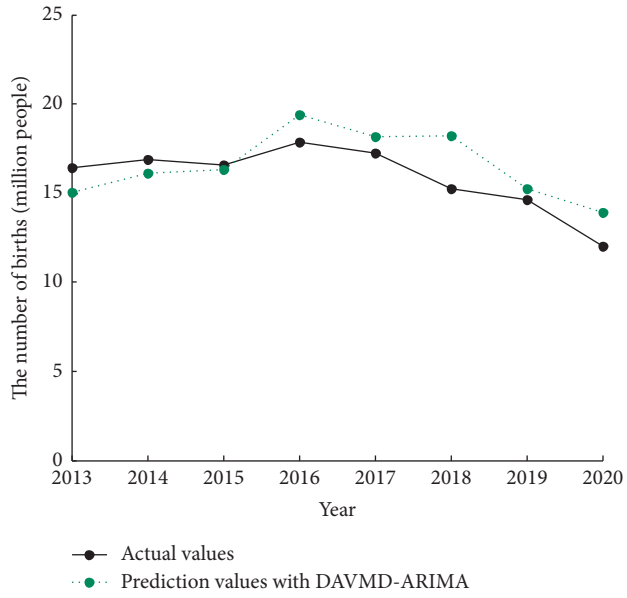


Figure 2: The actual China's birth population and the prediction results for China's birth population with DAVMD-ARIMA.

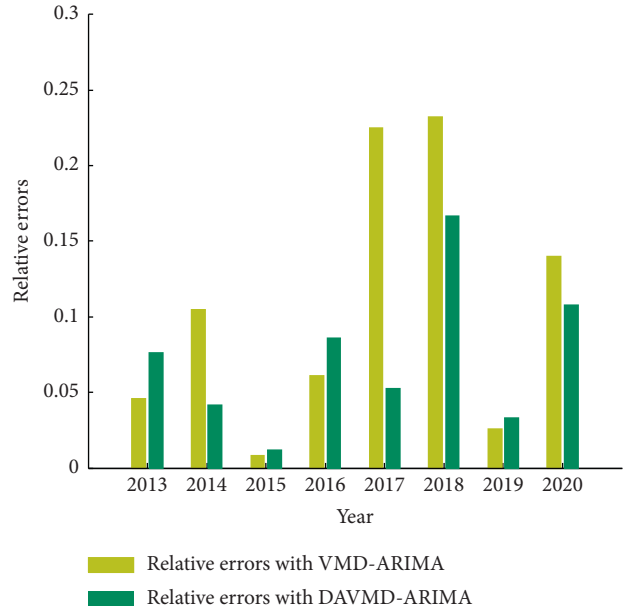


Figure 3: The relative errors for China's birth population prediction of VMD-ARIMA and DAVMD-ARIMA.

birth population prediction with VMD-ARIMA is 10.52%; however, the average relative error for China's birth population prediction with DAVMD-ARIMA is 7.2%; that is, the average relative error for China's birth population prediction with DAVMD is smaller than the average relative error for China's birth population prediction with VMD-ARIMA, so DAVMD-ARIMA is more suitable for China's birth population prediction than VMD-ARIMA. Accordingly, the decomposition performance of the adaptive variational mode decomposition technique with differential evolution algorithm is better than that of variational mode decomposition.

4. Conclusions

In order to solve the selection problem of the parameter combination of the decomposition number and the penalty factor of variational mode decomposition, an adaptive variational mode decomposition technique with differential evolution algorithm is proposed to obtain the variational mode decomposition model with superior performance. As the prediction for birth population is helpful for decision-making of related department, the decomposition performance of the adaptive variational mode decomposition technique with differential evolution algorithm is judged by the prediction results for China's birth population. The results are as follows:

(1) By the trend analysis of the prediction results for China's birth population with DAVMD-ARIMA, China's birth population takes on decreasing trend in recent years, which is in accord with actual situation. The trend of the prediction results for China's birth population with DAVMD-ARIMA is near to the trend of the actual China's birth population than the trend of the prediction results for China's birth population with VMD-ARIMA.

(2) The average relative error for China's birth population prediction with DAVMD-ARIMA is smaller than the average relative error for China's birth population prediction with VMD-ARIMA.

The conclusions are as follows:

(1) DAVMD-ARIMA is more suitable for China's birth population prediction than VMD-ARIMA

(2) Differential evolution algorithm can find the optimization combination of the decomposition number and the penalty factor of variational mode decomposition; then, the decomposition performance of the adaptive variational mode decomposition technique with differential evolution algorithm is better than that of variational mode decomposition

Data Availability

The data used to support the findings of the study are available from the author upon request.

Conflicts of Interest

The author declares no conflicts of interest with respect to the publication of this article.

References

[1] I. C. Yadav, S. Shahnawazuddin, and G. Pradhan, "Addressing noise and pitch sensitivity of speech recognition system through variational mode decomposition based spectral smoothing," *Digital Signal Processing*, vol. 86, pp. 55–64, 2019.

[2] H. Sharma, "Heart rate extraction from PPG signals using variational mode decomposition," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 1, pp. 75–86, 2019.

[3] B. Karan and S. Sekhar Sahu, "An improved framework for Parkinson's disease prediction using Variational Mode Decomposition-Hilbert spectrum of speech signal," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 717–732, 2021.

[4] I. Majumder, P. K. Dash, and R. Bisoi, "Variational mode decomposition based low rank robust kernel extreme learning machine for solar irradiation forecasting," *Energy Conversion and Management*, vol. 171, pp. 787–806, 2018.

[5] P. Chiradeja, S. Yoomak, and A. Ngaopitakkul, "Optimal allocation of multi-DG on distribution system reliability and power losses using differential evolution algorithm," *Energy Procedia*, vol. 141, pp. 512–516, 2017.

[6] W.-S. Lee, Yi-T. Chen, and Y. Kao, "Optimal chiller loading by differential evolution algorithm for reducing energy consumption," *Energy and Buildings*, vol. 43, no. 2-3, pp. 599–604, 2011.

[7] D. Chen, X. Zhang, Y. Xie, and H. Ding, "Precise estimation of cutting force coefficients and cutter runout in milling using differential evolution algorithm," *Procedia CIRP*, vol. 77, pp. 283–286, 2018.

[8] S. Mete, S. Ozer, and H. Zorlu, "System identification using Hammerstein model optimized with differential evolution algorithm," *AEU-International Journal of Electronics and Communications*, vol. 70, no. 12, pp. 1667–1675, 2016.

[9] S. Jazebi, S. H. Hosseinian, and B. Vahidi, "DSTATCOM allocation in distribution networks considering reconfiguration using differential evolution algorithm," *Energy Conversion and Management*, vol. 52, no. 7, pp. 2777–2783, 2011.

[10] A. Bagheri, O. E. Ozbulut, and D. K. Harris, "Structural system identification based on variational mode decomposition," *Journal of Sound and Vibration*, vol. 417, pp. 182–197, 2018.

[11] H. Zorlu, "Optimization of weighted myriad filters with differential evolution algorithm," *AEU-International Journal of Electronics and Communications*, vol. 77, pp. 1–9, 2017.

[12] K. Nandhini and S. R. Balasundaram, "Extracting easy to understand summary using differential evolution algorithm," *Swarm and Evolutionary Computation*, vol. 16, pp. 19–27, 2014.

[13] H.-K. Yu, N. Y. Kim, S. S. Kim, C. Chu, and M. K. Kee, "Forecasting the number of human immunodeficiency virus infections in the Korean population using the autoregressive integrated moving average model," *Osong Public Health and Research Perspectives*, vol. 4, no. 6, pp. 358–362, 2013.

[14] R. Wu, "Least absolute deviation estimation for general fractionally integrated autoregressive moving average time series models," *Statistics & Probability Letters*, vol. 94, pp. 69–76, 2014.

[15] Yi-S. Lee and L.-I. Tong, "Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming," *Knowledge-Based Systems*, vol. 24, no. 1, pp. 66–72, 2011.

Hindawi
Shock and Vibration
Volume 2021, Article ID 3272119, 10 pages
https://doi.org/10.1155/2021/3272119

Research Article

The Novel Sequence Distance Measuring Algorithm Based on Optimal Transport and Cross-Attention Mechanism

Yanmin Yu,[1,2] Yongcai Lai ,[1] Ping Yan,[2] and Haiying Liu[1,2]

[1]*Heilongjiang Academy of Agricultural Sciences Postdoctoral Programme, Harbin 150086, China*
[2]*Biotechnology Institute of Heilongjiang Academy of Agricultural Sciences, Harbin 150028, China*

Correspondence should be addressed to Yongcai Lai; sws@haas.cn

Received 9 July 2021; Revised 3 August 2021; Accepted 11 August 2021; Published 31 August 2021

Academic Editor: Chaoqun Duan

In this paper, we propose a novel sequence distance measuring algorithm based on optimal transport (OT) and cross-attention mechanism. Given a source sequence and a target sequence, we first calculate the ground distance between each pair of source and target terms of the two sequences. The ground distance is calculated over the subsequences around the two terms. We firstly pay attention from each the source terms to each target terms with attention weights, so that we have a representative source subsequence vector regarding each term in the target subsequence. Then, we pay attention from each representative vector of the term of the target subsequence to the entire source subsequence. In this way, we construct the cross-attention weights and use them to calculate the pairwise ground distances. With the ground distances, we derive the OT distance between the two sequences and train the attention parameters and ground distance metric parameters together. The training process is conducted with training triplets of sequences, where each triplet is composed of an anchor sequence, a must-link sequence, and a cannot-link sequence. The corresponding hinge loss function of each triplet is minimized, and we develop an iterative algorithm to solve the optimal transport problem and the attention/ground distance metric parameters in an alternate way. The experiments over sequence similarity search benchmark datasets, including text, video, and rice smut protein sequence data, are conducted. The experimental results show the algorithm is effective.

1. Introduction

1.1. Background. Sequence data is one of the most popular data type in real-world applications of machine learning and data mining [1–6]. For example, in natural language processing, a sentence is a sequence of words, and in computer vision, a video is a sequence of frames, while in bioinformatics, a protein structure is a sequence of amino acids in a polypeptide chain. Unlike the flat vector data of most machine learning problems, sequence data has the following inherent features:

(1) Sequence data is varying at the number of items. The flat feature is usually given at a fixed size, while the length of the sequences could be different, due to the sampling process to form the sequence.

(2) Sequence data has a temporal and relational nature. The order of the items in the sequence plays an important role in the understanding of the sequence. Given two sequences of the same items but with different orders, their meaning could be completely different. This is a critical, different nature different from the flat vector data, where the items of the vector are considered to be independent of each other and their orders are not important for the learning problem.

Given these two natures of the sequence data, the common machine learning methods are not necessarily applicable to the sequence data, such as the classification, similarity comparison, representation, and regression models. The most popular way to handle sequence data is to map a sequence to a flat vector and then apply the conventional methods. However, this methodology usually cannot capture the sequential feature of the data; thus, the results are not satisfying [4, 7–12].Comparing the similarity/

dissimilarity of a pair of sequences is a fundamental problem of sequence data analysis and understanding. The applications include the similarity search [13–16] and nearest neighbor-based classification [17, 18]. However, the similarity of the two sequences has an essential difference compared with the distance/similarity metrics of flat vectors, such as Euclidean distance, ℓp-norm distances, correlation, Mahalanobis distance, and all kinds of learned metrics. The calculation of the distance between a pair of sequences is more difficult than that of the flat vectors, due to the complex nature of the sequences as mentioned above. Two similar sequences may have different lengths because they are generated with different sampling rates, and encoding the temporal patterns and sequential relations of the items of sequences to distance measures is also difficult. To tackle these challenges, various solutions are proposed, such as dynamic time warping (DTW) [19–22] and optimal transport (OT) [23–26]. Most of the methods are based on the item-to-item ground distances of the item pairs of the two sequences and matching them accordingly. The ground distance is extremely important for these methods, but ground distance learning does not receive enough attention from the previous researches. In this paper, we study the problem of learning effective ground distance between the items of the two sequences for the purpose of sequence distance comparison.

1.2. Existing Works. In this section, we reviewed a few ground distance-based sequence distance learning methods.

(1) Villani [23] proposed to compare the distance between two sequences by OT. OT treats one sequence as a set of mass, while the other sequence as a set of demands. The effort to move one unit of mass from the ith item of the source sequence to the jth item of the target sequence is treated as the ground distance between the pairs (i, j). The purpose of OT is to move all the masses from the source sequence to the target sequence, with the minimum amount of effort. To this end, OT minimizes the overall effort of mass moving with regard to the amounts of mass moved from the ith source item to the jth target item for all pairs of (i, j). With the solution of the moved amounts, the overall effort is the distance between the sequences of OT.

(2) Su and Hua [4] improved the OT method to consider the positions of items of both source and target sequences. The thought behind this method is that the moved amount of mass from a source item to a neighboring target item should be larger than the other items. To this end, the two regularization objectives are imposed on the learning process of the moved amounts. The first one calculates a position similarity between each pair of source and target items and impose the corresponding moved amount to be large if the similarity is large. The second one firstly constructs a position distance between the pair of source and target items, converts it to the probability of positions being nearby, and finally minimizes the Kullback–Leibler (KL) divergence between the probability and moved amount of each pair.

(3) Su and Wu [7] developed a novel ground distance metric learning algorithm by firstly combining a sequence with its label to form a metasequence and then learn the ground distance to compare the sequence to the metasequence. A linear transformation function is designed to map the sequence to a new space, where the sequence items are calculated. With the ground distances of pairs, the OT method is applied to compare the sequence to the meta sequence. The linear transformation parameters and the transportation amount jointly in a minimization problem where a training set of sequences.

(4) Su et al. [5] designed a novel sequence representation and similarity learning method by using dimensionality reduction to the feature vectors of the items of sequences. It firstly maps the features of the items to a low-dimensional space so that the sequence classes are separated as much as possible. The class separability is measured by the sequence statistics, and different forms of statistics lead to different dimensional reduction methods. Two statistics are considered, which are model-based and distance-based. The model-based method explores the dynamical structure of the sequences, while the distance-based one explores the similarity of pairs of sequences.

1.3. Our Contribution. It has been proven that the OT-based sequence distance comparison is the most powerful method for the classification and retrieval of sequence data. The most critical factor of the OT-based method is actually the ground distance measure between the items. Although there are many studies on how to improve the OT-based sequence distance learning, however, most of them are focusing on learning the optimal parameters of OT, with a given ground distance metric. However, the ground distance is a critical component of the OT, and the quality of the ground distance directly affects the quality of OT-based distance methods. In this paper, we proposed a novel ground distance metric learning method, which employs the cross-attention mechanism [27–31]. To calculate the ground distance between two items from two sequences, respectively, we firstly represent each item by paying attention from itself to the neighbors of the other item and then paying attention back. The representation vector of one item is the linear combination of its neighbors weighted by the attention scores. The attention scores are the normalized similarity between a neighboring item and the target item. To learn the parameters, we build a unified learning framework to optimize the attention layers and the OT parameters, which are transported amounts. Our contributions of this work are listed as follows:

(1) We proposed a novel learning framework to learn the ground distance and OT parameters jointly. In this framework, the ground distance model is composed of cross-attention layers, and OT-based sequence distance is parameterized by the transport amounts. The learning framework allows the attention layers and the transport mounts to regularize the learning of each other. This is the first learning framework to guide the learning of attention layers by OT.

(2) We model the learning framework as a minimization problem and develop an iterative algorithm to solve it. In this algorithm, the attention weight parameters and the transport amounts are updated alternately until the algorithm converges. In each iterative step, we consider the optimization of the parameters one by one, while fixing the other parameters, by solving the suboptimization problems.

(3) We conducted extensive experiments over four benchmark datasets to compare our algorithm against the other sequence distance comparison algorithms. Experimental results show the advantage of the attention-based OT algorithm, and we also show the stable property of the algorithm regarding the change of the trade-off parameter and the iteration number.

1.4. Paper Organization. We organize the following parts of this paper as follows: in Section 2, we introduce the proposed algorithm of sequence distance comparison, in Section 3, we conduct experiments to compare our algorithm against the other popular sequence distance methods and also study the properties of the algorithm, and in Section 4, we give the conclusion of this paper and some future works of attention-based sequence distance learning.

2. Proposed Method

2.1. Problem Modeling. Suppose we have two sequences of items, denoted as $X = \{x1, \ldots, xn\}$ and $YY = \{y1, \ldots, yn\}$, where $xi \in R^{dx}$ is the vector of the ith item of X, and $yj \in R^{dy}$ is the vector of the jth item of Y. To calculate the distance between them, we firstly define an attention-based ground distance metric and then measure the optimal transport distance according to the ground metric.

2.1.1. Cross-Attention-Based Ground Distance. To calculate the ground distance between the ith item of X, xi, and the jth item of Y, yi, we firstly explore their neighboring items. For xi, we collect the h items in X before it, $x_{i-h}, \ldots, x_{i-1}$, and h items after it, $x_{i+1}, \ldots, x_{i+h}$, to form a subsequence around xi, denoted as Ni as the contextual sequence of xi. Similarly, we have the h items before and after yj from Y as its contextual sequence, Mj:

$$\begin{aligned} N_i &= \{x_{i-h}, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_{i+h}\}, \\ M_j &= \{y_{j-g}, \ldots, y_{j-1}, y_j, y_{j+1}, \ldots, y_{j+g}\}. \end{aligned} \tag{1}$$

In this way, the contextual and temporal information of each item is effectively encoded in the subsequences Ni and Mj. To compare the dissimilarity between xi and yj, we compare the two subsequences by the cross-attention mechanism.

2.1.2. Attention from Items of Ni to yj. Firstly, to compare the dissimilarity between the Ni and yj, we calculate the attention from items of Ni to yi. To estimate the attention weight, we firstly calculate the affinity between $xl \in Ni$ and $yk \in Mj$:

$$\omega_{lk} = f\left(\theta^{\top}\begin{bmatrix} xl \\ yk \end{bmatrix}\right), \quad \forall l: xl \in Ni, k: yk \in Mi, \tag{2}$$

where $f(\cdot)$ is a nonlinear activation function, such as hyperbolic tangent transformation, and $\theta \in R^{(dx+dy)}$ is the parameter of the affinity function. The attention weights are obtained by softmax normalization over the items of Ni:

$$\alpha_{lk} = \frac{\exp(\omega_{lk})}{\sum_{l':\, xl' \in Ni} \exp(\omega_{l'k})}. \tag{3}$$

With the attention weights from xl to yk, we calculate a representative vector of Ni with attention to yk:

$$z_k^i = \sum_{l:\, xl \in Ni} \alpha_{lk} x_l \in R^{dx}, \tag{4}$$

as the weighted sum of the items $xl \in Ni$.

2.1.3. Attention from z_k^i of Mj to Ni. We represent the subsequence Ni by averaging the vectors of the items as

$$\overline{x}_i = \frac{1}{|Ni|} \sum_{l:\, xl \in Ni} xl \in R^{dx}. \tag{5}$$

Again, we would like to pay attention from each z_k^i to xi. Similarly, we first calculate the affinity between them as follows:

$$\varpi_k^i = f\left(\phi^{\top}\begin{bmatrix} z_k^i \\ \overline{x}i \end{bmatrix}\right), \tag{6}$$

where $\phi \in R^{2dx}$ is the affinity function parameter. From the affinities between $\overline{x}i$ and $z_k^i | k: yk \in Mj$, we calculate the attention weights from xi to $yk \in Mj$ with a softmax function,

$$\beta_k^i = \frac{\exp\left(\varpi_k^i\right)}{\sum_{k':\, yk' \in Mj} \exp\left(\varpi_{k'}^i\right)}. \tag{7}$$

2.1.4. Cross-Attention-Based Ground Distance. To compare the distance between the representative vector z_k^i to and yk, we first perform a linear transformation over z_k^i by

$$W^{\top} z_k^i \in R^{dy}, \tag{8}$$

with $W \in R^{dx \times dy}$ as parameter. This transformation is to map z_k^i to the same space as yk. Then, we compare their distance by the squared Euclidean distance:

$$d\left(z_k^i, y_k\right) = \left\|W^\top z_k^i - yk\right\|_F^2. \tag{9}$$

The final distance between Ni and Mj is the attention-weighted sum of distances $d(z_k^i, yk)|k\colon k \in Mj$. The attention is calculated in equation (7), and the distance $d(Ni, Mj)$ is

$$d(Ni, Mj) = \sum_{k\colon yk \in Mj} \beta_k^i d\left(z_k^i, yk\right). \tag{10}$$

2.1.5. Optimal Transport Distance. With the ground distance, $d(Ni, Mj)$, between each pair of items $(xi, yj)|xl \in X, yk \in Y$ of two sequences, we can compute the transport distance. The ground distance between xi and yj is viewed as the effort to move one unit of mass from xi to yj. We define a variable, ηij, to denote the mount of mass moved from xi to yj, then the total effort to move the mass from X to Y is calculated as

$$\sum_{i,j\colon xi \in X, yj \in Y} \eta ij\, d(Ni, Mj). \tag{11}$$

Moreover, we define an amount of mass for each item xi of X to be moved out, γi. Thus, the constraint of the amounts moved out of xi is applied as

$$\sum_{j\colon yj \in Y} \eta ij = \gamma i \tag{12}$$

We also define an amount of mass to be received by each item yi of Y, δj, and accordingly

$$\sum_{i\colon xi \in X} \eta ij = \delta j. \tag{13}$$

The optimal transport distance between X and Y is achieved by solving the moved amounts to minimize the moving efforts with the above constraints:

$$\begin{aligned} d(X, Y) = \begin{cases} \min\limits_{tij\colon xi \in X, yj \in Y} & \sum\limits_{i,j\colon xi \in X, yj \in Y} \eta ij\, d(Ni, Mj) \\ \text{s.t.} & \eta ij \geq 0,\ \forall i, j\colon xi \in X,\ yj \in Y, \end{cases} \\ \sum_{j\colon yj \in Y} \eta ij = \gamma i, \quad \forall i\colon xi \in X, \\ \sum_{i\colon xi \in X} \eta ij = \delta j, \quad \forall j\colon yj \in Y. \end{aligned} \tag{14}$$

We rewrite the optimal transport distance as matrix form by defining the following matrices and vectors:

$$\begin{aligned} T &= [\eta ij] \in R_+^{n \times m}, \\ D &= [d(Ni, Mj)] \in R_+^{n \times m}, \\ \gamma &= [\gamma 1, \ldots, \gamma n]\top \in R_+^n, \\ \delta &= [\delta 1, \ldots, \delta m]\top \in R_+^m. \end{aligned} \tag{15}$$

We rewrite equation (14) as

$$\begin{aligned} &d(X, Y) = \min_{T \in \Theta} \operatorname{tr}\left(T^\top D\right), \\ &\text{where } \Theta = \left\{T | T \in R_+^{n \times m}, T1_n = \gamma, T^\top 1_m = \delta\right\}, \end{aligned} \tag{16}$$

where $tr(\cdot)$ is the trace of a matrix and 1_n is a vector of n ones.

2.1.6. Supervised Learning of Attention Parameters and Ground Distance Metric. In the distance measure of optimal transport, we need to learn the parameters of the two attention layers, θ and ϕ, and the parameter of the ground distance W. To learn these parameters, we have a training set of T triplets of sequences:

$$T = \left\{\left(X_t, Y_t^+, Y_t^-\right)\right\}\Big|_{t=1}^T. \tag{17}$$

The tth triplet is composed of an anchor sequence, Xt, a must-link sequence Y_t^+, and a cannot-link sequence Y_t^-. The must-link sequence is supposed to have a short distance to anchor sequence, while the cannot-link sequence is supposed to have a long distance to the anchor. In our scenario, we impose the cannot-link sequence has a longer distance to the anchor than the must-link one, with a margin of ε:

$$d\left(X_t, Y_t^-\right) > d\left(X_t, Y_t^+\right) + \varepsilon. \tag{18}$$

Accordingly, we define the hinge loss function as follows:

$$\begin{aligned} L\left(X_t, Y_t^+, Y_t^-; W, \theta, \phi\right) &= \max\left(0, d\left(X_t, Y_t^+\right) - d\left(X_t, Y_t^-\right) + \varepsilon\right) \\ &= \tau t \times \left(d\left(X_t, Y_t^+\right) - d\left(X_t, Y_t^-\right) + \varepsilon\right), \\ \text{where } \tau t &= \begin{cases} 1, & \text{if } d\left(X_t, Y_t^+\right) + \varepsilon > d\left(X_t, Y_t^-\right), \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{19}$$

The corresponding minimization problem is modeled to learn the parameters:

$$\min_{W,\theta,\phi} \left\{\frac{1}{T}\sum_{t=1}^T L\left(X^t, Y_t^+, Y_t^-; W, \theta, \phi\right) + C\left(\|W\|_F^2 + \|\theta\|_F^2 + \|\phi\|_F^2\right)\right\}. \tag{20}$$

In the objective, the first term is the average of the hinge losses over the training triplets. The second term is the squared $\ell 2$-norms of the parameters to reduce the complexity of the model. C is the trade-off parameter.

2.2. Problem Optimization. To solve the problem of equation (20), we substitute equations (16) and (19) into equation (20):

$$\begin{aligned} \min_{W,\theta,\phi} \Bigg\{ o(W, \theta, \phi) &= \frac{1}{T}\sum_{t=1}^T \tau t \\ &\times \left(\min_{T_t^+ \in \Theta} \operatorname{tr}\left(T_t^{+\top} D_t^+\right) - \min_{T_t^- \in \Theta} \operatorname{tr}\left(T_t^{-\top} D_t^-\right) + \varepsilon\right) \\ &+ C(\|W\|2F + \|\theta\|2F + \|\phi\|2F)\Big\}, \end{aligned} \tag{21}$$

where T_t^+ and T_t^- are the transport amount matrix of the positive and negative pairs of the tth training triplet, and D_t^+ and D_t^- are the corresponding ground distance matrix. In this optimization problem, the optimization of the transport amounts is coupled with the optimization of the parameters of the attention layer and the ground distance metric. The optimizations of (W, θ, ϕ) and $(T_t^+, T_t^-)|_{t=1}^T$ are dependent on each other, making the problem difficult to be solved directly. Instead of seeking the close solution of equation (21), we propose to solve the attention and ground distance metric parameters and the optimal transport variables jointly in an unified minimization problem:

$$\begin{cases} \min\limits_{W,\theta,\phi,(T_t^+,T_t^-)|_{t=1}^T} \left\{ o\left(W,\theta,\phi,(T+t,T-t)|_{t=1}^T\right) = \dfrac{1}{T}\sum\limits_{t=1}^{T} \tau t \times \left(\operatorname{tr}\left(T_t^{+\top} D + t\right) - \operatorname{tr}\left(T_t^{-\top} D_t^-\right) + \varepsilon\right) + C\left(\|W\|_F^2 + \|\theta\|_F^2 + \|\phi\|_F^2\right) \right\} \\ \text{s.t. } T + t \in \ominus,\ T - t \in \ominus,\ t = 1, \ldots, T. \end{cases} \tag{22}$$

In this optimization problem, both W, θ, ϕ and $(T_t^+, T_t^-)|_{t=1}^T$ are the variables of a joint objective function. The optimization of both variables are conducted simultaneously. To solve this problem, we use the alternate optimization method. In an iteration of an iterative algorithm, to optimize one parameter, we firstly fix the other parameters and then solve the suboptimization problem with regard to this parameter. The optimizations of these parameters are introduced as follows.

2.2.1. Optimization of W. By fixing the other parameters and only considering W, we have the following suboptimization problem:

$$\min_W \left\{ o1(W) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(\operatorname{tr}\left(T_t^{+\top} D_t^+\right) - \operatorname{tr}\left(T_t^{-\top} D_t^-\right)\right) + C\|W\|_F^2 \right\}. \tag{23}$$

We substitute equations (9) and (10) into equation (16) and rewrite the optimal transport distance between two sequences X and Y as

$$\begin{aligned} \operatorname{tr}\left(T^\top D\right) &= \sum_{i,j:\ xi\in X, yj\in Y} \eta i j\, d(Ni, Mj) \\ &= \sum_{i,j:\ xi\in X, yj\in Y} \eta i j \left(\sum_{k:\ yk\in Mj} \beta i k \|W\top z i k - \ yk\|_F^2 \right) \\ &= \sum_{i,j:\ xi\in X, yj\in Y} \eta i j \left(\sum_{k:\ yk\in Mj} \beta i k \times \left(\operatorname{tr}\left(W^\top z_k^i z_k^{i\top} W\right) \right.\right. \\ &\quad \left.\left. - 2\operatorname{tr}\left(W^\top z_k^i y_k^\top\right) + \ y_k^\top y_k \right)\right) \\ &= \operatorname{tr}\left(W^\top A_{X,Y,T} W\right) \\ &\quad - 2\operatorname{tr}\left(W^\top B_{X,Y,T}\right) + c_{X,Y,T}, \end{aligned} \tag{24}$$

where

$$\begin{aligned} A_{X,Y,T} &= \sum_{i,j:\ xi\in X, yj\in Y} \eta i j \sum_{k:\ yk\in Mj} \beta i k \left(z_k^i z_k^{i\top}\right), \\ B_{X,Y,T} &= \sum_{i,j:\ xi\in X, yj\in Y} \eta i j \sum_{k:\ yk\in Mj} \beta i k \left(z_k^i y_k^\top\right). \end{aligned} \tag{25}$$

Substituting equation (24) into equation (23), we rewrite the objective function as

$$\begin{aligned} o1(W) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \Big(& \left(\operatorname{tr}\left(W^\top A_{Xt,Y_t^+,Tt} W\right)\right. \\ & -2\operatorname{tr}\left(W^\top B_{Xt,Y_t^+,Tt}\right) + c_{Xt,Y_t^+,Tt}\Big) - \Big(\operatorname{tr}\left(W^\top A_{Xt,Y_t^-,Tt} W\right) \\ & \left. -2\operatorname{tr}\left(W^\top B_{Xt,Y_t^-,Tt}\right) + c_{Xt,Y_t^-,Tt}\right)\Big) + C \times \operatorname{tr}\left(W^\top W\right) \\ = \operatorname{tr}\left(W^\top E W\right) & - 2\operatorname{tr}\left(W^\top F\right) + c, \end{aligned} \tag{26}$$

where

$$\begin{aligned} E &= \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(A_{Xt,Y_t^+,Tt} - A_{Xt,Y_t^-,Tt}\right) + C \times I, \\ F &= \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(B_{Xt,Y_t^+,Tt} - B_{Xt,Y_t^-,Tt}\right), \\ c &= \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(c_{Xt,Y_t^+,Tt} - c_{Xt,Y_t^-,Tt}\right). \end{aligned} \tag{27}$$

The problem of minimizing $o1(W)$ of equation (23) has a closed-form solution. It is obtained by setting the derivative of $o1(W)$ with regard to W to zero:

$$\nabla W o1(W) = 2EW - 2F = 0 \Longrightarrow W^* = FE - 1. \tag{28}$$

2.2.2. Optimization of θ. We optimize the attention parameter of equation (2), θ. Fixing the other parameters and removing the irrelevant terms from the objective function, we have the following suboptimization problem for θ:

$$\min_{\theta}\left\{ o2(\theta) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(\mathrm{tr}\left(T_t^{+\top} D_t^{+}\right) - \mathrm{tr}\left(T_t^{-\top} D_t^{-}\right)\right) + C\|\theta\|_F^2 = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \sum_{i,j:\ xi\in Xt, yj\in Y_t^{+}} \eta_{ijt}^{+} + d(Ni, Mj) - \sum_{i,j:\ xi\in Xt, yj\in Y_t^{-}} \eta_{ijt}^{-} d(Ni, Mj) + C\|\theta\|_F^2 \right\}. \tag{29}$$

To solve this problem, we use the gradient descent algorithm as

$$\theta \leftarrow \theta - v\nabla_{\theta} o2(\theta), \tag{30}$$

where v is the descent step and $\nabla\theta o2(\theta)$ is the gradient function. To this end, we calculate the gradient of $o2(\theta)$ with regard to θ by the chain rule:

$$\nabla\theta o2(\theta) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(\sum_{i,j:\ xi\in Xt, yj\in Y_t^{+}} \eta_{ijt}^{+} \nabla_{\theta} d(\theta; Ni, Mj) - \sum_{i,j:\ xi\in Xt, yj\in Y_t^{-}} \eta_{ijt}^{-} \nabla_{\theta} d(\theta; Ni, Mj) \right) + 2C\theta, \tag{31}$$

where $\nabla\theta d(\theta; Ni, Mj)$ is the gradient of ground distance between Ni and Mj regarding θ. We substitute equation (9) into equation (10), and meanwhile rewrite the variables are function of θ, we have

$$\begin{aligned} d(\theta; Ni, Mj) &= \sum_{k:\ y_k\in Mj} \beta_k^i \left\| W\top z_k^i(\theta) - y_k \right\|_F^2, \\ \nabla_{\theta} d(\theta; Ni, Mj) &= 2\sum_{k:\ yk\in Mj} \beta_k^i W\left(W\top z_k^i(\theta) - y_k\right)\nabla_{\theta} z_k^i(\theta). \end{aligned} \tag{32}$$

Moreover, the derivatives of the functions of θ are

$$\nabla_{\theta} z_k^i(\theta) = \sum_{l:\ xl\in Ni} \nabla_{\theta}\alpha_{lk}(\theta) x_l. \tag{33}$$

2.2.3. Optimization of ϕ. To optimize ϕ, we have the following suboptimization problem:

$$\min\phi\left\{ o3(\phi) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(\mathrm{tr}\left(T_t^{+\top} D_t^{+}\right) - \mathrm{tr}\left(T_t^{-\top} D_t^{-}\right)\right) + C\|\phi\|_2^F \right\}. \tag{34}$$

Again, we use the gradient descent algorithm to update ϕ as

$$\phi\leftarrow\phi - v\nabla_{\phi} o3(\phi), \tag{35}$$

where $\nabla_{\phi} o3(\phi)$ is the gradient function of $o3(\phi)$ with regard to ϕ. According to the chain rule, we have

$$\nabla_{\phi} o3(\phi) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(\sum_{i,j:\ xi\in Xt, yj\in Y_t^{+}} \eta_{ijt}^{+} \nabla_{\phi} d(\phi; Ni, Mj) - \sum_{i,j:\ xi\in Xt, yj\in Y_t^{-}} \eta_{ijt}^{-} \nabla_{\phi} d(\phi; Ni, Mj) \right) + 2C\phi. \tag{36}$$

2.2.4. Optimization of $(T_t^{+}, T_t^{-})|_{t=1}^{T}$. To optimize the transport amounts, we have the simplified suboptimization problem as

$$\begin{cases} \min\limits_{(T_t^{+}, T_t^{-})|_{t=1}^{T}} & \left\{ o4\left((T_t^{+}, T_t^{-})|_{t=1}^{T}\right) = \frac{1}{T}\sum_{t=1}^{T} \tau t \times \left(\mathrm{tr}\left(T_t^{+\top} D_t^{+}\right) - \mathrm{tr}\left(T - t\top D - t\right) + \varepsilon\right) \right\} \\ \text{s.t.} & T + t \in \ominus,\ T - t \in \ominus,\ t = 1, \ldots, T. \end{cases} \tag{37}$$

According to the objective, the transport amount matrices $(T_t^{+}, T_t^{-})|_{t=1}^{T}$ are in $2T$ independent objectives, so their solutions are also independent to each other. Thus, we can decompose the optimization problem to $2T$ optimal transport problems. For the tth training triplet, we have the following two minimization problems of optimal transport:

$$\begin{cases} \min\limits_{T_t^{+}} & \frac{\tau_t}{T} \times \mathrm{tr}\left(T_t^{+\top} D_t^{+}\right) \\ \text{s.t.} & T_t^{-} \in \ominus, \end{cases} \qquad \begin{cases} \min\limits_{T_t^{-}} & -\frac{\tau_t}{T} \times \mathrm{tr}\left(T_t^{+\top} D_t^{-}\right) \\ \text{s.t.} & T_t^{-} \in \ominus. \end{cases} \tag{38}$$

Each one of the above problems can be solved as a line programming problem (LP).

2.3. Iterative Algorithm. With these optimization results, we design an iterative algorithm to update the parameters. In this algorithm, the parameters are firstly initialized as random variables. Then in a while loop, they are updated sequentially until a maximum iteration number is reached, or the objective value change is smaller than a given threshold. The algorithm is summarized in Algorithm 1.

3. Experiments

We conduct experiments over four benchmark sequence datasets to verify the performance of the proposed AGD algorithm. The experiments are performed from three aspects:

Input: training set of sequence triplets, $T = \{(Xt, Y_t^+, Y_t^-)\}|_{t=1}^{T}$, maximum iteration number κ, and objective difference threshold δ.
Initialization Initializing parameters (W, θ, φ) as random variables, iteration number $t = 0$.
While $t \leq \kappa$ or $\|o_t - o_{t-1}\|_1 \geq \delta$:
Repeat
(1) Update t according to equation (19) for each training triplet.
(2) Update W according to equation (28).
(3) Update θ by repeating the updating step in equation (30).
(4) Update φ by repeating the updating step in equation (35).
End repeat
Output: (W, θ, φ).

Algorithm 1: Iterative learning algorithm of attention-based ground distance (AGD).

(1) compare with the other sequence similarity/distance methods, (2) study the impacts of the trade-off parameter C, and (3) study the convergence of the iterative algorithm.

3.1. Datasets. In our experiments, we used four benchmark datasets of sequences:

(1) *Spoken Arabic Digits (SAD).* This dataset has 8,800 sequences [32]. Each sequence is a series of speech frames of a wave of a spoken Arabic digit. The vector of each item is the 13-dimensional Mel-frequency cepstrum coefficients feature vector. These sequences belong to 10 classes, and each class is a digit. Each class has 880 sequences. The number of items in each sequence is from 4 to 93.

(2) *NTU RGB + D (NTU).* This dataset has 56,880 sequences [33]. Each sequence is a Kinect video, and each item is a frame of the video. The sequences belong to 60 action classes. The feature vector of each item is constructed by combining the joint locations and the skeleton-based frame wide features.

(3) *Rice Blast Sequence (RBS).* This dataset has 66,153 protein sequences of rice genome proteins, collected from the MSU Rice Genome Database [34]. Each sequence is a sequence of amino acids, and each amino acid is represented by amino acid embedding. The embedding vectors are also learned as a parameter of the model. The sequences are tagged by rice blast disease or not.

(4) *Australian Sign Language (ASL) Signs.* This dataset is composed of 2,565 sequences of sign language signs [32]. The sequences are from 95 classes, and each class has 27 sequences. Each item of a sequence is presented by a 22-dimensional feature vector.

The summary of the statistics of the benchmark datasets is listed in Table 1.

3.2. Experimental Setting

(1) *Training.* To measure the quality of a distance/similarity measure of sequence, we perform the nearest neighbor classification over the sequence data. Given a dataset of sequences with their class labels, we first split the entire dataset by a 10-fold cross-validation protocol. Each fold is used as a test set, while the other folds are used as training folds. Within the training set, we use each sequence as an anchor sequence and randomly pick up another sequence of the same class as its must-link sequence, meanwhile pick up a sequence of a different class as its cannot-link sequence. In this way, we construct the training set of triplets of sequences. The model parameters are trained by the training set and then tested over the test set.

(2) *Testing.* With the trained sequence distance metric, we calculate the distance between each test sequence and each training sequence. The class label of the training sequence with the shortest distance to a test set is assigned to the test, as the classification result of the test sequence.

(3) *Performance Measure.* The accuracy of the test sequences is calculated as the performance. The accuracy rate is the percentage of the correctly classified test sequences over the total number of test sequences.

3.3. Experimental Results

3.3.1. Comparison to Other Methods. We compare the proposed AGD algorithm against the most popular sequence distance learning methods, including the optimal transport (OT) [23], the Order-Reserving Optimal Transport (OPOT) [4], the Regressive Virtual Sequence Metric Learning (RVSML) [7], and the Linear Sequence Discriminant Analysis (LSDA) [5]. The accuracy is reported in Figure 1. From this figure, we can observe that in all the benchmark datasets, the proposed AGD method always has the best performances. The differences between AGD and other methods vary from datasets. For example, in the NTU dataset, the AGD has much better accuracy than the others, while in the RBS dataset, it is only slightly better than the second-best method, LSDA. The main factor behind this phenomenon is the power of the attention mechanism, which embeds each item with its attention to the neighboring items from both the source and target sequences. In most cases, the LSDA is the second-best method, while the original OT method is the worst.

Table 1: Summary of datasets.

Dataset	Number of sequences	Number of classes
SAD	8,800	10
NTU	56,880	60
RBS	66,153	2
ASL	2,565	95

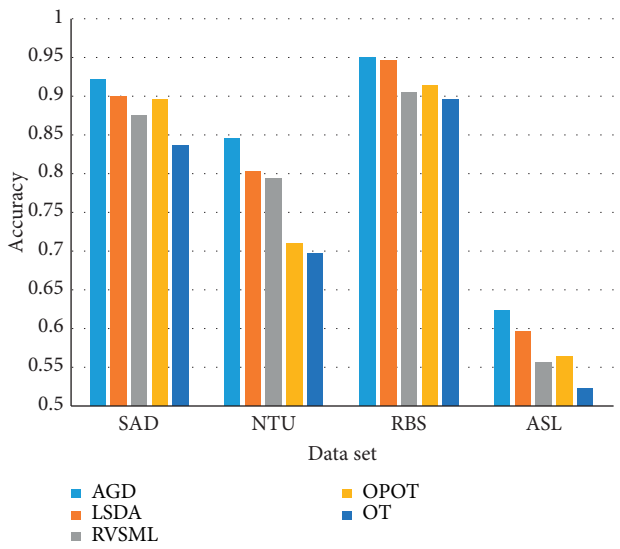


Figure 1: Accuracy of compared methods.

3.3.2. Sensitivity to Trade-Off Parameter. In the objective function of our method, there is only one trade-off parameter, C. It controls the regularization term's importance. We perform experiments with varying values of C and the results are shown in Figure 2. From the curves in the figure, we can see that the proposed AGD algorithm is stable to the changes of the trade-off parameter in most cases. The only exception is the results of the dataset NTU. But the change of the accuracy over the change of the value of C is acceptable. The overall conclusion is that AGD is not sensitive to C. Thus, the parameter tuning of C is easy for the users. One more observation is with the value of C increasing, the accuracy is slightly improving. This also verifies that the regularization term is also beneficial to the model.

3.3.3. Convergence Study. Since our algorithm is an iterative algorithm, we are also interested in the convergence of the algorithm. Thus, we plot the curve of accuracy versus the number of iterations. The curves are given in Figure 3. From this figure, we can see that with the iteration number increasing, the accuracy keeps improving until converge. The number of iterations for the convergence is around 50. The convergence of the algorithm is experimentally verified, and for the size of datasets comparable to our benchmark, the convergence iteration number is acceptable.

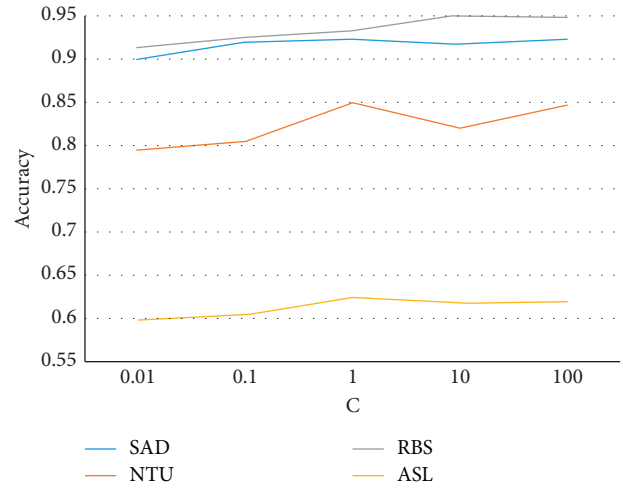


Figure 2: Sensitivity to the trade-off parameter C.

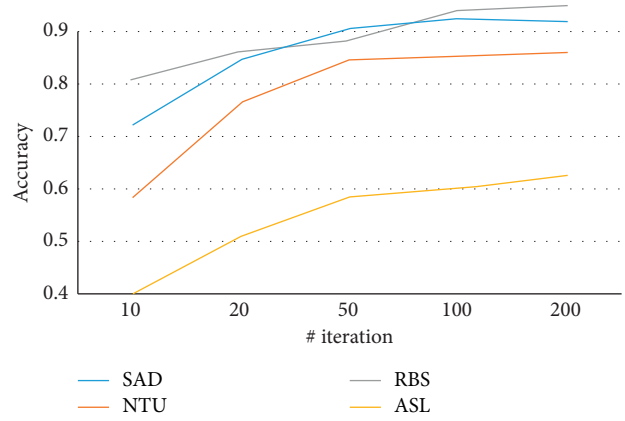


Figure 3: Convergence curves.

We test the significance of the convergence of the accuracy by the Ratio test, and the r values are reported in Table 2. According to these r values, all of them are smaller than 1, meaning all the curves are significantly converged.

3.3.4. Running Time. We also compare the running time of the proposed method. The running times over the four benchmark datasets are shown in Figure 4. From this figure, we have the following conclusions:

(1) Running time and data size are positively correlated. The largest dataset has the longest running time while the smaller one has shorter running time. This is natural since both the training and test processes scan the data points one by one, and more data points means more scanning time.

(2) Our algorithm is faster than the LSDA and RCSML algorithms, while it is slower than the OPOT and OT algorithms. This is acceptable given the significant improvement of the accuracy.

Table 2: r values of ratio test of the convergence of the accuracy.

Dataset	SAD	NTU	RBS	ASL
r value	0.701	0.674	0.833	0.891

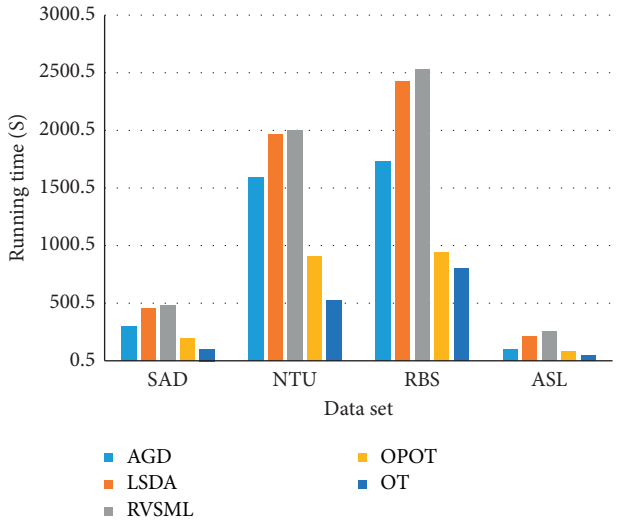


Figure 4: Running time analysis.

4. Conclusion

In this paper, we proposed a novel sequence distance measuring algorithm. This algorithm is based on OT, but its main focus is how to learn an effective ground distance measure for the two sequences. The ground distance learning falls to the framework of the cross-attention mechanism, and the attention layer parameters and the OT parameters are learned jointly. This design can use the OT to guild the learning of the attention layers. Thus, this framework can provide representation of the two sequences, the ground distance, and the OT simultaneously to optimize the model. The learning is also guided by the supervisor of the must-link and cannot-link triplets of the sequences. The parameters are optimized in an iterative algorithm, and the algorithm is tested over four sequence datasets. The experimental results show its advantage over the sequence comparison algorithms.

Data Availability

All the datasets used in this paper to produce the experimental results are publicly accessed online.

Conflicts of Interest

The authors declare that there are no potential conflicts of interest regarding the publication of this study.

Acknowledgments

This work was supported by the Project of "Breeding of New Varieties of High Quality and Anti-Resistant Rice" (Grant no. 2020ZX16B01013), Agricultural Science and Technology Innovation Spanning Project of Heilongjiang Academy of Agricultural Sciences (Grant no. HNK2019CX02), and the National Technology System for Modern Agricultural Industry "Wuchang Integrated Test Station" (Grant no. CARS-01-54).

References

[1] G. Dong and J. Pei, Sequence Data Mining, vol. 33, Springer Science & Business Media, Berlin, Germany.

[2] Yu Chung-Ching Yu and Y.-L. Yen-Liang Chen, "Mining sequential patterns from multidimensional sequence data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 1, pp. 136–140, 2005.

[3] B. Su and Y. Wu, "Learning meta-distance for sequences by learning a ground metricvia virtual sequence regression," *IEEE Annals of the History of Computing*, vol. 1, no. 1, p. 1, 2020.

[4] B. Su and G. Hua, "Order-preserving optimal transport for distances between sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2961–2974, 2018.

[5] B. Su, X. Ding, H. Wang, and Y. Wu, "Discriminative dimensionality reduction for multi-dimensional sequences," *IEEE Transactions on Pattern Analysis Andmachine Intelligence*, vol. 40, no. 1, pp. 77–91, 2017.

[6] C. Wang and H. Mo, "Learning deep attention network from incremental and decremental features for evolving features," *Scientific Programming*, vol. 2021, Article ID 1492828, 8 pages, 2021.

[7] B. Su and Y. Wu, "Learning meta-distance for sequences by learning a ground metricvia virtual sequence regression," *IEEE Transactions on Pattern Analysis and MachineIntelligence*, vol. 1, p. 1, 2020.

[8] B. Liu, C.-C. Li, and K. Yan, "DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks," *Briefings in Bioinformatics*, vol. 21, no. 5, pp. 1733–1741, 2020.

[9] K. James, M. Taylor, A. D. Steen, and A. Sadovnik, "Unaligned sequence similarity search using deep learning," in *Proceedings of the 2019 IEEE International Conferenceon Bioinformatics and Biomedicine (BIBM)*, pp. 1892–1899, IEEE, San Diego, CA, USA, November 2019.

[10] N. Paul, M. Versteegh, and M. Rotaru, "Learning text similarity with siamese recurrent networks," in *Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 148–157, Berlin, Germany, August 2016.

[11] G. Liang, H. Mo, Z. Wang, C.-Q. Dong, and J.-Y. Wang, "Joint deep recurrent network embedding and edge flow estimation," in *Proceedings of the International Conference on Intelligent Computing*, pp. 467–475, Springer, Shenzhen, China, August 2020.

[12] L. Yu, H. Wang, and H. Mo, "Estimating network flowing over edges byrecursive network embedding," *Shock and Vibration*, vol. 2020, Article ID 8893381, 7 pages, 2020.

[13] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching," *Nucleic Acids Research*, vol. 39, no. 2, pp. W29–W37, 2011.

[14] W. R. Pearson, "Flexible sequence similarity searching with the fasta3 program package," in *Bioinformatics Methods and Protocols*, Springer, Berlin, Germany, 2000.

[15] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," in *Proceedings of the International Conference On Foundations Of Data*

Organizationand Algorithms, pp. 69–84, Springer, Chicago, IL, USA, October 1993.
[16] S. Wiseman and A. M. Rush, "Sequence-to-sequence learning as beam-search optimization," 2016, https://arxiv.org/abs/1606.02960.
[17] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *ACM Sigkdd Explorations Newsletter*, vol. 12, no. 1, pp. 40–48, 2010.
[18] L. Neal, M. J. Zaki, and M. Ogihara, "Mining features for sequence classification," in *Proceedings of the 5th ACM SIGKDD international conference onKnowledge discovery and data mining*, pp. 342–346, San Diego, CA, USA, August 1999.
[19] M. Müller, *Dynamic Time Warping. Information Retrieval for Music and Motion*, Springer, Berlin, Germany, 2007.
[20] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns intime series," in *Proceedings of the KDD Workshop*, pp. 359–370, Seattle, WA, USA, 1994.
[21] S. Pavel, Dynamic Time Warping Algorithm Review, vol. 855, Information and Computer Science Department University of Hawaii at Manoa Honolulu, Honolulu, HI, USA.
[22] J. Eamonn and M. J. Pazzani, "Derivative dynamic time warping," in *Proceedings of the 2001 SIAM International Conference On Data Mining*, pp. 1–11, SIAM, Chicago, IL, USA, April 2001.
[23] C. Villani., *Optimal Transport: Old and New*, Vol. 338, Springer Science & BusinessMedia, Berlin, Germany, 2008.
[24] M. Cuturi, "Sinkhorn distances: lightspeed computation of optimal transport," *Advances in Neural Information Processing Systems*, vol. 26, pp. 2292–2300, 2013.
[25] P. Gabriel and M. Cuturi, "Computational optimal transport: with applicationsto data science," *Foundations and Trends R ○ in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.
[26] L. Ambrosio, "Lecture notes on optimal transport problems," in *Mathematical Aspectsof Evolving Interfaces*, Springer, Berlin, Germany, 2003.
[27] Y. Hao, Y. Zhang, K. Liu et al., "An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 221–231, Vancouver, Canada, January 2017.
[28] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: criss-cross attention for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 603–612, Seoul, Korea, 2019.
[29] K.-H. Lee, Xi Chen, G. Hua, H. Hu, and X. He, "Stacked crossattention for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 201–216, Munich, Germany, September 2018.
[30] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," *Advances in Neural Information Processing Systems*, vol. 2019, pp. 4003–4014, 2019.
[31] G. Liang, H. Mo, Y. Qiao, C. Wang, and J.-Y. Wang, "Paying deep attention to both neighbors and mt," in *Proceedings of the International Conference onIntelligent Computing*, pp. 140–149, Springer, Bari Italy, October 2020.
[32] A. Arthur and D. Newman, *Uci Machine Learning Repository*, University of California, Irvine, School of Information and Computer Sciences, Irvine, CA, USA, 2007.
[33] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: a largescale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference Oncomputer Vision And Pattern Recognition*, pp. 1010–1019, Las Vegas, NV, USA, June 2016.
[34] Y. Kawahara, M. De La Bastide, J. P. Hamilton et al., "Improvement of the oryza sativa nipponbare reference genome using next generation sequence and optical map data," *Rice*, vol. 6, no. 1, p. 4, 2013.

Hindawi
Shock and Vibration
Volume 2021, Article ID 2804539, 9 pages
https://doi.org/10.1155/2021/2804539

Research Article

Application of Boundary Local Feature Scale Adaptive Matching Extension EMD Endpoint Effect Suppression Method in Blasting Seismic Wave Signal Processing

Jing Wu,[1] **Li Wu**,[2] **Miao Sun**,[3] **Ya-ni Lu**,[1] **and Yan-hua Han**[1]

[1]*Faculty of Civil Engineering, Hubei Engineering University, Xiaogan 432000, China*
[2]*Faculty of Engineering, China University of Geosciences, Wuhan 430074, China*
[3]*College of Environment and Engineering, Hubei Land Resources Vocational College, Wuhan 430090, China*

Correspondence should be addressed to Li Wu; lwu@cug.edu.cn

Received 23 June 2021; Accepted 7 August 2021; Published 14 August 2021

Academic Editor: Chaoqun Duan

The intrinsic endpoint effect of empirical mode decomposition (EMD) will lead to serious divergence of the intrinsic mode function (IMF) at the endpoint, which will lead to the distortion of IMF and affect the decomposition accuracy of EMD. In view of this phenomenon, an EMD endpoint effect suppression method based on boundary local feature scale adaptive matching extension was proposed. This method can consider both the change trend of the signal at the endpoint and the change rule of the signal inside. The simulation results showed that the proposed method had better suppression effect on the intrinsic endpoint effect of EMD than the traditional EMD endpoint effect suppression method and achieved high-precision IMF. The endpoint effect suppression method of EMD based on boundary local feature scale adaptive matching extension was used to process the actual blasting seismic signal. The decomposition results showed that the method can effectively suppress the endpoint effect of EMD of blasting seismic signal and are helpful to extract the detailed characteristic parameters of blasting seismic signal.

1. Introduction

The endpoint effect is an unavoidable problem in most signal-processing methods. Empirical mode decomposition (EMD) [1] has been widely used as an adaptive algorithm that decomposes according to the characteristics of the data itself, so it is extremely important to solve the problem of endpoint effect of EMD [2–4].

Some methods have been put forward to solve the problem of endpoint effect of EMD, such as extremum extension method [5] and polynomial fitting extension method [6], which focus on the local change trend of signal endpoint but ignore the global signal feature. There are also a few methods that consider the global signal feature but ignore the change trend of signal endpoint [7].

In view of the above research status, an EMD endpoint effect suppression method based on boundary local feature scale adaptive matching extension is proposed, which can combine the local change trend of the signal endpoint with the feature of the original global signal itself. Compared with the traditional methods, this method has better effect and higher signal decomposition accuracy after it is processed by this method. Through the analysis of the simulation signal and the actual signal, it is proved that this method can not only effectively suppress the endpoint effect but also accurately extract signal feature parameters [8–11].

2. The Principle of Endpoint Effect

The intrinsic mode function (IMF) [12, 13] of the signal obtained by EMD needs to be screened many times. The essence of screening is to calculate the local mean value of the signal according to the upper envelope determined by all maximum points of the signal and the lower envelope determined by all minimum points of signal [14, 15]. However, the endpoint of the signal cannot be at the maximum or

minimum at the same time, and it is not necessarily the extreme point. Therefore, the upper and lower envelope may diverge at the endpoint, which distorts the EMD result [16, 17].

3. The Principle of Boundary Local Feature Scale Adaptive Matching Extension EMD Endpoint Effect Suppression Method

The boundary local feature scale adaptive matching extension EMD endpoint effect suppression method consists of two parts. The first part is boundary local feature scale extension (BLFSE), which considering the variation trend of the signal endpoint amplitude and the internal relationship between the global time of the signal and the interval relationship between the global time of the signal and the time interval the endpoint. The second part is adaptive matching, finding the time series with the highest matching degree in the global signal according to the extended local feature scale.

Taking the "boundary local feature scale" as the reference, a time series with the highest matching degree with "boundary local feature scale" will be found in the global signal, which is the result of the "boundary local feature scale adaptive matching extension EMD endpoint effect suppression method." Figure 1 shows the specific operation flow of the EMD endpoint effect suppression method based on the boundary local feature scale adaptive matching extension.

3.1. The Principle of Boundary Local Feature Scale Extension. The change trend of the signal is not only reflected at the endpoints but also reflected inside the signal [18, 19]. Therefore, there is an inherent connection between the global time of the signal and the interval time of the endpoints. According to this connection, time parameters corresponding to a maximum value point and a minimum value point can be calculated at the left and right endpoint of the signal. By importing the time parameters into the polynomial established by the maximum (minimum) value point of the endpoint, the amplitude parameters corresponding to the maximum (minimum) value point can be obtained.

3.1.1. The Time Parameter of the Boundary Local Feature Scale Extension. Take the left endpoint of the signal as an example, find the occurrence time of all the maximum points of the signal, and record them as $t_{\max 1}$, $t_{\max 2}$, …, $t_{\max i}$ ($i = 1,2,3, \ldots, M$). In the same way, find the occurrence time of all the minimum points of the signal and record them as $t_{\min 1}$, $t_{\min 2}$, …, $t_{\min i}$ ($i = 1,2,3, \ldots, N$). The maximum and minimum points which need to be extended are recorded as $t_{\max 0}$ and $t_{\min 0}$, respectively. The calculation of $t_{\max 0}$ and $t_{\min 0}$ are divided into the following four cases.

Case 1: $t_{\max 1} < t_{\min 1} \cap t_{\max M} < t_{\min N}$, $M = N$; $t_{\min 0}$ and $t_{\max 0}$ are solved with equations (1) and (2), respectively:

$$t_{\min 0} = \frac{\sum_{i=1}^{N-1} \left(t_{\max i+1} - t_{\min i}\right)}{N-1} - t_{\max 1}, \tag{1}$$

$$t_{\max 0} = \frac{\sum_{i=1}^{N} \left(t_{\min i} - t_{\max i}\right)}{N} + t_{\min 0}. \tag{2}$$

Case 2: $t_{\max 1} < t_{\min 1} \cap t_{\max M} > t_{\min N}$, $M = N + 1$; $t_{\max 0}$ is solved with the same equation (2) and $t_{\min 0}$ is solved with

$$t_{\min 0} = \frac{\sum_{i=1}^{N} \left(t_{\max i+1} - t_{\min i}\right)}{N} - t_{\max 1}. \tag{3}$$

Case 3: $t_{\max 1} > t_{\min 1} \cap t_{\max M} > t_{\min N}$, $M = N$; $t_{\min 0}$ and $t_{\max 0}$ are solved with equation (4) and equation (5), respectively:

$$t_{\min 0} = \frac{\sum_{i=1}^{N} \left(t_{\max i} - t_{\min i}\right)}{N} + t_{\max 0}, \tag{4}$$

$$t_{\max 0} = \frac{\sum_{i=1}^{N-1} \left(t_{\min i+1} - t_{\max i}\right)}{N-1} - t_{\min 1}. \tag{5}$$

Case 4: $t_{\max 1} > t_{\min 1} \cap t_{\max M} < t_{\min N}$, $M = N - 1$; $t_{\max 0}$ is solved with the same equation (5), and $t_{\min 0}$ is solved with

$$t_{\min 0} = \frac{\sum_{i=1}^{N-1} \left(t_{\max i} - t_{\min i}\right)}{N-1} + t_{\max 0}. \tag{6}$$

3.1.2. The Amplitude Parameter of the Boundary Local Feature Scale Extension. Take the left endpoint of the signal as an example, find all the amplitudes corresponding to the occurrence time of the maximum value points of the signal, and record them as $x_{\max 1}$, $x_{\max 2}$, …, $x_{\max i}$ ($i = 1,2,3, \ldots, M$). In the same way, find all the amplitudes corresponding to the occurrence time of the minimum value points of the signal and record them as $x_{\min 1}$, $x_{\min 2}$,…, $x_{\min i}$ ($i = 1,2,3, \ldots, N$). According to the amplitude variation trend near the endpoint, a maximum value point and a minimum value point near the left endpoint are extended.

The specific steps of the boundary local feature scale extension are as follows:

Step 1: take maximum points closest to the left endpoint of the signal, i.e., $x_{\max 1}, x_{\max 2}, \ldots, x_{\max a}$. The value of a is related to the sample size.

Step 2: polynomial fit is $x_{\max 1}, x_{\max 2}, \ldots, x_{\max a}$. Then, ($t_{\max 0}$, $x_{\max 0}$) can be obtained by taking $t_{\max 0}$ obtained in Section 3.1.1 into the fitting formula in "Step 1."

Similarly, ($t_{\min 0}$, $x_{\min 0}$) can be obtained.

3.2. The Principle of Adaptive Matching. The trend of the signal is reflected in the endpoint and in the whole signal. Therefore, it is possible to find a curve within the signal that has the highest matching degree with the "boundary local feature scale extension" obtained in Section 3.1.

Specific steps of adaptive matching are as follows:

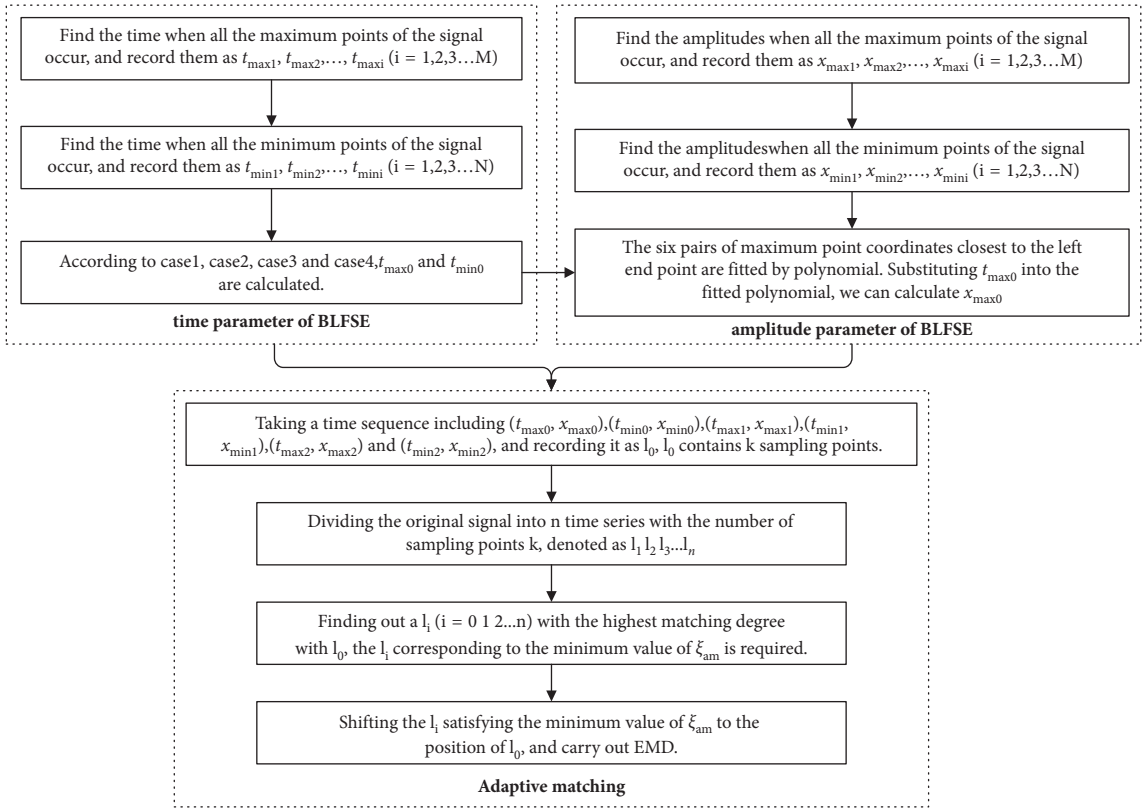


Figure 1: Boundary local feature scale adaptive matching extension EMD endpoint effect suppression method operation flow.

Step 1: take a time series that includes 6 extreme points, which contains the left point continuation results (t_{max0}, x_{max0}) and (t_{min0}, x_{min0}) and the two closest maximum and minimum value points (t_{max1}, x_{max1}), (t_{min1}, x_{min1}), (t_{max2}, x_{max2}), and (t_{min2}, x_{min2}). Record this time series as l_0, which contains k sampling points.

Step 2: divide the original signal into n time series with the number of sampling points k and record them as l_1, l_2, l_3, ...l_i ... l_n ($i = 0, 1, 2, ..., n$).

Step 3: calculate the adaptive matching coefficient (adaptive matching, ξ_{am}) [7] and find out l_i ($i = 0\ 1\ 2...n$) with the highest matching degree with l_0. The adaptive matching coefficient is calculated in equation (7), where $A = \max\{x_{max0}, x_{max1}, x_{max2}\}$ and $B = \max\{x_{min0}, x_{min1}, x_{min2}\}$. Find the minimum value of the adaptive matching coefficient, $\xi_{ammin} = \min\{\xi_{am1}, \xi_{am2}, \xi_{am3}, \ldots, \xi_{amn}\}$; the corresponding l_i is the desired time series, which has the highest matching degree with l_0:

$$\xi_{am} = \frac{\sqrt{\sum_{j=1}^{k} (l_0 - l_i)^2 / k}}{A - B}. \quad (7)$$

Step 4: shifting l_i satisfying the minimum value of ξ_{am} to the position of l_0, then carry out EMD.

In summary, the boundary local feature scale adaptive matching extension EMD endpoint effect suppression method can be realized by completing the above four steps.

4. Comparative Study of Multiple EMD Endpoint Effect Suppression Methods for Simulated Signals

The simulation signal is used to conduct a comparative study of multiple EMD endpoint effect suppression methods, and the correlation coefficient and standard deviation of error between the IMF and the original signal are analyzed to evaluate the EMD endpoint effect suppression.

Simulation signal $S(t)$ is composed of three sinusoidal signals with frequencies of 10 Hz, 30 Hz, and 60 Hz, that is, $S(t) = x_1(t) + x_2(t) + x_3(t)$, where $x_1(t) = \sin(2 \times \pi \times 10 \times t)$, $x_2(t) = \sin(2 \times \pi \times 30 \times t)$, and $x_3(t) = \sin(2 \times \pi \times 60 \times t)$. Sampling point $N = 200$, sampling time t is 0, $\pi/100$, $2\pi/100$, $3\pi/100$..... 2π, that is, on $[0, 2\pi]$, the middle point is taken with an interval of $\pi/100$.

EMD is performed on $S(t)$ directly, and the IMF is obtained as shown in Figure 2(a). It can be found that the two ends of IMF1, IMF2, and IMF3 have different degrees of divergence. With the decomposition, the divergence of IMF3 is the most serious and tends to develop into the data. The decomposition results obtained by boundary local feature scale adaptive matching extension EMD endpoint effect suppression method, extremum extension method, and polynomial fitting method are shown in Figures 2(b)–2(d), respectively. The correlation coefficient (r_{xy}) [14] and standard deviation of error (D_{sde}) [20] between IMF component and corresponding sinusoidal signal are calculated one by one. The equations of correlation coefficient and

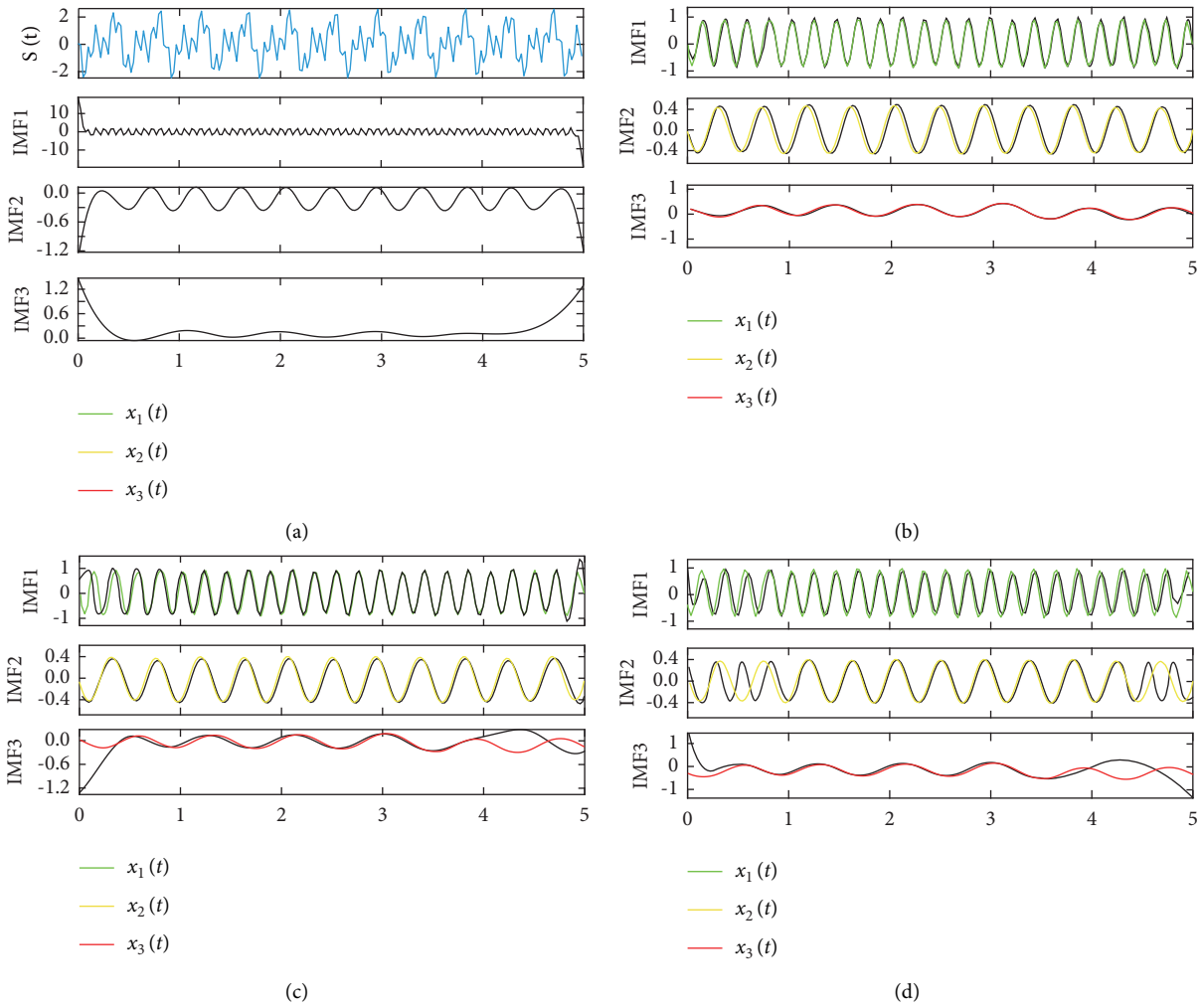


Figure 2: Simulation signal decomposition results.

standard deviation of error are shown in equations (8) and (9), respectively, where N is the number of sampling points, x corresponds to $x_1(t)$, $x_2(t)$, and $x_3(t)$, and y corresponds to IMF1, IMF2, and IMF3. The calculation results are shown in Table 1.

$$r_{xy} = \frac{\sum_{i=1}^{N}\sum_{k=1}^{3}\left(x_k^i(t) - \overline{x(t)}\right)\left(y_k^i(t) - \overline{y(t)}\right)}{\sqrt{\sum_{i=1}^{N}\sum_{k=1}^{3}\left(x_k^i(t) - \overline{x(t)}\right)^2\left(y_k^i(t) - \overline{y(t)}\right)}^2}, \quad (8)$$

$$D_{sde} = \sqrt{\frac{\sum_{i=1}^{N}\left[\sum_{k=1}^{3}\left(x_k^i(t) - y_k^i(t)\right)\right]^2}{N}}. \quad (9)$$

The following conclusions can be drawn from Figure 2 and Table 1:

(1) The three IMF components obtained by the boundary local feature scale adaptive matching extension EMD endpoint effect suppression method reflects the three sinusoidal signals contained in $S(t)$ and has high correlation and small error with the corresponding sinusoidal signals

(2) The IMF obtained by the boundary local feature scale adaptive matching extension EMD endpoint effect suppression method has the highest accuracy

(3) The effect of extremum extension method is slightly better than polynomial fitting method, especially, for the suppression of low frequency components

By further analysis, the IMF obtained in Figures 2(b)–2(d) are transformed by Hilbert transform, and the marginal spectrum-based boundary local feature scale adaptive matching extension EMD endpoint effect suppression method, extremum extension method, and polynomial fitting method are obtained, respectively, as shown in Figures 3(b)–3(d). Figure 3(a) is the marginal spectrum obtained by direct Hilbert transform of $x_1(t)$, $x_2(t)$, and $x_3(t)$. In Figure 3, the energy spectral density (ESD) is the ordinate.

It can be seen from Figures 3(b)–3(d) that the marginal spectral frequencies obtained by the boundary local feature scale adaptive matching extension EMD endpoint effect suppression method are 10.15 Hz, 30.11 Hz, and 60.12 Hz,

Table 1: Evaluation index of the endpoint effect suppression method.

Evaluating indicator		Boundary local feature scale adaptive matching extension method	Extremum extension method	Polynomial fitting method
r_{xy}	IMF1	0.9981	0.8969	0.8084
	IMF2	0.9979	0.9384	0.7603
	IMF3	0.9926	0.6837	0.6051
D_{sde}	IMF1	0.0081	0.2095	0.2438
	IMF2	0.0078	0.1563	0.2763
	IMF3	0.0091	0.3319	0.3919

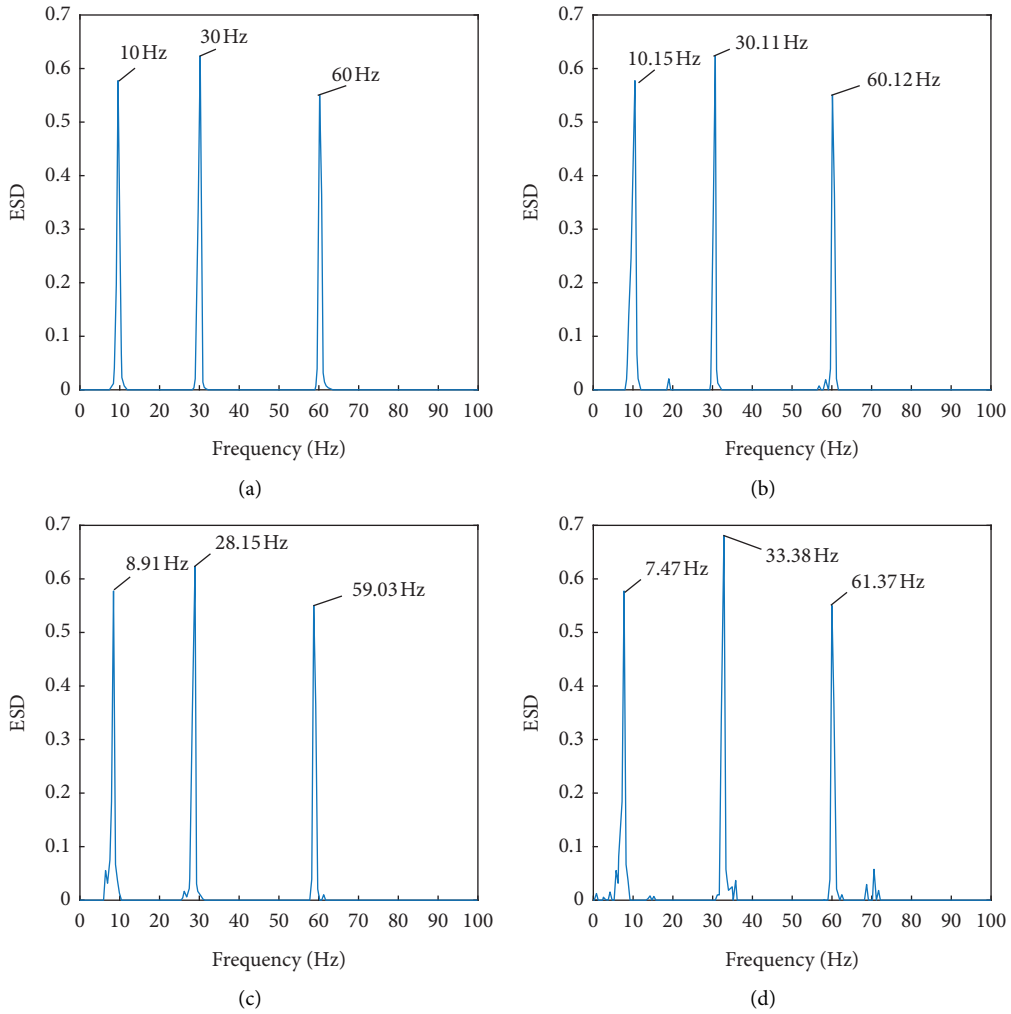


Figure 3: Marginal spectrum.

respectively; the marginal spectral frequencies obtained by the extremum extension method are 8.91 Hz, 28.15 Hz, and 59.03 Hz, respectively; the marginal spectral frequencies obtained by the polynomial fitting method are 7.47 Hz, 33.38 Hz, and 60.12 Hz, respectively. The marginal spectrum shows the frequency of simple harmonic wave contained in IMF by different endpoint effect suppression methods. The error between the marginal spectrum obtained by the boundary local feature scale adaptive matching extension EMD endpoint effect suppression method and the marginal spectrum obtained by $x_1(t)$, $x_2(t)$, and $x_3(t)$ direct Hilbert transform is the smallest, which indicates that the boundary local feature scale adaptive matching extension EMD endpoint effect suppression method can accurately detect the characteristic frequency contained in $S(t)$. The results are consistent with those in Figure 2 and Table 1, which further shows that the boundary local feature scale adaptive matching extension EMD endpoint effect suppression

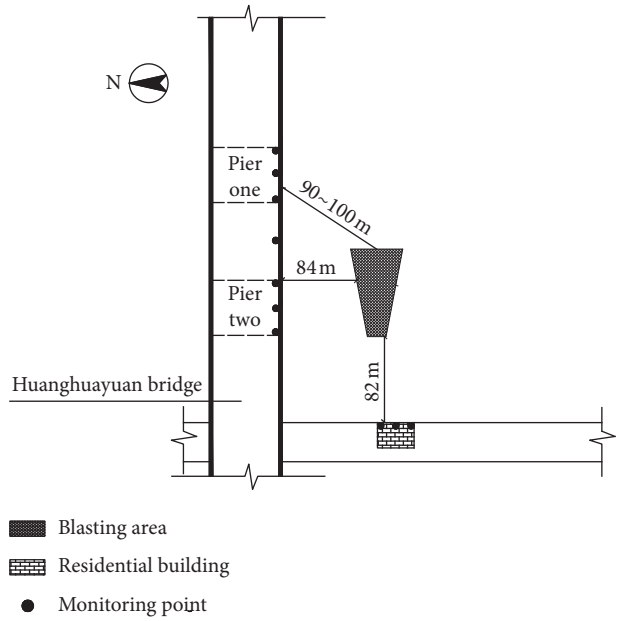


Figure 4: Blasting construction environment diagram.

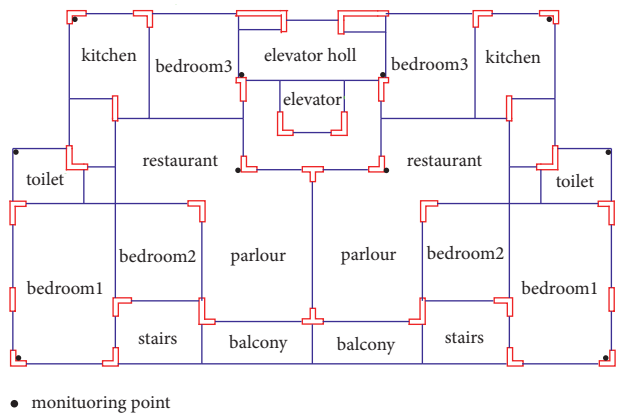


Figure 5: Layout of measuring points of houses.

method can effectively suppress the EMD endpoint effect, realize the extraction of the original signal detail feature parameters, and obtain high-precision IMF.

5. Application of Boundary Local Feature Scale Adaptive Matching Extension EMD Endpoint Effect Suppression Method of Blasting Seismic Wave Signal

The development of engineering blasting has greatly improved work efficiency and brought great convenience to the country's infrastructure construction. However, the impact of its seismic effects and air shock waves on the surrounding environment has become increasingly prominent. The main manifestations are destruction and cracking of existing buildings, slope instability and collapse, and fear of humans and animals. Among them, blasting seismic effect is considered to be the primary hazard of engineering blasting [21].

Especially, in the construction urban blasting engineering, the impact of blasting seismic effects on surrounding buildings is more prominent. Based on the blasting excavation project of the water intake tank of Huanghuayuan Bridge in Chongqing, the length of the water intake tank is 135 m, the upper and bottom width is 69 m, and the lower and lower width is 24 m. The layout plan of the blasting site is shown in Figure 4. It can be seen in Figure 4 that there is a residential building 82 meters away from the water intake tank. The residential building is the key monitoring object of this blasting construction, with 7 floors above the ground. The TC-4850 intelligent blasting vibration instrument is used to monitor the building. The layout of the measuring points is shown in Figure 5. Only the layout of the measuring points of the first floor is shown here, and the other floors are the same as the first floor.

The impact of the blasting seismic effect on surrounding buildings is studied through the EMD method. The boundary local feature scale adaptive matching extension

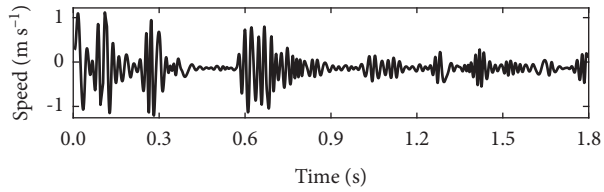


Figure 6: Seismic wave monitoring signal.

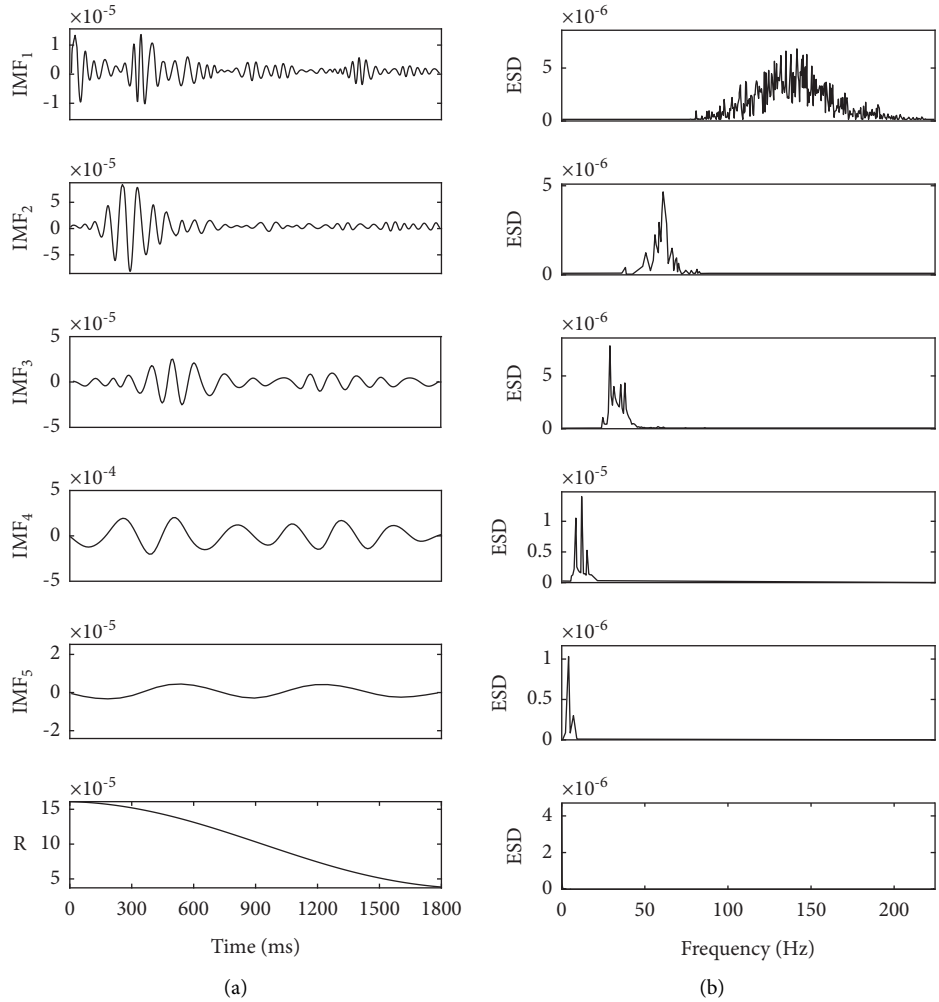


Figure 7: Seismic wave-monitoring signal decomposition diagram.

method is studied to suppress the EMD endpoint effect of the blasting seismic wave signal.

In the monitoring results, a typical seismic wave signal is selected as the research object, as shown in Figure 6. The sampling frequency of the signal is 4000 sps. According to the Nyquist sampling theorem [22], the Nyquist frequency of the measured blasting seismic wave signal is 2000 Hz, including 4096 sampling points.

The signal in Figure 6 is decomposed by the boundary local feature scale adaptive matching extension EMD endpoint effect suppression method. The decomposition results are shown in Figure 7(a). It can be found from Figure 7(a) that there is only slight divergence at the right end of IMF4, and the endpoint effect suppression of other components is well controlled. Further analysis is carried out to calculate the marginal spectrum of each IMF. The calculation results are shown in Figure 7(b). It can be found that each IMF carries a set of specific frequency signals of the blasting seismic wave signal, which once again shows that the boundary local feature scale adaptive matching extension EMD endpoint effect suppression method can realize the accurate extraction of signal feature parameters. The total marginal spectrum of the signal, as shown in Figure 8, is further obtained. Figure 8 shows that the energy of the underwater drilling blasting seismic wave is mainly concentrated in 0~50 Hz, which is consistent with the

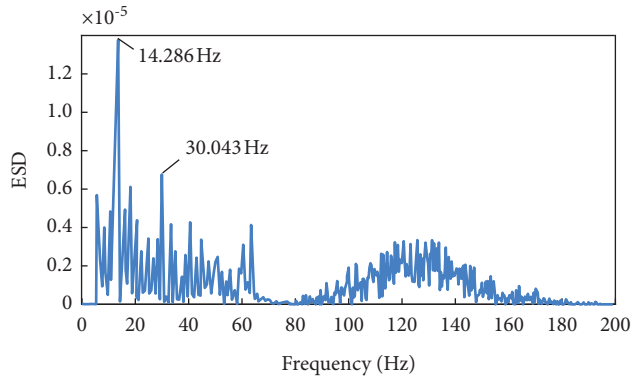


Figure 8: The total marginal spectrum of the signal.

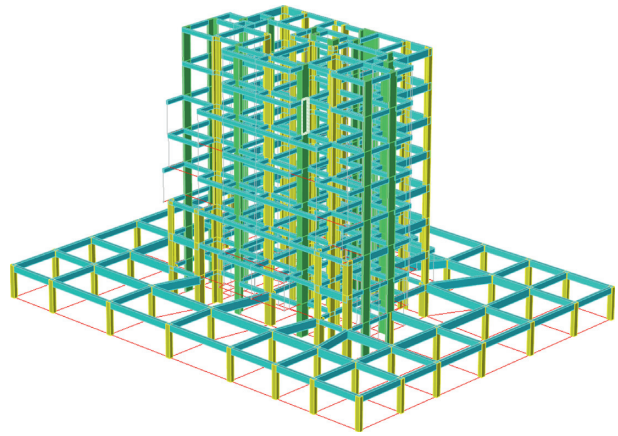


Figure 9: Three-dimensional model diagram of the residential building.

Table 2: The natural frequency corresponding to the first eight-order mode shape (unit: Hz).

Array	Natural frequency
1	8.234
2	13.806
3	18.806
4	25.023
5	29.920
6	32.156
7	42.714
8	44.976

conclusion drawn by [23]. The dominant frequency of the signal is the frequency corresponding to the maximum energy density [24], so the dominant frequency of this signal is 14.286 Hz. When the frequency of blasting seismic wave is the same as the natural frequency of the residential building, the amplitude of the structure will reach the maximum, thus inducing resonance harm.

Through the finite element analysis software of YJK, the three-dimensional model of the residential building is obtained, as shown in Figure 9, and the natural vibration frequency of the house is calculated. The natural vibration frequency of the first eight-order formation of the residential building is shown in Table 2. It can be found from the analysis of Table 2 that the second-order formation of the residential building is 13.806 Hz, and the dominant frequency of this blasting is 14.286 Hz, which are very close to each other. Therefore, the seismic wave generated by the blasting is very likely to cause the resonance of the residential building.

Therefore, the corresponding control measures must be taken in the actual construction to ensure the safety of the residential building. The conclusion also shows that the method proposed in this paper is not only helpful to suppress the EMD endpoint effect and obtain higher precision IMF but also can accurately extract the frequency parameters contained in the blasting seismic wave, which is helpful to control blasting vibration and provide basis for formulating scientific antiseismic measures.

6. Conclusions

(1) The boundary local feature scale adaptive matching extension method not only considers the local change trend of the signal at the endpoint but also retains the unique internal attributes of the signal through the global search ability of "adaptive matching," so as to retain the authenticity of the original signal to the greatest extent

(2) By comparing the decomposition results of simulation signals, it is found that the boundary local feature scale adaptive matching extension EMD endpoint effect suppression method can effectively suppress the EMD endpoint effect and obtain higher accuracy IMF

(3) The frequency energy information contained in the blasting seismic wave can be effectively extracted from the marginal spectrum of the IMF obtained by the boundary local feature scale adaptive matching extension EMD endpoint effect suppression method, which is helpful to identify the characteristic parameters of blasting seismic wave signal and control blasting vibration

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors would like to thank Ma Chengyang for his advice on the composition of the article. This work was supported by the National Natural Science Foundation of China (41672260 and 41907259) and the Scientific Research Program of Hubei Provincial Education Department (Q20202701).

References

[1] N. E. Huang, Z. Shen, S. R. Long, and M. L. C Wu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences*, vol. 454, no. 3, pp. 903–995, 1998.

[2] H. U. Aijun, "New process method for end effects of HILBERT-HUANG transform," *Chinese Journal of Mechanical Engineering*, vol. 44, no. 4, pp. 154–158, 2008.

[3] C. H. Bai, X. C. Zhou, D. C. Lin et al., "PSO-SVM method based on elimination of end effects in EMD," *Xitong Gongcheng Lilun yu Shijian/System Engineering Theory and Practice*, vol. 33, no. 5, pp. 1298–1306, 2013.

[4] Z. Liang, S. P. Peng, and J. Zheng, "LEMD endpoint extension method and its application in microseismic signals' denoising," *Zhendong yu Chongji/Journal of Vibration and Shock*, vol. 33, no. 21, pp. 155–160, 2014.

[5] L. Shen, X. J. Zhou, Z. G. Zhang, and W. G. Zhang, "Boundary-extension method in hilbert-huang transform," *Journal of Vibration and Shock*, vol. 28, no. 8, pp. 168–171, 2009.

[6] H. Liu, "Dealing with the end issue of EMD based on polynomial fitting algorithm," *Computer Engineering and Applications*, vol. 40, no. 16, pp. 84–86, 2004.

[7] J. Yang, G. Shi, T. Zhou, and F. Gao, "Waveform extension method based on similarity sequential detection for the end effects reduction of EMD," *Journal of Vibration and Shock*, vol. 37, no. 18, pp. 121–125, 2018.

[8] X. Cai, H. Zhao, S. Shang et al., "An improved quantum-inspired cooperative co-evolution algorithm with muli-strategy and its application," *Expert Systems with Applications*, vol. 2021, Article ID 114629, 2021.

[9] H. Zhao, H. Liu, Y. Jin, X. Dang, and W. Deng, "Feature extraction for data-driven remaining useful life prediction of rolling bearings," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, no. 99, pp. 1–10, 2021.

[10] W. Deng and J. Xu, "An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 99, pp. 1–10, 2020.

[11] W. Deng, S. Shang, X. Cai, H. Zhao, Y. Song, and J. Xu, "An improved differential evolution algorithm and its application in optimization problem," *Soft Computing*, vol. 25, no. 7, pp. 5277–5298, 2021.

[12] Xu Jia, S. Huang, and F. Ma, "The dynamic characteristics analysis for the large bridge based on the improved hilbert-huang transformation," *Geomatics and Information Science of Wuhan University*, vol. 35, no. 7, pp. 801–805, 2010.

[13] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2011.

[14] Z. He, Z. Zhu, and H. Xie, "Restrain boundary effect of EMD based on least square fitting," *Journal of System Simulation*, vol. 30, no. 9, pp. 3377–3385+3398, 2018.

[15] R. Hao and F. Li, "A new method to suppress the EMD endpoint effect," *Zhendong Ceshi Yu Zhenduan/Journal of Vibration, Measurement and Diagnosis*, vol. 38, no. 2, pp. 341–345, 2018.

[16] Z. N. Han, J. X. Gao, Y. X. Li, and W. T. Sun, "Dealing with end effect of EMD based on gray theory," *Advanced Materials Research*, vol. 228, pp. 1094–1099, 2011.

[17] D. Cao and X. Zhang, "Research on the comparison with methods of restraining ending effect of EMD and its application in fault diagnosis," *Journal of Mechanical Transmission*, vol. 37, no. 3, pp. 83–87, 2013.

[18] J. Qiu and J. J. Chen, "EMD in the research and application of deformation monitoring in embankment," *Applied Mechanics and Materials*, vol. 501, pp. 1868–1872, 2014.

[19] L. Wu, Y. Zhang, Y. Zhao, G. Ren, and S. He, "Mode mixing suppression algorithm for empirical mode decomposition based on self-filtering method," *Radioelectronics and Communications Systems*, vol. 62, no. 9, pp. 462–473, 2019.

[20] X. U. Ke, Z. H. Chen, and C. B. Zhang, "Rolling bearing fault diagnosis based on empirical mode decomposition and support vector machine," *Control Theory & Applications*, vol. 36, no. 6, pp. 915–922, 2019.

[21] S. Y. Qian, *Application of HHT in Blasting Vibration Signal Processing*, Central South University, Changsha, China, 2012.

[22] P. P. Vaidyanathan, "Generalizations of the sampling theorem: seven decades after Nyquist," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 9, pp. 1094–1109, 2002.

[23] M. Sun, L. Wu, Q. Yuan, and C. Ma, "Time frequency analysis of blasting seismic wave signal based on CEEMDAN," *Journal of South China University of Technology*, vol. 48, no. 3, pp. 76–81, 2020.

[24] X. Z. Shi, *Time-Frequency Analysis of Blasting Vibration Signal and Research on Characteristic Parameters of Blasting Vibration and Hazard Prediction*, Central South University, Changsha, China, 2007.

Hindawi
Shock and Vibration
Volume 2021, Article ID 8718523, 6 pages
https://doi.org/10.1155/2021/8718523

Research Article

Research on Improved Ray Casting Algorithm and Its Application in Three-Dimensional Reconstruction

ZheShu Jia,[1] **DeYun Chen**,[1] **and Bo Wang**[2]

[1]*School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China*
[2]*School of Automation, Harbin University of Science and Technology, Harbin 150080, China*

Correspondence should be addressed to ZheShu Jia; jzs_19@stu.hrbust.edu.cn

Received 25 June 2021; Revised 18 July 2021; Accepted 24 July 2021; Published 5 August 2021

Academic Editor: Chaoqun Duan

Spinal pathology treatment has become an urgent issue to be solved. How to effectively prevent and treat spinal pathology has become a research hotspot in the field of surgery. Aiming at the problem of too long volume rendering time caused by the trilinear interpolation sampling method in the reconstruction and visualization of the vertebra 3D model, an improved ray projection algorithm is proposed to quickly reconstruct a 3D vertebra model from medical CT vertebra images. This method first classifies CT data, assigns corresponding color values and opacity transfer functions to different types of data, and then uses inverse distance-weighted interpolation (IDWI) sampling to replace the trilinear interpolation sampling method for the voxel where the sampling point is located to accelerate the interpolation operation. The color value and opacity of the sampling points are obtained, and finally, the attributes of all the sampling points are synthesized and calculated to obtain the final rendering effect, and the reconstruction of the three-dimensional vertebra model is completed. Experimental results show that the proposed method not only can obtain higher quality rendered images but also has a certain improvement in rendering speed compared with traditional algorithms.

1. Introduction

Nowadays, spine-related diseases are a major problem in modern human society. Modern clinical medicine points out that spinal pathology has become an urgent problem to be solved. Pedicle screw placement is a relatively mature medical method for spinal pathology, and three-dimensional reconstruction of vertebrae is the premise of preoperative simulation and quantification. The three-dimensional vertebral model obtained by medical image visualization can provide more realistic anatomical structure for doctors, so as to help doctors accurately diagnose the disease and design treatment plan, which has important clinical practical value.

Medical image visualization [1] is the human body information obtained by digital imaging technology that is intuitively expressed as three-dimensional effect on the computer, so as to provide structural information that cannot be obtained by traditional means. Volume rendering has become an important method to realize the visualization of medical images because it can show the fine structure and small transformation of objects. As a classical method of volume rendering technology, ray casting algorithm has a wide range of applications in the field of 3D reconstruction visualization of medical images, but its rendering speed is slightly insufficient [2]. At present, the common methods to improve the rendering speed can be roughly divided into three types: hardware-based, software-based, and parallel mode [3]. Among them, the premise of hardware-based and parallel methods is to carry out on specific computer hardware, so its portability is greatly reduced. The acceleration method based on software type does not rely too much on the development of hardware, but improves from the angle of algorithm, which has high flexibility. However, this kind of method has some problems, such as too high

complexity to achieve ideal acceleration effect and not taking into account both rendering speed and reconstruction quality.

In order to speed up volume rendering and improve the quality of reconstruction, we proposed a ray casting algorithm based on inverse distance-weighted interpolation sampling to realize the 3D visualization of the vertebral CT image. Firstly, vertebrae and nonvertebrae are distinguished in the 3D volume data field, and the mapping from data to optical features is realized by transfer function. Then, the inverse distance-weighted interpolation method is used to replace the trilinear interpolation in the traditional ray casting algorithm to speed up the resampling process. Finally, the final color of each pixel is obtained by synthesizing the color value and opacity of all the sampling points, and finally, the three-dimensional reconstruction of the vertebral image is realized.

2. Related Work

With the development of computer hardware technology, hardware-based and parallel acceleration methods are first proposed. Cullip and Neumann first proposed a method to accelerate volume rendering using 3D texture hardware [4]. Zhang et al. used GPU-based fast ray casting method for volume rendering in CT 3D reconstruction to shorten the reconstruction time [5]. Ross et al. proposed a CPU-based volume rendering algorithm for 3D ultrasound images, which overcomes the difficulty of low adaptability of GPU-based algorithm [6]. Ma et al. proposed a parallel grid generation algorithm on GPU, which increased the average efficiency by 15 times [7]. Zhou et al. used CUDA to accelerate the implementation of ray casting algorithm, which was 70% faster than GPU [8]. Sans et al. compared the performance of OpenGL, OpenCL, and CUDA for different medical datasets [9].

However, both hardware-based and parallel-based acceleration methods are based on the premise of computer hardware, so they have some limitations. In contrast, the acceleration method based on software is more flexible and convenient, which can be transplanted between different machines quickly and has wider applicability. Mehaboobathunnisa et al. proposed a method of grouping rays projected by similar voxels to reduce the computational complexity of rendering algorithm, but the reconstruction result is not smooth enough due to artifacts [10]. Hadwiger et al. proposed the SparseLeap method which is a novel space hopping method and has been proved that it can avoid the problem of unnecessary space debris [11]. Based on the idea of spatial jump method, Deakin and Knackstedt used Chebyshev distance to guide how to effectively skip the blank area in the ray casting algorithm [12], and then they optimized this method and proposed an effective algorithm to generate anisotropic Chebyshev distance map for accelerating ray casting, but these two methods have some shortcomings in the final reconstruction effect [13]. Liu et al. reduced the amount of computation in the rendering process by adjusting the sampling frequency and adopting different interpolation strategies for the sampling points of different classification groups [14]. Bi et al. applied inverse distance-weighted interpolation algorithm to meteorological data visualization and achieved good rendering effect [15]. In order to improve the speed of ray casting algorithm, this paper proposes a ray casting algorithm based on inverse distance-weighted interpolation sampling, which improves the speed of volume rendering on the premise of meeting the quality requirements of 3D reconstruction.

3. Improved Ray Casting Algorithm

3.1. Traditional Ray Casting Algorithm. The basic idea of the traditional ray casting algorithm can be described as follows: firstly, collecting the sample points of all the voxels in the three-dimensional data field along the ray direction at equal intervals and then obtaining the corresponding color values and opacity values of the sample points, then synthesizing the color values and opacity values of all the sample points on the ray to obtain the final color of the pixel, and finally, calculating each pixel and obtaining the final two-dimensional image. The detailed steps are shown in Table 1.

In traditional ray casting algorithm, step 1 is to initialize the color value C and opacity value α. In step 2, all the voxels in the 3D volume data field are traversed, and the sample points are sampled by trilinear interpolation to obtain the color value and opacity value of the sample points. Finally, the color values and opacity values of all sampling points on the ray are combined in step 3 to get the final projection image.

3.2. Inverse Distance-Weighted Interpolation Method. Inverse distance-weighted interpolation (IDWI) is a kind of interpolation method which takes the distance between the sampling point and adjacent point as weight. Figure 1 shows the schematic diagram of the IDWI method.

The IDWI method considers that each adjacent point will have a certain influence on the sampling point, and the influence is closely related to the distance. The closer the sampling point is, the greater the weight is given to the adjacent point, and the weight contribution is inversely proportional to the distance [16].

Suppose that there are other neighboring points in the neighborhood of sampling point $A(x, y, z)$, denoted as $A_i(x_i, y_i, z_i)$, $i = 1, 2, \ldots, n$. Let $f(A)$ be the color and opacity value of A, then it can be described as follows:

$$f(A) = \sum_{i=1}^{n} \lambda_i f(A_i), \tag{1}$$

where λ_i is the weight of the distance from each adjacent point to the sampling point and can be calculated as follows:

$$\lambda_i = \frac{1/d}{\sum_{i=1}^{n} 1/d}, \tag{2}$$

where d_i is the Euclidean distance between adjacent points and sampling points and can be calculated as follows:

Table 1: Traditional ray casting algorithm (TRC).

Input:	3D volume data field D
Step 1:	Initial color value C and opacity per pixel α
Step 2:	**For** each voxel in volume data field D **Do**
2.1:	**If** current voxel contains sampling points, **then**
2.2:	The value C and α of each sampling point are obtained by trilinear interpolation
2.3:	**End if**
2.4:	**End for**
Step 3:	Compositely calculating the value of and C for each pixel
Output:	Opacity per pixel α, color value C

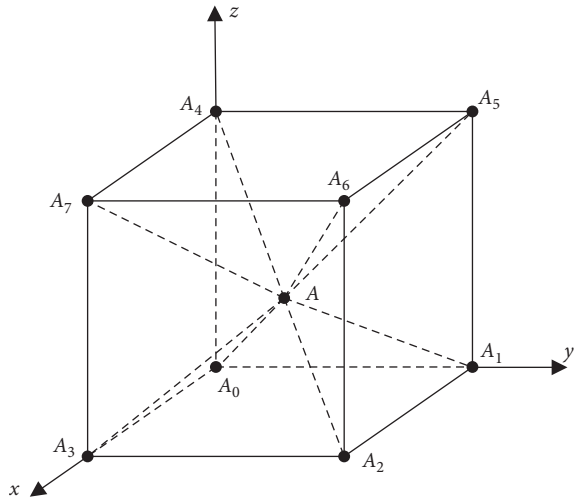


Figure 1: Inverse distance-weighted interpolation.

$$d_i = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2}. \quad (3)$$

Obviously, the sum of all weights is 1, that is,

$$\sum_{i=1}^{n} \lambda_i = 1. \quad (4)$$

3.3. Ray Casting Algorithm Based on IDWI Sampling. In order to improve the efficiency of the TRC, an improved ray casting algorithm is proposed in this paper. The basic idea is to use inverse distance-weighted interpolation instead of trilinear interpolation in interpolation sampling. The improved ray casting algorithm includes the following three steps:

(1) Classification of vertebrae and nonvertebrae: in this step, a three-dimensional volume data field is constructed, and a ray is emitted from the position of each pixel on the screen along the line of sight direction into the data field. According to the different characteristics of the data, the vertebral CT data are classified into different types, that is, vertebral and nonvertebral. In order to show the internal structure of 3D data field more intuitively, it is necessary to assign different color values and opacity to different types of data according to the classification results. The setting of color value and opacity needs to design different transfer functions. The function of transfer function is to complete the mapping from data to optical features (color, opacity, etc.). Formally, the transfer function T can be defined as follows:

$$T: \longrightarrow \{c, \alpha\}, x \in Rn, \quad (5)$$

where $\{c, \alpha\}$ is a binary set of color values and opacity, x is the attribute values of volume data, and n is the dimension of x and represents the number of attributes.

(2) Achieving the color and opacity values: after separating the region of vertebrae and nonvertebrae, the sample points of all the voxels in the 3D data field which are penetrated by light are collected at equal intervals along the ray direction, and the corresponding color and opacity values of the sample points are calculated according to the attribute values of the eight vertices of the voxels where the sample points are located. The sampling points are above the ray incident from the pixel to the 3D data field, usually in the voxel mesh they pass through, rather than just at the vertex of the voxel mesh. The inverse distance-weighted interpolation method is used to obtain the color value and opacity of the sampling points.

(3) Synthesizing the color values: in this step, the final color of the pixels is obtained by compositing the color values and opacity values of all the sampling points on the ray. In the process of synthesis, we use the front-to-back strategy. The formulas can be described as follows:

$$C_{\text{out}} = C_{\text{in}} + C_i\left(1 - C_{\text{in}}\right), \tag{6}$$

$$\alpha_{\text{out}} = \alpha_{\text{in}} + \alpha_i\left(1 - \alpha_{\text{in}}\right), \tag{7}$$

where C_{out} and α_{out} are the color values and opacity values after the cast ray passes through the sample point, respectively. C_{in} and α_{in} are the color values and opacity values before the incident sampling point i, respectively. C_i and α_i are the color values and opacity values of the current sampling point i, respectively. After the above iteration, each pixel is calculated to get the final three-dimensional image. The detailed steps are shown in Table 2.

3.4. Stability Analysis. In the process of image rendering using the classical ray casting algorithm, there are two kinds of situations in which the projected ray travels in the 3D volume data field: the ray travels in the empty voxel and the ray travels in the nonempty voxel. For analyzing the whole rendering process from the rendering data source, the calculation of empty voxels has no contribution to the final 3D reconstruction effect presented on the screen. However, if the same undifferentiated interpolation is applied to the empty voxels, the rendering time of the ray casting algorithm will be increased, and the real-time performance of the algorithm will be affected. Therefore, we use the spatial jump technique to skip the empty voxels in the direction of the projection ray and only interpolate the nonempty voxels in the 3D volume data field. For each empty voxel, first record the nearest distance from it to the opaque voxel, and then, no matter what the direction of the ray is, as long as the distance is a forward step, it will not intersect with the opaque voxel in this distance range when the ray is casting. Based on this, it can significantly reduce the redundant steps of ray casting, improve the efficiency of the algorithm, and has obvious advantages in computing resources and memory requirements.

4. Experiments and Analysis

4.1. Experimental Setup. The experimental data were obtained from the project cooperation Hospital of Xi'an Zhenwo 3D Technology Co., Ltd., with a total of 77 samples and a resolution of 512×512 DICOM format CT image data file, and each CT image interval is 2.5 mm. Hardware environment is Intel (R) core (TM) i5-4590 @ 3.30 GHz, 8 GB, AMD Radon (TM) r5340x, OS is windows10, and all programs are implemented in Python + VTK development environment. In this paper, five vertebrae of lumbar vertebrae are taken as the reconstruction object, and the DICOM format vertebral CT data is used for experiment. The time required by several ray casting algorithms for volume rendering is compared, and the PSNR and SSMI values are used as evaluation indexes to compare the reconstruction quality.

4.2. Reconstruction Results and Analysis. In the same experimental environment, two groups of experiments were carried out: the first group took all slices of lumbar CT data as input, compared with the traditional ray casting (TRC), the ray casting based on bounding box optimization (BRC) [17], the ray casting with viewpoint (VRC) [18], and our method (improved ray casting, IRC), the 3D reconstruction results and execute time are given; in the second group, part of CT slices in the data set was taken to reconstruct a single vertebra, and the time-consuming and reconstruction results of the four volume rendering methods were compared.

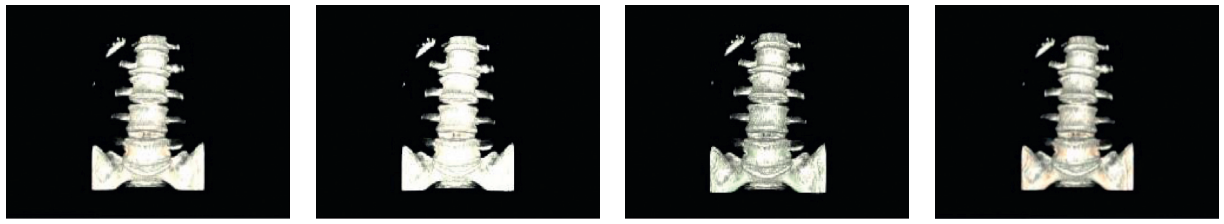
Figures 2 and 3 show the comparison of reconstruction effect of lumbar and single vertebral, respectively. In addition, Tables 3 and 4 show the comparison of reconstruction quality indexes to TRC of lumbar and single vertebra, respectively.

From Figures 2 and 3, we can see that our method IRC algorithm has no obvious difference in reconstruction quality compared with the other three algorithms. However, from the reconstruction quality indicators in Tables 3 and 4, we can see that the PSNR value of IRC is little different from the other algorithms, and the SSIM value is very close. This shows that our improved method not only ensures the quality of 3D reconstruction but also shows the better reconstruction model. Table 5 shows the running time of the four algorithms in two groups of experiments.

From Table 5, we can see that the running time of TRC, BRC, and VRC algorithm for single vertebra is 10.739 s, 8.254 s, and 7.851 s, respectively, which is slower than 6.473 s of our improved ray casting algorithm. Meanwhile, the running time of IRC for whole lumbar spine is 14.346 s, which is 9.979 s, 4.516 s, and 2.162 s faster than TRC (24.325 s), BRC (18.862 s), and VRC (16.508 s), respectively. Therefore, IRC is better than the other three algorithms in running speed for CT data files of the same experimental object. For each interpolation point, the trilinear interpolation method needs 21 times of multiplication and division and 28 times of addition and subtraction, while the inverse distance-weighted interpolation method only has 16 times of multiplication and division and 7 times of addition and subtraction. Because of the reduction of the amount of computation, the IRC shortens the running time.

Table 2: Ray casting algorithm based on IDWI sampling (RC-IDWIS).

Input:	3D volume data field D
Step 1:	Classification of vertebrae and nonvertebrae
1.1:	Along the line of sight, emitting a ray from the position of each pixel on the screen to enter the 3D volume data field D
1.2:	According to the different characteristics, the images are classified into vertebral and nonvertebral regions
1.3:	Set the color value and opacity through the transfer function by using formula (5)
Step 2:	Achieving the color and opacity values
2.1:	**For** each voxel in volume data field D **Do**
2.2:	**If** current voxel contains sampling points, **then**
2.3:	The value C and α of each sampling point are obtained by inverse distance-weighted interpolation sampling
2.4:	**End if**
2.5:	**End for**
Step 3:	Synthesizing the color value for each pixel
3.1:	**For** each pixel **Do**
3.2:	Calculate the value of C and α by using formulas (6) and (7)
3.3:	**End for**
Output:	Composite 3D image



(a) (b) (c) (d)

Figure 2: Comparison of the reconstruction effect of lumbar. (a) TRC. (b) BRC. (c) VRC. (d) IRC.

(a) (b) (c) (d)

Figure 3: Comparison of the reconstruction effect of single vertebra. (a) TRC. (b) BRC. (c) VRC. (d) IRC.

Table 3: Comparison of quality indicators to TRC of lumbar reconstruction.

Algorithm contrast	PSNR	SSIM
BRC and TRC	25.6313	0.9318
VRC and TRC	26.9164	0.9262
IRC and TRC	27.8933	0.9705

Table 4: Comparison of quality indicators of lumbar reconstruction.

Algorithm contrast	PSNR	SSIM
BRC and TRC	33.0852	0.9862
VRC and TRC	31.5443	0.9860
IRC and TRC	32.5044	0.9862

Table 5: The running time of the four algorithms in the two sets of experiments.

Experiment subject	TRC (s)	BRC (s)	VRC (s)	IRC (s)
Lumbar	24.325	18.862	16.508	14.346
Single vertebra	10.739	8.254	7.851	6.473

5. Conclusion

This paper proposed an improved ray casting algorithm to solve the problem of too long rendering time of traditional ray casting algorithm in the process of 3D reconstruction of vertebral CT image. According to the shortcomings of trilinear interpolation method in sampling speed, inverse distance-weighted interpolation sampling is used instead of

trilinear interpolation sampling. Compared with the other existing methods, our method can reduce the reconstruction time and improve the rendering speed without reducing the reconstruction quality. How to improve the existing methods so as to obtain more fine reconstruction effect of microstructure and further improve the reconstruction accuracy is the next research direction.

Data Availability

The datasets used in this paper are publicly available.

Conflicts of Interest

The datasets used in this paper are available from the corresponding author.

Acknowledgments

This paper was funded by the National Key Research and Development Program of China (no. 2017YFB1401800), the Philosophy and Social Sciences Research Planning Project of Heilongjiang Province (nos. 20GLB119 and 19GLB327), and the Talents Plan of Harbin University of Science and Technology: Outstanding Youth Project (no. 2019-KYYWF-0216).

References

[1] H. B. Almgoter and Z. T. Aldahan, "Development of a display system for visualization in medical applications," *Journal of Physics: Conference Series.IOP Publishing*, vol. 1660, Article ID 012099, 2020.

[2] M. Levoy, "Display of surfaces from volume data," *IEEE Computer Graphics and Applications*, vol. 8, no. 3, pp. 29–37, 1988.

[3] L. Lin and S. Chen, Y. Shao and Z. Gu, Plane-based sampling for ray casting algorithm in sequential medical images," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 874517, 5 pages, 2013.

[4] T. J. Cullip and U. Neumann, *Accelerating Volume Reconstruction With 3D Texture Hardware*, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 1993.

[5] Z. Zhang, S. Ghadais, O. R. Bingol, A. Krishnamurthy, and L. J. Bond, "A framework for 3D x-ray CT iterative reconstruction using GPU-accelerated ray casting," *AIP Conference Proceedings*, vol. 2102, no. 1, 2019.

[6] T. D. Ross, J. J. Bradley, L. J. Hudson, and M. P. O'Connor, "CPU-based volume rendering method for 3D ultrasound image," *Life Science Instruments*, vol. 17, no. 1, pp. 32–36, 2019.

[7] T. C. Ma, P. Li, and T. B. Ma, "A three-dimensional cartesian mesh generation algorithm based on the GPU parallel ray casting method," *Applied Sciences*, vol. 10, no. 1, p. 58, 2020.

[8] D. Zhou, H. Duand, F. Zhao, H. Kan, G. Li, and B. Qiu, "Improving efficiency for CUDA-based volume rendering by combining segmentation and modified sampling Strategies," *International journal of simulation systems, Science and Technology*, vol. 17, no. 42, pp. 1–9, 2016.

[9] F. Sans and R. Carmona, "Volume ray casting using different GPU based parallel APIs," in *Proceedings of the IEEE 2016 XLII Latin American Computing Conference (CLEI)*, Valparaiso, Chile, October 2016.

[10] R. Mehaboobathunnisa, A. A. H. Thasneem, and M. M. Sathik, "Fuzzy mutual information-based i grouped ray casting," *Journal of Intelligent Systems*, vol. 28, no. 1, pp. 77–86, 2019.

[11] M. Hadwiger, A. K. Al-Awami, J. Beyer, M. Agus, and H. Pfister, "SparseLeap: Efficient empty space skipping for large-scale volume rendering," *IEEE Transactions Visualization Computer Graphics*, vol. 24, no. 1, pp. 974–983, 2017.

[12] L. Deakin and M. Knackstedt, ""Accelerated volume rendering with Chebyshev distance maps," in *Proceedings of the SIGGRAPH Asia 2019 technical briefs*, pp. 25–28, Brisbane, Australia, November 2019.

[13] L. J. Deakin and M. A. Knackstedt, "Efficient ray casting of volumetric images using distance maps for empty space skipping," *Computational Visual Media*, vol. 6, no. 1, pp. 53–63, 2020.

[14] Y. Liu, H. J. Lu, and D. F. Chang, "Research on indoor smoke visualization based on improved ray-castingalgorithm," *Laser & Optoelectronics Progress*, vol. 58, no. 4, Article ID 0410005, 2021.

[15] S. B. Bi, Y. C. Gong, M. Y. Lu, H. Zhou, and M. Yuanziang, "Modeling and visualization on scalar fields of meteorological data," *Journal of System Simulation*, vol. 32, no. 7, p. 1331, 2020.

[16] C. Van Mierlo, M. G. R. Faes, and D. Moens, "Inhomogeneous interval fields based on scaled inverse distance weighting interpolation," *Computer Methods in Applied Mechanics and Engineering*, vol. 373, Article ID 113542, 2020.

[17] S. H. He, X. P. Wang, S. Wu, and S. Wen, "An improved ray casting method," *Chinese Journal of Stereology and Image Analysis*, vol. 18, no. 2, pp. 130–134, 2013.

[18] Y. H. Xie and Y. Ji, "Ray casting algorithm based on viewpoint correlation for 3D cloud visualization," *Semiconductor Optoelectronics*, vol. 40, no. 5, pp. 694–699, 2019.

Hindawi
Shock and Vibration
Volume 2021, Article ID 3920579, 11 pages
https://doi.org/10.1155/2021/3920579

Research Article

The New Method of Sensor Data Privacy Protection for IoT

Yue Wu,[1,2] **Liangtu Song**,[1,2] **and Lei Liu**[1,2]

[1]*Institute of Intelligent Machines, and Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, Anhui, China*
[2]*University of Science and Technology of China, Hefei 230026, Anhui, China*

Correspondence should be addressed to Yue Wu; wyw533k@mail.ustc.edu.cn

Received 24 May 2021; Revised 15 June 2021; Accepted 21 June 2021; Published 22 July 2021

Academic Editor: Chaoqun Duan

This article introduces the new method of sensor data privacy protection method for IoT. Asymmetric encryption is used to verify the identity of the gateway by the sensor. The IoT gateway node verifies the integrity and source of the data, then creates a block, and submits the block chain transaction. In order to avoid tracking the source of the data, a ring signature is used to anonymize the gateway transaction. The proxy re-encryption method realizes the sharing of encrypted data. On the basis of smart contracts, attribute-based data access control allows decentralized applications to finely control data access. Through experiments, the effects of sensor/gateway verification, transaction signatures, and sensor data encryption on performance are discussed. The results show that transaction delays are all controlled within a reasonable range. The system performance achieved by this method is also relatively stable.

1. Introduction

Various sensors have penetrated into all aspects of our lives, and many of them are continuously collecting our information [1]. In this context, privacy issues related to the Internet of Things, especially consumer Internet of Things, have become the focus of attention from all walks of life [2]. As a kind of microcomputer terminal, the Internet of Things device exists independently of the computer network but is closely connected with the Internet [3]. Today, IoT sensors have penetrated into various industries, from industrial automation to medical equipment and then to the financial industry [4]. In short, there will be sensors wherever humans live [5].

Privacy protection is an important issue in IoT systems but has different references in different scenarios. For example, in the smart grid, the leakage of household energy consumption data and other information may cause hidden dangers to the safety of household personnel and property [6–8]. In smart medical care, the leakage of health data generated by patients' wearable devices causes personal safety and ethical issues. Thus, following the above two cases, the privacy protection issue is considered related to the interests of the owner of the sensor data. This kind of privacy protection is called active privacy protection wherein the owners of sensor data have the opportunity to take necessary measures to protect their privacy [9, 10]. However, in public video surveillance applications, the privacy problem caused by face recognition is another situation. In this scenario, the issue of privacy protection is not an issue related to the interests of the owner of the sensor data (that is, the operator of the monitoring system) [11]. Whether to protect the privacy of the monitored person depends on the law and the level of the operator. Similarly, electronic license plate applications exist [12]. This kind of privacy protection is called passive privacy protection wherein the object whose privacy is compromised is powerless to solve this problem. For these two privacy protection problems, the problems to be solved and the directions to be considered are different. Thus, obvious differences are found in the solutions. For example, automatic coding method used in public video surveillance is rare in active privacy protection [13–15].

This article introduces how to solve the privacy protection about IoT sensor data based on blockchain [16]. The ring signature realizes the anonymization of gateway transactions, prevents data sources from being tracked, and

solves the anonymization problem of blockchain-based IoT users [17]. Through asymmetric encryption, the sensor verifies the identity of the gateway and encrypts the sensor data [18]. The gateway can create a block and submit a blockchain transaction after verifying the integrity and source of the data. Finally, combined with data access control and data encryption sharing, decentralized applications perform fine-grained control over data access.

2. Method Architecture

The method architecture is shown in Figure 1. The system consists of sensors, gateways, blockchain, and various decentralized applications. First of all, the sensor and the gateway need to perform bilateral authentication, which can not only prevent the sensor from accessing the fake gateway, but the gateway also filters out offensive sensor data. After the identity verification is passed, Hash-based message authentication code is used to verify the source and integrity of the data collected by the sensor. The sensor uploads the data to the gateway. In order to protect the privacy of the sensor data, the gateway will first encrypt the data with a public key, use a ring signature to verify the anonymity of the transaction, and finally submit it to the blockchain. Based on smart contracts, an attribute-based control method is adopted. The encrypted data on the blockchain are decrypted and used by decentralized applications by proxy re-encryption.

2.1. Sensor-Gateway Authentication. To avoid the leakage of private data caused by the sensor's access to the wrong gateway, this study adopts the Elliptic Curve Diffie–Hellman (ECDH) protocol combined with an asymmetric encryption method to realise the sensor's authentication of the gateway's identity and negotiate a shared key in the process.

ECDH is a variant of the Diffie–Hellman (DH) protocol that uses elliptic curve cryptography. The difference between the two is that ECDH is based on the elliptic curve discrete logarithm problem, whereas the DH protocol is based on the discrete logarithm problem. Similar to the DH protocol, the two parties in ECDH communication use their elliptic curve key pairs to negotiate a shared key on an insecure channel. This key can be used for the symmetric encryption of subsequent communications between the two parties. The negotiation process is shown in Figure 2.

The sensor and the gateway share a set of elliptic curve domain parameters (p, a, b, G, n, h). The sensor generates a random number $d_{a,g}$, $(d_g \in [2, n-1])$, then the sensor sends Q_s to the gateway, and the gateway sends Q_g to the sensor. Finally, both parties obtain the same shared key (i.e., $K = K_s = K_g$).

The ECDH-based IoT device authentication to the gateway and the key negotiation process is shown in Figure 3.

(i) The sensor side generates a random number d_s, encrypts Q_s with the public key of the gateway, and sends it

(ii) The gateway side generates a random number, d_g, and sends back $Q_s \| Q_g$ in plaintext

(iii) If the sensor successfully parses out Q_s, the gateway identity is correct

The above process shows that, when the system is initialised, an ECC key pair needs to be allocated to the gateway, and the sensor needs to know the public key that it needs to access the gateway.

2.2. Sensor Data Encryption. The gateway creates and submits a blockchain transaction after verifying the integrity and source of the sensor data. To protect the privacy of sensor data when constructing a transaction, this article first uses the ECC public key to encrypt sensor data. The principle of public key encryption is as follows:

(i) The 32-bit input data m are converted into point M on the elliptic curve, where f represents the selected elliptic curve equation:

$$M = (m, f(m)). \tag{1}$$

(ii) A random number $r \in [2, n-1]$ is taken, where n represents the order of the selected elliptic curve equation.

(iii) The first part $C1$ of the encrypted output is calculated, where G is the Genrator of the elliptic curve:

$$C1 = r \times G. \tag{2}$$

(iv) The second part $C2$ of the encrypted output is calculated, where K represents the public key used for encryption:

$$C2 = M + r \times K. \tag{3}$$

(v) The encrypted outputs $C1$ and $C2$ are obtained.

The process of decrypting with the ECC private key is as follows:

(i) M is calculated according to the following formula, where k is the private key used for decryption:

$$M - C2 - k \times C1. \tag{4}$$

(ii) The original information m is extracted from M. For example, M is taken to obtain the x coordinate. The above process is only applicable to the encryption and decryption processing of the input message m with a length of 32 bytes. The message m of any length should be divided if the 32-byte fragments $m_1, \ldots, m_n$ are viewed. The encryption process is performed on each fragment at a time, and the encryption result is obtained.

$$C = [C1, C2_1, \ldots, C2_n]. \tag{5}$$

When decrypting, $C1$ is used to process $C2_1, \ldots, C2_n$ sequentially and obtain the result.

2.3. Anonymization of Transaction Sources. To avoid tracking and identifying the source of the data (gateway), the Borromean ring signature method is used to anonymize gateway transactions. Ring signature is a special group signature method. The difference is that a ring signature does not require an additional group manager. With ring signature, the real signer can be hidden behind a set of public keys (address), thus realising the transaction anonymity of the true initiator.

Assuming that the message to be signed is m, the signer's private key is S_s, and the selected ring members are $P_1, P_2, \ldots, P_r$, the calculation process of the ring signature is as follows:

(i) The hash of the message m is calculated as the symmetric key k as follows:

$$k = h(m). \tag{6}$$

(ii) The signer chooses a random value $v \in [0, 1]$.

(iii) The signer chooses a random value for other members (i.e., $x_i \in [0, 1], 1 \le i \le r, i \ne s$).

(iv) y_s is solved, which makes the following formula true:

$$C_{k,v}(y_1, y_2, \ldots, y_k) = v,$$
$$C_{k,v}(y_1, y_2, \ldots, y_k) = E_k(y_r \oplus E_k(y_{r-1} \oplus E_k(\ldots \oplus E_k(y_1 \oplus v)))). \tag{7}$$

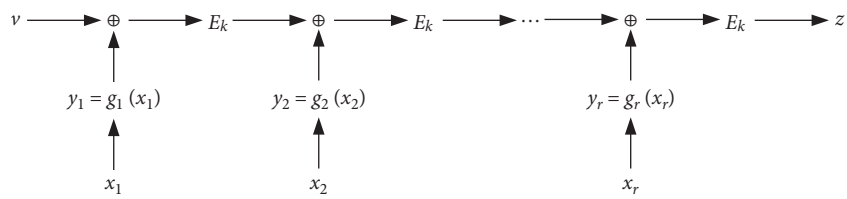
(v) Signer trapdoor permutation and inversion are used:

$$x_s = g_x^{-1}(y_s). \tag{8}$$

(vi) The ring signature is the output:

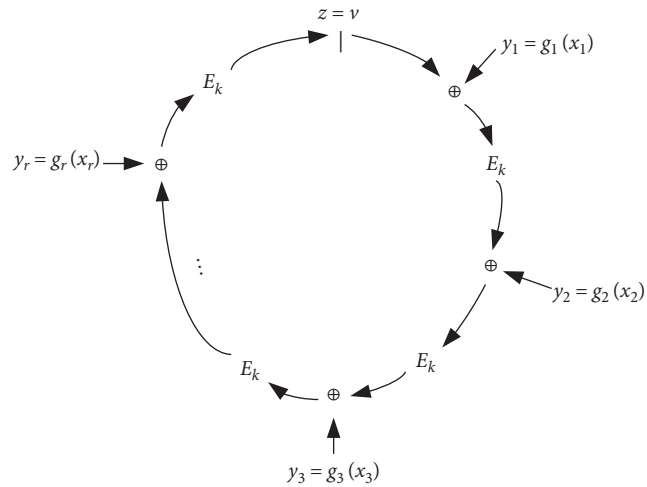
$$(P_1, P_2, \ldots, P_r; v; x_1, x_2, \ldots, x_r). \tag{9}$$

The equation calculated in step 4 can be shown in the figure below, where E_k is the symmetric encryption function:



(10)

The verification process of the ring signature is as follows:



(11)

The equation can be expressed as a ring as follows:

(i) Calculate y_i as follows:

$$y_i = g_i(x_i). \tag{12}$$

(ii) Calculate the encryption key k as follows:

$$k = h(m). \tag{13}$$

(iii) Verify the ring equation as follows:

$$C_{k,v}(y_1, y_2, \ldots, y_k) = v. \tag{14}$$

2.4. Sensor Data Shared. The encrypted sensor data can be decrypted and used by the encryptor. Third-party use must be considered in many cases. However, directly sharing the private key of the encryptor is not safe. Thus, this article uses a proxy re-encryption method to realise the sharing of encrypted data. The proxy re-encryption process is shown in Figure 4. In this article, the blockchain node acts as a re-encryption agent.

Users A and B hold key pairs (sK_A, pK_A) and (sK_B, pK_B), respectively. User A uses his public key pK_A to encrypt the inscription data m to obtain the ciphertext C_A on the chain. When user A needs to share his data with user B, user A generates a re-encryption key $rK_{A \longrightarrow B}$ for user B and provides it to the blockchain node to re-encrypt the specified ciphertext to C_B. After receiving it, user B uses his private key. The key sK_B can be decrypted.

(i) b: user B's private key

(ii) a: user A's private key

(iii) q: order

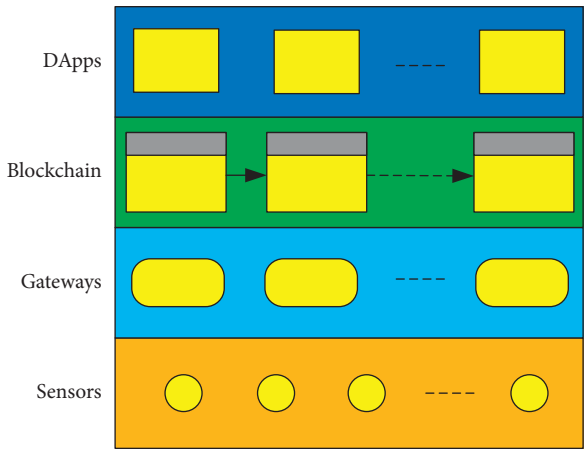


Figure 1: Method architecture.

Given that two sets of private keys are required to generate the re-encryption key and to avoid using B's private key, user A needs to generate an additional temporary private key and provide it to user B after encrypting it with B's public key.

To access the data of the specified sensor, decentralized applications need to request sensor data access permission from the gateway to which the sensor belongs, that is, to send the following message to the specified gateway. The fields of the message are shown in Table 1.

If the gateway agrees for decentralized applications to access the specified sensor, the gateway submits an authorisation transaction to the PAP contract on the chain, granting the requester the access permission to the specified sensor. The transaction parameters are as follows:

(i) Target contract: PAP
(ii) Contract action: grant
(iii) Action parameters:
(iv) Sensor: authorised sensor ID
(v) User: requester ID
(vi) Rk: re-encryption key generated for the requester
(vii) Sk: decryption key generated by the requester, which is encrypted with the requester's public key

After the transaction is confirmed on the chain, the pap contract is triggered to modify the access strategy of the specified sensor. The process is shown in Figure 5.

After the above transaction is confirmed, the requesting party can encrypt and decrypt the specified sensor data to obtain plaintext data. Decentralized applications request the latest sensor data by sending the following message to the blockchain node. The fields of the message are shown in Table 2.

After the node receives the above request, it will first check whether the requester has the permission to access the requested sensor. If the permission is granted, it will return the latest data (encrypted form) of the sensor and the key authorised by the gateway to the requester. The fields of the message are shown in Table 3.

The sequence diagram of the above process is shown in Figure 6.

2.5. Data Access Control. ABAC can finely control access to resources and mainly includes four components, namely, policy enforcement point (PEP), policy decision point (PDP), policy administration point (PAP), and policy information point (PIP). The data access process is shown in Figure 7.

(i) PEP is responsible for receiving user requests, invoking PDP permission evaluation and determining whether to allow access to specified resources based on PDP evaluation results
(ii) PDP evaluates the access request based on the rule base and returns the evaluation result (i.e., denying or allowing access)
(iii) PAP is the management interface of rules provided for administrators, such as adding new access policies and updating designated access policies
(iv) PIP provides out-of-core attribute information for PDP

3. Experimental Evaluation

3.1. Verification of Encrypted Data Utilisation Process. The experimental configuration is as follows:

(i) Synthesis of sensor data: open
(ii) Verification of sensors and gateway devices: open
(iii) Encryption of transactions: open
(iv) Data encryption method: ECC

The data chaining process in the experiment is described as follows:

(i) The sensor verifies the identity of the gateway and negotiates a shared key
(ii) The sensor submits data to the IoT gateway
(iii) The IoT gateway node verifies the integrity and source of the data and rejects the data if the verification fails
(iv) The gateway encrypts sensor data, generates new transactions, and performs ring signatures
(v) After the blockchain node verifies that the transaction is correct, the node will queue the transaction up in the buffer pool
(vi) Blockchain nodes generate blocks to confirm transactions

The access authorisation process in the experiment is as follows:

(i) The gateway submits an authorisation transaction to the blockchain node and grants the data access rights of 80000# sensor to decentralized applications.

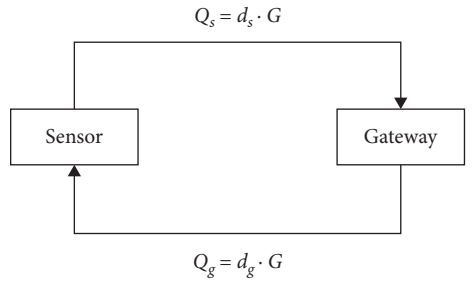


Figure 2: Key agreement.

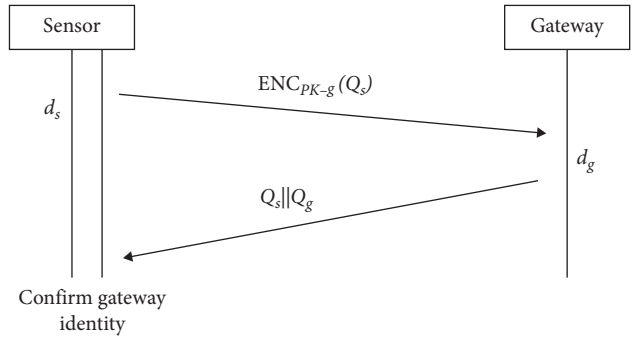


Figure 3: IoT device authentication to the gateway and key agreement.

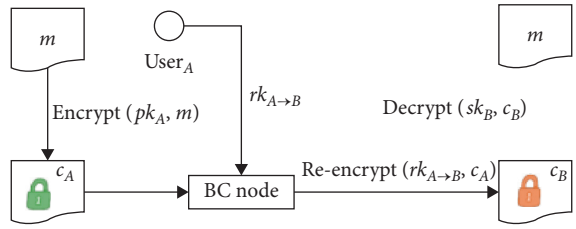


Figure 4: Proxy re-encryption.

(ii) Decentralized applications regularly submit query requests to blockchain nodes and obtain encrypted sensor data and the proxy re-encryption group.

(iii) Decentralized applications perform re-encryption first, decrypts the data, and displays the decrypted plaintext. The experimental results are in line with expectations. The screenshot is shown in Figure 8.

3.2. The Impact of Encryption on Performance. The program is implemented in Python language, and the experiment is mainly carried out in the following aspects: the impact of transaction signatures on performance and the impact of sensor data encryption on performance.

The system has set up 50 sensor nodes, 10 gateways, and 4 blockchain nodes. The system can set parameters in advance, set the block generation cycle of the blockchain node to 5 s, set the sensor report data cycle to 1 s, enable sensor data integration, configure whether to verify the sensor/gateway, and configure transaction signatures and sensor encryption methods.

3.2.1. The Impact of Transaction Signatures on Performance. We use different signature methods to discuss the impact on system performance. The parameters are as follows:

(i) Cycle of block generation: 5 s

(ii) Cycle of sensor data submission: 1 s

(iii) Synthesis of sensor data: open

(iv) Verification of sensors and gateway devices: close

(v) Encryption of transactions: close

(vi) Data encryption method: none/ECC/ring

The transaction delay time statistics of the no-signature method and the ECC signature method are shown in Figures 9–12

Figures 9–12 show the influence of ECC signature on transaction delay time. When no signature is added, all transactions are confirmed within 5 s of the block generation period. When the ECC signature is added, the maximum transaction confirmation time is delayed to more than 7 s. That is, given that ECC signature requires a certain processing time, the simulation system has been overloaded

Table 1: The format of the message sent by decentralized applications.

#	Field	Type	Explanation
1	Type	String	Message type and value: sensor_rights_request
2	Payload	Bytes	The message payload is as follows: (i) Requestor: requester ID (ii) Requestor_pk: requestor public key (iii) Sensor: target sensor number

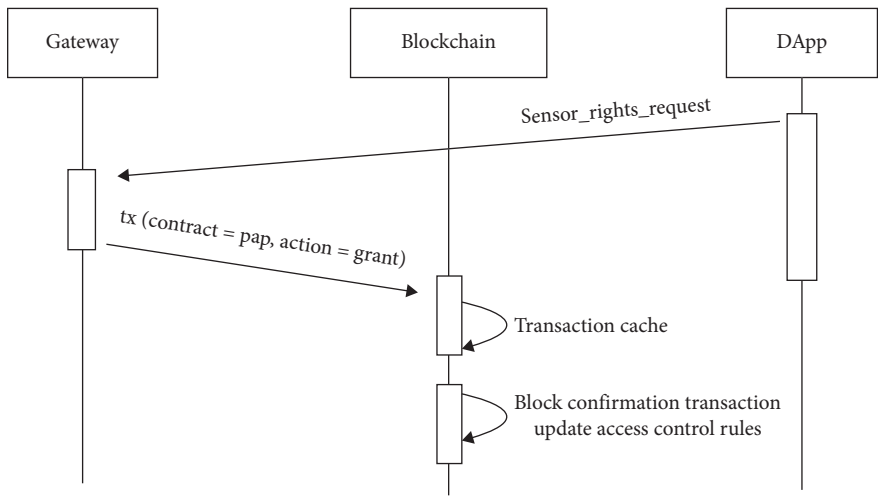


Figure 5: Transaction process.

Table 2: The format of the request sensor data sent by the decentralized applications.

#	Field	Type	Explanation
1	Type	String	Message type and value: state_request
2	Payload	Bytes	The message payload is as follows: (i) Target contract: SCADA (ii) Contract view: sensor_latest (a) View parameters: (b) Requestor: requestor ID (c) Sensor: requested sensor ID

Table 3: Node returns sensor data.

#	Field	Type	Explanation
1	Type	String	Message type and value: state_response
2	Payload	Bytes	The message payload is as follows: (i) Encrypted sensor data (ii) Data re-encryption key (iii) Data decryption key (encrypted by the requester's public key)

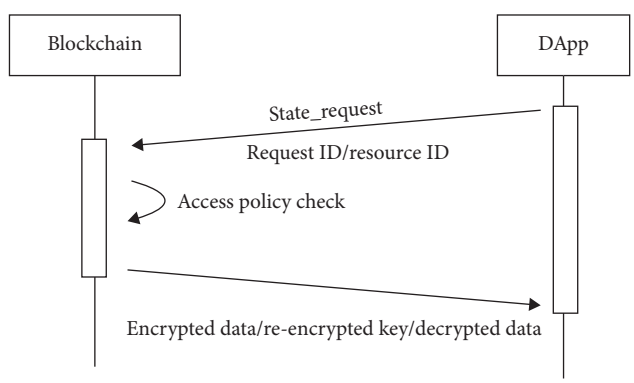


Figure 6: Access policy check.

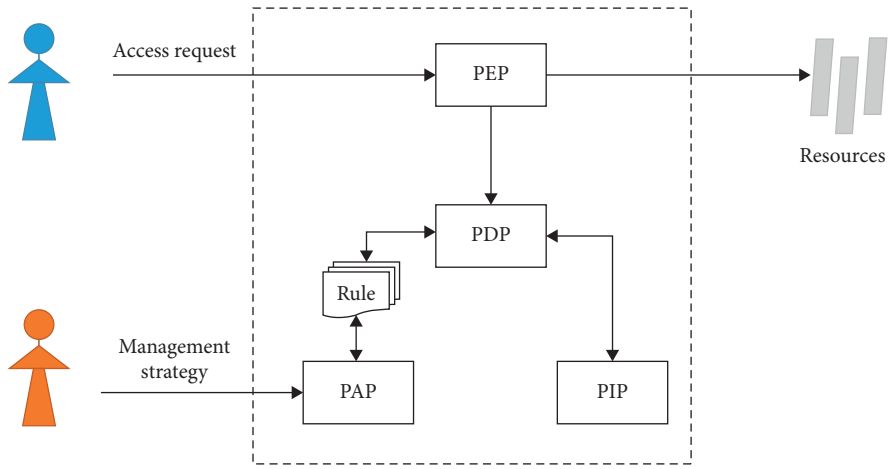


Figure 7: Data access control.

Figure 8: System operation result.

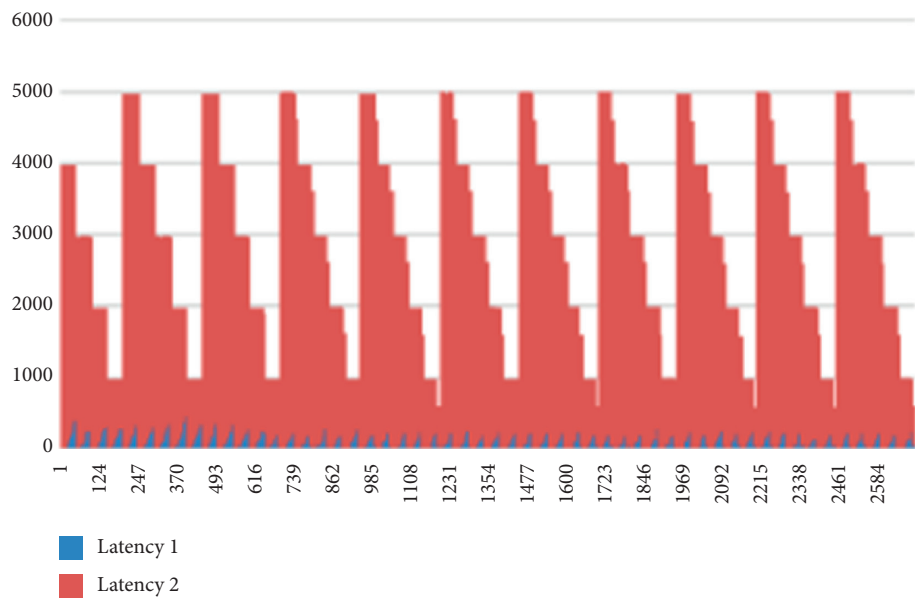


Figure 9: Transaction delay (no signature, $T = 5$ s).

when the block generation period is set to 5 s. When the ring signature method is used, this overload phenomenon is more obvious because the ring signature requires more processing time. Figures 13 and 14 show that the maximum transaction delay time has exceeded 20 s.

The block generation period is set to 30s, and the experiment of the ring signature method is re-run. The results are shown in Figures 15 and 16

The transaction delays are kept within the block generation period, and the system is operating normally.

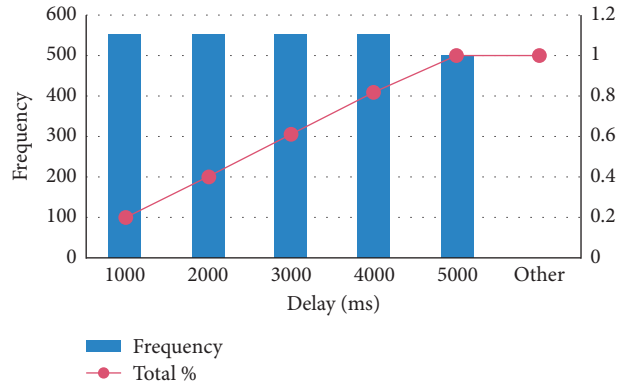


Figure 10: Transaction delay histogram (no signature, $T = 5$ s).

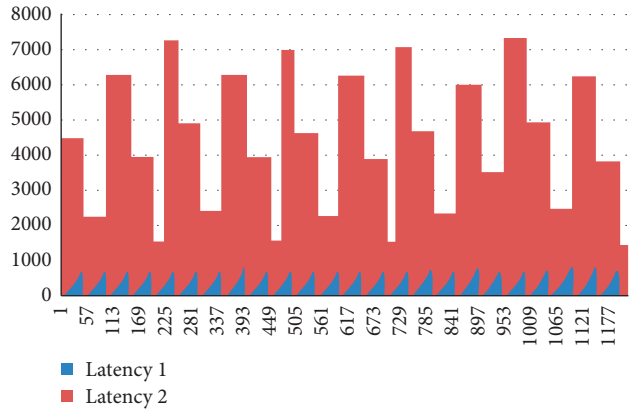


Figure 11: Transaction delay (ECC signature, $T = 5$ s).

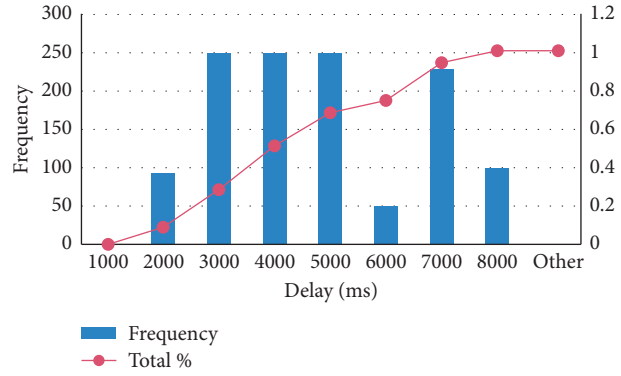


Figure 12: Transaction delay histogram (ECC signature, $T = 5$ s).

3.2.2. The Impact of Sensor Data Encryption on Performance. Sensor data encryption and nonencryption methods are used to investigate the impact of this link on performance. The experimental parameters are as follows:

(i) Cycle of block generation: 5 s

(ii) Cycle of sensor data submission: 1 s

(iii) Synthesis of sensor data: open

(iv) Verification of sensors and gateway devices: close

(v) Encryption of transactions: close

(vi) Data encryption method: none/ECC/ring

The experimental results are shown in Figures 17 and 18.

Figures 17 and 18 show that the encryption of sensor data has little effect on transaction delay. Figure 17 shows the corresponding transaction delay histogram statistics and transaction delay cumulative ratio statistics when sensor data encryption is not enabled. Figure 18 shows the statistics of the transaction smoking, eating, and releasing graphs and

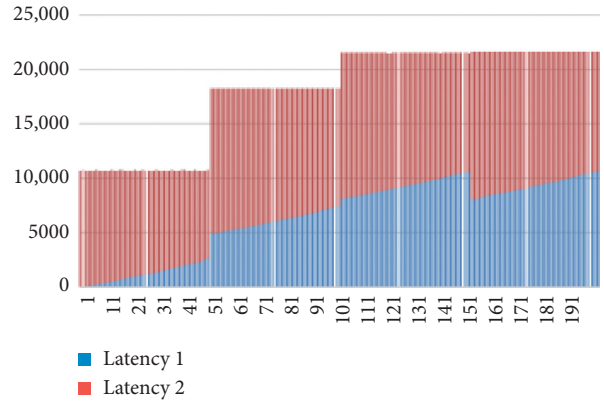


Figure 13: Transaction delay (ring signature, $T = 5$ s).

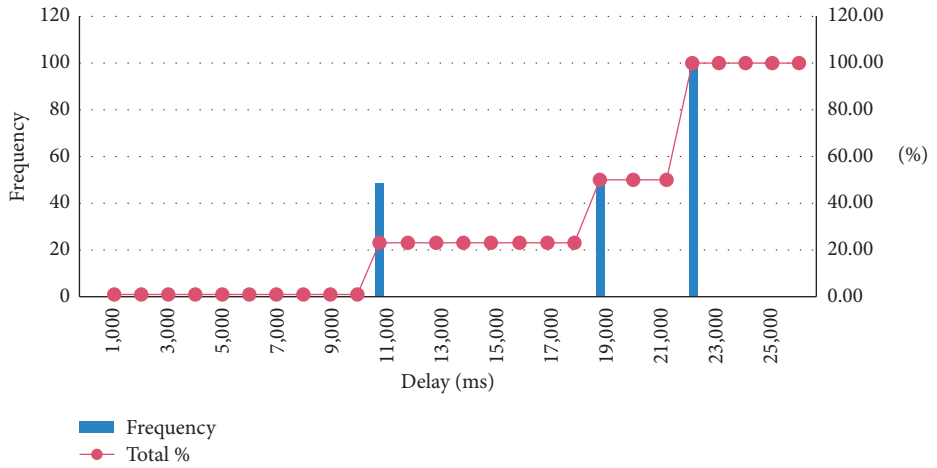


Figure 14: Transaction delay histogram (ring signature, $T = 5$ s).

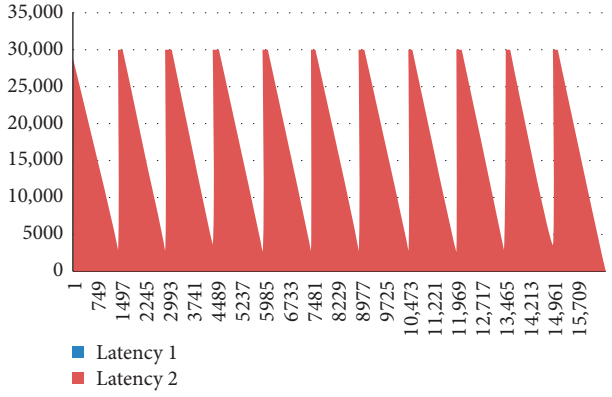


Figure 15: Transaction delay (ring signature, $T = 30$ s).

the cumulative ratio of transaction delays after the sensor data encryption is turned on. In Figures 17 and 18, the blue histogram counts the number of transactions that fall in each delay interval, and the red line graph calculates the proportion of the number of delays corresponding to the current interval in the total number of transactions, which we call it the cumulative proportion of delay in this interval is displayed on the ordinate on the right. From the two figures, we can see that after the sensor data encryption is turned on, it will affect the total number of transactions. As can be seen from Figure 18, the number of transaction delays in each interval has been reduced. However, after the sensor data encryption is turned on, the transaction delay time will not be affected. From the red line chart, we can see that all transaction delays are still below 5000 ms, and the cumulative proportion accounts for almost 100%.

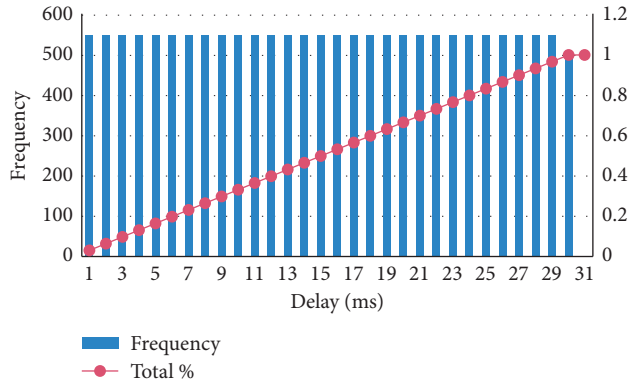


Figure 16: Transaction delay histogram (ring signature, $T = 30$ s).

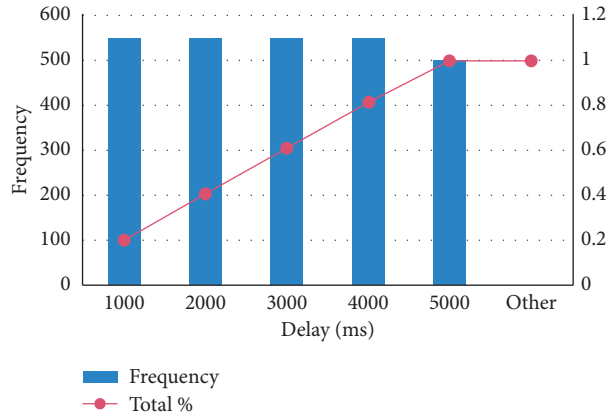


Figure 17: No sensor data encryption.

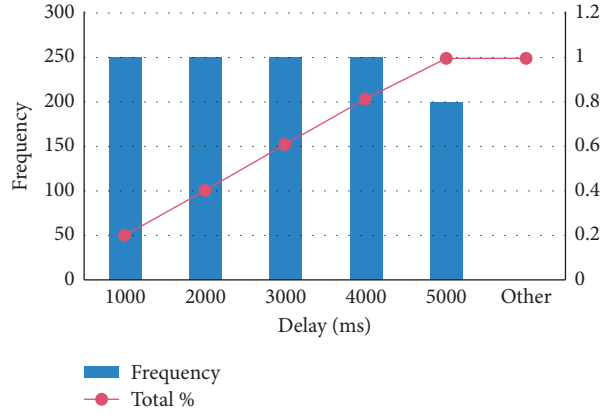


Figure 18: Enable sensor data encryption.

4. Conclusion

This article introduces sensor data privacy protection method for IoT based on blockchain technology. The ring signature realizes the anonymization of gateway transactions, prevents data sources from being tracked, and solves the anonymization problem of blockchain-based IoT users. Through asymmetric encryption, the sensor verifies the identity of the gateway and encrypts the sensor data. The gateway can create a block and submit a blockchain transaction after verifying the integrity and source of the data. Finally, combined with data access control and data encryption sharing, decentralized applications can finely control data access. Through experiments, the impact of transaction signatures on performance and the impact of sensor data encryption on performance are analyzed. The

results show that transaction delays are all controlled within a reasonable range. The system performance achieved by this method is also relatively stable.

Data Availability

The data that support the findings of the research are available from the corresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the Key Research and Development Plan of Anhui Province (201904a06020056). The authors are very grateful for the research foundation of the conference paper "IoT Data Privacy Protection Scheme Based on Blockchain" in IOP Science.

References

[1] M. A. Khan and K. Salah, "IoT security: review, blockchain solutions, and open challenges," *Future Generation Computer Systems*, vol. 82, pp. 395–411, 2018.

[2] R. Mitchell and I.-R. Chen, "A survey of intrusion detection in wireless network applications," *Computer Communications*, vol. 42, pp. 1–23, 2014.

[3] L. Atzori, A. Iera, and G. Morabito, "The internet of things: a survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[4] A. A. Khan, M. H. Rehmani, and A. Rachedi, "Cognitive-radio-based internet of things: applications, architectures, spectrum related functionalities, and future research directions," *IEEE wireless communications*, vol. 24, no. 3, pp. 17–25, 2017.

[5] J. Granjal, E. Monteiro, and J. Sa Silva, "Security for the internet of things: a survey of existing protocols and open research issues," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1294–1312, 2015.

[6] S. M. H. Bamakan, N. Faregh, and A. ZareRavasan, "Di-ANFIS: an integrated blockchain-IoT-big data-enabled framework for evaluating service supply chain performance," *Journal of Computational Design and Engineering*, vol. 8, no. 2, pp. 676–690, 2021.

[7] S. Cirani, G. Ferrari, and L. Veltri, "Enforcing security mechanisms in the IP-based internet of things: an algorithmic overview," *Algorithms*, vol. 6, no. 2, pp. 197–226, 2013.

[8] Y. Liu, K. Wang, Y. Lin, and W. Xu, "Lightweight blockchain system for industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3571–3581, 2019.

[9] A. Reyna, C. Martín, J. Chen, E. Soler, and M. Díaz, "On blockchain and its integration with IoT. Challenges and opportunities," *Future Generation Computer Systems*, vol. 88, pp. 173–190, 2018.

[10] A. Dorri, S. S. Kanhere, R. Jurdak et al., "Blockchain for IoT security and privacy: the case study of a smart home," in *Proceedings of the IEEE Percom Workshop on Security Privacy and Trust in the Internet of Thing*, pp. 618–623, IEEE, Kona, HI, USA, March 2017.

[11] P. Patil, M. Sangeetha, and V. Bhaskar, "Blockchain for IoT access control, security and privacy: a review," *Wireless Personal Communications*, vol. 117, pp. 1–20, 2020.

[12] D. Minoli and B. Occhiogrosso, "Blockchain mechanisms for IoT security," *Internet of Things*, vol. 1-2, pp. 1–13, 2018.

[13] Y. Qian, Y. Jiang, J. Chen et al., "Towards decentralized IoT security enhancement: a blockchain approach," *Computers & Electrical Engineering*, vol. 72, pp. 266–273, 2018.

[14] R. Di Pietro, X. Salleras, M. Signorini et al., "A blockchain-based trust system for the internet of things," in *Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies*, pp. 77–83, Indianapolis, IN, USA, June 2018.

[15] H. Si, C. Sun, Y. Li, H. Qiao, and L. Shi, "IoT information sharing security mechanism based on blockchain technology," *Future Generation Computer Systems*, vol. 101, pp. 1028–1040, 2019.

[16] D. Johnson, A. Menezes, and S. Vanstone, "The elliptic curve digital signature algorithm (ECDSA)," *International Journal of Information Security*, vol. 1, no. 1, pp. 36–63, 2001.

[17] N. P. Smart, "The exact security of ECIES in the generic group model," in *Proceedings of the IMA International Conference on Cryptography and Coding*, pp. 73–84, Springer, Cirencester, UK,, December 2001.

[18] L. Zhou, L. Wang, Y. Sun, and P. Lv, "BeeKeeper: a blockchain-based IoT system with secure storage and homomorphic computation," *IEEE Access*, vol. 6, pp. 43472–43488, 2018.

Hindawi
Shock and Vibration
Volume 2021, Article ID 2108457, 6 pages
https://doi.org/10.1155/2021/2108457

Research Article

A Novel Fault Diagnosis Method for Motor Bearing Based on DTCWT and AFSO-KELM

Yan Lu and Peijiang Li

Department of Information Engineering, Quzhou College of Technology, Quzhou 324000, China

Correspondence should be addressed to Yan Lu; luyan@qzct.net

Received 12 May 2021; Revised 29 May 2021; Accepted 8 June 2021; Published 17 June 2021

Academic Editor: Chaoqun Duan

Aiming at the defects of wavelet transform-based feature extraction and extreme learning machine-based classification, a novel fault diagnosis method for motor bearing, based on dual tree complex wavelet transform and artificial fish swarm optimization-kernel extreme learning machine (DTCWT-AFSO-KELM), is proposed in this paper. Firstly, the dual tree complex wavelet transform instead of the discrete wavelet transform is used to decompose the motor bearing signal; then, the features with large differentiation of motor-bearing fault are extracted; finally, the states of motor bearing are classified by using artificial fish swarm optimization-kernel extreme learning machine. In order to better prove the superiority of this method, four kinds of state data of motor bearing under the conditions of 0 HP (horsepower) load, 1 HP load, 2 HP load, and 3 HP load are used to test. The experimental results indicate that the diagnosis accuracies of DTCWT-AFSO-KELM are obviously better than those of discrete wavelet transform and artificial fish swarm optimization-kernel extreme learning machine (DWT-AFSO-KELM) or discrete wavelet transform and extreme learning machine (DWT-ELM) under different loads.

1. Introduction

Fault diagnosis of motor bearing is very significant to ensure the normal operation of the motor. At present, fault diagnosis methods of motor bearing include artificial neural network, support vector machine, and extreme learning machine. [1–3], among which extreme learning machine [4–9] has a wide application in classification and prediction fields due to its advantages of fast network training speed and good generalization performance, and extreme learning machine is a promising fault diagnosis method of motor bearing. However, it is necessary to set the number of hidden layer nodes in the training process of extreme learning machine. The existence of multicollinearity in data samples may lead to singularity, resulting in inconsistent input weights of the hidden layer, which affects the generalization performance of extreme learning machine. In this paper, kernel mapping is used to replace the random mapping of extreme learning machine, and the extreme learning machine is denoted as kernel extreme learning machine (KELM).

The penalty factor and kernel parameter of the kernel extreme learning machine need to be optimized due to the influence of the penalty factor and kernel parameter on the classification performance of the kernel extreme learning machine. Because of the shortcomings of ant colony algorithm and bee colony algorithm, artificial fish swarm optimization (AFSO) algorithm is used to optimize the parameters of kernel extreme learning machine to realize the parameter optimization of kernel extreme learning machine effectively. AFSO algorithm has the characteristics of fast search speed, high precision, and avoiding local minimum. Therefore, the kernel extreme learning machine optimized by artificial fish swarm optimization algorithm has better classification performance than ordinary extreme learning machine.

Feature extraction plays an important role in fault diagnosis of motor bearing, and signal decomposition algorithm is the key to accurately extract the running features of motor bearing. The dual tree complex wavelet transform (DTCWT) is a new wavelet transform method with many

excellent characteristics developed in recent years. It has the advantages of approximate translation invariance, antiband aliasing, and complete reconstruction. Therefore, this paper uses dual tree complex wavelet transform to replace the commonly used discrete wavelet transform (DWT) [10–12] to decompose the motor bearing signal, which is helpful to improve the feature differentiation of different states of motor bearing in fault diagnosis of motor bearing.

Aiming at the defects of wavelet transform-based feature extraction and extreme learning machine-based classification, a novel fault diagnosis method for motor bearing based on dual tree complex wavelet transform and artificial fish swarm optimization-kernel extreme learning machine is proposed in this paper. Firstly, the dual tree complex wavelet transform is used to decompose the signal of motor bearing, and the features of large differentiation of different states of motor bearing are extracted, and the kernel extreme learning machine optimized by artificial fish swarm optimization is used to classify the state of motor bearing. The motor bearing data set of Case Western Reserve University is used as the experimental data. The common faults of motor bearing include inner ring fault, outer ring fault, and ball fault. In order to better prove the superiority of this method, four kinds of state data of motor bearing under the conditions of 0 HP load, 1 HP load, 2 HP load, and 3 HP load are used to test.

Firstly, the dual tree complex wavelet transform is introduced; secondly, the kernel extreme learning machine optimized by artificial fish swarm optimization is described; thirdly, the detailed experimental results and analysis are described; finally, conclusions are described.

2. Dual Tree Complex Wavelet Transform

Dual tree complex wavelet transform is a new wavelet transform method developed in recent years with many excellent characteristics, such as approximate translation invariance, antialiasing, and complete reconstruction. DTCWT adopts two parallel DWTs with different low-pass and high-pass filters. In order to obtain better symmetry, the filter length of one branch tree is odd and the other branch tree is even. There is a delay of one sampling interval between two branch tree filters, and the real part tree is missing. The sample value can be acquired by the imaginary part tree without losing the hidden information contained in the original signal hide information.

The reconstruction algorithm of dual tree complex wavelet coefficients is described as [13–15]

$$d_i(t) = 2^{(i-1)/2}\left[\sum_m d_i^{\mathrm{Re}}(k)\phi_h\left(2^i t - m\right) + \sum_n d_i^{\mathrm{Im}}(k)\phi_g\left(2^i t - n\right)\right],$$
$$c_j(t) = 2^{(j-1)/2}\left[\sum_m c_j^{\mathrm{Re}}(k)\varphi_h\left(2^j t - m\right) + \sum_n d_i^{\mathrm{Im}}(k)\varphi_g\left(2^i t - n\right)\right], \quad (1)$$

where m and n denote the lengths of the real and imaginary part filters of the dual tree.

This paper uses the dual tree complex wavelet transform to replace the commonly used discrete wavelet transform to decompose the motor-bearing signal, which is helpful to improve the feature differentiation of fault diagnosis of motor bearing. The number of subsignals of the motor-bearing signal is determined according to the amplitude-frequency contrast algorithm for the purpose of achieving the separation of signal frequencies accurately.

3. Kernel Extreme Learning Machine Optimized by Artificial Fish Swarm Optimization

3.1. Kernel Extreme Learning Machine. The classification function of extreme learning machine is described as

$$f(x) = h(x)\beta, \quad (2)$$

where $h(x)$ is the feature mapping function matrix and $\beta = [\beta_1, \ldots, \beta_L]^T$ is the weight vector connecting the hidden layer and the output layer.

The training objective of extreme learning machine is to calculate the output weight vector β, $\beta = H^{-1}T$, where $H = [h(x_1), \ldots, h(x_N)]^T$ is the hidden layer feature mapping matrix and the training objective matrix.

For KELM, (I/C) is added to the main diagonal of the unit diagonal matrix, and the KELM weight matrix is described as

$$\beta = H^T\left(\frac{I}{C} + HH^T\right)^{-1}T, \quad (3)$$

where C is the penalty factor and I is the identity matrix.

KELM introduces the kernel function instead of the characteristic matrix, and the corresponding KELM classification function is

$$f(x) = \begin{bmatrix} K(x, x_1) \\ K(x, x_2) \\ \cdots \\ K(x, x_N) \end{bmatrix}^T \left(\frac{I}{C} + \Delta\right)^{-1} T^T, \quad (4)$$

where $\Delta = K(x_i, x_j)$ is the kernel function, KELM selects the Gaussian radial basis function, and $K(x_i, x_j) = \exp(-\alpha\|x_i - x_j\|^2)$, where α is the kernel parameter.

The penalty factor C and radial basis function kernel parameter α of KELM need to be optimized. The suitable intelligent optimization algorithm needs to be selected to optimize the penalty factor C and radial basis function kernel parameter α of KELM.

3.2. Parameter Optimization of Kernel Extreme Learning Machine Based on Artificial Fish Swarm Optimization Algorithm. Artificial fish swarm optimization algorithm makes up for the shortcomings of ant colony algorithm and bee colony algorithm [16] and has the

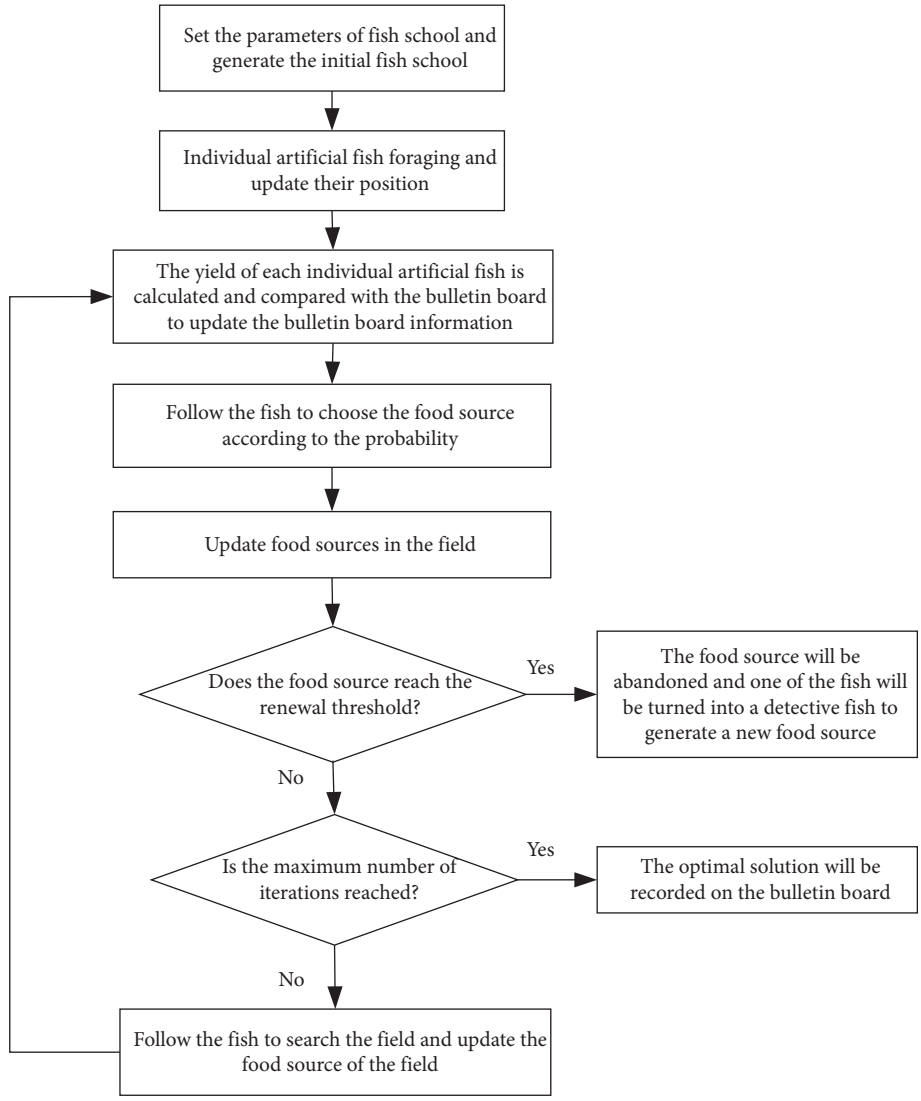


Figure 1: Parameter optimization of kernel extreme learning machine based on artificial fish swarm optimization algorithm.

characteristics of fast search speed, high precision, and avoiding local minimum. Therefore, the parameters of kernel extreme learning machine can be effectively optimized by artificial fish swarm optimization algorithm. The process of optimizing kernel extreme learning machine by artificial fish swarm is shown in Figure 1, and the optimization process of kernel extreme learning machine by artificial fish swarm optimization is described as follows:

Step 1: KELM's penalty factor C and radial basis function kernel parameter α constitute the individual position of the artificial fish, and the parameters of the fish group are set, including the size of the fish group m, the maximum number of iterations, and the moving step; m artificial fish individuals are randomly generated as the initial fish group;

Step 2: individual artificial fish foraging and update their position according to the following formula:

$$Y_{i\text{next}} = Y_p + \left(\text{step} \cdot \frac{Y_c - Y_i}{\|Y_c - Y_i\|}\right) + \left(\text{step} \cdot \frac{Y_{\max} - Y_i}{\|Y_c - Y_i\|}\right), \tag{5}$$

where Y_c is the center position of the whole artificial fish group, $Y_{\max}$ is the optimal position of the whole artificial fish group at present, Y_p is the position of the artificial fish after foraging behavior, and Y_i is the current position of the artificial fish.

Step 3: the yield of each individual artificial fish is calculated and compared with the bulletin board to update the bulletin board information.

Step 4: follow the fish to choose the food source according to the probability.

Step 5: update food sources in the field.

Step 6: check the update threshold of the food source. If the update threshold is reached, the food source will be abandoned and one of the fish will be turned into a

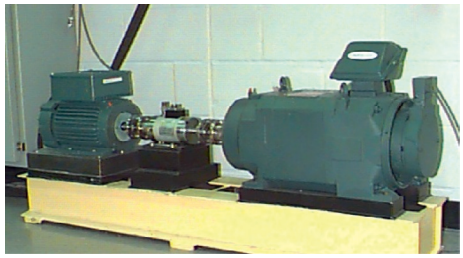


Figure 2: The experimental device.

detective fish to generate a new food source; otherwise, go to step 7.

Step 7: check whether the maximum number of iterations is satisfied. If it is satisfied, the optimal solution will be recorded on the bulletin board, namely, the penalty factor C and the radial basis function kernel parameter α of the optimal KELM are output, and the algorithm ends; otherwise, calculate the fish yield, update the bulletin board information, and go to step 3.

4. Experimental Results and Analysis

The motor bearing data set of Case Western Reserve University is used as the experimental data. The experimental device is shown in Figure 2 [17]. Motor-bearing faults are usually inner ring fault, outer ring fault, and ball fault. In order to better prove the superiority of this proposed method, four kinds of state data of motor bearing under 0 HP load, 1 HP load, 2 HP load, and 3 HP load are used to test. Firstly, 280 samples (70 samples expressing normal state, 70 samples expressing inner ring fault, 70 samples expressing outer ring fault, and 70 samples expressing ball fault) of motor bearing under 0 HP load were used as the training samples and another 200 samples (50 samples expressing normal state, 50 samples expressing inner ring fault, 50 samples expressing outer ring fault, and 50 samples expressing ball fault) of motor bearing under 0 HP load were tested. Secondly, 280 samples (70 samples expressing normal state, 70 samples expressing inner ring fault, 70 samples expressing outer ring fault, and 70 samples expressing ball fault) of motor bearing under 1 HP load were used as the training samples and another 200 samples (50 samples expressing normal state, 50 samples expressing inner ring fault, 50 samples expressing outer ring fault, and 50 samples expressing ball fault) of motor bearing under 1 HP load were tested. Thirdly, 280 samples (70 samples expressing normal state, 70 samples expressing inner ring fault, 70 samples expressing outer ring fault, and 70 samples expressing ball fault) of motor bearing under 2 HP load were used as the training samples and another 200 samples (50 samples expressing normal state, 50 samples expressing inner ring fault, 50 samples expressing outer ring fault, and 50 samples expressing ball fault) of motor bearing under 2 HP load were tested. Finally, 280 samples (70 samples expressing normal state, 70 samples expressing inner ring fault, 70 samples expressing outer ring fault, and 70 samples expressing ball fault) of motor bearing under 3 HP load were used as the training samples and another 200 samples (50 samples expressing normal state, 50 samples expressing inner ring fault, 50 samples expressing outer ring fault, and 50 samples expressing ball fault) of motor bearing under 3 HP load were tested. The decomposition signal of motor bearing based on the dual tree complex wavelet transform is shown in Figure 3.

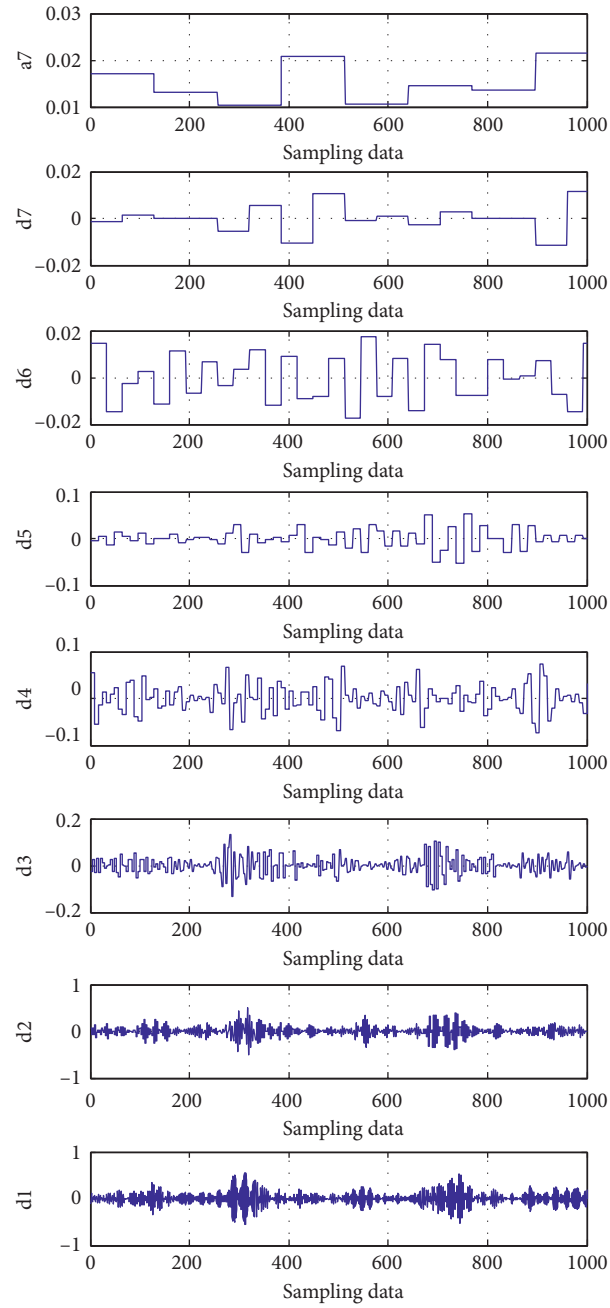


Figure 3: The decomposition signal of motor bearing based on the dual tree complex wavelet transform.

Figure 4 shows the diagnosis results of DTCWT-AFSO-KELM, DWT-AFSO-KELM, and DWT-ELM under different load conditions. As shown in Table 1, under 0 HP load condition, the diagnosis accuracy of motor bearing by using DTCWT-AFSO-KELM is 99.5%, the diagnosis accuracy of

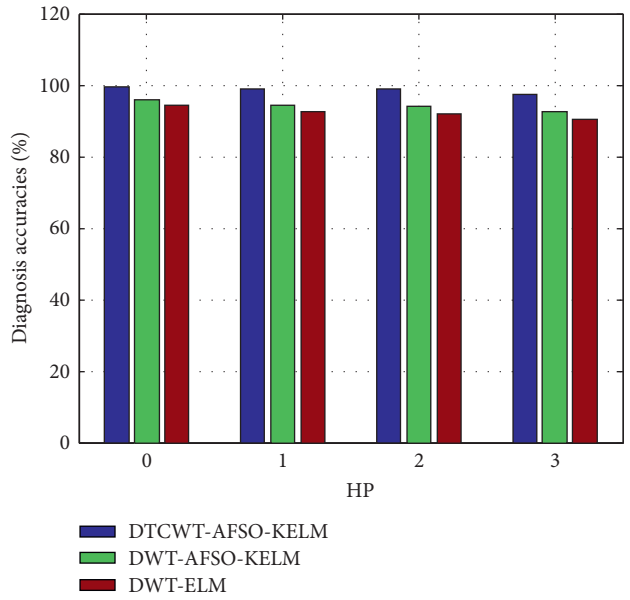


Figure 4: The diagnosis results of DTCWT-AFSO-KELM, DWT-AFSO-KELM, and DWT-ELM under different load conditions.

Table 1: The diagnosis accuracies of DTCWT-AFSO-KELM, DWT-AFSO-KELM, and DWT-ELM under different load conditions.

Load	Diagnosis method	Diagnosis accuracy (%)
0 HP	DTCWT-AFSO-KELM	99.5
	DWT-AFSO-KELM	96
	DWT-ELM	94.5
1 HP	DTCWT-AFSO-KELM	99
	DWT-AFSO-KELM	94.5
	DWT-ELM	92.5
2 HP	DTCWT-AFSO-KELM	99
	DWT-AFSO-KELM	94
	DWT-ELM	92
3 HP	DTCWT-AFSO-KELM	97.5
	DWT-AFSO-KELM	92.5
	DWT-ELM	90.5

motor bearing by using DWT-AFSO-KELM is 96%, and the diagnosis accuracy of motor bearing by using DWT-ELM is 94.5%. Under 1 HP load condition, the diagnosis accuracy of motor bearing by using DTCWT-AFSO-KELM is 99%, the diagnosis accuracy of motor bearing by using DWT-AFSO-KELM is 94.5%, and the diagnosis accuracy of motor bearing by using DWT-ELM is 92.5%. Under 2 HP load condition, the diagnosis accuracy of motor bearing by using DTCWT-AFSO-KELM is 99%, the diagnosis accuracy of motor bearing by using DWT-AFSO-KELM is 94%, and the diagnosis accuracy of motor bearing by using DWT-ELM is 92%. Under 3 HP load condition, the diagnosis accuracy of motor bearing by using DTCWT-AFSO-KELM is 97.5%, the diagnosis accuracy of motor bearing by using DWT-AFSO-KELM is 92.5%, and the diagnosis accuracy of motor bearing by using DWT-ELM is 90.5%. It can be seen that the diagnosis results of motor bearing by using DTCWT-AFSO-KELM are obviously better than those by using DWT-AFSO-KELM or DWT-ELM.

5. Conclusions

In this paper, a novel fault diagnosis method for motor bearing based on dual tree complex wavelet transform and artificial fish swarm optimization-kernel extreme learning machine is proposed because of the defects of wavelet transform-based feature extraction and extreme learning machine-based classification. The contributions of this paper are reduced as follows:

(1) The dual tree complex wavelet transform is used to replace the commonly used discrete wavelet transform to decompose the motor bearing signal, which is helpful to improve the feature differentiation of different states of motor bearing in fault diagnosis of motor bearing.

(2) The parameters of kernel extreme learning machine are optimized by artificial fish swarm optimization algorithm which has the characteristics of fast search speed, high precision, and avoiding local minimum, and the artificial fish swarm optimization-kernel extreme learning machine is used to classify the state of motor bearing. The experimental results show that the diagnosis accuracies of DTCWT-AFSO-KELM are obviously better than those of DWT-AFSO-KELM or DWT-ELM. The future research and development focus on the improvement for artificial fish swarm optimization algorithm to obtain the kernel extreme learning machine with more excellent classification performance.

Data Availability

The data that support the findings of the research are available from the corresponding author.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Basic Public Welfare Research Project of Zhejiang Province (LGG18E050003) and the Second Batch of Teaching Reform Research Projects in the 13th Five-Year Plan of Zhejiang Province (jg20190891).

References

[1] D.-T. Hoang and H.-J. Kang, "Rolling element bearing fault diagnosis using convolutional neural network and vibration image," *Cognitive Systems Research*, vol. 53, pp. 42–50, 2019.

[2] X. Yan and M. Jia, "A novel optimized svm classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing," *Neurocomputing*, vol. 313, pp. 47–64, 2018.

[3] J. Zheng, Z. Dong, H. Pan, Q. Ni, T. Liu, and J. Zhang, "Composite multi-scale weighted permutation entropy and

extreme learning machine based intelligent fault diagnosis for rolling bearing," *Measurement*, vol. 143, pp. 69–80, 2019.

[4] Y. Li, Y. Zeng, Y. Qing, and G.-B. Huang, "Learning local discriminative representations via extreme learning machine for machine fault diagnosis," *Neurocomputing*, vol. 409, pp. 275–285, 2020.

[5] Y. Qing, Y. Zeng, Y. Li, and G.-B. Huang, "Deep and wide feature based extreme learning machine for image classification," *Neurocomputing*, vol. 412, pp. 426–436, 2020.

[6] S. Shukla and B. S. Raghuwanshi, "Online sequential class-specific extreme learning machine for binary imbalanced learning," *Neural Networks*, vol. 119, pp. 235–248, 2019.

[7] F. Olivetti de Franca and M. Zabuscha de Lima, "Interaction-transformation symbolic regression with extreme learning machine," *Neurocomputing*, vol. 423, pp. 609–619, 2020.

[8] G. S. Nandini and A. P. S. Kumar, "Dropout technique for image classification based on extreme learning machine," *Global Transitions Proceedings*, vol. 2, no. 1, pp. 111–116, 2021.

[9] A. Law and A. Ghosh, "Multi-label classification using a cascade of stacked autoencoder and extreme learning machines," *Neurocomputing*, vol. 358, pp. 222–234, 2019.

[10] R. Kamgar, M. Dadkhah, and H. Naderpour, "Seismic response evaluation of structures using discrete wavelet transform through linear analysis," *Structure*, vol. 29, pp. 863–882, 2021.

[11] K. Gopala Krishnan and P. T. Vanathi, "An efficient texture classification algorithm using integrated discrete wavelet transform and local binary pattern features," *Cognitive Systems Research*, vol. 52, pp. 267–274, 2018.

[12] J. Nobre and R. F. Neves, "Combining principal component analysis, discrete wavelet transform and xgboost to trade in the financial markets," *Expert Systems with Applications*, vol. 125, pp. 181–194, 2019.

[13] N. Aishwarya and C. Bennila Thangammal, "Visible and infrared image fusion using dtcwt and adaptive combined clustered dictionary," *Infrared Physics & Technology*, vol. 93, pp. 300–309, 2018.

[14] Z. He, Z. Tang, Z. Yan, and J. Liu, "DTCWT-based zinc fast roughing working condition identification," *Chinese Journal of Chemical Engineering*, vol. 26, no. 8, pp. 1721–1726, 2018.

[15] I. J. Kadhim, P. Premaratne, and P. J. Vial, "High capacity adaptive image steganography with cover region selection using dual-tree complex wavelet transform," *Cognitive Systems Research*, vol. 60, pp. 20–32, 2020.

[16] K. P. Kumar, B. Saravanan, and K. S. Swarup, "Optimization of renewable energy sources in a microgrid using artificial fish swarm algorithm," *Energy Procedia*, vol. 90, pp. 107–113, 2016.

[17] Case Western Reserve University, *Bearing Data Center[DB/OL]*, Case Western Reserve University, Cleveland, OH, USA, 2017, https://csegroups.case.edu/bearingdatacenter.