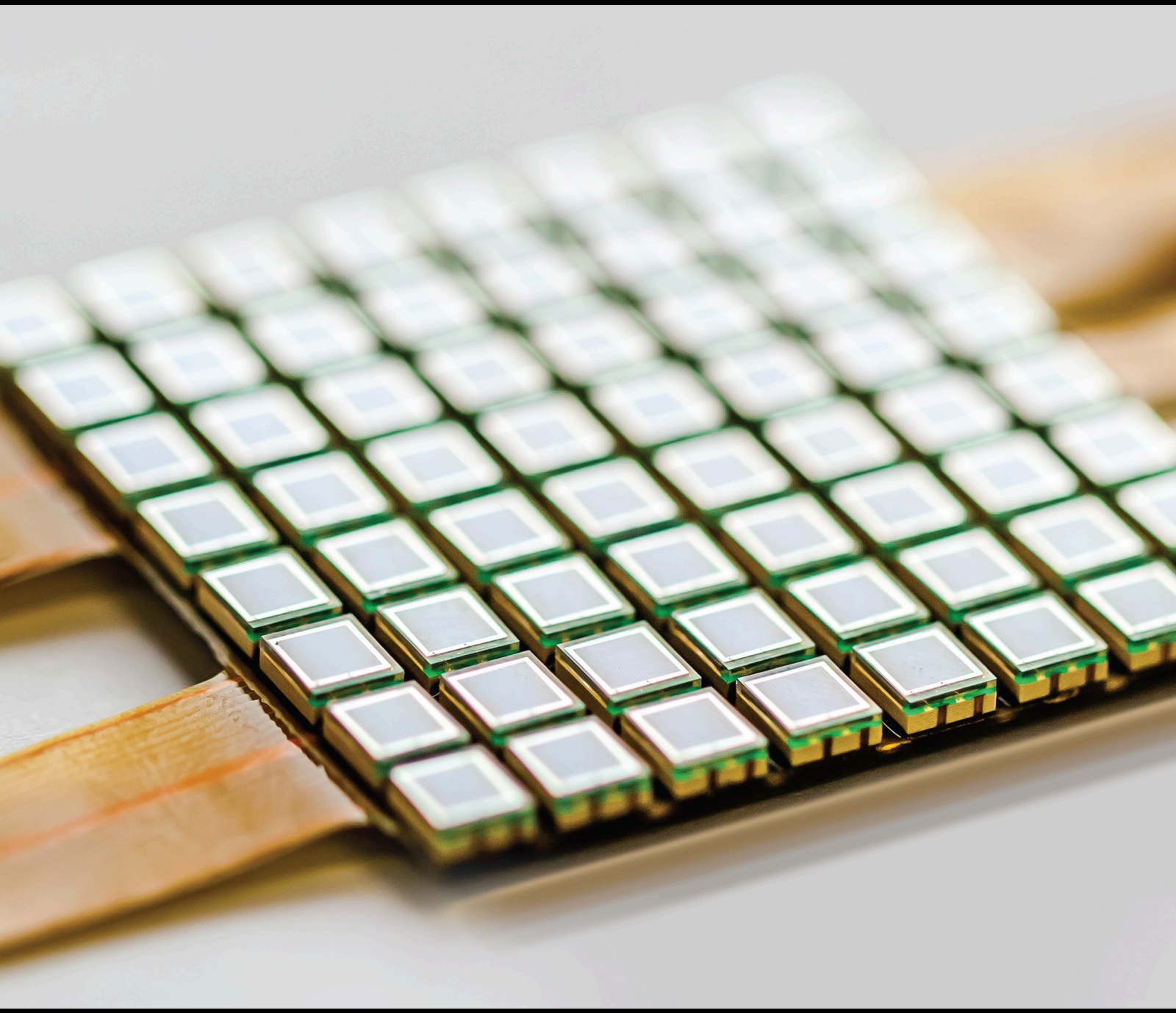
Journal of Sensors

Sensors, Signal, and Artificial Intelligent Processing

Special Issue Editor in Chief: Bin Gao
Guest Editors: Wai Lok Woo and Guiyun Tian



Sensors, Signal, and Artificial Intelligent Processing

Journal of Sensors

Sensors, Signal, and Artificial Intelligent Processing

Special Issue Editor in Chief: Bin Gao
Guest Editors: Wai Lok Woo and Guiyun Tian

Contents

Contents

Hindawi
Journal of Sensors
Volume 2022, Article ID 9793204, 5 pages
https://doi.org/10.1155/2022/9793204

Editorial

Sensors, Signal, and Artificial Intelligent Processing

Bin Gao,[1] **Wai Lok Woo**,[2] **and Guiyun Tian**[3]

[1]*School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China*
[2]*Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK*
[3]*School of Electrical and Electronic Engineering, Newcastle University, Newcastle upon Tyne NE1 8ST, UK*

Correspondence should be addressed to Bin Gao; bin_gao@uastc.edu.cn

Received 14 June 2022; Accepted 14 June 2022; Published 23 August 2022

Sensors and signal processing are fully applied in modern industry. However, with the rapid expanding of sensing applications, this requires sensors with multifunctionality, intelligence, and quick response of decision. Thus, they have substantially conducted things in both complexity and sophistication to procure a technically and economically key parts in current works. Sensors are functional infrastructure that convert physical variables into digitally meaningful signals. However, this is crucial since sensors can have extra functionalities which consist of low power, multimodality, high-speed processing, physically small, and intelligence. On the other hand, signals from multiple sensors can be integrated and combined to make decisions for solving specific problems. This consists of several examples such as building electromagnetic sensors for nondestructive testing to automatically detect cracks, corrosion, debonding and etc.; hybrid of infrared and vibration sensors for monitoring of mechanical failure is investigated. In particular, multiple functionality sensing structure is integrated in many applications such as the hot topic in unmanned vehicle. Simulation work of sensing in multiphysics is starting to be very attractive to researchers. All the above contents have confirmed that sensor intelligence, multiphysics, and multifunctionality presents a significant role in modern engineering.

Sensors, Signal, and Artificial Intelligent Processing (SSAIP) crosses the boundary of multidisciplinary research concentrating on the physics-mathematical foundation and practical applications of sensors, sensing principle signal processing, and machine learning algorithms that enable a system to intelligently learn, reason, and process. This topic bridges the gap between theory and application, generating novel sensing methodologies for both longstanding and emergent industry applications. The core of SSAIP targets the physical mechanism in sensors, signal processing, sensing machine learning/deep learning, and sensor integration. SSAIP introduces a new theoretical framework that enables interpretation of sensing using physics-based statistical signal processing. In addition, novel developments of a variety sensors that integrate the processing of signals including audio, chemical, biosignals, electromagnetic, thermography, multiphysics signals, images, multispectral, and video can be expected from this framework. More form of intelligence learning algorithms will be involved into sensor framework to augment extra solving ability. These algorithms have the capacity to generalize and discover knowledge for themselves. In particular, it is capable of learning new information whenever unseen sensor signal is captured. This *Special Issue* includes more than twenty works focused on sensor signal and information processing based on diverse technologies for different applications.

1. Wireless Sensor Network

Wireless sensor network-based research has attracted in plenty of applications. In the application of wireless power transmission, the main impact factors include inductance and quality factors. These factors will affect the transmission efficiency. Once the requirement of high-power wireless charging is acquired, ferrite bricks can be selected to increase the self-inductance of the coils as well as transmission efficiency. Y. Zhu et al. [1] have investigated several effects on wireless transmission in real applications. Both theoretical

and simulation have been conducted, and the wireless transmission system has been generated. In addition, both obtained high power and efficiency transmission for EVs. On the other side, the water pollution attracted the researchers to build sensor networks in optimizing water quality monitoring by developing new underwater sensor coverage technique. Since sensor network exists limitation within the monitoring range, 3D target coverage of heterogeneous is required to be optimized since multisensors are essential for fusion. To achieve this goal, in [2], a chaotic parallel fish swarm algorithm has been proposed. The chaotic selection with the combination of the global search capabilities has been applied. The diagnosis is a key technique to ensure the reliability for wireless sensor networks. In [3], the fault diagnosis of sensor nodes in WSN has been proposed where kernel lined learning machine is proposed. This method can be optimized by artificial bee colony algorithm in which it can solve regression issues.

In addition, intelligent transportation WSN takes important role. This method deploys remote sensing sensor nodes within high yield and low energy consumption for complex traffic parameter coordination. Z. Zha et al. [4] have proposed a modified clone genetic algorithm with adaptive ability to solve task allocation in ITWSNs. It employs operator of clonal expansion to speed up the convergence rat, while adaptive operator is updated to improve the global search capability. In particular of industrial WSN, it usually uses a huge number of sensors for monitoring. This costs redundant nodes. Y. Liu et al. [5] have proposed a quantum clone grey wolf optimization method to improve the usage of IWSNs. It combines the idea of quantum computing as well as the clone operation.

Notwithstanding the above, environmental monitoring is important, and it is used to monitor temperature and oxygen. B. Liu et al. [6] have proposed a chaotic based evolutionary algorithm for clustering of the low power in environmental monitoring. Through simulation experiments, this improved node energy usage efficiency. J. Xiao et al. [7] have designed an optimization algorithm with adaptive whale strategy which decreases the energy consumption with QoS. The simulation results suggest that the CAWOA-based routing algorithm got better performance in terms of routing energy consumption, convergence speed, and optimization ability. Another hot research field is passive target sensing in wireless case. In [8], a passive moving target localization system in single access point is proposed. The multiple antenna access point has been generated to form an antenna array where its localization can reach 1.087 m.

Thanks to wireless communication capability added to the sensors, another advanced sensing technology emerged and became popular in numerous manufactures and institutions for SHM applications, that is, RFID sensors [9]. Among the different types of RFID sensors available, passive sensors are the one that received a lot of attention in health monitoring area due to the fact that they have the potential to offer various advantages from low-cost solution and battery-less to long lifetime system perspective [9–13]. The most common RFID based sensors developed for SHM applications are given in [14]. Although wireless technologies offer many advantages, there are still some limitations that remain to be solved, such as power consumption, bandwidth constraints, transmission range, and possible security issues. Over the last two decades, numerous WSN-based SHM systems have been proposed in the literature. Current development of WSN for SHM systems proposed in the literature is summarized in [14–19].

2. Multimodality Sensing in Nondestructive Testing

Multimodality sensing techniques have been proposed in nondestructive testing. Crack can be treated as crucial for safety assessment. Epoxy resin has been used in [20], and the active sensing technique using piezoelectric ceramics is used to monitor cracks. Once wavelet method is applied, the relationship between the wavelet signal and bearing capacity after grouting is thus established. These results indicate that the sensing techniques are able to evaluate the strength. Electromagnetic sensing is an effective stress detection method according the piezoresistive effect. Y. Yu et al. [21] have proposed a nondestructive approach that applied eddy current mechanism of finite element technique. The results have illustrated the connection between the applied force and the magnetic field. In addition, numerical simulations have been undertaken to bridge the relationship between the magnetic flux density and the stress information. R. Wu et al. [22] has investigated the challenges associated with corrosion detection of metal under coating. The authors presented a detailed investigation of various techniques based on ultrasonic, acoustic, electromagnetic, radiographic, and thermographic. In [23], Y. He et al. investigated the angular MBN affected by the residual stress. The residual stress was closely correlated to the magneto elastic anisotropy energy. J. Capó-Sánchez et al. have used the angular distribution of MBN energy to predict the magnetic easy axis, which has successfully indicated that the applied uniaxial stress gave origin to a continuous rotation of the magnetic easy axis. M. Neslušan et al. [24] have found a remarkable decrease of Barkhausen noise near the true yield stress which can be used to alert the high risk of incoming breakage. On a separate hand, X. Kleber and A. Vincent [25] have investigated the dependence of Barkhausen noise on elastic and plastic deformations in Armco iron and a low carbon steel to explain the effect of residual internal stresses through magnetoelastic coupling and dislocation-domain wall interaction.

In [26], principal component analysis (PCA) and Tucker decomposition are developed and compared to assess the performance of microcrack detection. Here, specimens with different fatigue microcrack are detected by using the eddy current pulsed thermography (ECPT). In addition, the potential correspondence between crack closure and temperature change has been established. In high-speed train safety inspection, [27] develops a vision technique based on two convolutional neural networks to detect defects. The authors have presented networks which are capable of inherently detecting differences between two images and thus further

identifying the changes by using a pair of images. Notwithstanding the above, M. He et al. [28] have found a method for suppressing the effect of uneven surface emissivity of material in the moving mode of eddy current thermography. Y. Gao et al. [29] have reported a ferrite yoke based on ECPT to enhance the detectability of multiple cracks 39. K. Li et al. [30] have illustrated a Helmholtz-coil-based ECPT configuration for the state detection and characterization of bond wire lift-off in IGBT modules. Z. Liu et al. [31] have proposed an L-shaped sensor to diagnose natural cracks in a static system. In [32], M. Goldammer et al. show how NDT can be automated using as an example of industrial applications at the Siemens sector energy.

3. Image and Video Processing

While reservoir fractures are essential locations to gather oil and gas, imaging logging technology has become a mainstream method for obtaining stratigraphic information. In [33], W. Zhang et al. have proposed an optimal path search strategy to effectively identify and extract the fracture information in well logging images. The logging image is first transformed into the optimal path search, and this is followed by the identification of reservoir fractures. Video surveillance systems are often deployed at places such as airports and train stations. However, these systems are prone to interference with cluttered backgrounds. C.-H. Tseng et al. [34] have proposed a person retrieval method to extract the attributes of a masked image using an instance segmentation module. The reported experimental results shows that the retrieval system can achieve effective retrieval performance for multi-camera surveillance systems. In similar line, motion capture technology plays an important role in the production field of film and television. In [35], kinematic constraints (KC) and cyclic consistency (CC) network are employed to study the methods of kinematic style migration. Cycle-Consistent Adversarial Network (CCycleGAN) is developed in [35], and the motion style migration network of convolutional self-encoder has been used as a generator to establish the cyclic consistent constraint. In order to normalize the movement generation to solve the problems such as jitter and sliding step, the kinematic constraints are concurrently used.

In application field of power system, traditional manual inspection methods of a power transmission line (PTL) suffer from the issue of supplying the demand for high quality and dependability for power grid maintenance. The authors in [36] have presented a review of technologies for three-dimensional (3D) reconstruction, object detection, and visual servo of PTL inspection.

In weak-light environments, images suffer from low contrast. In [37], a simple and novel correction method has been proposed based on adaptive local gamma transformation and color compensation. The proposed method converts the source image into YUV color space, and the Y component is estimated using a fast guided filter.

4. Wearable Sensing

The growing application of body area networks (BANs) in different fields makes the low energy clustering a paramount issue. A clustering optimization algorithm in BANs is a fundamental scheme to guarantee that the essential collected data can be forwarded in a reliable path and improve the lifetime of BANs. However, the classical clustering method leads to high cost when constraints such as large overall energy consumption are undertaken. Hence, a binary immune hybrid artificial bee colony algorithm (BIHABCA), a randomized swarm intelligent scheme, is applied in BANs [38]. Furthermore, it designs the formulation that considers both distances between two nodes and the length of bits. The results show that the energy cost of the network optimized by the proposed BIHABCA method decreased the energy cost of transmitting and receiving data in BANs.

In recent years, with the development of wearable sensor devices, research on sports monitoring using inertial measurement units has received increasing attention. J. Fan et al. [39] have designed a sensor fusion basketball shooting posture recognition system based on convolutional neural networks. It has collected 12,177 sensor fusion basketball shooting posture data entries of 13 Chinese adult male subjects aged 18–40 years and with at least 2 years of basketball experience without professional training. The intertest achieved an average recall rate of 89.8%, an average precision rate of 91.1%, and an accuracy rate of 89.9%. In [40], the wearable sensors consist of a neoprene band that contains circuitry for measuring electrodermal activity (EDA), 3-axis motion, temperature, and electrocardiogram (ECG). In [41], for the assessment of emotions, anxiety, mood, depression, and stress, a head-mounted type was proposed by using electroencephalogram (EEG), chest band heart rate variability (HRV), and skin conductance (SC) sensors. In [42], the eyeball is tracked by a wearable device, and the mental state of the person is detected by analyzing the activity of the eyeball. In order to objectively evaluate the human mental health, integrated sensors of mobile phone are applied [43–51, 56].

5. Different Sensing Techniques and Instruments

The ring laser gyro inertial measurement unit has many systematic error terms which seriously affects the stability time and accuracy. A system-level temperature modeling and compensation method is proposed based on the relevance vector regression method [52]. H. Xu et al. [53] have proposed a new indoor people detection and tracking system using a millimeter-wave (mmWave) radar sensor. The recursive Kalman filter tracking algorithm is used to track multiple people simultaneously. The method is lightweight for scalability and portability. For most high-precision power analyzers, calibration before measurement is important to ensure accuracy. W. Zhou et al. [54] have proposed a nonlinear calibration method based on sinusoidal excitation and DFT transformation. In particular, through Fourier transform, the phase value at the initial moment of the fundamental frequency is calculated.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

The guest editors of this *Special Issue* thank the staff of Sensors for the trust shown and the good work done.

Bin Gao
Wai Lok Woo
Guiyun Tian

References

[1] Y. Zhu, Z. Wang, X. Cao, and L. Wu, "Design of high-power high-efficiency wireless charging coils for EVs with MnZn ferrite bricks," *Journal of Sensors*, vol. 2021, Article ID 9931144, 18 pages, 2021.

[2] L. Cao, Y. Yue, and Y. Zhang, "A novel fault diagnosis strategy for heterogeneous wireless sensor networks," *Journal of Sensors*, vol. 2021, Article ID 6650256, 18 pages, 2021.

[3] Z. Jie, G. Qi, and C. Liu, "A chaotic parallel artificial Fish swarm algorithm for water quality monitoring sensor networks 3D coverage optimization," *Journal of Sensors*, vol. 2021, Article ID 5529527, 12 pages, 2021.

[4] Z. Zha, C. Li, J. Xiao et al., "An improved adaptive clone genetic algorithm for task allocation optimization in ITWSNs," *Journal of Sensors*, vol. 2021, Article ID 5582646, 12 pages, 2021.

[5] L. Yang, J. Xiao, C. Li, H. Qin, and J. Zhou, "Sensor duty cycle for prolonging network lifetime using quantum clone grey wolf optimization algorithm in industrial wireless sensor networks," *Journal of Sensors*, vol. 2021, Article ID 5511745, 13 pages, 2021.

[6] L. Bao, R. Yang, M. Xu, and Z. Jie, "A chaotic elite niche evolutionary algorithm for low-power clustering in environment monitoring wireless sensor networks," *Journal of Sensors*, vol. 2021, Article ID 5558643, 12 pages, 2021.

[7] J. Xiao, L. Yang, C. Li, and J. Zhou, "A novel QoS routing energy consumption optimization method based on clone adaptive whale optimization algorithm in IWSNs," *Journal of Sensors*, vol. 2021, Article ID 5579252, 14 pages, 2021.

[8] X. Yang, J. Wang, W. Nie, Y. Wang, W. Nie, and Y. Wang, "Passive localization of moving target with channel state information," *Journal of Sensors*, vol. 2021, Article ID 6140914, 9 pages, 2021.

[9] S. D. Glaser, L. Hui, M. L. Wang, O. Jinping, and J. Lynch, "Sensor technology innovation for the advancement of structural health monitoring: a strategic program of US-China research for the next decade," *Journal of Smart Structures and Systems*, vol. 3, no. 2, pp. 221–244, 2007.

[10] Y. Lee, B. Phares, M. V. Jayselan, and S. A. Osman, "Recent trends in bridge health monitoring," *International Journal of Civil and Structural Engineering Research*, vol. 4, no. 1, pp. 347–356, 2016.

[11] H. Chung, T. Enomoto, M. Shinozuka et al., "Real time visualization of structural response with wireless MEMS sensors," in *13th World Conference on Earthquake Engineering*, Vancouver, B.C, Canada, August 2004no. 121.

[12] K. J. Loh, J. P. Lynch, and N. A. Kotov, "Passive wireless strain and pH sensing using carbon nanotube-gold nanocomposite thin films," *Proceedings of SPIE, Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems*, vol. 6529, pp. 1–12, 2007.

[13] G. Imam and Y. Tian, "Enhanced sensitivity of low frequency (LF) RFID sensor signal for structural health monitoring (SHM) in high temperature environment," in *9th World Conference on Non-Destructive Testing*, Munich, Germany, 2016.

[14] J. P. Lynch and K. J. Loh, "A summary review of wireless sensors and sensor networks for structural health monitoring," *The Shock and Vibration Digest*, vol. 38, no. 2, pp. 91–130, 2006.

[15] S. Cho, C. Yun, J. P. Lynch, and T. Nagayama, "Smart wireless sensor technology for structural health monitoring of civil structures," *International Journal of Steel Structures*, vol. 8, no. 4, pp. 267–275, 2004.

[16] Y. Wang, J. P. Lynch, and K. H. Law, "Validation of an integrated network system for real-time wireless monitoring of civil structures," in *Proceedings of the 5th International Workshop on Structural Health Monitoring*, pp. 12–14, Stanford, CA, September 2005.

[17] S. Kim, S. Pakzad, D. Culler et al., "Health monitoring of civil infrastructures using wireless sensor networks," in *Proceedings of the 6th International Symposium on Information Processing in Sensor Networks*, pp. 254–263, New York, 2007.

[18] P. Wang, Y. Yan, G. Y. Tian, O. Bouzid, and Z. Ding, "Investigation of wireless sensor networks for structural health monitoring," *Hindawi Journal of Sensors*, vol. 2012, article 156329, 7 pages, 2012.

[19] Y. Lim and J. Park, "Networking strategies for structural health monitoring in wireless sensor networks," *International Journal of Energy, Information and Communications*, vol. 6, no. 3, pp. 11–18, 2015.

[20] H. Meng, W. Yang, and X. Yang, "Real-time monitoring of timber-surface crack repair using piezoelectric ceramics," *Journal of Sensors*, vol. 2021, Article ID 8201780, 15 pages, 2021.

[21] Y. Yu, F. Yuan, H. Li, C. Ulianov, and G. Tian, "Prediction of inhomogeneous stress in metal structures - a hybrid approach combining eddy current technique and finite element method," *Journal of Sensors*, vol. 2021, Article ID 6647093, 9 pages, 2021.

[22] R. Wu, H. Zhang, R. Yang, W. Chen, and G. Chen, "Non-destructive testing for corrosion evaluation of metal under coating," *Journal of Sensors*, vol. 2021, Article ID 6640406, 16 pages, 2021.

[23] Y. He, M. Mehdi, H. Liu, E. J. Hilinski, and A. Edrisy, "Angular magnetic Barkhausen noise of incline- and cross-rolled non-oriented electrical steel sheets," *Materials Characterization*, vol. 177, p. 111200, 2021.

[24] M. Neslušan, M. Jurkovič, T. Kalina, M. Pitoňák, and K. Zgútová, "Monitoring of S235 steel over-stressing by the use of Barkhausen noise technique," *Engineering Failure Analysis*, vol. 117, p. 104843, 2020.

[25] X. Kleber and A. Vincent, "On the role of residual internal stresses and dislocations on Barkhausen noise in plastically deformed steel," *Ndt & E International*, vol. 37, no. 6, pp. 439–445, 2004.

[26] X. Zhang, J. Peng, D. Luquan et al., "Detection of fatigue microcrack using eddy current pulsed thermography," *Journal of Sensors*, vol. 2021, Article ID 6647939, 8 pages, 2021.

[27] Z. Wang, J. Peng, W. Song et al., "A convolutional neural network based classification and decision-making model for visible defects identification of high-speed train images," *Journal of Sensors*, vol. 2021, 17 pages, 2021.

[28] M. He, L. Zhang, J. Li, and W. Zheng, "Methods for suppression of the effect of uneven surface emissivity of material in the moving mode of eddy current thermography," *Applied Thermal Engineering*, vol. 118, pp. 612–620, 2017.

[29] Y. Gao, G. Y. Tian, K. Li, J. Ji, P. Wang, and H. Wang, "Multiple cracks detection and visualization using magnetic flux leakage and eddy current pulsed thermography," *Sensors & Actuators A Physical*, vol. 234, pp. 269–281, 2015.

[30] K. Li, G. Y. Tian, L. Cheng, A. Yin, W. Cao, and S. Crichton, "State detection of bond wires in IGBT modules using Eddy current pulsed thermography," *IEEE Transactions on Power Electronics*, vol. 29, no. 9, pp. 5000–5009, 2014.

[31] Z. Liu, B. Gao, and G. Y. Tian, "Natural crack diagnosis system based on novel L-shaped electromagnetic sensing thermography," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 11, pp. 9703–9714, 2020.

[32] M. Goldammer, H. Mooshofer, M. Rothenfusser, J. Bass, and J. Vrana, "Automated induction thermography of generator components," *AIP Conference Proceedings*, vol. 1211, no. 1, pp. 451–457, 2010.

[33] W. Zhang, T. Wu, Z. Li, Y. Li, A. Qiu, and Y. Shi, "Automatic detection of fractures based on optimal path search in well logging images," *Journal of Sensors*, vol. 2021, Article ID 5577084, 10 pages, 2021.

[34] C.-H. Tseng, C.-C. Hsieh, D.-J. Jwo, J.-H. Wu, R.-K. Sheu, and L.-C. Chen, "Person retrieval in video surveillance using deep learning–based instance segmentation," *Journal of Sensors*, vol. 2021, 12 pages, 2021.

[35] H. Wang, D. Dandan, J. Li, W. Ji, and L. Yu, "A cyclic consistency motion style transfer method combined with kinematic constraints," *Journal of Sensors*, vol. 2021, Article ID 5548614, 17 pages, 2021.

[36] M. Chen, Y. Tian, S. Xing et al., "Environment perception technologies for power transmission line inspection robots," *Journal of Sensors*, vol. 2021, Article ID 5559231, 16 pages, 2021.

[37] W. Wang, X. Yuan, Z. Chen, X. J. Wu, and Z. Gao, "Weak-light image enhancement method based on adaptive local gamma transform and color compensation," *Journal of Sensors*, vol. 2021, Article ID 5563698, 18 pages, 2021.

[38] M. Xu and J. Zhou, "A biologically inspired algorithm for low energy clustering problem in body area network," *Journal of Sensors*, vol. 2021, 12 pages, 2021.

[39] J. Fan, S. Bi, G. Wang, L. Zhang, and S. Sun, "Sensor fusion basketball shooting posture recognition system based on CNN," *Journal of Sensors*, vol. 2021, Article ID 6664776, 16 pages, 2021.

[40] R. R. Fletcher, S. Tam, O. Omojola, R. Redemske, and J. Kwan, "Wearable sensor platform and mobile application for use in cognitive behavioral therapy for drug addiction and PTSD," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1802–1805, Boston, MA, USA, 2011.

[41] J. B. Wang, L. A. Cadmus-Bertram, L. Natarajan et al., "Wearable sensor/device (Fitbit One) and SMS text-messaging prompts to increase physical activity in overweight and obese adults: a randomized controlled trial," *Telemedicine and e-Health*, vol. 21, no. 10, pp. 782–792, 2015.

[42] M. Vidal, J. Turner, A. Bulling, and H. Gellersen, "Wearable eye tracking for mental health monitoring," *Communications*, vol. 35, no. 11, pp. 1306–1311, 2012.

[43] H. Lu, D. Frauendorfer, M. Rabbi et al., "StressSense: detecting stress in unconstrained acoustic environments using smartphones," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, pp. 351–360, New York, 2012.

[44] J. A. Naslund, K. A. Aschbrenner, and S. J. Bartels, "Wearable devices and smartphones for activity tracking among people with serious mental illness," *Mental Health & Physical Activity*, vol. 10, pp. 10–17, 2016.

[45] F. H. Wilhelm and Grossman, "Emotions beyond the laboratory: theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment," *Biological Psychology*, vol. 84, no. 3, pp. 552–569, 2010.

[46] M. Schmid Mast, D. Gatica-Perez, D. Frauendorfer, L. Nguyen, and T. Choudhury, "Social sensing for psychology: automated interpersonal behavior assessment," *Current Directions in Psychological Science*, vol. 24, no. 2, pp. 154–160, 2015.

[47] P. Kakria, N. K. Tripathi, and P. Kitipawang, "A real-time health monitoring system for remote cardiac patients using smartphone and wearable sensors," *International Journal of Telemedicine and Applications*, vol. 2015, 2015.

[48] A. Gaggioli and G. Riva, "From mobile mental health to mobile wellbeing: opportunities and challenges," *Studies in Health Technology & Informatics*, vol. 184, pp. 141–147, 2013.

[49] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, "Activity-aware mental stress detection using physiological sensors," *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol. 23, no. 13, pp. 211–230, 2010.

[50] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. Al'absi, and S. Shah, "AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field," in *Proceedings of the 9th ACM conference on embedded networked sensor systems*, pp. 274–287, New York, 2011.

[51] V. W. Tseng, M. Merrill, F. Wittleder, S. Abdullah, M. H. Aung, and T. Choudhury, "Assessing mental health issues on college campuses: preliminary findings from a pilot study," in *ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct ACM*, New York, 2016.

[52] H. Liang, Y. Tao, M. Wang, Y. Guo, and X. Zhao, "System-level temperature compensation method for the RLG-IMU based on HHO-RVR," *Journal of Sensors*, vol. 2021, Article ID 6613574, 16 pages, 2021.

[53] H. Xu, H. Cheena, A. Thomas, and J. K. P. Tsoi, "Indoors detection and tracking of people using mmWave sensor," *Journal of Sensors*, vol. 2021, Article ID 6657709, 14 pages, 2021.

[54] W. Zhou, S. Yang, L. Wang, H. Sheng, and Y. Deng, "A nonlinear calibration method based on sinusoidal excitation and DFT transformation for high-precision power analyzers," *Journal of Sensors*, vol. 2021, 9 pages, 2021.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5578361, 9 pages
https://doi.org/10.1155/2021/5578361

Research Article

A Nonlinear Calibration Method Based on Sinusoidal Excitation and DFT Transformation for High-Precision Power Analyzers

Wenjian Zhou, Sheng Yang, Li Wang, Hanmin Sheng, and Yang Deng

School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China

Correspondence should be addressed to Li Wang; colorsky@uestc.edu.cn

Received 4 February 2021; Revised 8 May 2021; Accepted 11 September 2021; Published 4 October 2021

Academic Editor: Antonio Lazaro

For most high-precision power analyzers, the measurement accuracy may be affected due to the nonlinear relationship between the input and output signal. Therefore, calibration before measurement is important to ensure accuracy. However, the traditional calibration methods usually have complicated structures, cumbersome calibration process, and difficult selection of calibration points, which is not suitable for situations with many measurement points. To solve these issues, a nonlinear calibration method based on sinusoidal excitation and DFT transformation is proposed in this paper. By obtaining the effective value data of the current sinusoidal excitation from the calibration source, the accurate calibration process can be done, and the calibration efficiency can be improved effectively. Firstly, through Fourier transform, the phase value at the initial moment of the fundamental frequency is calculated. Then, the mapping relationship between the sampling value and the theoretical calculation value is established according to the obtained theoretical discrete expression, and a cubic spline interpolation method is used to further reduce the calibration error. Simulations and experiments show that the calibration method presented in this paper achieves high calibration accuracy, and the results are compensation value after calibration with a deviation of $\pm 3 \times 10^{-4}$.

1. Introduction

The calibration of a high-precision power analyzer is a key function in the signal measurement process. The calibration accuracy directly affects the accuracy and reliability of the subsequent measurement of voltage, current, power, harmonics, and other parameters [1]. There is a strict linear relationship between the input and output of ideal instrument, and no time lag or distortion exists. However, in the practical engineering scenario of power measurement, the relationship between input and output is always nonlinear due to the inherent and unchangeable characteristics of analog channel and sensor probe. In order to compensate or eliminate the nonlinearity of the instrument, the entire system needs to be nonlinearly calibrated; thus, the correctness of subsequent parameter calculations can be ensured [2–4]. The key to calibration is how to establish a mapping relationship between sampling theoretical values and actual values for each sampling point. This mapping relationship is essentially a mathematical relationship expression which needs to be designed and adjusted according to actual situation [5].

Among the common calibration methods, the hardware compensation method is usually a method that uses both digital and analog circuits for compensation. Li [6] proposed an accurate online calibration system for current transformers, and the accuracy can reach 0.05 level. Luo [7] designed an improved calibration system based on direct current (DC) negative feedback for the calibration of current transformers. The calibration uncertainty of this system reaches 0.038% within the measuring range. The hardware compensation method circuit [8, 9] is usually complicated, and the circuit design process costs much. Besides, the calibration range is small, and the accuracy cannot be guaranteed within the entire range. In addition, the design of the hardware circuit and the zero drift of the electronic device will also reduce the accuracy of the calibration [10, 11]. Therefore, this method is generally not used in scenes that require precise calibration.

The principle of the mathematical model calibration algorithm is utilizing a limited number of sample information to establish a mathematical model of the measurement signal according to the principle of minimum error. In article [12], Jin put forward a calibration method based on OC-SVM. This method can detect the change points in the time series and obtain better accuracy using less training data. However, it is difficult to establish a corresponding mathematical model for nonlinear systems. Wang, Kong [13], and others classified the errors of the acoustic vector sensor array and designed an optimization model and error self-calibration algorithm for the acoustic vector sensor array. This algorithm can perform quite well in parameter estimation, but when the mathematical model is established, the iterative calculation of coefficients still needs much work. Therefore, this method is generally not used in actual projects which require a large number of data calculation [14].

The nonlinear segmented calibration method divides the uncalibrated data into segments and then linearizes these sections. Wang, Peng [15], and Chengxian [16] both chose this method to do calibration work, because it can achieve high accuracy though the accuracy often depends on the experience of the calibrator. Moreover, if the calibration results fail to meet the standard, it is necessary to perform the segmented calibration again. In most cases, the calibration efficiency is not high enough; thus, the calibration workload is relatively large. To solve this problem, a calibration algorithm based on discrete Fourier transform (DFT) is provided in paper [17, 18]. This method directly carries on the Fourier transform processing to the sampled data sequence, which has the advantages of fast operation speed and less computation. However, the algorithm is mainly suitable for harmonic measurement, and due to the frequency spectrum leakage under the frequency shift condition, the algorithm error is large.

In this paper, a nonlinear calibration method based on sinusoidal excitation and DFT transform is presented. This method uses the initial phase calculated by DFT to establish the relationship of the original dense set and then establishes the mapping relationship between the actual sampling value and the theoretical calibration value in the selected calibration interval. After that, by interpolating the data, an ideal calibration curve is obtained.

2. Fundamental Knowledge of the Proposed Method

2.1. Algorithm Analysis. To implement this algorithm, it is necessary to determine the mapping relationship between the original sampled value and the theoretical value. For the purpose of determining the mapping relationship, the signal is sampled with a fixed sampling rate f_s. Then, Fourier transform is performed on the obtained discrete sequence of sampling points to calculate the phase φ_0 at the initial moment of the fundamental frequency. Next, the initial phase can be used to calculate the theoretical discrete expression of the original signal. The theoretical value corresponding to the sampled value is calculated through the theoretical discrete expression, and the mapping relationship between the original sampled values and the theoretical values is established. Then, determine the minimum calibration interval. In order to make the calibration interval include the maximum range of the signal amplitude, the interval can be calibrated from the trough of the signal to the peak of the wave, which is half a period. According to the initial phase φ_0 and the set standard source effective value A_m, the mapping relationship between the theoretical value and the actual sampling value in the calibration interval is established, and the denoising process is performed. Finally, a smooth calibration curve is obtained by spline interpolation for the calibration points mapped in the two-dimensional coordinate system, and we can obtain the theoretical valued of other sampling points from the calibration curve. The overall flow of the algorithm is shown in Figure 1.

Take a power signal as an example for specific description. Performing fixed frequency sampling on the measured original signal, the sampling rate f_s is 25600 Hz, the sampling period N is 10, and the number of sampling points in each period is 512; then, the fundamental frequency of the original signal is $f_b = 1/T_s \times M = f_s/M$, and it is 50 Hz.

2.2. Solve the Initial Phase by DFT. Collect N cycles of data on the original voltage or current signal and do DFT on the collected discrete time series $\{x_k\}(k = 0, 1, \cdots, N \times M - 1)$, $M = 512$, as shown in formula (1).

$$X(f_b) = \sum_{k=0}^{N\times M-1} x_k \times e^{-j\frac{2\pi f_b}{N\times M}k}. \tag{1}$$

The calculation result $X(f_b)$ is a complex number, which can be expressed as $X(f_b) = X_R(f_b) + X_I(f_b) \times i$. The expressions for the real and imaginary parts are

$$\begin{aligned} X_R(f_b) &= \sum_{k=0}^{N\times M-1} x_k \cos\left[2\pi k f_b/(N\times M)\right], \\ X_I(f_b) &= -\sum_{k=0}^{N\times M-1} x_k \sin\left[2\pi k f_b/(N\times M)\right]. \end{aligned} \tag{2}$$

According to the real and imaginary parts of the complex number, the initial phase $\varphi_0 = \arctan\left(X_I(f_b)/X_R(f_b)\right)$ of the current signal with the fundamental frequency f_b can be calculated, and the continuous expression of the original signal is

$$y = A_m \times \cos\left(wt + \varphi_0\right). \tag{3}$$

Since the sampling points are discrete, continuous expressions need to be converted into discrete expressions. The discrete sequence $\{x_k\}$ is sampled and extracted at time t; so, the relational expression between time t and subscript k is

$$t = f_b \times T_s \times k. \tag{4}$$

According to formula (4), convert the continuous

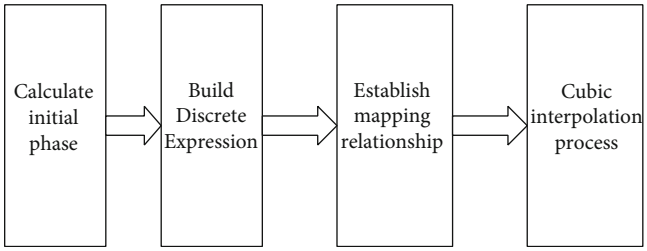


Figure 1: Schematic diagram of calibration algorithm.

expression of the signal into a theoretical discrete expression:

$$y_k = A_m \times \cos(2\pi f_b \times T_s \times k + \varphi_0), \quad (5)$$

where A_m is the amplitude of the waveform output by the calibration source. Since there is a mapping relationship between y_k and k and x_k and k in the theoretical discrete expression, there is also a one-to-one correspondence between y_k and x_k. The $\{x_k, y_k\}$ mapping relationship of $N \times M$ sampling points can be obtained in N sampling periods, and the original dense set is established.

2.3. Calibration of Calibration Curve. In order to ensure that the maximum range of the measured signal can be covered, the calibration interval is determined to be $[A_m \times \cos(\pi), A_m \times \cos(2\pi)]$ according to the theoretical discrete expression of the original signal, which is the maximum range. Next, take the left end point $A_m \times \cos(\pi)$ of the calibration interval as the starting point and calculate the subscript k_0 corresponding to the sampling point. The calculation formula of k_0 is as follows:

$$k_0 = (\pi - \varphi_0) \times M/(2\pi). \quad (6)$$

The k_0 calculated by the above formula is not necessarily a positive integer. If k_0 is a positive integer, then the sampled value x_{k_0} of the sampling point is recorded as x_0', and the theoretical value y_{k_0} is calculated by the theoretical discrete expression $y_{k_0} = A_m \times \cos(2\pi f_b \times T_s \times k_0 + \varphi_0)$ and recorded as y_0'. If k_0 is not an integer, the sampling value x_{k_0} of the nearest sampling point from the starting point is recorded as x_0', and the theoretical value y_{k_0} is recorded as y_0'.

According to formula (6) and the relationship between signal fundamental frequency f_b and fixed frequency sampling frequency f_s, $f_b = f_s/M$, as the frequency f_b increases, the sampling point M decreases; so, the subscript of the sampling point k_0 also decreases. According to formula (5), when the amplitude A_m and f_b change, the corresponding theoretical value of calibration y_k will change, which means that the calibration coefficient will change. The initial phase φ_0 is obtained by DFT calculation, which will not affect the calibration process.

Taking $A_m \times \cos(\pi)$ as the starting point, calculate the subscript k of subsequent sampling points in the calibration interval. Because x_0' corresponds to the point x_{k_0} with the subscript k_0 in the original sequence $\{x_k\}$, the original sequence $\{x_k\}$ starts from the subscript k_0 and takes a sampling point every ΔM points as the new sequence $\{x'\}$. The points in are marked as $x_1', x_2' \cdots x_k' \cdots x_{M/\Delta M}'$, and then the original sequence $\{x_i\}$ and the new sequence $\{x_i'\}$ have the following mapping relationship:

$$x_k' = x_{\Delta M \times k + k_0}. \quad (7)$$

Substituting the expression of x_k' in formula (7) into the theoretical discrete expression, the corresponding y_k' is

$$y_k' = A_m \times \cos[2\pi f_b \times T_s \times (\Delta M \times k + k_0) + \varphi_0]. \quad (8)$$

$M/\Delta M$ calibration points can be taken in each cycle, and $N \times M/\Delta M$ sampling points are taken as calibration points in a total of N sampling cycles, and a mapping relationship of $\{x_k', y_k'\}$ is established.

In N sampling cycles, the sampling points in each cycle are repeated periodically and the sampling points in each half cycle in a cycle are mirror symmetrical. Therefore, it is necessary to average the repeated $2N$ sampling points. The process of averaging can be regarded as the process of smoothing and removing noise. The calculation formula is as follows:

$$x_k'' = \frac{x_k' + x_{M/\Delta M - k - 1}' + x_{k+(M/\Delta M)}' \cdots + x_{k+N\times(M/\Delta M)}' + x_{N\times(M/\Delta M)-k-1}'}{2N},$$
$$y_k'' = \frac{y_k' + y_{M/\Delta M - k - 1}' + y_{k+(M/\Delta M)}' \cdots + y_{k+N\times(M/\Delta M)}' + y_{N\times(M/\Delta M)-k-1}'}{2N}. \quad (9)$$

After averaging, the mapping relationship of $\{x_k'', y_k''\}$ with subscript k from 0 to $M/2\Delta M - 1$ totaling $M/2\Delta M$ sampling points is obtained. The obtained $M/2\Delta M$ sampling points are the required calibration points.

There are many ways to establish the calibration relationship for $\{x_k'', y_k''\}$ of $M/2\Delta M$ sampling points. The available methods include straight line fitting, polynomial fitting, and interpolation. Straight line fitting can only guarantee the continuity in the interval, but cannot guarantee the smoothness in the calibration curve. In the polynomial curve fitting, if there is a large deviation of some data points, the fitting accuracy will decrease as the order increases. So, the spline interpolation method is used in this article.

2.4. Cubic Spline Interpolation. The spline interpolation method is a method that draws a curve of all points in the form of variable splines [19–21]. Every two adjacent points can determine the polynomial of each segment; so, the spline interpolation is composed of a series of polynomials. Cubic spline interpolation is a widely used spline interpolation method, and each segment is a cubic polynomial. This method has several advantages, the piece-wise low-order interpolation polynomials are easier to solve and that can improve the smoothness of the interpolation function as good as high-order spline interpolation. Meanwhile, the compensation effect at adjacent frequency points is better than straight line fitting.

The calculation method of cubic spline interpolation used in this paper is explained below. The mapping between

the original value and the spline value is as follows:

$$\begin{matrix} x : a = x_0 < x_1 < \cdots < x_n = b \\ y : y_0 y_1 \cdots y_n \end{matrix}. \tag{10}$$

The cubic spline function $S(x)$ is a piece-wise cubic equation, with intervals and $n+1$ data points. The cubic equation for each interval obeys the following conditions:

(1) In each interval$[x_i, x_{i+1}]$, $S(x) = S_i(x)$ is a cubic polynomial

$$S_i(x_i) = y_i (i = 0, 1, \cdots, n). \tag{11}$$

(2) The first derivative $S'(x)$ and the second derivative $S''(x)$ of the cubic spline function $S(x)$ are continuous in $[a, b]$, and $S(x)$ is smooth and continuous

Therefore, the cubic polynomial created for each interval can be written as

$$\begin{matrix} S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \\ i = 0, 1, \cdots, n-1. \end{matrix} \tag{12}$$

The derivation process of calculating these unknown coefficients a_i, b_i, c_i, and d_i is as follows:

(1) Calculate the step length of each segment $h_i = x_{i+1} - x_i (i = 0, 1, \cdots, n-1)$

(2) The formula $a_i = y_i$ can be derived from the formula $S_i(x_i) = y_i$

(3) Derived from the formula $S_i(x_{i+1}) = y_{i+1} (i = 0, 1, \cdots, n-1)$:

$$a_i + h_i b_i + h_i^2 c_i + h_i^3 d_i = y_{i+1}. \tag{13}$$

(4) According to the differential continuity of the spline $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}) (i = 0, 1, \cdots, n-2)$,

$$b_i + 2c_i h_i + 3d_i h_i^2 - b_{i+1} = 0. \tag{14}$$

Similarly, according to $S_i''(x_{i+1}) = S_{i+1}''(x_{i+1}) (i = 0, 1, \cdots, n-2)$,

$$2c_i + 6d_i h_i - 2c_{i+1} = 0. \tag{15}$$

(5) $m_i = S_i''(x_i) = 2c_i$, b_i, c_i, and d_i can be expressed by m_i, then using the Cubic spline interpolation method to get the following results:

$$\begin{aligned} a_i &= y_i, \\ b_i &= \frac{y_{i+1} - y_i}{h_i} - \frac{h_i}{2} m_i - \frac{h_i}{6}(m_{i+1} - m_i), \\ c_i &= \frac{m_i}{2}, \\ d_i &= \frac{m_{i+1} - m_i}{6h_i}. \end{aligned} \tag{16}$$

(6) Substitute b_i, c_i, and d_i into formula (14):

$$h_i m_i + 2(h_i + h_{i+1}) m_{i+1} + h_{i+1} m_{i+2} = 6\left[\frac{y_{i+2} - y_{i+1}}{h_{i+1}} - \frac{y_{i+1} - y_i}{h_i}\right]. \tag{17}$$

When there are $n-1$ equations and $n+1$ unknown m values to be solved, two additional formulas are needed to solve this equation. Therefore, the boundary conditions are used to limit the differential values of the two endpoints x_0 and x_n [22], that is, the second-order differential $S'' = 0$, which is expressed as $m_0 = 0$ and $m_n = 0$. The equation to be solved can be expressed as

$$\begin{bmatrix} 1 & 0 & 0 & & \cdots & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & 0 & & \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & 0 & \\ 0 & 0 & h_2 & 2(h_2 + h_3) & h_3 & \vdots \\ \vdots & & & \ddots & \ddots & \ddots \\ & & 0 & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & \cdots & & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} m_0 \\ m_1 \\ m_1 \\ m_3 \\ \vdots \\ m_3 \end{bmatrix} = 6 \begin{bmatrix} 0 \\ \frac{y_2 - y_1}{h_1} - \frac{y_1 - y_0}{h_0} \\ \frac{y_3 - y_2}{h_2} - \frac{y_2 - y_1}{h_1} \\ \frac{y_4 - y_3}{h_3} - \frac{y_3 - y_2}{h_2} \\ \vdots \\ \frac{y_n - y_{n-1}}{h_{n-1}} - \frac{y_{n-1} - y_{n-2}}{h_{n-2}} \\ 0 \end{bmatrix}. \tag{18}$$

So, the following step is to solve the equation to get m_i and then calculate the values of all unknown parameters b_i, c_i, and d_i using m_i, and the expression of the spline curve $S_i(x)$ can be finally obtained.

According to the interpolation method, a spline curve is drawn for the mapping relationship of $\{x_k'', y_k''\}$ of $M/2\Delta M$ points after averaging. The sequence $\{x_i''\}$ contains the maximum range of the sampled value, which means the abscissa

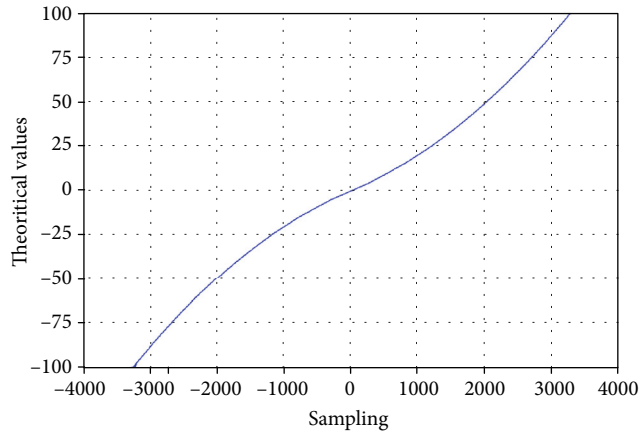


Figure 2: Calibration curve.

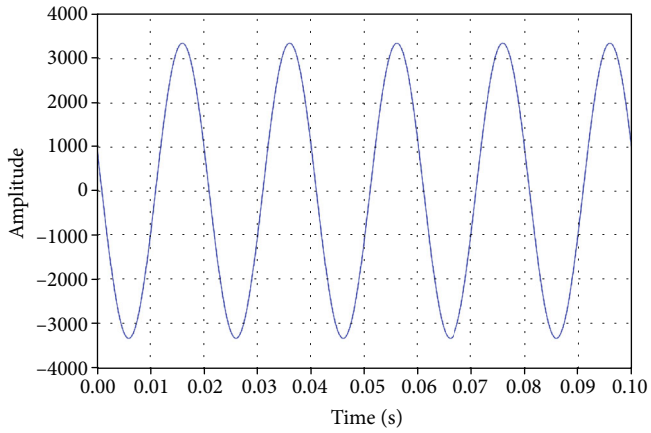


Figure 3: The original sequence waveform.

of the spline curve's range of the sampled value is maximized as well. There are still work to be done to deal with other sampling points to be calibrated: first, determine which interval of the spline curve $[x_i'', x_{i+1}'']$ the sampling value falls within and then substitute the sampling value y_i'' into the corresponding piece-wise function $S_i(x)$ to calculate the corresponding theoretical value.

3. Implementation of the Proposed Method

3.1. Simulation. The key point of this calibration method is to accurately establish the mapping relationship between the measured signal sampling value and the theoretical value.

The realization process of this method has been theoretically deduced above. Now, carry out a simulation experiment on this method and compare the final calibration curve obtained by the calibration algorithm proposed above with the calibration curve of y_k and x_k of the given hypothesis. Calculate the errors of the two calibration curves. If the error meets the accuracy requirements, which means the calibration curve obtained by the algorithm is close enough to the real calibration curve, so the calibration algorithm can be considered feasible.

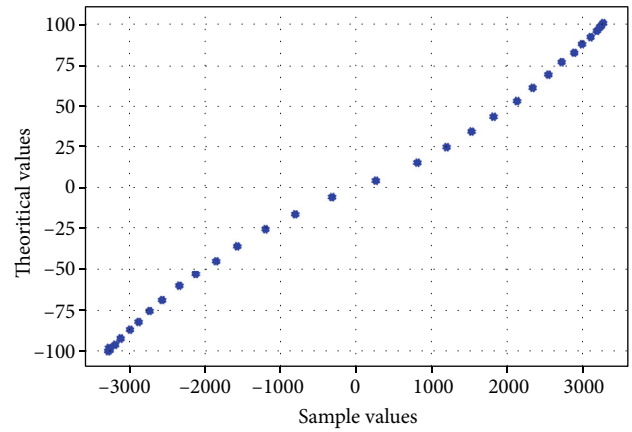


Figure 4: Calibration diagram of calibration points.

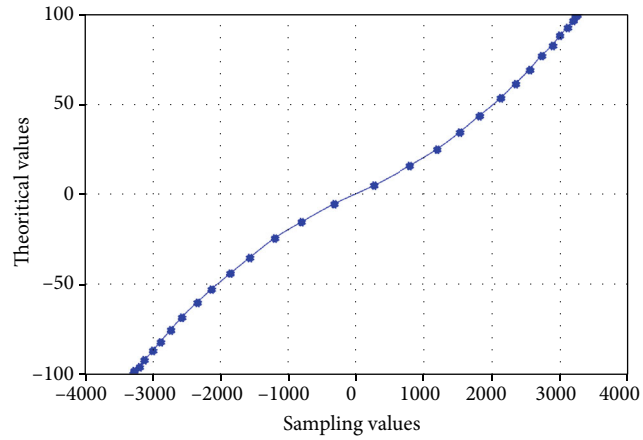


Figure 5: Calibration curve of cubic spline interpolation.

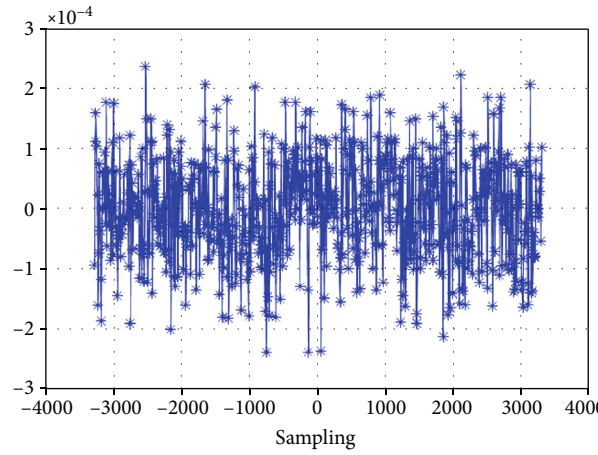


Figure 6: Relative error of calibration curve.

Assuming that the frequency of the given original signal is $f = 50Hz$, the amplitude of the signal A_m is 100, the number of sampling points per period N is 512, the initial phase φ_0 is given as 60°, and the theoretical discrete expression of the original signal is $y_k = 100 \times \cos(\pi k/256 + \pi/3)$. According to the characteristics of the sensor, the calibration

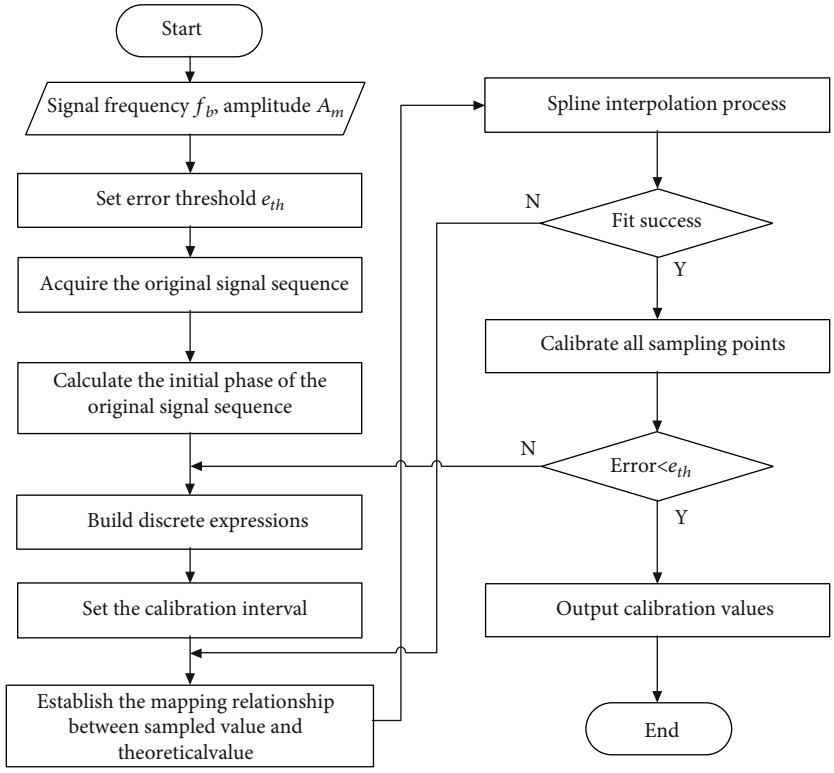


Figure 7: Calibration algorithm flow.

relationship between the sampled value x_k of the original signal and the theoretical value x_k is as follows:

$$y_k = f(x_k) = \begin{cases} \frac{1}{2}x_k + \frac{1}{2}x_k^2 (x_k \geq 0) \\ \frac{1}{2}x_k - \frac{1}{2}x_k^2 (x_k < 0) \end{cases}. \quad (19)$$

The calibration curve of the available calibration relationship is shown in Figure 2:

Knowing the relationship between the theoretical discrete expression y_k and k and also the actual nonlinear relationship between the sampled value x_k and the theoretical value y_k, x_k can be inversely deduced according to $x_k = f^{-1}(y_k)$. The simulated waveform of the original sequence of x_k is shown in Figure 3, the abscissa represents time, and the ordinate represents amplitude:

Using the method mentioned above, the mapping relationship between the sampling value x_k and the theoretical value y_k is established through the theoretical discrete expression, and the 32 sampling points obtained are calibrated, as shown in Figure 4:

Use cubic spline interpolation to make a continuous smooth calibration curve $S(x)$ from 32 calibration points. Figure 5 shows the calibration curve obtained according to cubic spline interpolation.

Compared with the calibration curve $y = f(x)$ given in Figure 2, the calibration curve made by spline interpolation is very close to the given calibration curve.

Substitute all the sampled values in a period into the calibration curve and the actual calibration curve, respectively, and calculate the relative error between them. The abscissa of Figure 6 is the sampled value, and the ordinate is the error of the true value minus the compensation value after calibration, which can be seen from the figure is that the relative error is less than $\pm 3 \times 10^{-4}$. Therefore, the calibration algorithm proposed in this paper is feasible.

3.2. Software Verification. The principle and simulation of the calibration algorithm based on sinusoidal excitation and DFT transform are described above. The calibration method firstly sets the standard source output frequency f_b and the sine signal with the maximum value of A_m, collects the original signal sequence $\{x_k\}$, calculates the initial time phase φ_0 of $\{x_k\}$, and obtains the discrete expression of the original signal according to φ_0. Then, determine the calibration interval of the original signal and establish the mapping relationship between the sampling value and the theoretical value in the calibration interval. Finally, for the processed calibration points, use cubic spline interpolation to make a calibration curve and substitute all the sample values to be calibrated into the calibration curve to calculate the theoretical value. Figure 7 is a specific flow chart of the algorithm.

4. Experiment Result and Analysis

To decide whether the accuracy of the calibration algorithm on the high-precision power analyzer meets the design requirements, an experimental test platform is built. By

Figure 8: Experimental platform.

Table 1: Voltage measurement comparison.

Standard value	AC voltage (V) signal frequency 100 Hz					
	The old method			The proposed method		
	Phase A	Phase B	Phase C	Phase A	Phase B	Phase C
30.00	30.01	29.98	29.97	30.00	29.96	29.95
60.00	59.96	59.98	60.02	60.05	59.98	59.94
100.00	100.02	100.04	100.01	100.08	99.96	99.96
220.00	220.11	220.13	220.05	220.16	219.92	219.90
380.00	379.90	380.28	379.86	380.24	379.88	379.85
500.00	500.04	499.77	499.85	500.31	500.47	499.97
660.00	660.39	660.56	659.67	660.67	659.64	659.78
800.00	801.18	801.28	800.79	800.85	800.83	799.69
950.00	950.95	951.58	950.62	951.52	949.50	948.89

Table 2: 100 Hz current measurement comparison.

Standard value	AC current (A) signal frequency 100 Hz					
	The old method			The proposed method		
	Phase A	Phase B	Phase C	Phase A	Phase B	Phase C
0.10	0.10	0.10	0.10	0.10	0.10	0.10
0.50	0.49	0.50	0.49	0.49	0.49	0.49
1.00	0.99	0.99	0.99	0.98	0.99	0.98
5.00	5.00	4.99	5.00	4.97	4.98	4.99
10.00	10.04	9.98	10.01	9.99	9.98	10.01
25.00	25.09	24.81	25.11	24.90	24.98	24.94
40.00	38.98	39.01	38.64	39.85	40.04	39.96
55.00	52.18	51.95	52.02	54.24	54.43	54.32

comparing the measurement data of the power analyzer equipped with our algorithm and other high-precision testing equipment, the analysis results are obtained.

The specific experimental platform of this project is shown in Figure 8. On the right is the Fluke standard source 6003A as standard input, on the left is the power analyzer equipped with this calibration method, and on the lower left is Yokogawa's WT1800 high-precision power analyzer. The actual product is shown below. Connect the input signal of the standard source to the power analyzer to be tested and the Yokogawa power meter, respectively, and compare the measurement data of the two.

Table 3: 500 Hz current measurement comparison.

Standard value	AC current (A) signal frequency 500 Hz					
	The old method			The proposed method		
	Phase A	Phase B	Phase C	Phase A	Phase B	Phase C
0.10	0.10	0.10	0.10	0.10	0.10	0.10
0.50	0.49	0.50	0.49	0.49	0.49	0.49
1.00	0.98	0.97	0.98	0.98	0.99	0.98
5.00	5.01	4.99	5.00	4.99	5.01	4.99
10.00	10.04	9.97	10.02	9.97	9.98	10.01
25.00	25.25	24.76	25.21	24.95	24.95	24.90
40.00	38.68	38.89	38.35	39.88	40.02	39.97
55.00	51.84	51.72	51.88	54.34	54.45	54.40

Table 4: 1000 Hz current measurement comparison.

Standard value	AC current (A) signal frequency 1000 Hz					
	The old method			The proposed method		
	Phase A	Phase B	Phase C	Phase A	Phase B	Phase C
0.10	0.10	0.10	0.10	0.10	0.10	0.10
0.50	0.49	0.50	0.49	0.49	0.49	0.49
1.00	0.99	0.99	0.99	0.99	0.99	0.98
5.00	5.01	4.97	5.00	4.98	4.97	4.99
10.00	10.11	9.88	9.86	9.99	9.98	10.01
25.00	24.72	24.68	25.18	24.92	24.94	24.91
40.00	38.62	39.28	38.46	39.79	40.08	39.89
55.00	51.78	51.65	51.71	54.24	54.31	54.29

Before the measurement experiment, the high-precision power analyzer needs to be calibrated. The frequency measurement range of the power meter is 10 Hz-1 kHz, the voltage measurement range is 0.1 V-1000 V, and the current range is 0.1A-80A. The voltage measurement accuracy is 0.2% of range, and current measurement accuracy is 0.1% of range add current sensor accuracy. With fluke standard source as input, the instrument is calibrated separately with two methods, the traditional segmented calibration method and the calibration method proposed in this article. After calibration, the measured signals of the two are shown and compared in the following Tables 1–4.

As can be seen from Table 1, the effect of the two calibration methods on voltage measurement is basically the same, because the linearity of the voltage sensor in the actual project is better. The nonlinearity of current sensors is usually poor, and the calibration method proposed in this article is usually used in the current calibration process to achieve high-accuracy. From Tables 2–4, the proposed calibration method is obviously better than the traditional segmented calibration method when measuring large current and high frequency signals. When using this method to measure voltage, the error of the measured value is the largest at 950 V, the error is $(951.52 - 950)/950 = 0.16\% < 0.2\%$, and the voltage accuracy meets the requirements. When the current is

measured at 55 A, the error of the current sensor has exceeded 2%, but the maximum error shown in Tables 2–4 is $(55 - 54.24)/55 = 1.38\%$, which meets the requirements of current accuracy. The comparison of experimental results verifies that the calibration method proposed in this paper is effective in the application of nonlinear systems and has high calibration accuracy.

5. Conclusion

In this paper, a nonlinear calibration algorithm based on sinusoidal excitation and DFT transformation is proposed. This algorithm overcomes the shortcomings of traditional methods, like it is difficult to determine the segment turning point and segment range in old methods, and multiple manual calibrations are cumbersome; also, the calibration accuracy can decline within the overall measurement range. In addition, this method only needs to obtain the effective value data of the current calibration source. Even if the segment turning point is increased, the calibration of the instrument can be accurately completed by obtaining the effective value data only once, which not only improves the calibration accuracy but also avoids repeated operation of the calibration source, thus greatly improves the calibration efficiency. The simulation experiment verifies the feasibility and accuracy of the algorithm, and the voltage and current parameters are measured by a high-precision power analyzer equipped with the algorithm. The experimental results show that the measured values of the voltage and current after calibration within the range meet the accuracy requirements.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is partly supported by the Fundamental Research Funds for the Central Universities under Grant ZYGX2019J063 and the second batch of industry-university cooperation collaborative education projects of the Ministry of Education in 2019 (201902059007).

References

[1] H. E. van den Brom and D. Hoogenboom, "Power quality measurements-the importance of traceable calibration," *Electrical Power Quality and Utilisation Journal*, vol. 16, no. 1, pp. 25–29, 2013.

[2] H. Zhiyuan, J. Haibin, Y. Li, and P. Cheng, "Research on calibration method of wideband power analyzer," in *2014 XXXIth URSI General Assembly and Scientific Symposium (URSI GASS)*, pp. 1–4, Beijing, 2014.

[3] L. Zuliang, W. Lei, H. Hongtao et al., "A proposal to establish a system for calibration of harmonic power analyzers," in *2016 Conference on Precision Electromagnetic Measurements (CPEM 2016)*, pp. 1-2, Ottawa, ON, 2016.

[4] Y. Zhao, H. Wang, Y. Zheng, Y. Zhuang, and N. Zhou, "High sampling rate or high resolution in a sub-Nyquist sampling system," *Measurement*, vol. 166, no. 15, pp. 108175–1081758, 2020.

[5] P. M. Tzvetkov, I. N. Kodjabashev, and K. S. Galabov, "Comparison of approaches for calibration of electrical power quality analyzers," in *2019 XXIX International scientific symposium "metrology and metrology assurance" (MMA)*, pp. 1–4, Sozopol, Bulgaria, 2019.

[6] Z. Li, S. Yan, W. Hu, Z. X. Li, and Y. C. Xu, "High accuracy online calibration system for current transformers based on clamp-shape Rogowski coil and improved digital integrator," *Mapan*, vol. 31, no. 2, pp. 119–127, 2016.

[7] P. Luo, Z. Li, and H. Li, "A high-current calibration system based on indirect comparison of current transformer and Rogowski coil," *Measurement Science and Technology*, vol. 24, no. 12, article 125005, 2013.

[8] L. Wang, L. Guo, J. Jiang, and D. Qiu, "A Hilbert-transform-based method to estimate and correct timing error in time-interleaved ADCs," *Journal of Electronic Testing*, vol. 31, no. 3, pp. 291–299, 2015.

[9] A. Brandolini, M. Faifer, and R. Ottoboni, "A simple method for the calibration of traditional and electronic measurement current and voltage transformers," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 5, pp. 1345–1353, 2009.

[10] Y. Zhou, Q. Huang, J. Li, and L. Yao, "A calibration method of the zero-drift for testing system of electronic transformer characteristics," in *2015 International Conference on Smart Grid and Clean Energy Technologies (ICSGCE)*, pp. 97–102, Offenburg, 2015.

[11] W. Ren, Y. Yuan, X. Hu, Z. Yang, and Y. Zhang, *Steady-state error calibration technology for electronic instrument transformer*, pp. 1–6, 2012.

[12] B. Jin, Y. Chen, D. Li, K. Poolla, and A. Sangiovanni-Vincentelli, "A one-class support vector machine calibration method for time series change point detection," in *2019 IEEE International conference on prognostics and health management (ICPHM)*, pp. 1–5, San Francisco, CA, USA, 2019.

[13] P. Wang, Y. Kong, and M. Zhang, "Error self-calibration algorithm for acoustic vector sensor array," *Journal of Sensors*, vol. 2019, Article ID 9052547, 10 pages, 2019.

[14] B. Friedlander and A. J. Weiss, "Effects of model errors on waveform estimation using the music algorithm," *IEEE Transactions on Signal Processing*, vol. 42, no. 1, pp. 147–155, 1994.

[15] L. Wang, J. Wang, J. Zhao, and J. Wu, "Line-element based nonlinear adaptive piecewise compensating correction for LVDT sensors," *Journal of Beijing Institute of Technology*, vol. 22, no. 4, pp. 497–503, 2013.

[16] D. Chengxian, "Based-on nonlinear compensation self-tuning PID control," in *Proceedings of the 3rd World Congress on Intelligent Control and Automation (Cat. No. 00EX393)*, vol. 5, pp. 3104–3106, Hefei, China, 2000.

[17] S. Yang, A. Y. K. Yan, and C. M. N. Ng, "High accuracy and traceable power quality instrument calibration using high-speed digitizing technique," in *2016 Conference on Precision Electromagnetic Measurements (CPEM 2016)*, pp. 1-2, Ottawa, ON, 2016.

[18] Q. Guo, J. Wu, H. Jin, and C. Peng, "An innovative calibration scheme for interharmonic analyzers in power systems under asynchronous sampling," *Energies*, vol. 12, no. 1, p. 121, 2019.

[19] X. Yang, X. Meng, T. Jiang, and A. Husnain, "An error correction method based on polynomial fitting to improve the accuracy of the EM indoor positioning system," in *2016 Sixth International conference on Instrumentation & Measurement, computer, communication and control (IMCCC)*, pp. 932–935, Harbin, 2016.

[20] L. Mâțiu-Iovan, "Some aspects of implementing a cubic spline interpolation algorithm on a DSP," in *2012 10th International Symposium on Electronics and Telecommunications*, pp. 291–294, Timisoara, 2012.

[21] L. Lin, H. Wang, and Y. Bai, "Nonlinear error correct of intelligent sensor by using genetic algorithms and cubic spline interpolation," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 455–460, Boston, MA, 2005.

[22] T.-L. Tsai and J.-Y. Chen, "Investigation of effect of endpoint constraint on time-line cubic spline interpolation," *Journal of Mechanics*, vol. 25, no. 2, pp. 151–160, 2009.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5577084, 10 pages
https://doi.org/10.1155/2021/5577084

Research Article

Automatic Detection of Fractures Based on Optimal Path Search in Well Logging Images

Wei Zhang[1], **Tong Wu**[1], **Zhipeng Li**[1], **Yanjun Li**[1], **Ao Qiu**[1,2] and **Yibing Shi**[1]

[1]*School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*
[2]*Welltech Research and Design Institute of China Oilfield Services Co., Ltd., Langfang 065000, China*

Correspondence should be addressed to Yanjun Li; yjli@uestc.edu.cn

Received 14 January 2021; Revised 23 January 2021; Accepted 9 August 2021; Published 13 September 2021

Academic Editor: Bruno Ando

Reservoir fractures are essential locations to gather oil and gas. Recently, imaging logging technology has become a mainstream method for obtaining stratigraphic information. This paper proposed a combined optimal path search strategy to effectively identify and extract the fracture information in well logging images. Specifically, the threshold segmentation of logging images is used to obtain the binary image. In addition, the identification of connected fractures in the logging image is transformed into the optimal path search, and the identification and extraction of reservoir fractures are realized by constructing the optimal path between the two ends of fractures. Finally, an improved ant colony algorithm is applied to filter irrelevant information and extract fractures automatically by recording all the ants' exploration trajectories in the ant colony. Compared with previous approaches, the proposed method can eliminate irrelevant background features and merely reserve pixels corresponding to fractures. Simultaneously, relative to the conventional strategy, the time consumption is reduced by more than 98%. The findings of this study can help for better extracting fractures automatically and reducing manual workload.

1. Introduction

Formation fractures are discontinuous profiles widely distributed in different lithologies and gradually formed through diagenesis or tectonic deformation. Naturally fractured reservoirs store over 50% of the known petroleum and gas reserves worldwide [1]. Accordingly, the detection of fractures has attracted plenty of researchers because they are vital indicators of the admissibility of petroleum and gas reservoirs in tight formations [2] and affect reservoir properties, enrichment, and gas reservoir development [3–6]. For instance, microfractures in shale are considered important fluid transport networks as well as oil and gas migration pathways and are key factors in forming oil and gas reservoirs [7]. So far, there are many technologies for discovering potential fractures in petroleum reservoirs, such as subsurface electromagnetic technology [8], tiltmeter [9], downhole microseismic fracture monitoring [10], and radiotracer diagnosis [11]. Compared with these methods, the imaging logging analysis technology [12] originated in 1986 has become the mainstream method of oil and gas reservoir exploration, which wins the favor of considerable engineers for its intuitive and accurate display of wellbore and stratigraphic structure. Ultrasonic imaging logging [13, 14], one of the representative imaging logging technologies, can characterize the geometric features of fractures and distinguish various geological features more clearly. Figure 1 is a schematic diagram of ultrasonic imaging logging. Each pixel in the image corresponds to one arrival time of the ultrasonic signal reflected from the borehole wall. A bright area indicates that the arrival time of echo measured in this area is short. Conversely, a dark area indicates long arrival time or no reflection exists. Therefore, this paper's task is to design a filter which extracts pixels corresponding to fractures as shown in Figure 2.

Due to the complexity of reservoir, it is a time-consuming process to accurately identify fractures manually. Accordingly, automatic identification of fractures in logging images is of great significance [15], which prompts many researchers to explore proper approaches. Initially, the

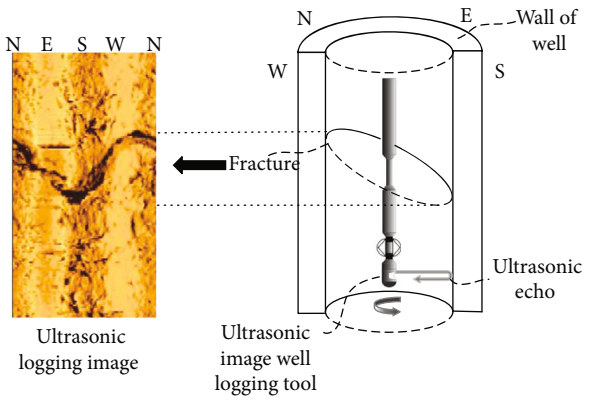


Figure 1: The principle of ultrasonic image logging.

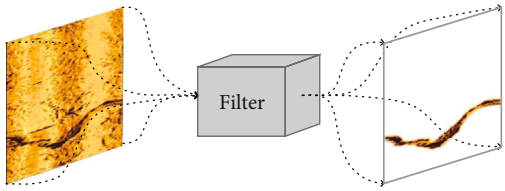


Figure 2: The object of this paper: filtering irrelevant noise from background and reserving fractures.

scholars established mathematical models of the fractures to characterize the characteristics. Changchun et al. regarded the shape of fracture as a sine curve and utilized two-dimensional Hough Transform (2D-HT) to detect fractures with fixed patterns in the image [16]. Taiebi et al. used a multi-scale technique based on directional filtering and Hough Transform (HT) to extract fractures in the logging images further [17]. Directional filtering can enhance the contrast between the target fracture and the imaging logging background, which is positive to conduct HT for fracture segmentation. Liu et al. used ant colony algorithm to ascertain the edges of fractures. Then, the sinusoidal fracture was found by HT [18]. Jianping et al. improved the HT according to the characteristics of fractures in imaging well logging. For such a sine curve with fixed periods, the initial detection step is divided into two steps: the voting mechanism was used to determine the baseline position of sinusoids; and then the 2D-HT determined the amplitude and initial phase parameters of the sinusoidal curve [19]. Based on the baseline position determined by voting mechanism, Yingming et al. leveraged genetic algorithms to perform nonlinear fitting of sine-shaped scattered points in the relevant area to realize pixel extraction of sinusoidal regions such as fracture, bedding, and layer boundaries of imaging logging images [20]. Since squeeze friction in the actual drilling process will cause the borehole wall damaged, the fracture shape is not necessarily in standard sinusoidal shape. Hence, the method based on mathematical models is not effective in fitting and identifying nonstandard sinusoidal fractures.

Recently, with the rapid development of computer vision, researchers have tried to use image segmentation to separate pixels corresponding to fractures from the background of ultrasonic logging images and then realize the extraction of fractures. Wang utilized the valley edge algorithm to determine local dark area to accomplish rock fracture detection without determining a threshold [21]. Xu et al. combined the classification into a cascade system using the K-nearest neighborhood (KNN) classifier, which can classify rock structures and extract the expected features in the ultrasonic well logging images [22]. This method divides the image into superpixel blocks with uniform size and moderate compactness, which provides the basis for subsequent fracture recognition. Sorncharean et al. used grid unit analysis chain and fracture unit verification to eliminate false detection of shadow boundaries on the image, so as to realize the detection of fractures on road images with uneven illumination and strong texture [23]. Zhang used the clustering-minimum spanning tree method to extract the crack area by edge extraction and threshold segmentation and then used the minimum spanning tree method to detect the crack in asphalt pavement [24]. However, in actual images, simply using low-level content information such as color, brightness, and texture of pixels is not enough to generate a good segmentation effect, and it is easy to produce wrong segmentation results. To sum up, all mentioned approaches have the following issues: they fail to identify irregular fractures because of low extraction accuracy, and the identification and extraction efficiency is not efficient. Recognition results often remain some noncrack information.

To address above issues, we have analysed the fractures' semantic information in logging images and utilized the pattern of fracture regions to determine fracture region. Different from the above methods, connectivity of fracture regions in logging images is considered a key standard where pixels corresponding to reservoir fractures connected two ends of an image [25]. On the other hand, heuristic algorithms have played an increasingly vital role in similar optimization problems. For example, Deng et al. [26] proposed multiple strategies for global optimization problems. And Deng et al. [27] further used an improved quantum-inspired differential evolution algorithm to avoid premature convergence and improve the global search ability. Therefore, based on the above inference, this paper proposes a method that combined optimal path search strategy (COPS) to identify and extract fractures in ultrasonic logging images, and the main contributions of this research are as follows:

(1) Construct a mathematical model describing this task. And convert fracture recognition to path search

(2) Preprocess logging images, and build a path search space by OTSU (OTSU) threshold segmentation algorithm

(3) Use the ant colony algorithm (ACA) to search potential paths connecting two ends of fractures, which extracts the entire fracture region

(4) Improve ant colony algorithm by designing a parallel searching strategy to accelerate path search dramatically

The paper is organized by following steps: first, the proposed methods are presented by detailed computational

procedure. Then, experiments are performed to analyse our method's performances related to accuracy and operation rate. Besides, according to experimental results, advantages and potential disadvantages are discussed to point out the direction for further research. Finally, we summarize the entire research.

2. Methods

2.1. Mathematical Model. In the stratum, different structural planes show certain differences on account of different morphological and physical characteristics. In fractured reservoirs, the expansion zone generated by structural fractures is where a large amount of oil and gas transfer and accumulate. Consequently, the method designed in this paper is mainly aimed at the identification and extraction of such fractures in ultrasonic logging images. This type of fracture is a fracture zone surrounding the wellbore and formed along the borehole wall. The differences mainly manifest in the color depth, shape difference, width change of the band-shaped sinusoid, and the irregularity of the rock spot and texture characteristics around the curve [28]. The fracture presents a three-dimensional ellipse shape on the wellbore wall, and if the cylindrical well wall is cut along its axis, the three-dimensional elliptical fracture will be a sinusoidal curve after unfolding on the two-dimensional image [29–32], as shown in Figure 3.

The function of the fracture can be described by Equation (1) [33]:

$$y = A \sin(\omega x + \beta) + y_0, \tag{1}$$

in the formula, y_0 is the baseline position of the sinusoidal crack curve, A is the amplitude of the curve, β is the initial phase, and ω is the angular velocity.

For reservoir fractures distributed along the circumference of the well wall, the shape of fractures can be approximately considered to be sinusoidal. However, in practice, due to the mechanical vibration in drilling process and the influence of geological activities, the fractures spreading in the logging images may not show a standard sinusoidal shape, and sometimes, it is even far from a sinusoidal form. Hence, in the process of image processing, if a sinusoidal shape is regarded as a prior condition of the fracture morphology, the algorithm will have a high probability to generate false recognition of fractures or fail to fit the shape of fractures well.

According to the fractures distributed around the borehole wall, a feature that is obviously different from other interference information is that the fractures connect the left and right sides in the logging image. Therefore, we suppose the ultrasonic logging image is G, containing an ordered pair (V, E), where V is a vertex set composed of all pixels and E is a set of disordered point pairs composed of elements in V, denoted as edge set, whose elements are edges, and a same point pair can appear multiple times in E, denoted as $G = (V, E)$. As a result, for the identification and extraction of fractures in the logging image, the problem is transformed into finding a subset V_{subset} of vertices in graph G that can connect the regions on both sides of the graph; this V_{subset} can correspond to the area where the fracture is located.

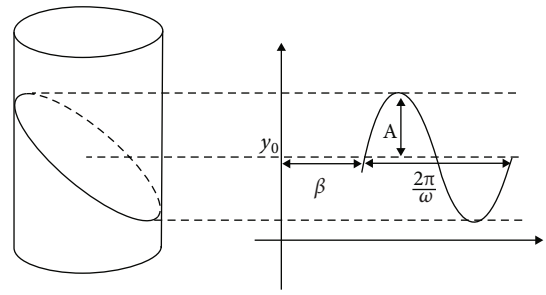


Figure 3: General form of fractures distributed along the shaft wall.

2.2. Image Preprocessing. According to the principle of ultrasonic logging images, fractures are darker in the image and appear dark gray sinusoidal shapes. Factors like the unnatural etching of the strata, extrusion deformation caused by pressure, and the damage to the borehole wall by the drill bit during the drilling process will leave a geometric shape similar to the fracture or a similar gray value on the ultrasonic logging image. Therefore, before performing fracture identification and extraction, it is necessary to perform image segmentation to filter unnecessary interference information and enhance the fracture characteristics at the same time.

As mentioned in introduction, based on the OTSU algorithm, this paper performs preliminary threshold segmentation on the original ultrasonic logging image to extract effective fracture area information. The OTSU algorithm, also known as the maximum between-class variance method, is an algorithm to determine threshold. This threshold is used to perform fixed threshold binarization of the image to maximize the variance between classes. According to the gray characteristics of the image, it is divided into two parts: the background and the foreground. The segmentation with the largest variance between the classes means the smallest probability of misclassification. Suppose the size of a single ultrasonic logging image is $X \times Y$, grayscale range $R = [0,255]$, $R \in N^*$, and the probability of the gray level i is

$$p_i = \frac{R_i}{X \times Y}. \tag{2}$$

Define the fracture area C_t, the background area C_b, and a gray-scale threshold T; then assign all the pixels in the logging image to C_t and C_b according to this threshold, the probabilities of C_t and C_b are θ_t and θ_b, respectively; so the mean gray value of the fracture area μ_t and the mean value of the background area μ_b are

$$\mu_t = \sum_{i=0}^{T} i \bullet p(i \mid C_t) = \sum_{i=0}^{T} \frac{ip_i}{\omega(t)}, \tag{3}$$

$$\mu_b = \sum_{i=T+1}^{255} i \bullet p(i \mid C_b) = \frac{\sum_{i=0}^{255} ip_i - \sum_{i=0}^{T} ip_i}{1 - \omega(t)}, \tag{4}$$

in the equation, $\omega(t) = \sum_{i=0}^{t} p_i$, so, in the two categories, the variances are

$$\sigma_t^2 = \sum_{i=0}^{T} (i - \mu_t)^2 p(i \mid C_t), \tag{5}$$

$$\sigma_b^2 = \sum_{i=T+1}^{255} (i - \mu_b)^2 p(i \mid C_b). \tag{6}$$

According to the literature [34], this paper adopts an easily calculated interclass variance evaluation function σ_c^2 as the standard of the threshold segmentation, which is defined as follows:

$$\sigma_c^2 = \frac{[\mu_T \omega(t) - \mu(t)]^2}{\omega(t)[1 - \omega(t)]}, \tag{7}$$

among them, $\mu_T = \sum_{i=0}^{255} i p_i$ and $\mu(t) = \sum_{i=0}^{T} i p_i$. Finally, find the best threshold to transformed into

$$T^* = \arg \max_{0 \le t \le 255} \{\sigma_b^2\}. \tag{8}$$

As shown in Figure 4, the ultrasonic logging image is transformed into a binary image after threshold segmentation. Compared with the maximum entropy threshold segmentation algorithm [35], the OTSU algorithm has a better extraction effect in response to ultrasonic imaging logging fracture identification and minimizes the interference of background noise.

2.3. Optimal Path Search Based on OTSU Threshold Segmentation Algorithm and Ant Colony Algorithm. As mentioned above, the search of fracture area in logging images can be transformed into seeking the connected area on the left and right sides of the connection image $G = (V, E)$. Therefore, the search space is established according to the binary image segmented by a threshold value. In a binary image, the gray value of the potential fracture area is 0, which is a passable region; the gray value of the background area is 1, which is an obstacle region. In this paper, the serial number method [36] is used to establish the grid set corresponding to the binary ultrasonic logging image, and the effect is shown in Figure 5.

After rasterizing the image and establishing the search space, it is necessary to search the area connecting the left and right sides to determine the fracture area. For path search using the grid method, the search efficiency is positively related to the size of the space, so an efficient and stable search algorithm is needed. Compared with other algorithms, the ACA has strong robustness and adaptability and has achieved good results in solving path planning [37]. Accordingly, the ACA is used to search for the fracture area.

The ACA is a swarm intelligence algorithm that simulates the foraging behavior of ants in nature. By releasing pheromones on the foraging path, the ant colony will tend to walk on the path with higher pheromone concentration and releasing pheromones simultaneously. Pheromone can attract more ants and form a feedback loop. Eventually, the entire ant colony will find the most suitable path. The search steps of fractures area based on ACA are as follows:

(1) Suppose the number of ants in the ant colony is m, the number of pixels in the potential fracture area is n, and the distance between pixels V_i and V_j in the fracture area is $d_{ij}(i, j = 1, 2, \cdots, n)$, the pheromone concentration between any target pixel points V_i and V_j in the same fracture region at time t is $\tau_{ij}(t)$, and the initial pheromone concentration is $\tau_{ij}(0) = \tau_0$. The pixel points V_{io} and V_{jo} are randomly selected on the left and right sides of the fracture area as the starting and ending points of the path search

(2) Let $P_{ij}^k(t)$ be the probability of the ant k in the ant colony moving from the pixel point i to the pixel point j; the calculation formula is

$$P_{ij}^k(t) = \begin{cases} \dfrac{\left[\tau_{ij}(t)\right]^a \bullet \left[\eta_{ij}(t)\right]^b}{\sum_{r \in \text{allow}_k} \left[\tau_{ir}(t)\right]^a \bullet \left[\eta_{ir}(t)\right]^b}, & r \in \text{allow}_k, \\ 0, & r \notin \text{allow}_k, \end{cases} \tag{9}$$

among them, $\eta_{ij}(t)$ is the heuristic function, $\text{allow}_k\, k = (1, 2, \cdots, k)$ is the set of pixels to be searched by ant k, a is the pheromone importance factor, and b is the importance factor of the heuristic function. Through the roulette wheel selection method, according to the transition probability of the remaining pixels, the ant k will go to the next pixel

(3) When an ant has completed a traversal and there is no way to go, the pheromone concentration $\tau_{ij}(t)$ on the path between the traversed pixels will be updated, as shown in

$$\begin{cases} \tau_{ij}(t+1) = (1 - \rho), \tau_{ij}(t) + \Delta\tau_{ij}, & 0 < \rho < 1, \\ \Delta\tau_{ij} = \sum_{k=1}^{n} \Delta\tau_{ij}^k, & 0 < \rho < 1, \end{cases} \tag{10}$$

$$\Delta\tau_{ij}^k = \begin{cases} \dfrac{Q}{L_k}, & \text{the } k\text{th ant goes from pixel } i \text{ to pixel } j, \\ 0, & \text{others}, \end{cases} \tag{11}$$

where $\Delta\tau_{ij}^k$ represents the pheromone released by the ant k on the connection path between the pixel V_i and V_j, ρ denotes pheromone volatile factor, $\Delta\tau_{ij}$ represents the sum of the pheromone concentration accumulated by all ants on the connecting path between the pixel V_i and V_j, Q is a constant, and L_k is the length of the path that the ant k passes

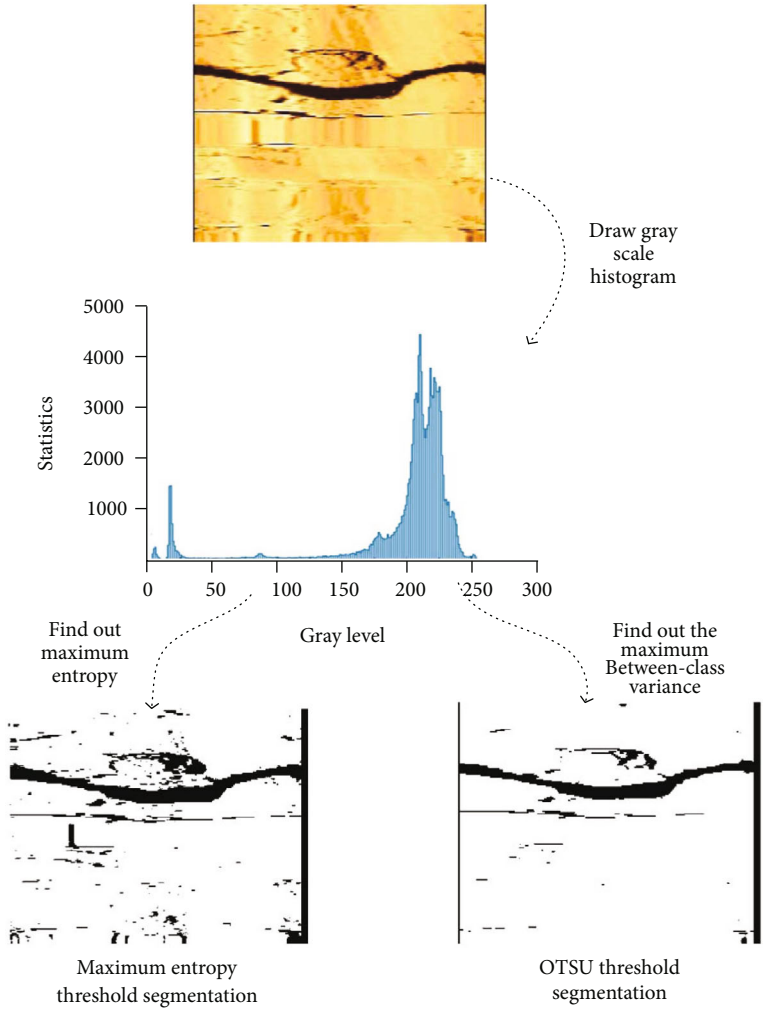


Figure 4: Results of threshold segmentation by maximum entropy algorithm and OTSU algorithm.

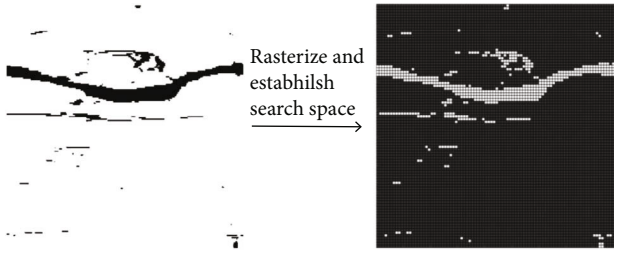


Figure 5: Image rasterization to establish search space.

(4) When the maximum number of iterations is reached, the path search ends and the search path is printed out

Through the optimal path search (OPS) algorithm based on the OTSU algorithm and ACA, the connected areas at both ends of the connecting image G can be traversed by different ants in the ant colony. When all ants' footprints are recorded, the fracture areas in the ultrasonic logging image can be identified and extracted. Simultaneously, since there are no ants traversing the nonfracture areas, these areas can be filtered out, and only information about the fracture areas can be retained.

2.4. Combined Optimal Path Search Strategy. Through strategies mentioned above, the fracture area in the ultrasonic logging image can be correctly identified and extracted. However, due to the large resolution of a single logging image, the algorithm could converge slowly and easily fall into a locally optimal solution if the optimal path search is performed directly. Therefore, this paper proposes the COPS given the morphological characteristics of fracture regions and ultrasonic logging images.

The current common ACA application is designed for the shortest path search. The pheromone concentration $\Delta \tau_{ij}^{k}$ is calculated according to the total length of the ant's path, which means that the shorter path ant colonies select to travel, the higher the pheromone concentration will be left. Then, the overall ant colonies tend to the shortest path. The object of this paper is the overall search of the fracture area; that is, we hope the ant colony can traverse the entire fracture area of the logging image as much as possible. To

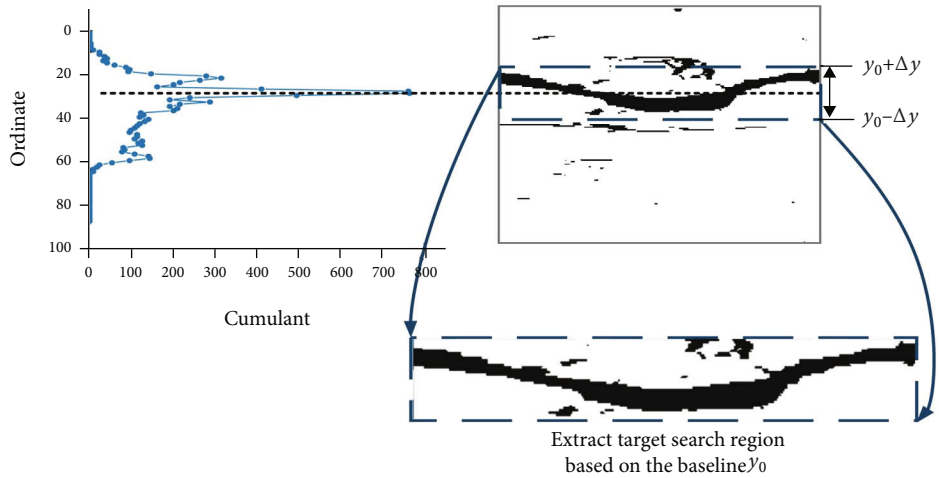


Figure 6: Location of fracture area based on voting accumulation.

this end, change Equation (11) to

$$\Delta\tau_{ij}^{k} = \begin{cases} \dfrac{Q}{d_{ij}}, & \text{The } k\text{th ant goes from pixel } i \text{ to pixel } j, \\ 0, & \text{others}, \end{cases} \tag{12}$$

where d_{ij} represents the distance between the pixel points V_i and V_j passed by the ant. This formula uses local information of the path that ants pass through to calculate the released pheromone concentration. Therefore, it can increase the probability of the ant going to the different pixels in fracture areas and as many pixels as possible can be traversed.

It can be seen from Equation (1) that the approximate position of the sinusoidal fracture curve in the logging image can be determined by the baseline y_0. Thus, a specific path search range can be determined by seeking the baseline position, and the scale of the path search can be reduced. For any point s_1 on the sine curve, another point s_2 on the curve can be determined, which meets $X(s_1) - X(s_2) = T/2$, where T is the period of the sine curve, and the midpoint $S_0 = (x_{0,}\, y_0)$ between s_1 and s_2 must fall on the baseline y_0. By searching for such point pairs and using the voting accumulation mechanism, the ordinate information of all midpoints is counted to determine the baseline position. Baseline positioning is performed on the binarized ultrasonic logging image in Figure 5, and the area where the baseline is located is cropped to get the target search area G_s, and the result shown in Figure 6 is obtained.

In order to further improve the efficiency of ACA and accelerate the convergence speed, the ACA in this paper is improved by adopting a strategy searching single area path parallelly. For the target search region G_s, the fracture area is connected through left and right. Therefore, G_s is divided into l different subregions G_{si} $(i = 1, 2, \cdots, l)$ along the longitudinal direction, and the fracture area within each subregion G_{si} is still connected from left to right. The improved

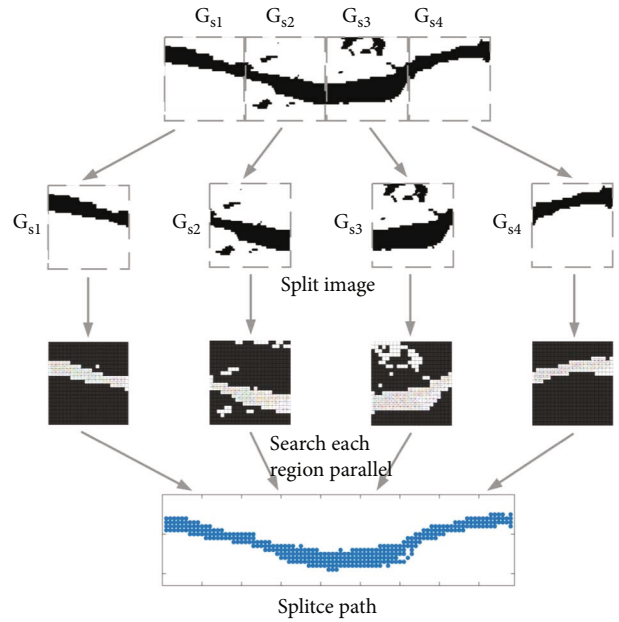


Figure 7: Parallel search strategy of small area path.

ACA is utilized to seek the path in each subarea G_{si} simultaneously, and then all the subfracture areas searched in the subarea are spliced to complete the identification and extraction of the fractures in the logging image. The calculation process is shown in Figure 7. Using the strategy of searching a single region path parallelly, the region that the algorithm needs to traverse simultaneously is only $2\Delta y/Yl$ of the original region, which can accelerate the search speed and make the algorithm converge rapidly. Since the search range is reduced and each area is searched by an independent ant colony, it can also effectively improve the situation that the ant colony falls into optimal local results.

3. Experiments and Results

3.1. Experiment Platform. For verifying the performance of the optimal path search strategy proposed in this paper, we

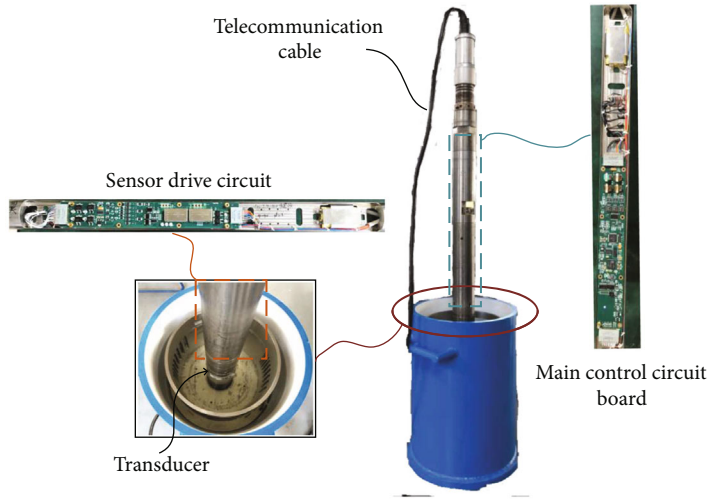


Figure 8: Ultrasonic image well logging system.

tested the algorithm in real logging images. Figure 8 depicts the overall structure of the ultrasonic image well logging instrument. The well logging tool consists of rotary ultrasonic transducer driven by motor, main control board receiving ultrasonic signal and conducting signal processing, sensor drive board driving ultrasonic transducer, and cable for communication with a host computer. 250 points are collected evenly by a transducer rotating one circle, and a logging image is drawn according to the arrival time and maximum amplitude of the echo.

The test object is the ultrasonic logging image G_Z from 4423 meters to 4559 meters downhole of the Zhanjiang production oil well. Images are provided by the Oilfield Technology Research Institute of China Oilfield Services Co., Ltd. (COSL). Logging image G_Z is cut into 72 subimages with a resolution of 352×352. The image processing algorithm runs on a hardware computing platform based on an i9-9900k processor with 64GB memory, and the programming language is MATLAB R2020a. Figure 9 shows the operation of the circumferential ultrasonic imaging logging tool in the Zhanjiang production well.

3.2. Test Results. In order to further verify the recognition accuracy of the proposed algorithm for different forms of fractures and different background noise interference, five typical fracture regions in G_Z were selected for testing and compared with the results of the common threshold segmentation algorithm and the artificially marked fracture regions as ground truth. The results are shown in Figure 10. As shown in Figure 10, whether it is a sinusoidal fracture, a horizontal fracture, or a fracture with obvious background noise, the combined optimal search strategy algorithm can eliminate the residual background noise and more accurately preserve fracture information while compared with the traditional direct threshold algorithm.

Confusion matrix, which is shown in Figure 11, is an index for evaluating classification performance. According to pixels' classification in ground truth and segmentation results of two algorithms, we calculate average "true positive" (TP), "false positive" (FP), "true negative" (TN), and "false negative" (FN) in the above images to draw the confusion matrices. Consequently, precision and recall which determine performances of classification can be calculated as

Figure 9: Logging operation in Zhanjiang.

$$\begin{cases} \text{Precision} = \dfrac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Recall} = \dfrac{\text{TP}}{\text{TP} + \text{FN}}. \end{cases} \tag{13}$$

Table 1 shows the corresponding F1 score, mean Intersection over Union (IoU), precision, and recall generated by two mentioned methods. Obviously, toward fracture segmentation in well logging images, the IoU can be increased from 20.72 to 43.21, and the F1 score can be increased by 0.26. Precision and recall are commonly contradictory when assessing the same segmentation result, and higher recall from Maximum Entropy Threshold represents that it classifies more pixels as fractures, but its precision is reduced. In summary, all results demonstrate that the fracture segmentation generated by the proposed method is closer to the fracture information marked manually.

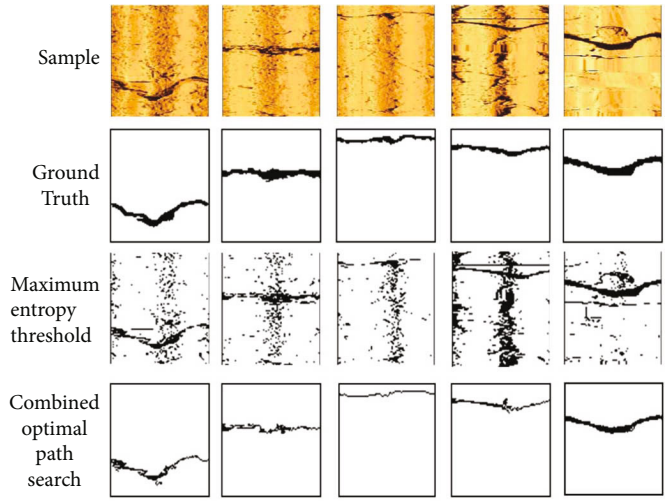


Figure 10: Comparison of recognition effect.

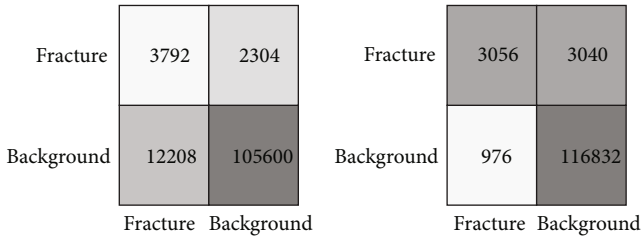


Figure 11: Confusion matrices: Maximum Entropy Threshold (a) and ours (b).

Table 1: Accuracy assessment for segmentation results.

Parameters	Precision	Recall	F1 score	Mean IoU (%)
Maximum Entropy Threshold	0.2370	0.6620	0.3432	20.72
Ours	0.7579	0.5013	0.6035	43.21

3.3. Performance Analysis. Compared with the conventional ACA, this paper proposes the COPS to perform target decomposition and then single path search in each independent area. Thus, it can greatly improve the search efficiency. Figure 12 shows the time-consuming comparison of fracture identification and extraction of the above five typical logging images using conventional ACA and the algorithm proposed in this paper to perform path search on the computing platform of this paper. Both colony sizes are 20, and the number of iterations is 100 while pheromone importance factor a is 1, importance factor of the heuristic function b is 7, and pheromone volatile factor ρ is 0.3. It can be seen from Figure 12 that the algorithm proposed in this paper has a significant efficiency improvement.

For exploring the factor affecting segmentation further, we conducted comparative tests by adjusting iterations as the number of iterations determines whether the ant colony could traverse all possible paths. We use parameters mentioned above with different numbers of iteration to carry out experiments. The average time consumed and IoUs are shown in Table 2. By observing the outcomes, while 100 is determined as the iteration number, the proposed method can acquire a balance between time consumed and visual results. Accordingly, we performed 100 iterations when using our method.

4. Discussion

In this paper, we propose an effective approach coping with fracture recognition and extraction. Relative to conventional methods, i.e., threshold segmentation and Hough Transform, not only can the method identify fracture regions accurately, but it can also extract pixels corresponding to fracture regions to show intact shape of fracture clearly.

Compared with the conventional ant colony algorithm, one advantage is apparent acceleration in the rate of path search by parallel search strategy as search space is reduced by 4 times. By analysing the principle of our method, another reason of acquiring remarkable results based on our method is that ant colonies tend to search paths with high concentration of pheromone. This indicates that when an ant reaches the end of the path, the rest of ant colonies are likely to follow it to avoid invalid searches, which have a positive impact on filtering irrelevant information. However, due to the same reason that ants tend to search paths with high concentration of pheromone, all fractures may not be extracted completely when more than one fracture exists in logging images.

On the other hand, based on prior knowledge of fracture connectivity, the proposed method cannot extract the complete fracture regions if actual fractures in logging images are discontinuous and disconnected, which is intractable in our completed study. Therefore, we will focus on the task of identifying and extracting multiple fractures in the next stage.

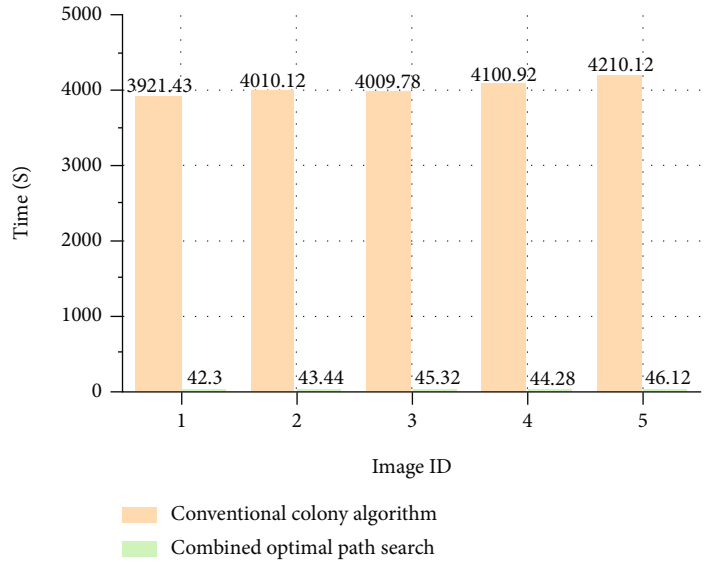


Figure 12: Time-consuming comparison.

Table 2: Effect of iteration number on time consumed and IoU.

Iterations	Time consumed (s)	Mean IoU (%)
10	6.31	39.89
50	21.65	40.40
75	33.79	41.92
100	45.32	43.21
200	87.89	43.19

5. Summary and Conclusions

In this paper, the COPS algorithm is proposed which filters background interference information and automatically extracts fractures in ultrasonic logging images. The method's key is to convert the target recognition in the fracture area into optimal path search and search the path according to the fracture's feature that connected on both sides of the image. The main remarks of this study are as follows:

(1) Establish a mathematical model of the fracture, determine the reference position of the fracture through the voting accumulation mechanism, and appropriately cut the area where the crack is located according to the reference position

(2) The images obtained by cutting are further cut along the vertical direction to obtain each subregion with partial fracture information, and the ACA was used to search and extract the fractures in each subspace

(3) The fractures extracted from each subspace are spliced to obtain a complete fracture

Experiments and analysis have been conducted to illustrate our method's advantages and potential disadvantages. In the next stage, we will concentrate on improving multiple fractures extraction in a well logging image.

Data Availability

The well logging data used to support the findings of this study were supplied by Ao Qiu under license and so cannot be made freely available. Requests for access to these data should be made to Wei Zhang (weizhang@uestc.edu.cn).

Conflicts of Interest

The authors declare that they have no conflict of interest.

Acknowledgments

The research is supported by the National Science Foundation of China under the Grant number 61201131 and development and industrial application of ultra-high temperature and high-pressure wireline logging system from the Science and Technology Project of CNOOC under the Grant number CNOOC-KJ ZDHXJSGG YF 2019-02. We gave our sincere thanks to those who provided their valuable comments on the writing of this paper. Especially, we would like to thank the Oilfield Technology Research Institute of China Oilfield Services Co., Ltd., for providing the logging images.

References

[1] H. Saboorian-Jooybari, M. Dejam, Z. Chen, and P. Pourafshary, "Comprehensive evaluation of fracture parameters by dual laterolog data," *Journal of Applied Geophysics*, vol. 131, pp. 214–221, 2016.

[2] J. Gartner and J. Suau, "Fracture detection from well logs," *The Log Analyst*, vol. 2, 1980.

[3] S. X. Wu, H. T. Li, S. X. Long, Z. L. Liu, C. L. Wang, and J. T. Zhang, "A study on characteristic and diagenesis of carbonate reservoirs in the middle Triassic Leikoupo Formation in western Sichuan Depression," *Oil & Gas Geology*, vol. 32, no. 4, pp. 542–550, 2011.

[4] F. Dongjun, T. Zhu, and L. Hongtao, "Geological features and control factors of reservoir forming in middle Triassic Leikoupo Formation in western Sichuan Basin," *Journal of Xi'an Shiyou University (Natural Science Edition)*, vol. 28, no. 6, pp. 1–7, 2013.

[5] Y. Tang, "Characterization of the sedimentation of the Leikoupo Formation and the weathering crust reservoirs at the top of the formation in the western Sichuan Basin," *Oil & Gas Geology*, vol. 34, no. 1, pp. 42–47, 2013.

[6] Z. Xiangyuan, X. Hu, X. Kaihua, and J. Yuewei, "Characteristics and major control factors of natural fractures in carbonate reservoirs of Leikoupo Formation in Pengzhou area, western Sichuan Basin," *Oil & Gas Geology*, vol. 39, no. 1, pp. 30–39, 2018.

[7] J. W. Carey, Z. Lei, E. Rougier, H. Viswanathan, and H. Viswanathan, "Fracture-permeability behavior of shale," *Journal of Unconventional Oil & Gas Resources*, vol. 11, pp. 27–43, 2015.

[8] H. Sun, Y. Shi, and W. Zhang, "Time-domain modeling analysis of pulsed eddy current testing on ferromagnetic casing," *Review of Scientific Instruments*, vol. 91, no. 9, article 094702, 2020.

[9] C. A. Wright, E. J. Davis, G. M. Golich et al., "Downhole tiltmeter fracture mapping: finally measuring hydraulic fracture dimensions," in *SPE Western Regional Meeting*, Bakersfield, California, 1998.

[10] J. A. Quirein, J. Grable, B. Cornish, R. Stamm, and T. Perkins, "Microseismic fracture monitoring," in *SPWLA 47th Annual Logging Symposium*, Veracruz, Mexico, 2006.

[11] W. E. Kline, E. M. Kocian, and W. E. Smith, "Evaluation of cementing practices by quantitative radiotracer measurements," in *IADC/SPE Drilling Conference*, Dallas, Texas, 1986.

[12] O. Serra, *Formation Micro Scanner Image Interpretation*, Schlumberger Educational Service, Houston, Texas, 1989.

[13] K. W. Winkler and R. D'Angelo, "Ultrasonic borehole velocity imaging," *Geophysics*, vol. 71, no. 3, pp. F25–F30, 2005.

[14] H. Tian, J. Zhao, and S. Deng, "Application of ultrasonic imaging well logging technology in the casing well," *Journal of University of Jianghan Oil Worker*, vol. 11, 2005.

[15] J. Kherroubi, "Automatic extraction of natural fracture traces from borehole images," in *19th International Conference on Pattern Recognition (ICPR)*, pp. 978–981, Tampa, 2008.

[16] Z. Changchun and S. Ge, "A Hough transform-based method for fast detection of fixed period sinusoidal curves in images," in *International Conference on Signal Processing*, Beijing, China, 2002.

[17] F. Taiebi, G. Akbarizadeh, and E. Farshidi, "Detection of reservoir fractures in imaging logs using directional filtering," in *2017 5th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, pp. 150–154, Qazvin, Iran, 2017.

[18] Q. Liu, L. Xue, B. Pan et al., "Fracture detecting based on ant colony algorithm," *Global Geology*, vol. 16, no. 2, pp. 94–98, 2013.

[19] Y. Jianping, C. Jingong, S. Xiangyun, and C. Yumao, "Intelligent picking method of fracture information in imaging logging images," *Natural Gas Industry*, vol. 29, no. 3, pp. 51–53 +136, 2009.

[20] L. Yingming, Q. Yingshu, and F. Qingfu, "The automatic pickup method of imaging logging images showing sinusoidal geological structure," *Logging Technology*, vol. 37, no. 5, pp. 523–526, 2013.

[21] W. Wang, "An edge-based segmentation algorithm for rock fracture tracing," in *International Conference on Computer Graphics, Imaging and Visualization (CGIV'05)*, pp. 43–48, Beijing, China, 2005.

[22] X.-C. Yin, Q. Liu, H.-W. Hao, Z. B. Wang, and K. Huang, "FMI image based rock structure classification using classifier combination," *Neural Computing and Applications*, vol. 20, no. 7, pp. 955–963, 2011.

[23] S. Sorncharean and S. Phiphobmongkol, "Crack detection on asphalt surface image using enhanced grid cell analysis," in *4th IEEE International Symposium on Electronic Design, Test and Applications (delta 2008)*, pp. 49–54, Hong Kong, China, 2008.

[24] Z. Dongbing, "Research on asphalt pavement crack detection method based on clustering-minimum spanning tree," *Journal of Sun Yat-sen University: Natural Science Edition*, vol. 56, no. 4, pp. 68–74, 2017.

[25] F. Taibi, G. Akbarizadeh, and E. Farshidi, "Robust reservoir rock fracture recognition based on a new sparse feature learning and data training method," *Multidimensional Systems and Signal Processing*, vol. 6, 2019.

[26] W. Deng, J. Xu, X. Z. Gao, and H. Zhao, "An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020.

[27] W. Deng, H. Liu, J. Xu, H. Zhao, and Y. Song, "An improved quantum-inspired differential evolution algorithm for deep belief network," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7319–7327, 2020.

[28] Y. Cao, E. Yan, and D. Hu, "Calculation methods of rock mass discontinuity orientation measured by borehole camera technology and technology reliability," *Journal of China University of Geosciences: Earth Science*, vol. 39, no. 4, pp. 473–480, 2014.

[29] K. Glossop, P. J. G. Lisboa, P. C. Russell, A. Siddans, and G. R. Jones, "An implementation of the Hough transformation for the identification and labelling of fixed period sinusoidal curves," *Computer Vision and Image Understanding*, vol. 74, no. 1, pp. 96–100, 1999.

[30] A. Shahinpour, *Borehole image log analysis for sedimentary environment and clay volume interpretation, [M.S. thesis]*, Institutt for petroleumsteknologi og anvendt geofysikk, 2013.

[31] M. Seifallahi, B. Tokhmechi, A. Soleimani, and A. Ahmadi Fard, "A novel methodology for fracture extraction from borehole image logs," in *The First International Conference Oil, Gas, Petrochemical And Power Plant*, Tehran: SID, 2013.

[32] F. Khoshbakht, M. Mohammadnya, and A. M. Bagheri, "Fractures analysis and identify stress in hydrocarbon reservoirs using imageing logs," in *The Third Conference on Iranian Rock Mechanics*, University of Amir Kabir, 2007.

[33] S. A. Wong, *Automatic Feature Detection from Wellbore Images*, Texas A & M University. Libraries, 1989.

[34] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems Man & Cybernetics*, vol. 9, no. 1, pp. 62–66, 2007.

[35] T. Pun, "A new method for grey-level picture thresholding using the entropy of the histogram," *Signal Processing*, vol. 2, no. 3, pp. 223–237, 1980.

[36] E. Shi, M. Chen, J. Li, and Y. Huang, "Research on global path planning method of mobile robot based on ant colony algorithm," *Transactions of the Chinese Society of Agricultural Machinery*, vol. 45, no. 6, pp. 53–57, 2014.

[37] J. Ming, F. Wang, G. Yuan, and S. Longlong, "Research on path planning of mobile robot based on improved ant colony algorithm," *Chinese Journal of Scientific Instrument*, vol. 40, no. 2, pp. 113–121, 2019.

Hindawi
Journal of Sensors
Volume 2021, Article ID 6647093, 9 pages
https://doi.org/10.1155/2021/6647093

Research Article

Prediction of Inhomogeneous Stress in Metal Structures: A Hybrid Approach Combining Eddy Current Technique and Finite Element Method

Yating Yu,[1,2] **Fei Yuan**,[1] **Hanchao Li**,[1] **Cristian Ulianov**,[3] **and Guiyun Tian**[3]

[1]*School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, 611731 Chengdu, China*
[2]*Institute of Electronic and Information Engineering of UESTC in Guangdong, 523808 Dongguan, China*
[3]*School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, UK*

Correspondence should be addressed to Yating Yu; yuyating-uestc@hotmail.com

Received 25 October 2020; Accepted 15 June 2021; Published 7 July 2021

Academic Editor: Fanli Meng

Concentrated stresses and residual ones are critical for the metal structures' health, because they can cause microcracks that require emergency maintenance or can result in potential accidents. Therefore, an accurate approach to the measurement of stresses is key for ensuring the health of metal structures. The eddy current technique is an effective approach to detect the stress according to the piezoresistive effect. However, it is limited to detect the surface stress due to the skin effect. In engineering, the stress distribution is inhomogeneous; therefore, to predict the inhomogeneous stress distribution, this paper proposes a nondestructive approach which combines the eddy current technique and finite element (FE) method. The experimental data achieved through the eddy current technique determines the relationship between the applied force and the magnetic flux density, while numerical simulations through the FE method bridge the relationship between the magnetic flux density and the stress distribution in different directions. Therefore, we can predict the inhomogeneous stress nondestructively. As a case study, the applied stress in a three-point-bending simply supported beam was evaluated, and the relative error is less than 8% in the whole beam. This approach can be expected to predict the residual stress in metal structures, such as rail and vehicle structures, if the stress distribution pattern is known.

1. Introduction

Stress concentration, or residual stress, is the main cause of the micro cracks in metal components and structures (such as oil/gas pipeline [1, 2], airfoil [3], steel bridge [4], hoisting equipment [5], and polycrystalline solids [6]). Consequently, the microcracks in such key components and subassemblies can suddenly cause structure failure when operating under alternating loads; this may lead to economic and environmental losses, as well as human casualties [7, 8], if there is no appropriate maintenance strategy. Therefore, the evaluation approach to the stress distribution is essential to control the early stage quality and to extend the lifetime by structural health monitoring. However, it is typically difficult to measure or predict the residual stress [9].

Non-Destructive Testing (NDT) techniques play a vital role in assuring safety and serviceability of a variety of key infrastructure assets and facilities [10–24]. The commonly applied techniques in stress or residual stress measurement are X-ray diffraction, ultrasonic testing, magnetic memory method (MMM), magnetic Barkhausen noise (MBN), and eddy current testing (ECT) [11–19]. The X-ray diffraction technique is suitable for the surface stress measurement [6]; however, it is not efficient for measuring the stress in depth. In engineering, it is usually combined with other destructive techniques to extend the measurement depth to overcome this problem [11]. The ultrasonic technique has a higher measurement depth than the X-ray diffraction technique; however, it has high requirements on the material surface and its measurement accuracy is low [12]. MMM utilizes

the geomagnetic field and magnetostrictive effect to detect the stress or residual stress; therefore, it is suitable for the online and the real-time measurement, but its detection signal is too weak to easily be interfered by the environment noise, so the detection accuracy is not very high [13, 14]. Barkhausen noise can measure the surface and subsurface stress but is limited to magnetic material, because the Barkhausen effect only exists in the ferrite material [15, 16]. In addition to this, for the crack bridging stress, Greene et al. adopted the Raman microprobe technique to measure bridging stresses for fatigue samples. However, the calibration is necessary because the Raman shift can be affected by changes in chemical composition [17]. The eddy current technique has a low cost and low sensitivity to environmental influence (e.g., moisture and dust) and is suitable for quantitative surface/subsurface measurements (e.g., crack detection [18–21], magnetic permeability and electrical conductivity characterization [22], and displacement measurement [23]) for the conductive material. Due to the piezoresistive effect of conductive material, the eddy current technique appears to be more advantageous over other techniques for stress evaluation in metal structures. Some investigations [24] indicated that the eddy current response is sensitive to stress changes in metals.

Therefore, more and more research works for stress measurement are focused on the use of the eddy current technique. In 2001, Ricken et al. employed a Giant Magneto Resistive (GMR) sensor and an eddy current sensor to characterize the axial stress of steel wire. The investigations indicate that the magnetic flux density detected by the GMR sensor is related to the applied load, while the resistance and the inductance of the eddy current sensor are a function of the mechanical stress [25]. However, the plasticity [24, 25] and aging/heat treatment [26] also affect the electrical conductivity of metals as well as applied stress. Therefore, they have a combined influence on the eddy current response. Morozov and Tian [27] deeply investigated the eddy current response of samples of aluminum alloys (AA-1050, AA-2024, AA-5083, and AA-7075) with different levels of plastic deformation and different heat treatments. The aluminum alloys subjected to elastic uniaxial loading were monitored by the circular and directional eddy current probes. The experimental results indicated that the stress coefficients are generally positive and depend on the annealing (heat treatment) condition as well as the level of prior plastic work; the rectangular probe is much more sensitive when oriented normally to the tensile stress. After that, they continued to study the response of the eddy current on the change of electrical conductivity due to the elastic and plastic stress. They pointed out that the real part of the EC response is sensitive mainly to elastic stress, while the imaginary part of the EC response is sensitive mainly to plastic stress [28]. Zhou et al. [29] proposed a pulse electromagnetic method (PEM) with a U-shaped sensor to detect the unidirectional tension stress in ferromagnetic metals.

The existing literatures illustrate that the eddy current technique is an effective nondestructive technique for measuring stress. The stress distribution is generally inhomogeneous under the complex load in manufacturing and in service. However, only limited literatures considered the inhomogeneous stress. Nagy et al. [30–32] proposed the residual stress assessment for nickel-based superalloys after the shot-peened processing by the eddy current technique. While their investigation also discovered that the relationship between the electrical conductivity profile and the sought residual stress profile is very sensitive to the sample's state of precipitation hardening [33] and thermoplastic effect [34]. Furthermore, the eddy current can detect the surface stress due to the skin effect. Therefore, the prediction accuracy for the inhomogeneous stress should be improved further by combining other methods. Ahn et al. [35] combine the X-ray diffraction and FEM together to the prediction of welding residual stresses in fibre laser-welded AA2024-T3, while Zhu et al. [36] proposed an approach to determine the residual stress in metal by combining slot milling method and finite element approach, and the determination result by the proposed method is verified by the X-ray diffraction. However, the slot milling method is a destructive method while X-ray diffraction is expensive and is harmful for the operator's health.

Therefore, a prediction approach is proposed by combination the eddy current technique and finite element method (FEM) to evaluate the inhomogeneous distribution stress in metals accurately and nondestructively. The paper is structured as follows: the fundamental theories for the stress evaluation by the eddy current technique are presented in Section 2; the inhomogeneous stress distribution evaluation approach based on the eddy current technique and FE numerical simulation is proposed in Section 3; a case study, i.e., measurement of stress distribution in a simply supported beam with a three-point-bending deformation, is shown in Section 4. Section 5 draws conclusions and makes recommendations for future work.

2. Fundamental Theory

2.1. Piezoresistive Effect. The geometry and resistivity of metals change when mechanical loads are applied and, consequently, cause the change of the metal's resistance. The change of material's resistivity due to applied loads is called the piezoresistive effect. Considering the cuboid metal as an example, its resistance can be expressed as

$$R = \frac{\rho \cdot l}{w \cdot th}, \tag{1}$$

where ρ is the resistivity of the metal (Ωm), l is the length of the cuboid (mm), w is the width of the cuboid (mm), and th is the thickness of the cuboid (mm).

According to the piezoresistive effect, if the metal is stretched in length, ρ, l, w, and th are all changed. The change of the resistance of cuboid metal due to the changes of ρ, l, w, and th can be expressed as

$$\frac{\mathrm{d}R}{R} = \frac{\mathrm{d}l}{l} + \frac{\mathrm{d}\rho}{\rho} - \frac{\mathrm{d}w}{w} - \frac{\mathrm{d}th}{th}. \tag{2}$$

Assuming that the strain along the length is ε, which

equals to dl/l, the strain in width and thickness can be expressed as

$$\begin{cases} \dfrac{\mathrm{d}w}{w} = -\nu\varepsilon, \\ \dfrac{\mathrm{dth}}{\mathrm{th}} = -\nu\varepsilon, \end{cases} \tag{3}$$

where ν is the Poisson's ratio of the metal.

Therefore, Equation (2) can be simplified to

$$\frac{\mathrm{d}R}{R} = (1 + 2\nu) \cdot \varepsilon + \frac{\mathrm{d}\rho}{\rho}, \tag{4}$$

where $(1 + 2\upsilon).\varepsilon$ is the change in resistance due to the change of the geometry of the cuboid and $d\rho/\rho$ is the change in resistance due to the piezoresistive effect.

Therefore, the resistance change in metals stems from the change of their geometry and the change of the resistivity resulting from the applied mechanical stress. For some metals, such as platinum alloy, the resistance change due to piezoresistivity is much larger than that due to geometry change.

2.2. Operation Principle of Eddy Current Method for Stress Measurement. The operation principle of the eddy current testing instrument is shown in Figure 1. An alternating current I in the driving coil creates an alternating magnetic field H_1, which is the primary magnetic field and induces current I_2 in the sample. The eddy currents simultaneously generate a secondary magnetic field H_2, which resists the variation of the primary magnetic field and changes the resultant magnetic field H.

The geometric parameters of coil, such as the number of turns N, the inner radius r_1, the outer radius r_2, and the height h_c, are key factors to the primary magnetic field H_1; the lift-off l, sample's electrical conductivity σ_{ele}, and sample's relative magnetic permeability μ_r affect the secondary magnetic field H_2; the excitation current I and the excitation angular frequency ω have influence on the primary magnetic field H_1 as well as on the secondary magnetic field H_2.

Therefore, H is dependent on factors such as the lift-off, excitation frequency, sample electrical conductivity, sample relative magnetic permeability, and probe coil geometry. The Z-component of magnetic flux density B_z is commonly used as a detection signal due to its detectability and strength. Therefore, B_z can be expressed as Equation (5)

$$B_z \sim B_z(N, I, r_1, r_2, l, h_c, \sigma_{ele}, \mu_r, \omega). \tag{5}$$

According to Equations (4) and (5), the applied mechanical stress can cause the change of electrical conductivity of the sample due to geometry change and piezoresistivity effect. Consequently, the change of sample's electrical conductivity can induce the change of resultant magnetic flux density. Therefore, the eddy current technique can theoretically reflect the applied stress in metals.

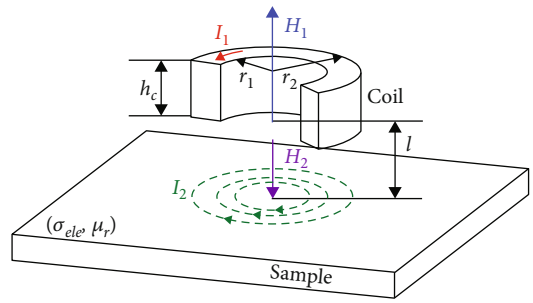


Figure 1: The operation principle of the eddy current technique.

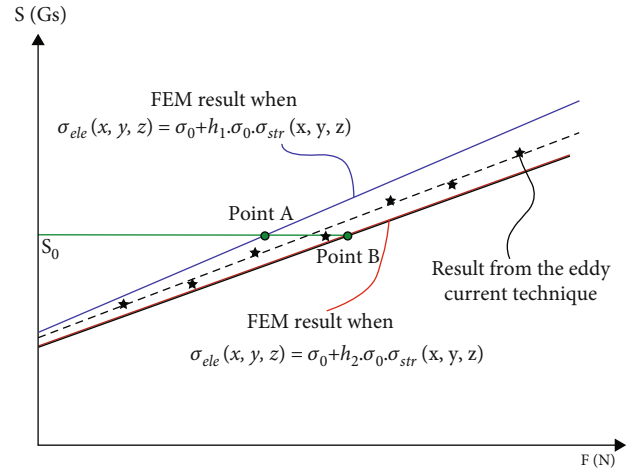


Figure 2: Schematic graph for the approach to measuring of inhomogeneous stress.

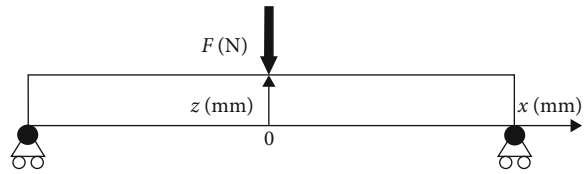


Figure 3: Three-point-bending simply supported beam.

The actual size of the sample is much larger than the distribution range of the magnetic field in the eddy current method, so the stress detected by the eddy current technique can reflect the piezoresistivity effect accurately in metals.

3. Hybrid Approach for Inhomogeneous Distribution Stress

Ricken et al. [25] investigated the relationship of between the resistance and inductance of the coil of the eddy current sensor and the stress, and the results indicated that the impedance of the coil is linear changing with the stress. Therefore, the linear relationship between the eddy current detection signal S and the applied force F is assumed as Equation (6), which can be symbolically presented as the dotted line in Figure 2.

In Figure 2, ★ represents symbolically the data from the experiment by the eddy current technique. The coefficient

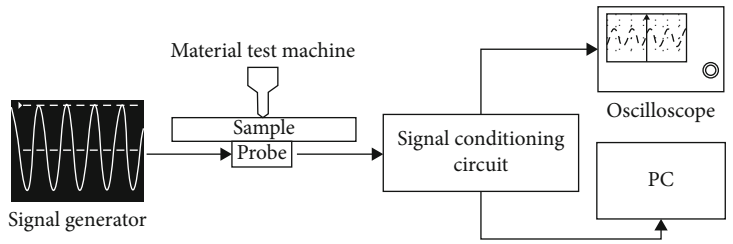


Figure 4: Schematic diagram of the experiment setup.

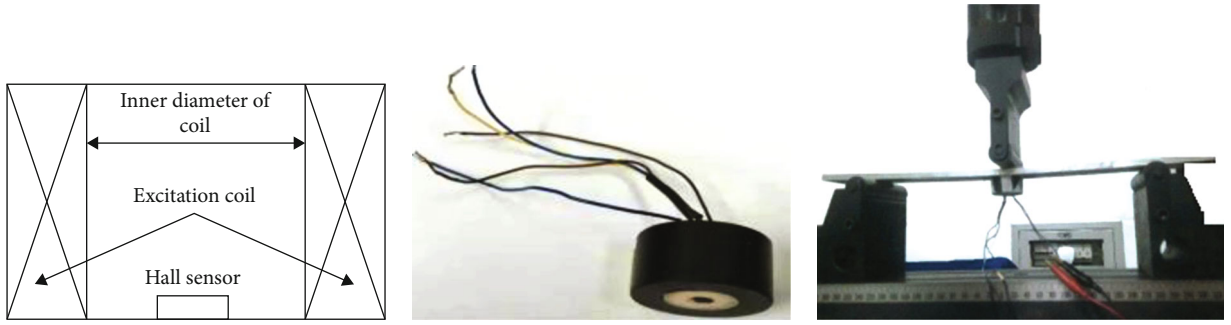


Figure 5: The probe (a) the structural diagram of probe, (b) the real probe, and (c) the location of the probe in the experiment.

of determination of the dotted line with the experiment data is denoted as R_{s0}.

$$S = KF + a, \tag{6}$$

where S is the detection signal of the eddy current technique, F is the applied force, and K and a are the slope and the intercept of the linear relationship between S and F, respectively, which can be determined by experiment.

According to the mechanics of material, when the force F is applied on the conductive sample, the generated internal force can be calculated as $G(F)$, and the stress distribution on the cross-section of the conductive sample can be generalized as

$$\sigma_{\text{str}}(x, y, z) = \frac{G(F) \cdot C(x, y, z)}{P_{\text{geom}}(w, l, th)}, \tag{7}$$

where $C(x, y, z)$ are the coordinates of the detection point in x-axis, in y-axis, and in z-axis; $P_{\text{geom}}(w, l, \text{th})$ is constant when the geometry of the sample is determined; w, th is width and height of the sample cross-section, and l is the length of the sample.

Gong et al. [37] found that the change of resistivity of the conductive sample is linear with the change of strain of the sample. According to the constitutive relationship of stress and strain, the electrical conductivity in the cross-section $\sigma_{\text{ele}}(x, y, z)$ can be assumed as

$$\sigma_{\text{ele}}(x, y, z) = \sigma_0 + h \cdot \sigma_0 \cdot \sigma_{\text{str}}(x, y, z), \tag{8}$$

where σ_0 is the electrical conductivity of the sample without stress and h is a conversion coefficient between the stress and electrical conductivity of the sample.

Table 1: The influence of the applied force on the signal from the hall sensor.

Applied force (N)	Voltage from the hall sensor (mV)	B_z (Gs)
40	4080.828	70.58
80	4085.593	71.59
120	4090.988	72.74
160	4093.198	73.22
200	4096.094	73.83
240	4102.996	75.31

As shown in Figure 2, the range of coefficient h is $h_2 < h < h_1$, where h_1 is the upper limit and h_2 is the lower limit. The coefficient h_2 can be obtained according to the following steps:

Step 1. $n = 1$.

Step 2. $h = n \times h_0$, where h_0 is a tiny positive value.

Step 3. Set the conductivity distribution in the numerical model according to Equation (8).

Step 4. Obtain the relationship of $F \sim S$, and calculate its coefficient of determination R_s.

Step 5. If $|R_s - R_{s0}| < \lambda$ (λ is a tiny positive value), we denote the $h_2 = h$; if no, set n as $n + 1$, and turn to Step 2.

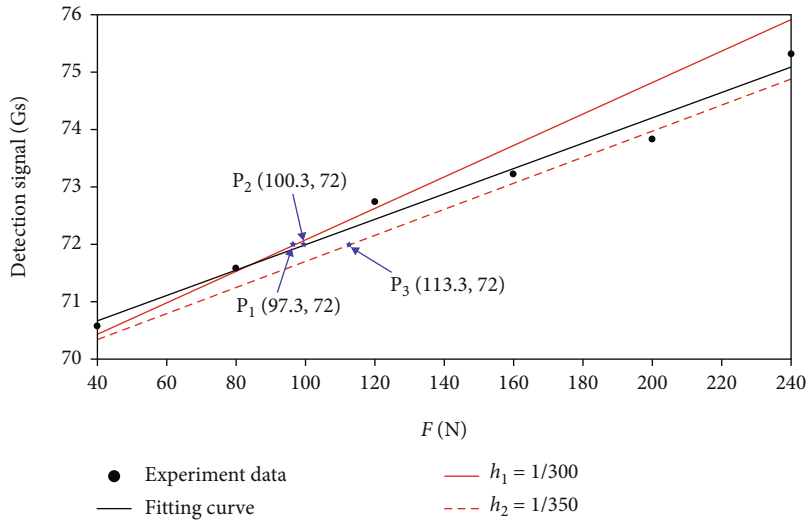


Figure 6: Prediction of the coefficient h by experimental study.

In order to get the upper limit h_1, we can represent $h = n \times h_0$ in step 2 as $h = h_1 + n \times h_0$, as well as represent $h_2 = h$ in step 5 as $h_1 = h$. Therefore, the range of coefficient h can be determined as $h_2 < h < h_1$.

As shown in Figure 2, if we know the detection signal is S_0 , we can get the point A and point B. Point A is the intersection point of S_0 and the line $h = h_1$ while Point B is the intersection point of S_0 and line $h = h_2$. Therefore, we can get the applied forces for point A and point B, and we denote the applied force for point A is F_1 and denote that for point B is F_2. After that, we obtained the applied force F_m as the average value of F_1 and F_2. Finally, we can get the stress distribution according to Equation (7).

As illustrated above, in Step 3, we input the conductivity distribution in the numerical model according to Equation (8) and can get the detection signals under different applied loads. Then, in Step 4 we can obtain the relationship of $F \sim S$ and calculate the coefficient of determination R_s between the $F \sim S$ and experimental data from the eddy current technique. Therefore, the electromagnetic data from the eddy current technique and the mechanical model by numerical simulation are combined.

4. Case Study

In the case study, the continuous and inhomogeneous stress distribution in a simply supported beam for three-point-bending deformation is considered. The beam material is aluminum alloy 7075 and its size is 250 mm × 35 mm × 6 mm. The sketch of three-point-bending experiment is shown in Figure 3.

4.1. Experimental Investigation on Relationship of $F \sim B_z$

4.1.1. Experimental Setup. The experimental setup (shown in Figure 4) includes the power supply for the eddy current probe, the signal generator, the signal conditioning circuit, aluminum alloy 7075, signal acquisition system, and material test machine (SANS CM75105). The probe consists of the excitation coil and the hall sensor SS495A. The Hall sensor SS495A is a linear magnetic sensor, which can reflect B_z by the output voltage signal linearly. The structural diagram of the probe and the real probe is shown in Figure 5. In the experiment, the probe is located below the applied force, as shown in Figure 5(c), and a 10 V voltage signal of 5 kHz is input the probe as the excitation signal. The impedance of the coil can increasing and then decreasing with the excitation frequency in range of 10 Hz~10 MHz [36], while when the excitation frequency is too low, the skin depth is deep; the stress we detect is a resultant stress in the range of skin depth instead of the surface stress; therefore, 0.5 kHz is select as the excitation frequency.

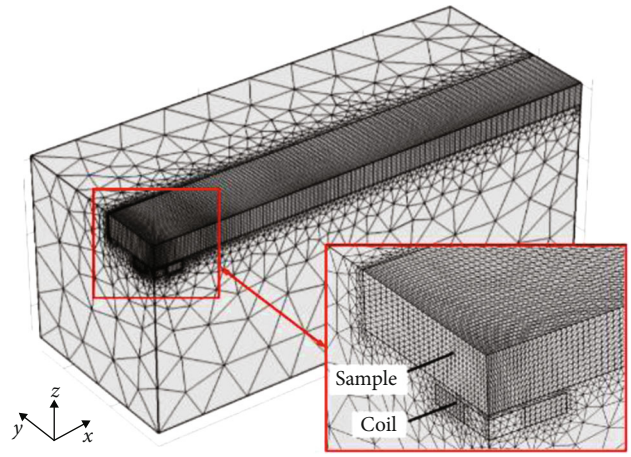


Figure 7: The FE model.

To ensure the elastic deformation in the experiment, the maximal stress of the sample is required to be below the admissible stress. The admissible stress of aluminum alloy 7075 is 200 MPa [38]. Therefore, the maximal force from the material test machine applied on the sample was calculated as 280 N. Therefore, in the experiment, the load from the material test machine varies in the range from 0 N to 240 N, with steps of 40 N.

Table 2: The geometrical and electrical parameters in the FE model.

Coil	Value	Sample	Value	Air	Value
Inner radius (mm)	3.5	Height (mm)	6	Height (mm)	100
Outer radius (mm)	7.5	Width (mm)	35	Width (mm)	100
Height (mm)	2	Length (mm)	250	Length (mm)	250
Electrical conductivity (MS/m)	55.85	Electrical conductivity (MS/m)	Equation (11)	Lift-off (mm)	0.5

4.1.2. Analysis of Experimental Results. Table 1 lists the voltage signal from the Hall sensor for the different applied forces. The hall sensor SS495 in the experiment is a type of linear hall sensor. The output voltage is linearly changed with the magnetic flux density in the range of -640 Gs ~ +640 Gs. Therefore, the magnetic flux density can be calculated for different applied forces.

As can be seen in Table 1, the voltage signal and B_z are increasing with the applied force. The potential reason is that, with the increasing applied force, the stress in the beam is linearly increasing, thereafter the resistivity of the sample is increasing and the conductivity decreases. The decreasing conductivity can result in the increasing resultant magnetic flux density, according to the principle of the eddy current technique shown in Figure 1. The stronger the B_z, the stronger the voltage signal from hall sensor.

The relationship of $F \sim B_z$ is shown in Figure 6 (black full line). The expression of $B_z \sim F$ is fitted as Equation (9)

$$B_z = 0.022F + 69.79, \tag{9}$$

where the coefficient of determination R_{s0} is 0.98.

4.2. Determination of Conductivity Distribution in the Cross-Section of the Beam. For the three-point-bending simply supported beam shown in Figure 3, the stress at the point (x, z) can be expressed as

$$\sigma_{str}(x, z) = \frac{l}{4Iz} F \cdot z - \frac{1}{2Iz} F \cdot x \cdot z, \tag{10}$$

where l is the length of the beam and I_z is the inertia moment.

According to Equation (9), the electrical conductivity distribution in the three-point-bending simply supported beam can be expressed as

$$\sigma_{\text{ele}} = \sigma_0 + h \cdot \sigma_0 \cdot \sigma_{\text{str}} = \sigma_0 \left(1 + \frac{l}{4I_z} \text{hFz} - \frac{1}{2I_z} \text{hFxz}\right), \tag{11}$$

where h is the conversion coefficient about the stress and electrical conductivity of sample, which is determined by the numerical simulation in Section 4.3.

4.3. Determination of Coefficient h by Finite Element Method

4.3.1. Numerical Modelling by Finite Element Method. According to the symmetry in geometry and boundary conditions, a 1/4 structural model was adopted by COMSOL Multiphysics, one of the typical commercial softwares for FEM. The model was constructed according to the mesh strategy and boundary conditions investigated in [39] and shown in Figure 7. The parameters in the finite element (FE) model are listed in Table 2.

4.3.2. The Coefficient h. Considering the electrical conductivity distribution from Equation (11) in the FE model, the relationship curve between the applied force and the magnetic flux density in z-component for the different h_i coefficients can be obtained. If $\lambda = 0.01$, it will result in $h_1 = 1/300$ and $h_2 = 1/350$. The relationship curve between the applied force and the B_z for $h = h_1$ and $h = h_2$ is shown in Figure 6.

4.4. Applied Stress Evaluation for the Three-Point-Bending Simply Supported Beam. In three-point-bending deformation, if B_z detected by the eddy current probe shown in Figure 5 is 72 Gauss, the applied force can then be inversely calculated as $F_1 = 97.3$ N when $h = h_1$ and $F_2 = 113.3$ N when $h = h_2$, as shown in Figure 6. Therefore, the measured applied force can be expressed as F_m as Equation (12)

$$Fm = \frac{F1 + F2}{2} = 105.3\text{N}. \tag{12}$$

Furthermore, according to Figure 6, the calibration applied force F_c is 100.3 N when $B_z = 72$ Gauss. After obtaining the applied force, the stress at any point $P(x, y, z)$ in the structure can be evaluated according to Equation (10).

When the magnetic flux density B_z detected by the hall sensor is 72 Gauss, the continuous stress distribution is calculated with Equation (10) according to the measured applied force $F_m = 105.3$ N in Equation (12) and shown in Figure 8(a). Figure 8(b) shows the stress distribution under the calibration applied force $F_c = 100.3$ N calculated by FEM. Figure 8(c) shows the absolute error of the continuous stress distribution in Figures 8(a) and 8(b).

As can be seen from Figure 8, besides the locations of the applied force and the supports, the absolute error in the whole simply supported beam is lower than 4 N/mm2 and the relative error is lower than 8%. However, the absolute error at the locations of the applied force and the supports is obvious because the stress concentration effect is taken into account in FEM, while it is not considered in the proposed method. The potential reason for the relative error of 8% in the proposed approach is the stress distribution in experiment has little difference with Equation (7), because the sample is not strictly the continuous, homogeneous, and isotropous solid.

The case study indicated that the proposed method can predict the inhomogeneous stress distribution in the simply supported beam of the metal structures by FEM according to the surface stress detected by the eddy current sensor.

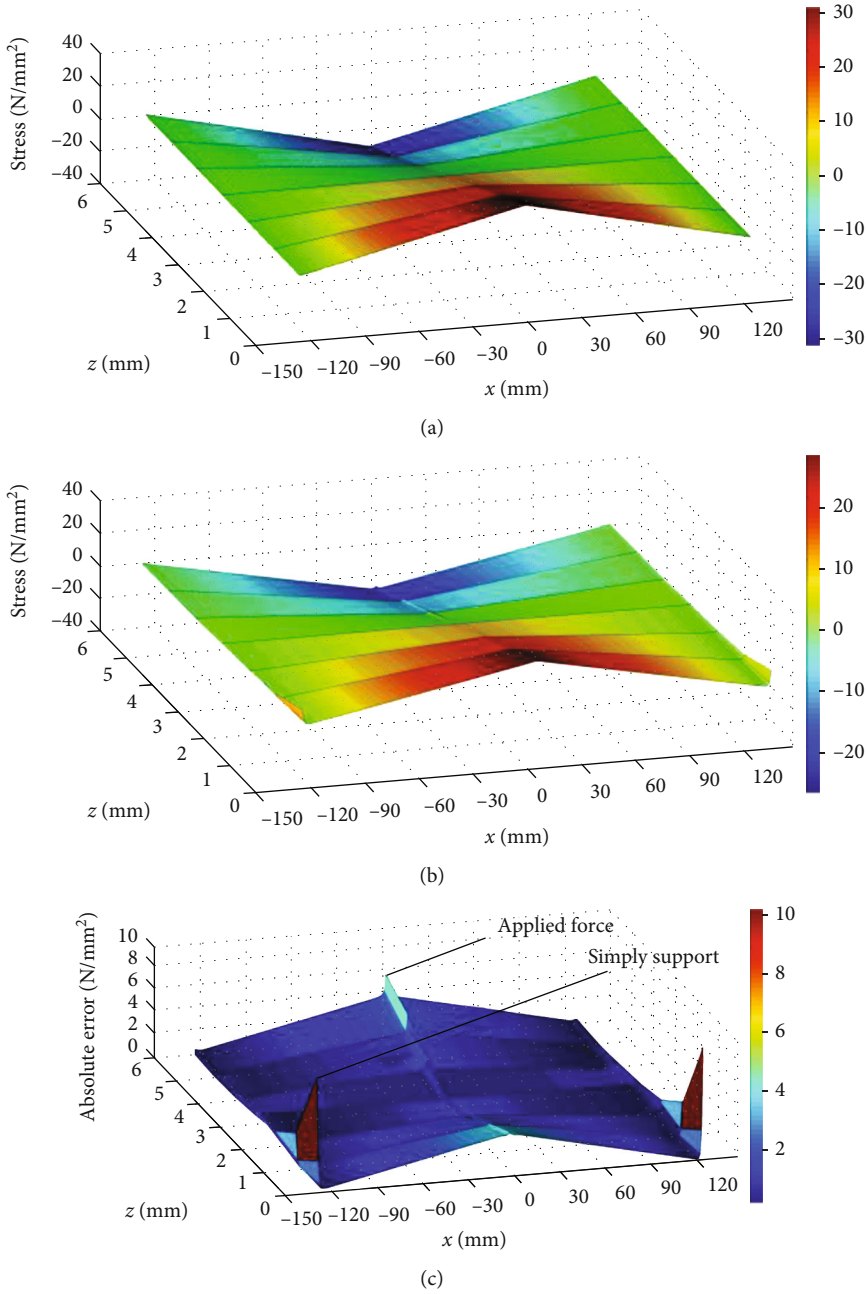


FIGURE 8: Continuous stress distribution (a) Stress distribution for $F_m = 105.3$ N calculated with Equation (10). (b) Stress distribution for $F_c = 100.3$ N by FEM. (c) The absolute error of the stress distribution between $F_c = 100.3$ N and $F_m = 105.3$ N when $B_z = 72$ Gauss.

5. Conclusions

This paper presented a hybrid approach combining the eddy current technique and FE method, which uses the piezoresistive effect, to predict the stress distribution in early stage damages in metal structures. The surface stress can be obtained through the eddy current technique, while the FEM method can describe the relationship between surface stress and in-depth stress, so that the inhomogeneous stress can be predicted. The main conclusions are drawn below.

(1) The results of the experiment indicate that the detection signal of the eddy current technique linearly changes with the applied force on the metal structures for the bending deformation, which is consistent with the conclusion in [28, 36]

(2) A new hybrid approach to determine the coefficient h between the stress and the electrical conductivity is proposed. The coefficient h bridges the gap between the magnetic flux density detected through the eddy current technique and the inhomogeneously distributed stress, which is key for inhomogeneous stress evaluation. It also provides an approach to approximate the piezoresistive coefficient of unknown material without damage

(3) The case study shows that the proposed approach can readily measure the inhomogeneous stress distribution in bending deformation with high accuracy, which provides an example for the application of the proposed approach for inhomogeneous stress measurement under other conditions. Such as if the distribution pattern of the residual stress is known under certain conditions, this approach can be further extended to evaluate residual stress under specific working conditions for key structures, such as high-speed rail, oil/gas pipelines, airfoils, and rail and road vehicle structures

This proposed approach provides a feasible approach to predict the inhomogeneous stress or even for residual stress. However, the influence of the sample surface, material electromagnetic features, and the excitation frequency on the evaluation accuracy and the sensitivity still need to be further investigated.

Data Availability

The [DATA TYPE] data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 61960206010, Grant No. 51675087), the National Nature Science Foundation of Guangdong Province (Grant No. 2018A030313893), and also the Fundamental Research Funds for the Central Universities (Grant No. ZYGX2018J067).

References

[1] L. R. Lothhammer, M. R. Viotti, A. Albertazzi, and C. L. N. Veiga, "Residual stress measurements in steel pipes using DSPI and the hole-drilling technique," *International Journal of Pressure Vessels and Piping*, vol. 152, pp. 46–55, 2017.

[2] S. Taheri and A. Fatemi, "Fatigue crack behavior in power plant residual heat removal system piping including weld residual stress effects," *International Journal of Fatigue*, vol. 101, pp. 244–252, 2017.

[3] D. M. Spasic, S. N. Stupar, A. M. Simonovic, D. Trifkovic, and T. D. Ivanov, "The failure analysis of the star-separator of an aircraft cannon," *Engineering Failure Analysis*, vol. 42, pp. 74–86, 2014.

[4] W. M. Chen, L. Liu, P. Zhang, and S. R. Hu, "Non-destructive measurement of the steel cable stress based on magneto-mechanical effect," *Health Monitoring of Structural and Biological Systems*, vol. 7650, 2010.

[5] D. K. Zhang, S. R. Ge, and Y. Qiang, "Research on the fatigue and fracture behavior due to the fretting wear of steel wire in hoisting rope," *Wear*, vol. 255, no. 7-12, pp. 1233–1237, 2003.

[6] K. Bobzin, W. Wietheger, M. A. Knoch et al., "Comparison of residual stress measurements conducted by X-ray stress analysis and incremental hole drilling method," *Journal of Thermal Spray Technology*, vol. 29, no. 6, pp. 1218–1228, 2020.

[7] Z. Wang, F. Chen, P. Li, and Y. Liu, "Research on the effect of material performances ε f and $\Delta\varepsilon c$ on the strain fatigue life predication accuracy," *International Journal of Damage Mechanics*, vol. 22, no. 5, pp. 737–751, 2013.

[8] F. Y. Wang, K. M. Mao, and B. Li, "Prediction of residual stress fields from surface stress measurements," *International Journal of Mechanical Sciences*, vol. 140, pp. 68–82, 2018.

[9] P. J. Withers and H. K. D. H. Bhadeshia, "Residual stress. Part 1 – measurement techniques," *Metal Science Journal*, vol. 17, no. 4, pp. 355–365, 2001.

[10] P. Zhang, Y. D. Tan, W. X. Liu, and W. X. Chen, "Methods for optical phase retardation measurement: a review," *Science China Technological Sciences*, vol. 56, no. 5, pp. 1155–1164, 2013.

[11] A. Alawadi and H. Abdolvand, "Measurement and modeling of micro residual stresses in zirconium crystals in three dimension," *Journal of the Mechanics and Physics of Solids*, vol. 135, p. 103799, 2020.

[12] H. Kim, T. Kim, D. Morrow, and X. Jiang, "Stress measurement of a pressurized vessel using ultrasonic subsurface longitudinal wave with 1-3 composite transducers," *IEEE Transactions on Ultrasonics Ferroelectrics and Frequency Control*, vol. 67, pp. 158–166, 2019.

[13] H. H. Huang, J. Y. Yao, Z. W. Li, and Z. F. Liu, "Residual magnetic field variation induced by applied magnetic field and cyclic tensile stress," *NDT & E International*, vol. 63, pp. 38–42, 2014.

[14] J. W. Wilson, G. Y. Tian, and S. Barrans, "Residual magnetic field sensing for stress measurement," *Sensors And Actuators A-Physical*, vol. 135, no. 2, pp. 381–387, 2007.

[15] V. Vengrinovich, D. Vintov, A. Prudnikov, P. Podugolnikov, and V. Ryabtsev, "Magnetic Barkhausen effect in steel under biaxial strain/stress: influence on stress measurement," *Journal of Nondestructive Evaluation*, vol. 38, no. 2, pp. 1–8, 2019.

[16] P. Wang, X. L. Ji, X. M. Yan et al., "Investigation of temperature effect of stress detection based on Barkhausen noise," *Sensors and Actuators A-Physical*, vol. 194, pp. 232–239, 2013.

[17] R. B. Greene, S. Gallops, S. Funfschilling et al., "A direct comparison of non-destructive techniques for determining bridging stress distributions," *Journal of the Mechanics and Physics of Solids*, vol. 60, no. 8, pp. 1462–1477, 2012.

[18] Y. T. Yu and J. Guan, "Investigation of signal features of pulsed eddy current testing technique by experiments," *Insight*, vol. 55, no. 9, pp. 487–492, 2013.

[19] Y. T. Yu, Y. Yan, F. Wang, G. Y. Tian, and D. J. Zhang, "An approach to reduce lift-off noise in pulsed eddy current nondestructive technology," *NDT & E International*, vol. 63, pp. 1–6, 2014.

[20] Q. Ma, B. Gao, G. Y. Tian, C. Yang, L. Xie, and K. Chen, "High sensitivity flexible double square winding eddy current array for surface micro-defects inspection," *Sensors and Actuators A: Physical*, vol. 309, p. 111844, 2020.

[21] Z. W. Liu, B. Gao, and G. Y. Tian, "Natural crack diagnosis system based on novel L-shaped electromagnetic sensing thermography," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 11, pp. 9703–9714, 2020.

[22] D. Desjardins, T. W. Krause, and L. Clapham, "Transient eddy current method for the characterization of magnetic permeability and conductivity," *NDT & E International*, vol. 80, pp. 65–70, 2016.

[23] Y. T. Yu, T. Yang, and P. A. Du, "A new eddy current displacement measuring instrument independent of sample electromagnetic properties," *NDT & E International*, vol. 48, pp. 16–22, 2012.

[24] L. M. Li, L. Q. Zhong, and X. Chen, "Residual stress caused magnetic field abnormal change upon arc welding joints," *International Journal of Applied Electromagnetics and Mechanics*, vol. 33, no. 3-4, pp. 1295–1301, 2010.

[25] W. Ricken, J. Liu, and W. J. Becker, "GMR and eddy current sensor in use of stress measurement," *Sensors and Actuators A-Physical*, vol. 91, no. 1-2, pp. 42–45, 2001.

[26] M. J. Starink and X. M. Li, "A model for the electrical conductivity of peak-aged and overaged Al-Zn-Mg-Cu alloys," *Metallurgical and Materials Transactions A-Physical Metallurgy And Materials Science*, vol. 34a, no. 4, pp. 899–911, 2003.

[27] M. Morozov, G. Y. Tian, and P. J. Withers, "The pulsed eddy current response to applied loading of various aluminium alloys," *NDT & E International*, vol. 43, no. 6, pp. 493–500, 2010.

[28] M. Morozov, G. Y. Tian, and P. J. Withers, "Elastic and plastic strain effects on eddy current response of aluminium alloys," *Nondestructive Testing and Evaluation*, vol. 28, no. 4, pp. 300–312, 2013.

[29] D. Q. Zhou, M. Pan, Y. Z. He, and B. L. Du, "Stress detection and measurement in ferromagnetic metals using pulse electromagnetic method with U-shaped sensor," *Measurement*, vol. 105, pp. 136–145, 2017.

[30] M. P. Blodgett and P. B. Nagy, "Eddy current assessment of near-surface residual stress in shot-peened nickel-base superalloys," *Journal of Nondestructive Evaluation*, vol. 23, no. 3, pp. 107–123, 2004.

[31] F. Yu and P. B. Nagy, "Simple analytical approximations for eddy current profiling of the near-surface residual stress in shot-peened metals," *Journal of Applied Physics*, vol. 96, no. 2, pp. 1257–1266, 2004.

[32] F. Yu, M. P. Blodgett, and P. B. Nagy, "Eddy current assessment of near-surface residual stress in shot-peened inhomogeneous nickel-base superalloys," *Journal of Nondestructive Evaluation*, vol. 25, no. 1, pp. 17–28, 2006.

[33] B. A. Abu-Nabah, W. T. Hassan, D. Ryan, M. P. Blodgett, and P. B. Nagy, "The effect of hardness on eddy current residual stress profiling in shot-peened nickel alloys," *Journal of Nondestructive Evaluation*, vol. 29, no. 3, pp. 143–153, 2010.

[34] F. Yu and P. B. Nagy, "Dynamic piezoresistivity calibration for eddy current nondestructive residual stress measurements," *Journal of Nondestructive Evaluation*, vol. 24, no. 4, pp. 143–151, 2005.

[35] J. Ahn, E. He, L. Chen et al., "FEM prediction of welding residual stresses in fibre laser-welded AA 2024-T3 and comparison with experimental measurement," *International Journal of Advanced Manufacturing Technology*, vol. 95, no. 9-12, pp. 4243–4263, 2018.

[36] R. Zhu, Q. Zhang, H. Xie, X. Z. Yu, and Z. W. Liu, "Determination of residual stress distribution combining slot milling method and finite element approach," *Science China Technological Sciences*, vol. 61, no. 7, pp. 965–970, 2018.

[37] L. Gong, Y. Luo, X. M. Zou, and Y. He, "Influence factors of resistance for the metal material under the action of stress," in *Proceedings of the 4th college physics experiment teaching seminars of China*, pp. 114–117, Chongqing, China, 2006.

[38] D. X. Cheng, *Mechanical design handbook*, Chemical Industry Press, Beijing, China, 6th edition, 2016.

[39] Y. T. Yu, X. H. Li, A. Simm, and G. Y. Tian, "Theoretical model-based quantitative optimisation of numerical modelling for eddy current NDT," *Nondestructive Testing and Evaluation*, vol. 26, no. 2, pp. 129–140, 2011.

Hindawi
Journal of Sensors
Volume 2021, Article ID 6640406, 16 pages
https://doi.org/10.1155/2021/6640406

Review Article

Nondestructive Testing for Corrosion Evaluation of Metal under Coating

Ruikun Wu,[1,2] Hong Zhang,[1,2] Ruizhen Yang,[1,3] Wenhui Chen,[1,2] and Guotai Chen[1,2]

[1]*Key Laboratory of Nondestructive Testing Technology, Fujian Polytechnic Normal University, Fujian Province University, China*
[2]*School of Electronic and Mechanical Engineering, Fujian Polytechnic Normal University, China*
[3]*College of Civil Engineering, Changsha University, China*

Correspondence should be addressed to Hong Zhang; zhhgw@hotmail.com

Received 13 November 2020; Accepted 2 June 2021; Published 30 June 2021

Academic Editor: Antonio Lazaro

The use of steel has grown rapidly over the past decades. However, corrosion under coating detection still presents challenges for nondestructive testing (NDT) techniques. One of such challenges is the lift-off introduced by complex structures. Inaccessibility due to structure leads corrosion to be undetected, which can lead to catastrophic failure. Furthermore, lift-off effects reduce the sensitivities. The limitations of existing NDT techniques heighten the need for novel approaches to the characterization of corrosion. This paper begins with a discussion of the challenges associated with corrosion detection of metal under coating. Secondly, reviews are given of the most NDT methods used for the detection of corrosion under coating. The different techniques based on nondestructive testing methods such as ultrasonic, acoustic, electromagnetic, radiographic, and thermographic have been detailed out. This review presents the significance and advantages provided by the emerging NDT techniques. In the end, the trends and identified problems are summarized.

1. Introduction

Nondestructive testing (NDT) refers to the implementation of the defect detection of material, which at the same time does not affect the material's future performance. Corrosion is the deterioration in material properties due to interaction with the environment [1] and materials which corrode including metals and alloys, nonmetals, woods, ceramics, plastics, and composites [2]. For many applications, the preferred metal is still mild steel with its virtues of relatively low cost, mechanical strength, and ease of fabrication. The fact that it corrodes easily is the main drawback, which means that it rapidly loses strength which readily leads to structural failure. Therefore, steel or other metal structures such as vessels are typically coated to control corrosion. The purpose of coating is to prevent corrosion from occurring by inserting a barrier between the environment and the metal surface. Although coating provides a high level of protection against corrosion, the metal is still prone to corrosion. This form of corrosion occurs on steel under coating, and it is very difficult to detect due to the corrosion being concealed by the coating layer. The development of undetected corrosion can lead to serious failure. Serious consequences include the risk to the safety of persons, damage to the environment, and economic impacts.

Undetected corrosion may develop beneath coating which can lead to failure without curb. When steel is in contact with water and oxygen, it is likely to corrode. In order to improve the efficiency of examination, various NDT methods have been adopted to detect corrosion under coating without removing [3–5], each of which has different capabilities. These NDT techniques are then used to target problem areas where further inspection is needed.

In order to improve the reliability of coated metal, NDT approaches are used to investigate the existence of corrosion. The overall structure of this paper include the following: Section 2 of this paper will present the challenges posed by CUI. Section 3 reviews the development of nondestructive techniques with case studies. Then, the comparison and discussion have been provided in Section 4. Trends are in Section 5. Finally, the conclusions are outlined.

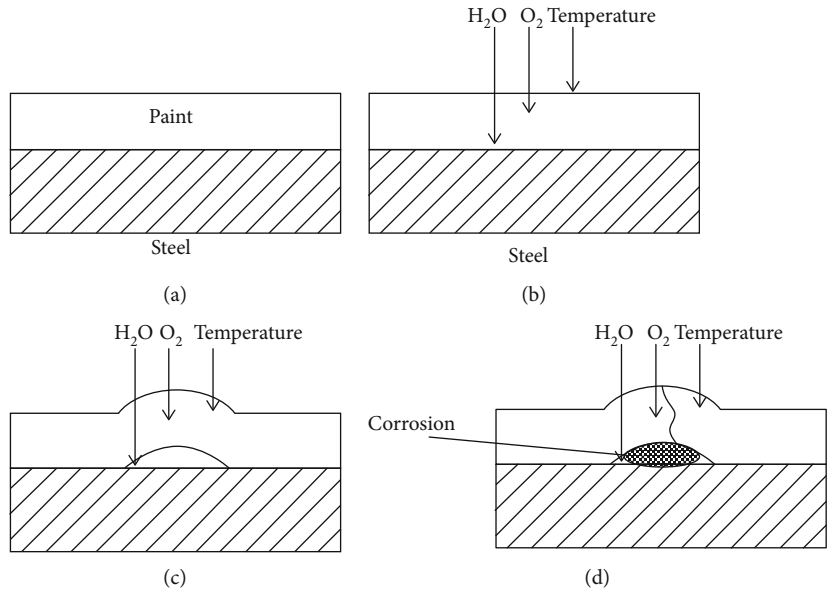


Figure 1: The mechanism of corrosion formation.

2. Challenges Posed by Corrosion under Coating

Unfortunately, metals are susceptible to corrosion. The occurrence of reactions among the metal surface, oxygen and water will result in the loss of electrons and the transformation of metal atoms into metal ions. Examples of the mechanisms of corrosion are [1]:

$$\begin{aligned} \text{The anode} &: \mathrm{Fe} \longrightarrow \mathrm{Fe}^{2+} + 2\mathrm{e}^{-} \\ \text{The cathode} &: \mathrm{O}_2 + 2\mathrm{H}_2\mathrm{O} + 4\mathrm{e}^{-} \longrightarrow 4\mathrm{OH}^{-} \end{aligned} \quad (1)$$

Corrosion is formed by the interaction of positively charged Fe^{2+} ions with negatively charged OH^{-} ions, as in $Fe^{2+} + 2OH^{-} \longrightarrow Fe(OH)_2$. The hydroxide is insoluble and separates from the electrolyte. A more familiar name for Fe $(OH)_2$ is a rust, white-green precipitate. With further access to oxygen, $Fe(OH)_2$ oxidizes to ferric hydroxide $Fe(OH)_3$, which in turn converts to Fe_2O_3 (reddish-brown rust) and H_2O:

$$4\mathrm{Fe(OH)}_2 + \mathrm{O}_2 \longrightarrow 2\mathrm{Fe}_2\mathrm{O}_3 + \mathrm{H}_2\mathrm{O}\,(\text{rust}) \quad (2)$$

Different types of corrosion, such as Fe_3O_4 (black magnetite), $\gamma - Fe_3O_4$ (brown rust), and γ(FeOOH) (yellow rust), are observed depending on the nature of the interaction in the environment, and the general mechanism of corrosion formation is shown in Figure 1. Metal can be affected by corrosion in many different ways, depending on the nature of corrosion and the prevalence of specific environmental conditions.

Various techniques such as electrochemical impedance spectroscopy [6, 7] can be used to determine the condition of the metal and determine the level/severity of any corrosion. Challenges of NDT for detection under coating are diverse. Along with the usual challenges of accurately detecting and quantifying corrosion, the task is made more difficult due to the large distance between the sensor and the metal surface introduced by the inaccessible coating layer. This distance is known as the lift-off. The immediate effect of a large lift-off is the reduction in sensitivity to small changes accompanying corrosion, such as variation in thickness or loss of mass. Meanwhile, the thickness of the coatings may be different, resulting in variations in lift-off leading to lift-off effects. This lift-off effect will cause errors in detection.

In addition, the challenges associated with the characterization of metal corrosion require an understanding of the microstructural and physical changes occurring prior to corrosion initiation and growth. In most cases of corrosion, changes in the intrinsic material properties are dominant in the early stages. Physical damages such as defects will be observable when the accumulation of these changes exceeds a critical limit. Therefore, there is no universal method for the detection of corrosion under coating because the behaviour of corrosion is affected by the mix of these diverse factors.

To detect corrosion on the metal surface under coating, NDT is a powerful tool. In the next section, the state-of-the-art NDT techniques for detecting corrosion in metal are reviewed. The advantages and disadvantages of each method are summarized, followed by the challenges for the evaluation and monitoring of metal underlying coating.

3. State-Of-The-Art NDT Techniques

Several NDT techniques have been applied for corrosion detection, and each has certain advantages and disadvantages. For example, eddy current-based techniques have been used for corrosion inspection, where a relatively small probe is employed and no physical contact with the specimen is needed [8]. However, these methods can only be used to detect corrosion on the surface or near-surface. Furthermore, they are sensitive not only to variations in conductivity and

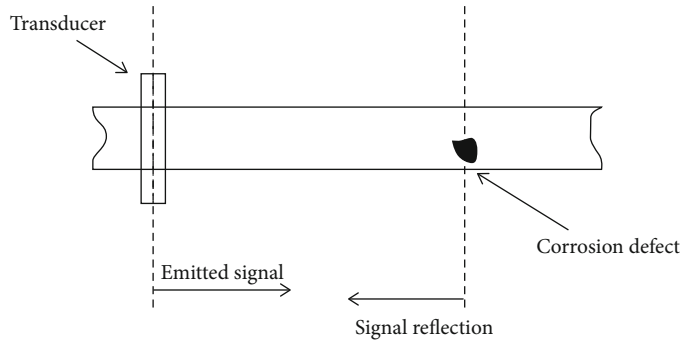


Figure 2: Ultrasonic testing for corrosion detection [10].

magnetic permeability of the specimen, but also to lift-off variations.

3.1. Ultrasonic Testing (UT). Ultrasonic testing is based on generating and detecting mechanical waves or vibrations within samples under test. These samples are not limited to solids. A majority of ultrasonic techniques adopt 1 MHz to 100 MHz as operating frequencies. The term ultrasonic refers to those sound frequencies beyond human hearing restrictions. The traveling speed of ultrasonic waves in a material is dependent on the material's density and elastic modulus. Therefore, ultrasonic methods are very suitable for the characterizing properties of materials. Furthermore, changes in material properties will strongly reflect ultrasonic waves at the boundaries. Thus, ultrasonic methods are often used for the measurement of thickness and corrosion monitoring [9].

Although UT has the capability to detect corrosion, these methods suffer from a difficulty in distinguishing between reflections from surface/near-surface corrosion and reflections from multiple material surfaces. Another disadvantage is that they require coupling media such as water or gel to acoustically couple pulses from the transducer to the material. This makes traditional UT unsuitable for certain situations due to the requirements for surface preparation. Much work has been done in recent years to develop noncontact ultrasonic techniques for defect characterization without requiring surface preparation. Air-coupled techniques include adaptations of traditional piezoelectric transducers [10], which are more suitable for most inspection conditions. However, an appropriate angle for the introduction of the ultrasound into samples to be inspected is a strict requirement. As shown in Figure 2, an experimental set to measure corrosion through guided acoustic waves in pipe has been reported. Guided acoustic waves are emitted by a ring transducer which run through the pipe. When a corrosion is encountered by waves on the pipe walls, they are reflected and returned to transducer.

Electromagnetic acoustic transducers (EMATs) have also been developed, in which electromagnetic noncontact transducers are used to generate and receive acoustic signals, although this does mean that EMATs are limited to electrically conductive materials. A modified variational mode decomposition (VMD) linked wavelet method is proposed by Si et al. [11] for EMAT denoising with a large lift-off detection condition. High-frequency narrowband noise in EMAT signals can be suppressed too. By using an ultrasonic B-scan method, short-time Fourier transform (STFT) is used by Le et al. to analyze the ultrasonic signal for pitting corrosion detection in a multilayer structure [12].

Nowadays, developments in the field of ultrasonic techniques has led to phased array ultrasonics in instruments which can be portable [13]. The precise tailoring of ultrasonic waves is introduced into the sample by using the phased firing of ultrasonic arrays in one transducer. To and Dang [14] used a phased array ultrasonic probe for the detection and sizing of stress corrosion cracks in fuel tank. This method employs a single phased array probe with focal laws for two kinds of sectorial scans (S-scans) separately based on transverse and longitudinal wave velocities. The transverse wave S-scan with an angle beam transverse wave method is used for the detection of stress corrosion cracks. The longitudinal wave S-scan uses a multiple beam method to determine the size of stress corrosion cracks.

3.2. Laser Ultrasonic. Given recent advances in laser ultrasonic techniques, lasers can be used to generate and detect ultrasonic waves [15]. This noncontact technique has been used for the measurement of material thickness, flaw detection, and materials characterization. A laser ultrasonic system is composed of a laser ultrasonic generator and an interferometric sensor.

Liu et al. [16] employed laser ultrasonics for the detection and locating of internal corrosion in hollow metallic components. 125 MHz piezoelectric transducers and broad band laser-ultrasonic are adopted to generate ultrasonic waves which will interact with corrosion, and then changes in generated wave-modes are examined using time-frequency analysis techniques. In a laser-ultrasonic method, as shown in Figure 3, laser beams are generated and detected at the front side where reflection mode is utilized. A pulsed Nd:YAG laser is used to generate ultrasound. A laser ultrasonic receiver (TEMPO) is used to detect the reflected ultrasound. A time-varying analog voltage is produced by the detector which is proportional to the instantaneous displacement at ultrasonic frequencies.

The geometric images of corrosion can be provided by scanning the samples for corrosion identification. The limitation of this technique is that access to the surface of the

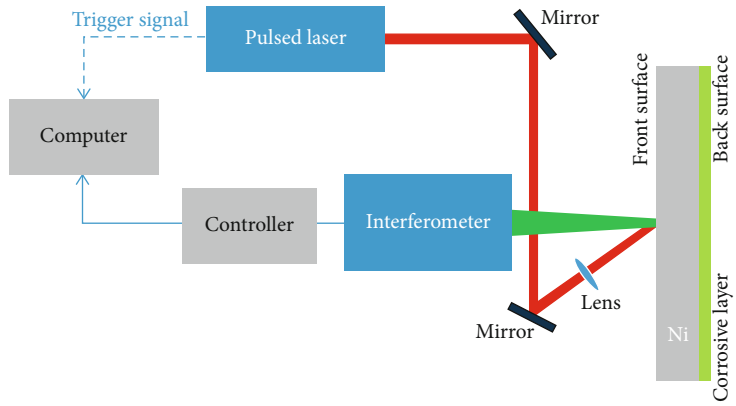


Figure 3: Schematic for laser-ultrasonic corrosion detection [16].

sample under test is required. Furthermore, this technique is sensitive to surface-breaking defects, and thus, its scope is limited.

3.2.1. Acoustic Emission (AE). AE is defined as strain energy suddenly released within or on the surface of a material, which generates a transient elastic wave. Therefore, the dynamic process associated with the degradation of a structure can be detected by AE. When an external stimulus, such as a change in pressure, load, or temperature, is applied to a structure, an energy released causes localized sources to form stress waves, and these waves then propagate to the surface, where sensors are used to record them.

In the majority of studies, AE techniques have been used to detect pitting corrosion. Zhang et al. [17] evaluated acoustic emission waveform for stress corrosion cracking monitoring in 304 stainless steel. As shown in Figure 4, the AE signals were collected by a mounted wideband AE sensor (Physical Acoustics Co.), and then through a preamplifier, an acquisition device was used to capture the AE signals. Wu and Byeon [18] used an AE technique to monitor progression of pitting corrosion in austenitic stainless steels. They confirmed that AE can be applied in fundamental research on pitting corrosion because it offers many potential advantages over other techniques. Zaki et al. [19] demonstrated the effectiveness of AE in the corrosion detection of concrete structures at an early stage.

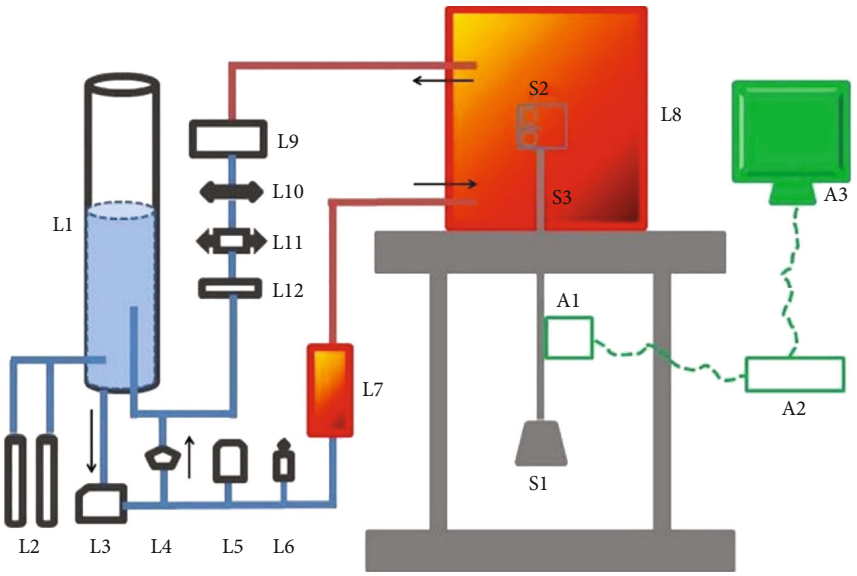
The AE technique can be used to detect corrosion occurring in real time giving it an advantage over other NDT method. Unfortunately, AE systems can only be used for qualitative testing. Additional NDT methods are required to obtain quantitative results in regard to the size and depth of corrosion. Furthermore, environment noise affects the AE signals received. Therefore, the use of signal discrimination and noise reduction techniques is essential during real-world applications.

3.3. Eddy Current (EC). Eddy current technique is one of the most effective methods for the detection and characterization of surface defects and corrosion in conductive samples. This technique is based on holding a conducting coil with alternating currents close to the sample. A primary magnetic field is established in an axial direction around the coil. This electrical current then creates its own secondary magnetic field, which is opposite in direction at all times and opposes the coil's magnetic field in accordance with Lenz's Law, as illustrated in Figure 5 [20]. The interaction between the magnetic field generated by the coil and the magnetic field generated by eddy currents is then examined with sensors or coils.

EC methods can be very effective and have been adopted to detect the presence of corrosion on the surface of metal samples. Thus, EC methods are the most common NDT methods which have become extremely portable and relatively inexpensive. Raude et al. employed advanced eddy current array technology for stress corrosion cracking inspection [21]. However, conventional EC techniques have difficulties detecting and quantifying small metal loss due to corrosion in multilayer structures. This is because the ability of EC techniques to detect subsurface defects is largely determined by the skin effect phenomenon. Most of the current flow occurs on the surface of a conductor due to the skin effect, exponentially decaying with increasing depth.

The development of EC technology has led to the introduction of the pulsed eddy current (PEC). This is a natural evolution of EC method and has been developed to improve penetration depth. With conventional EC methods, a fixed frequency sinusoidal current is used to generate eddy currents on the surface of a conductor. However, in PEC, the shape of the excitation current is a square pulse or step function. Looking at the Fourier transform of a step function, it is clear that it contains a continuum of frequency components compared to just one for sinusoidal EC. Because penetration depth is dependent on operating frequency, the PEC response signal will contain information from multiple depths, which is thus equivalent to multiple-frequency EC. The detection capabilities of PEC have been demonstrated in corrosion characterization [22]. PEC can be automated, and it has the advantages of greater penetration, the ability to locate corrosion, and only moderate cost.

Grosso et al. [23] have presented the results of a method based on a multifrequency eddy current along with signal processing to characterize iron oxide in a petrochemical storage tank. The thicknesses of the individual layers of the lap joint have been mapped with this technique. The result shows that corrosion can be quantified with an error of less than 5% for different corrosion thickness. Furthermore, recent developments



L1: Water tank
L2: Gas tank
L3: High pressure pump
L4: DO sensor
L5: Buffer tank
L6: Safety valve
L7: Pre heater
L8: Autoclave
L9: Condensing tank
L10: Filter
L11: Regulator
L12: Flow meter
S1: Load cell
S2: Specimen
S3: Load rod
A1: AE sensor
A2: AE preamplifier
A3: AE acquisition

Figure 4: Experimental setup of in situ AE corrosion monitoring system [17].

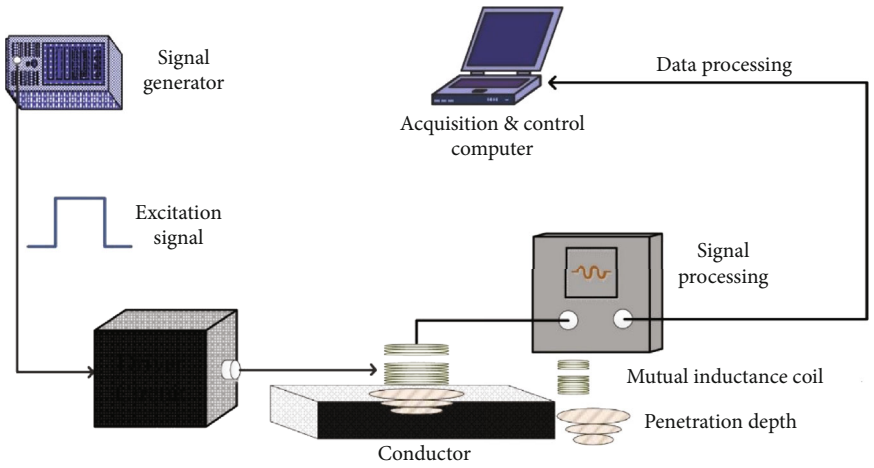


Figure 5: Schematic for eddy current corrosion detection [20].

in eddy current technology have led to multichannel portable instruments which allow the faster inspection of larger areas. Meanwhile, new magnetic sensors have been developed to replace coils [24], such as giant magneto resistive (GMR) sensors. Bailey et al. have investigated GMR sensor array to characterize corrosion of pipes under coating [25]. Rifai et al. have reviewed and described the implementation of GMR sensors in detail [26]. However, EC-based methods are limited to electrically conducting materials. Furthermore, these methods are very sensitive to lift-off effects, and the surface of the material must be accessible.

3.4. Magnetic Flux Leakage (MFL). MFL is a derivative of magnetic particle inspection (MPI). It is based on measuring the leakage of magnetic flux caused by the presence of corrosion. In practice, magnetisation is provided by a permanent magnet or an electromagnet by DC, AC, or pulsed excitation. The difference between MPI and MFL is that the latter measures flux leakage using magnetic field sensors such as Hall devices or magneto-resistive sensors. The inspection system is mainly composed of magnetic signal sensor, computer, serial port server, and three-axis transmission device, as shown in Figure 6 [27]. Qu et al. proposed a spontaneous

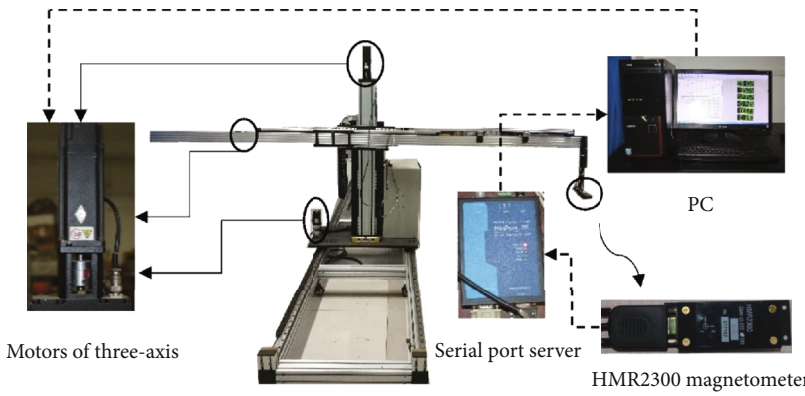


Figure 6: Schematic for magnetic flux leakage (MFL) inspection [27].

magnetic flux leakage (SMFL) method for the corrosion width predicting of cables. Three-axis transmission device can provide three mutually perpendicular scanning paths which are driven by three motors. The current position of the magnetic sensor is recorded. The Honeywell HMR2300 magnetic sensor is used to collect three-dimensional magnetic flux leakage signal. It is a giant magnetoresistance sensor with a range of ±2 gauss. The resolution is about 70 micro-gauss. Thus, the data acquired can be processed using computer-based analysis, signal processing, and quantitative assessment. However, MFL is only appropriate for the characterization of corrosion in ferromagnetic-material.

MFL techniques are popular in the inspection of pipelines. Azizzadeh and Safizadeh [28] employed an adaptive filter and a wavelet-based denoising technique for pipeline inspection. Xia et al. [29] used a high-resolution self-magnetic flux leakage (SMFL) to generate magnetic fields to quantitative measure corrosion. With three growth models (logistic model, exponential model, and linear model) and magnetic dipole model, leakage and mass loss can also be detected with SMFL. The precise location of corrosion can also be provided. Ege and Kuramik [30] adopted MFL with KMZ51 AMR sensors to examine the speed variable for corrosion detection in metal pipelines. Both penetrating depth and detecting sensitivity have been improved due to the abundant components of the speed variable signal.

3.5. Microwave NDT (MNDT). Microwave frequency range is between 300 MHz and 300 GHz. Unlike ultrasound signals, dielectric coating materials can be easily penetrated by microwave signals without suffering from high attenuation and then internal structures of materials can interact with these microwave signals. These microwave signals would then totally reflect at the metal surface. Therefore, these signals travel twice through the areas of corrosion and defects, which increases the possibility of detecting them under coating. With the measurement of transmitted or reflected microwave signals, microwave NDT techniques examine magnitude or phase in inspecting the specimen. Furthermore, reflection and transmission properties are influenced by lift-off and the frequency of operation during inspection. The experimental setup for microwave NDT has been depicted in Figure 7 [31]. A waveguide probe is placed above sample under test with a specified lift-off. A vector network analyzer are is used to provide excitation signals and to obtain the reflected signals' frequency spectrum information. A PC is used to control the vector network analyzer and acquire measurement data transmitted to the PC through GPIB (General Purpose Interface Bus). An X-Y scanner was connected to a controller through a parallel port and controlled by PC. A MATLAB program is used to make the X-Y scanner and vector network analyzer working collaboratively during measurement.

Zoughi [32] employed 3D microwave camera to detect steel corrosion on concrete up to 500 mm. Kharkovsky and Zoughi [33] gave an overview of microwave- and millimetre wave-based NDT&E methods. A wide range of applications was discussed, which included detecting corrosion and the precursors of pitting in insulated structures backed by aluminium and steel. Adhvaryu et al. [34] demonstrated 2.4 GHz apertured EBG-based microwave patch antenna for steel rebar corrosion characterization in civil structures, and good resolution has been achieved at about 14 mm depth. Far-field and near-field microwave NDT approaches detect corrosion through the magnitude and phase variation of the reflection coefficient [35, 36]. From the standpoint of high resolution, signal interpretation, and insensitivity to relative position between sample and antenna, a far-field mode is preferable; but it requires large-aperture antennas to achieve good spatial resolution. In most situations, the use of large antennas is generally impractical and inconvenient. Moreover, near-field mode can be performed indoors, eliminating influences due to weather, electromagnetic interference, etc. Near-field microwave imaging techniques with open-ended rectangular waveguide are commonly used for NDT fields. For producing image, a microwave synthetic aperture radar (SAR) is scanned over sample under test, and the measured reflected signals are used to form a 2D intensity raster image [37]. Mukherjee et al. [38] used a split-ring resonator (SRR) sensor for composite imaging with super resolution capability. The difference between the phase and magnitude of reflected signals was used to produce high-resolution image for pit dimension evaluation. Qaddoumi et al. [39] demonstrated an open-ended rectangular waveguide sensor

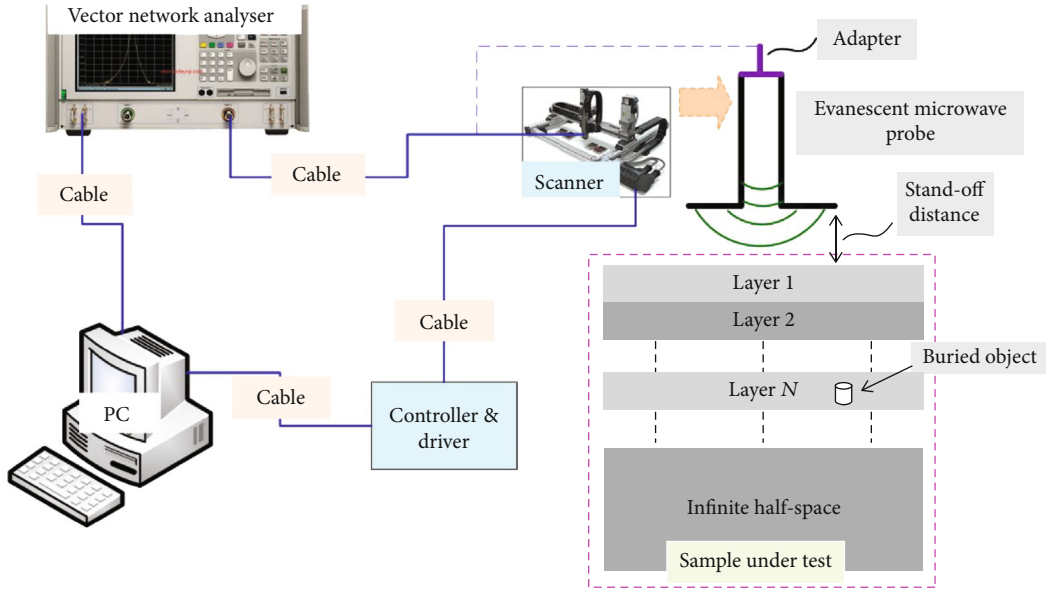


Figure 7: Schematic for microwave NDT [31].

operating in the near-field at a frequency of 24 GHz for defect detection and classification in nonceramic insulators. Defects were detected and classified by using a novel artificial neural network.

For corrosion detection under coating, the use of microwave NDT can be extended to the detection of water in the coating layer, since water is a cause of corrosion [40]. Meanwhile, samples with coating-related corrosion can be created to determine whether or not blisters, delamination, and other coating defects can be distinguished with microwave NDT [41]. These different types of defects may look similar when analyzing the features of microwave signal. Therefore, new features are required as well as looking into how the responses in coating to corrosion change over time, such as if blisters may appear as sudden sharp changes compared to defects on metal. For the microwave scanning process over large areas, compressive sensing could also be adopted to reduce scanning time [42]. A future work will also involve looking into methods to obtain quantitative information about conditions under the surface of steel, such as variations in physical parameters. This can be achieved by correlation using advanced feature extraction methods [43]. Zhang and others proposed a K-band sweep frequency microwave imaging system with a waveguide aperture [44]. Figures 8(a) and 8(b) show the images of coated samples with 1- and 6-month corrosions, respectively. Images were obtained using the averaged magnitude of the reflection coefficient.

However, microwaves cannot efficiently penetrate through conductive materials, which means that only surface corrosion can be sensed and it is difficult to detect subsurface one.

3.6. Terahertz (THz) Technology. THz refers to electromagnetic waves with frequencies ranging from 0.1 THz to 10 THz. Wavelengths of THz radiation are between the microwave and infrared spectra which are approximately from 0.03 mm to 3 mm. A THz wave with known wavelength is used to illuminate a sample during inspection. This THz wave is examined at or near the radiation source after interaction with the sample under test. The inner structure of the sample is determined by analyzing changes in the THz signal, because the dielectric characteristics of the sample or a discontinuity will affect the THz signal. Figure 9 illustrates a typical THz system for detection of defects in SOFI (Spray on Foam Insulation) layers for space shuttle, and researches have shown that THz-based NDT can provide an effective evaluation for shuttle fuel tank under insulation materials. THz-based imaging has been chosen by NASA which will be used for future launch inspection.

Since the mid-1980s, THz-based methods have made important advances. THz wave has a better penetration through most dry, nonmetallic materials such as foams, ceramics, glass, resins, coating, rubber, and composite materials [45]. Therefore, these methods have been applied in the NDT&E fields and can be divided into continuous THz and pulse THz methods. THz-based NDT has unique advantages in detecting inner defects in nonmetallic materials compared to other NDT techniques. The THz wave can penetrate nontransparent materials and evaluate inner defects. Moreover, THz-based NDT has been used to inspect insulated materials. The ability of THz-based imaging for corrosion under coating has been studied by the U.S. Army Research Laboratory and NASA. Corrosion under coating leads nominally smooth surfaces to become rough and irregular, and erosion can be detected by THz-based imaging [46].

Tu et al. demonstrated a THz-based system for manned spacecraft imaging, in which a high-speed time domain is employed for nondestructive evaluation [47]. Moreover, many studies have been shown that signal processing methods can be adopted for THz-based NDT. You et al. adopted a two-dimensional continuous wavelet transform approach to extract defect information from responses, which overcomes the limitations of traditional indirect methods, where reflection signals from a metal base are used

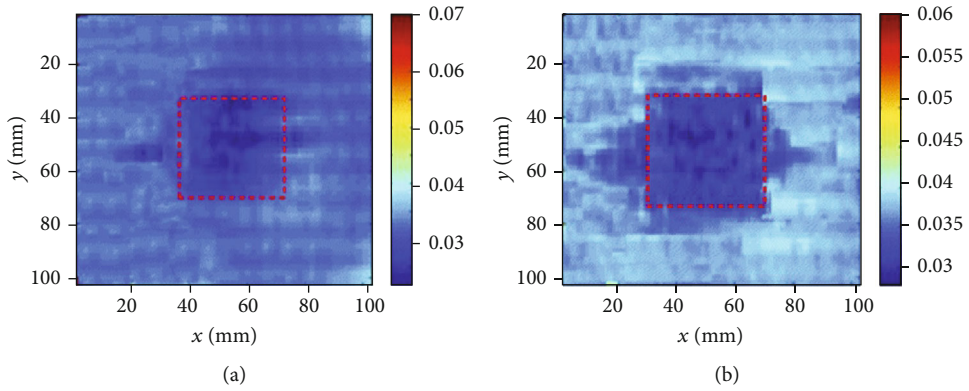


Figure 8: Images formed by the averaged magnitude of the reflection coefficient for (a) 1-month and (b) 6-month corrosions [44].

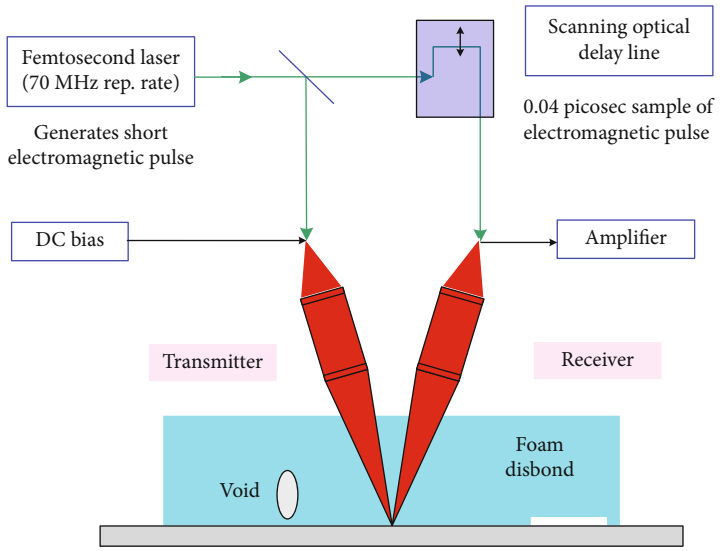


Figure 9: A reflected pulse THz system used for corrosion detection.

for further analysis [48]. Cao et al. adopted reflected terahertz pulse echoes to infer four layers of coatings on metallic substrates [49]. However, the THz wave cannot penetrate metallic material, which limits the scope of its application. Furthermore, THz-based methods have disadvantages of high costs and strong water absorption.

3.7. Thermography Testing. Thermography testing measure thermal variance for a characterized sample which undergoes a response to a stimulus. Improvements in IR cameras have led to more advanced forms of thermography. The advantages of thermography are that it is real time and noncontact and a large area can be inspected in a short time. With an IR camera, a thermal image is produced by infrared light which is invisible to the human eye and which is emitted from objects due to their thermal condition. Infrared thermography detection is based on differences in temperature conditions. There are two types of thermography: active and passive [50]. Active thermography (AT) is defined as the application of a stimulus to heat up the target to allow a wide range of its characteristic to be determined. These obtained characteristics can be defects or corrosion. Passive thermography (PT) is defined as measuring temperature differences among target material, surrounding materials, and ambient temperature conditions. Normally, active thermography-based methods are the most commonly used. Sfarra et al. [51] employed an infrared thermography for the cellular structure detection in honeycomb structures. Maierhofer et al. [52] studied the use of infrared thermography for voids and cellular hollow testing. One drawback of the IR method is the high cost of quality thermal cameras, but recent developments have led them to become significantly less expensive.

3.8. Radiograph Methods. Radiograph is one of the most common NDT methods and is based on differences in the attenuation of penetrating radiation in materials depending on radiation energy and material density and thickness. Different thicknesses and types of materials give different attenuation coefficients, and variations in transmitted radiation intensity are caused by corrosion. Therefore, the value of radiographs for detecting defects, corrosion, and welds in metals has been proven [53]. For the investigation of corrosion in coated mild steel, as shown in Figure 10, an X-ray tube

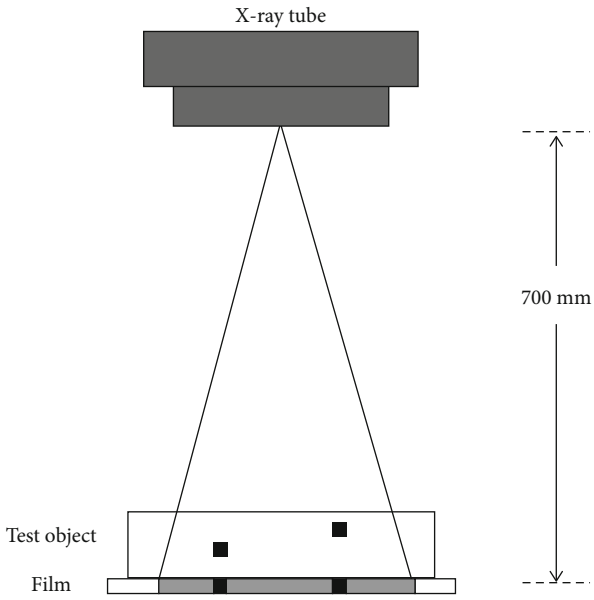


Figure 10: Schematic of radiography setup for concealed corrosion detection [54].

has been used with a 160 kV generator and 1.5 mm × 1.5 mm focal spot [54]. The voltage and current of tube are set to 150 kV and 10.7 mA. For mapping corroded areas in 11 mm thick specimens, exposure time is set to 190 s. Film is AGFA D7. During the radiographic examination, single wall single image technique is used to obtain the concealed corrosion. The film are kept intact with the sample under test. To minimize distortion (due to beam divergence) and geometric unsharpness, the distance between the film and X-ray source is set to 700 mm. The radiographic image sensitivity of 2% is achieved using the appropriate image quality indicator. By using the film digitizer Model Array 2905, radiographs are digitized with 50-micron resolution.

There are several different radiographic techniques, including profile radiography, digital radiography, flash radiography, and real-time radiography [55]. These techniques are based on the use of either gamma rays or X-rays to image the profile of a structure or to provide information about the thickness of an inner structure. In many cases, discontinuities in insulated metals are readily detected. Radiography is still widely used in spite of its expense and the fact that ionising radiation poses health and safety risks. Recent developments in digital radiography have helped to eliminate the use of film, thus reducing costs. Stannard et al. adopted synchrotron X-ray tomography for corrosion fatigue crack analyses [56]. Barlow et al. used X-ray microprobe for corrosion characterization at the buried polymer-steel interface [57].

Apart from the above issues, however, there are several notable limitations on the use of radiography. For example, it is not suitable for the detection of surface corrosion and it is also not possible to extract quantitative information for the estimation of corrosion's depth.

3.9. Eddy Current Pulsed Thermography (ECPT). The configuration of an ECPT system is shown in Figure 11. ECPT involves an application of a high-frequency electromagnetic wave (typically 50 kHz–500 kHz) at a high current around 256 A to 380 A to the material under inspection for a short period of typically 20 ms–1 s [59, 60]. Induced eddy currents are forced to divert when they encounter a discontinuity, which leads to increases and decreases in the density of the eddy current in that area. Areas with increased density of the eddy current are exhibiting higher levels of Joule (Ohmic) heating, and thus, corrosion can be obtained from sequenced thermograms during the heating and cooling periods. It consists of an induction heating system which induces eddy currents in the sample under inspection and generates a heat; the generated heat is recorded by an IR camera to form digital data; then, these digital data will be displayed on a monitor and stored in PC.

He et al. [61, 62] reported an application of ECPT for the detection of corrosion blisters in mild steel under coating. At 50 and 200 ms, Figures 12(a) and 12(b) show thermograms of a coated sample with 3-month corrosion, respectively. The shape of corrosion can be seen in Figure 12(a). A wide range of defects can be considered based on interactions between the distribution of eddy current density and heat conduction. The corrosion blister areas were easily detected using sequenced thermograms from an IR camera during the experimental study. Yang et al. used electromagnetic induction thermography for coating imaging and nondestructive visualization evaluation on coated mild steel [63]. The complex influence of parameter variation on temperature has been eliminated by using phase analysis, and the corrosion height can be estimated. However, ECPT has the disadvantages of a limited ability to be used to inspect conductive material, and furthermore, the heat inducing equipment is very bulky.

3.10. Microwave Thermography (MWT). As shown in Figure 13, basic principles and types of MWT have been reviewed by Zhang et al. [64]. MWT exhibits a great potential including fast heating, high resolution, fast inspection and high sensitivity, no contact requirement, and better detectability for the inner defect.

Foudazi and others proposed the microwave thermography for corroded reinforced steel bar detection and characterization [65]. They employed a $14 \times 24\,cm^2$ horn antenna to illuminate steel bars with 50 W of microwave signal for 10s. Because of the relatively low thermal conductivity of corroded steel, heat dissipates quickly in uncorroded steel. Moreover, these temperature differences between the corroded areas indicated that different amounts of corrosion absorb different amounts of microwave energy. With a preliminary simulation and experimental study of the microwave thermography for corrosion detection in steel bars, it demonstrated that a higher excitation microwave frequency will lead to a higher temperature. Pieper and others demonstrated the active microwave thermography for large areas corrosion inspection on reinforcing steel bars for cement-based structures [66]. During an experimental study, two steel (AISI 1008) bars (each with a length of 150 mm and a radius of 4.8 mm) were measured which have been parallel embedded in a concrete block ($170 \times 150 \times 50\,mm^3$). One of them was light corroded along half of its length, the other was significantly corroded on the order of 1-4 mm of its

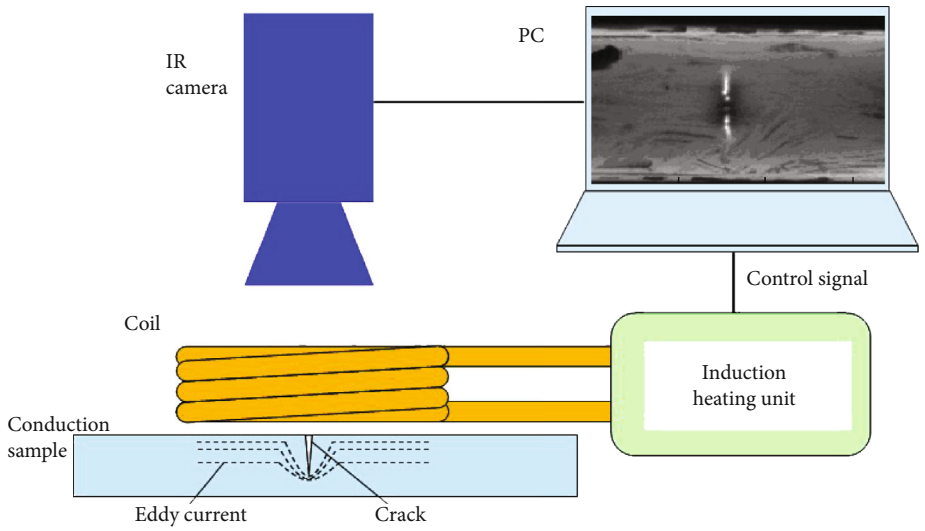


FIGURE 11: A basic configuration of ECPT system [58].

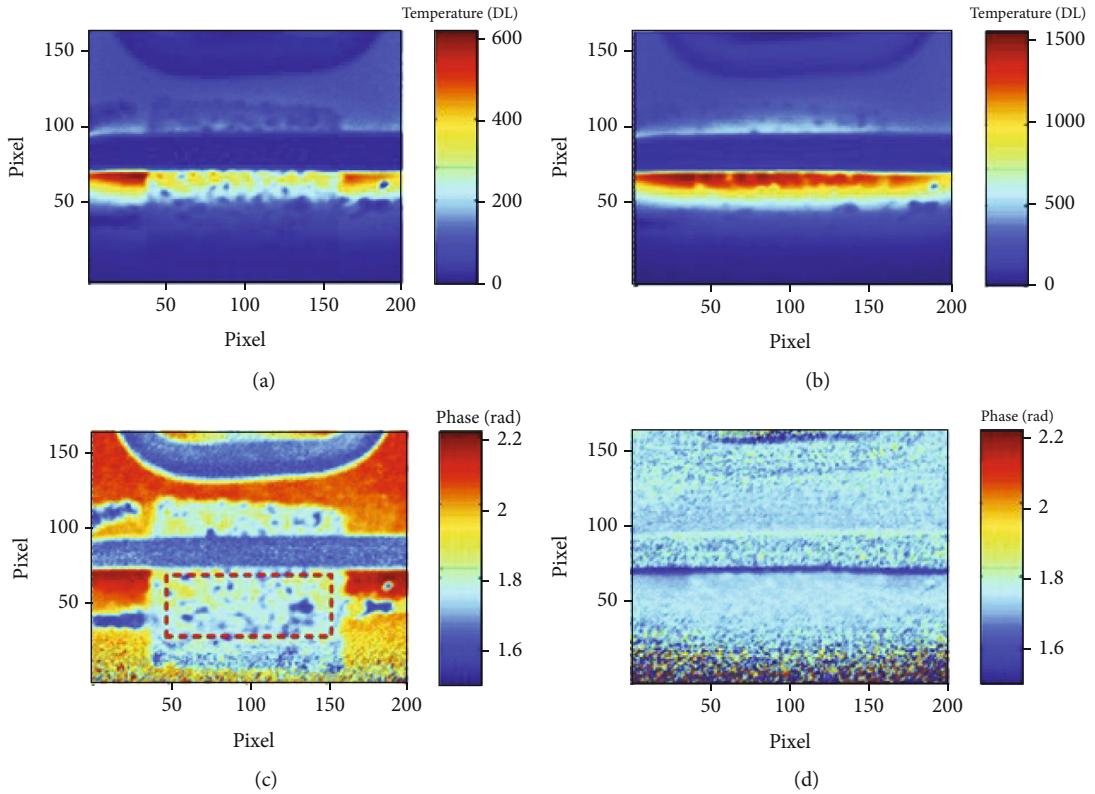


FIGURE 12: Thermal images of a coated sample with 3-month corrosion at (a) 50 and (b) 200 ms. Phase images of a coated sample with 3-month corrosion at (c) 4 and (d) 10 Hz [61].

length. The sample was heated for 5 sec by a microwave oven operating at 2.45 GHz. Keo and others presented a microwave thermography to detect steel in reinforced concrete wall [67]. During an experimental study, a commercial magnetron operating at 2.45 GHz was associated with a pyramidal horn antenna to illuminate a maximum 800 W microwave energy. The thermograms were recorded at 1 image per sec by using the ALTAIR software with a computer. The maximum temperature areas correspond to the presence of steel reinforcements in the specimen.

4. Discussion

Nine NDT techniques have been discussed. A comparison of these technologies is provided in Table 1, giving an overview

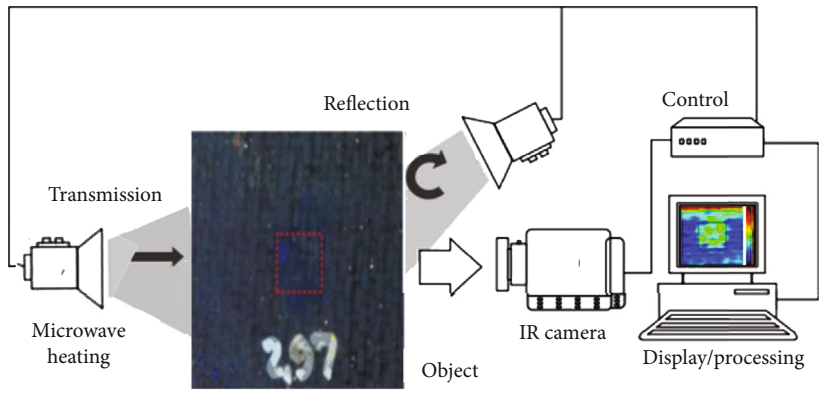


Figure 13: MWT setup for corrosion detection.

Table 1: Advantages and limitations of NDT technologies for corrosion characterization.

NDT technologies	Advantages	Limitations
Ultrasonic	Fast; inspect large area; penetrate deeply in materials; excellent for corrosion detection; can be automated.	Requires coupling material and contacting with surface; reference standards are required; surface needs to be smooth.
Radiography	Broad range of materials and thicknesses can be inspected; inspection film can be recorded.	Requires a minimum intensity difference; radiation safety requires precautions; expensive; requires to access both sides of the structure.
MFL	Portable; inexpensive; sensitive to surface and near-surface flaws and corrosion.	Bulk size; limited to ferromagnetic materials; requires postinspection and surface preparation.
EC	Quick to perform; moderate cost; no probe contact required.	Limited to materials with electrically conducting; penetration depth is limited; surface must be accessible and smooth.
PEC	Penetrating deeper without altering the coating; noncontact.	Limited inspection area; inability to detect localized corrosion.
Thermography testing (including ECPT)	Good for surface corrosion; high sensitivity; remote sensing; fast; inspect large area.	Expensive; requires heating and cooling of the system; reference standards required; poor resolution on thick sections.
THz	Noncontact; good resolution; inspect coating layer properties; real-time; one-side manner; high sensitivity.	Expensive; complex wave interactions; sensitive to environments (moisture, etc.).
MWT	Very sensitive; quick to perform; time dependent data is available; can detect nonvisible damaged areas.	Expensive; complex data analysis for quantification; localized detection requires knowledge of damaged area.

of each method and identifying the advantages and limitations of current techniques.

Corrosion effects are a complex combination of multiple factors, including variations in conductivity, permeability, and permittivity and changes in thickness. The probability of corrosion detection is affected by these factors. There is no universally applicable method for corrosion detection, due to the complex combination of these different factors. Therefore, none of the NDT techniques discussed above can address all the challenges of detecting corrosion under coating, since coating layers induce a large lift-off between sensors and the inspected surface.

Selection of an NDT technique requires consideration of more than the detection capabilities. The application, portability of equipment, inspection schedule, inspection area, types of materials, accessibility, costs, and expected corrosion types are also important. Some techniques provide good quantitative information but perform poorly when coating is introduced. Therefore, new methods are required to monitor corrosion which can be used for long-term operation and minimal volume and at lower cost and risk.

5. Trends

5.1. New Principles and Methods. A future work with the RFID and microwave NDT system will be geared towards improvements in handling other challenges related to

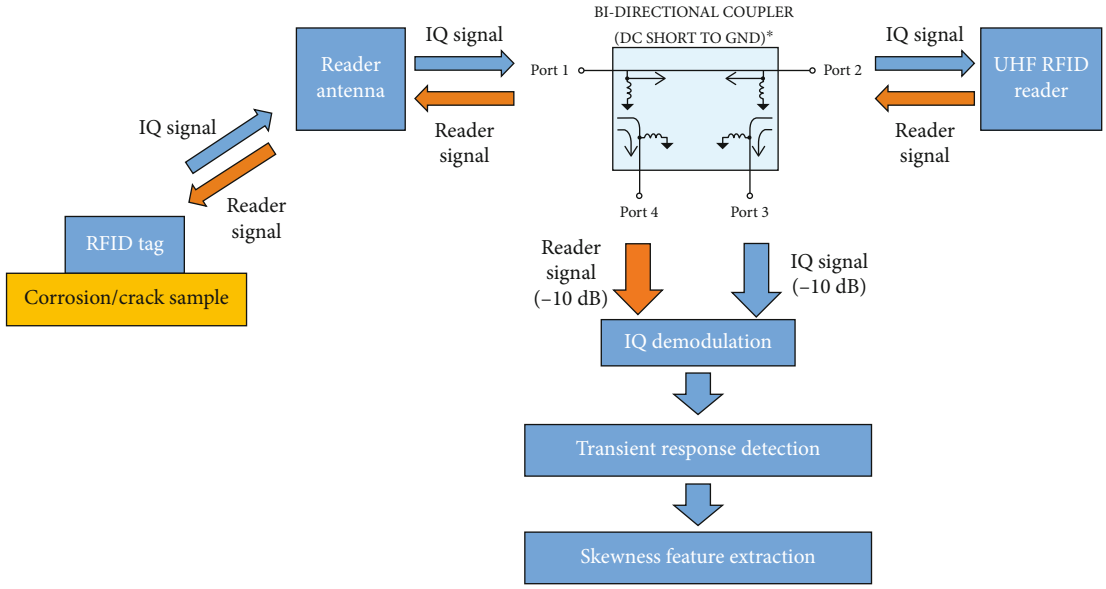


Figure 14: RFID-based sensing system diagram for corrosion detection [71].

insulated materials. These improvements will also be important steps towards commercial feasibility.

The inability of the RFID system to monitor large areas limits its potential field of application. Solving this problem requires the redesign of the tag unit [68, 69]. Additionally, the inability of the RFID sensing system to handle large lift-off requires investigations into whether or not carefully placed ferrite material near the tags will improve reading range. New tags employ circular three arm (CTA) element; a parasitic element has been added into the centre of CTA to improve the reliability [70].

As shown in Figure 14, an RFID-based sensing platform comprises RFID reader and tag [71]. The communication between the reader and tag is an asymmetric bidirectional link: the direct link and the reverse link. In the direct link, the reader emitted the power to power up the tag. Then, in the reverse link, the tag modulates backscattering signal and reflects the waves. During measurement, the tag acted as a sensor; the tagged object and the nearby environment can be exploited clearly with characterization of tag antenna's radiation. The power-based parameters have been investigated for stress and defect sensing. In addition to signal strengths and phases, IQ signal in RFID systems is investigated to improve sensitivity and robustness for corrosion sensing which contains high order information.

HF passive RFID-based monitoring networks can be formed by combining passive RFID tag with SAW (surface acoustic wave) [72]. With cost-efficient and lower power consumption, RFID-based approaches may represent a revolutionary solution for the condition and intelligent structural health monitoring of railways, in-service nuclear power plants, and aerospace applications [73]. Zhang and others [74] employed a UHF RFID antenna for structure health monitoring. Zhao and others employed a T-shape antenna UHF RFID to increase the gain of the miniaturized antenna and the sensitivity [75]. As well as cost-effective passive RFID tags with low power consumption, other sensors could easily be connected for multiple-purpose sensing. In addition, new features of RFID tags based on advanced signal processing can be used for human body temperature and other measurements.

For corrosion detection under insulation, the use of microwave NDT can be extended to the detection of water in the insulation layer, since water is a cause of corrosion [40]. Meanwhile, samples with insulation related corrosion can be created to determine whether or not blisters, delamination, and other insulation defects can be distinguished with microwave NDT [39]. These different types of defects may look similar when analyzing the features of microwave signal. Therefore, new features are required as well as looking into how the responses in insulation to corrosion change over time, such as if blisters may appear as sudden sharp changes compared to defects on metal. For the microwave scanning process over large areas, compressive sensing could also be adopted to reduce scanning time [42].

5.2. Signal Processing Algorithms. A future work will also involve looking into methods to obtain quantitative information about conditions under the surface of steel, such as variations in physical parameters. This can be achieved by correlation using advanced feature extraction methods [76]. To extract useful features from the captured thermal images, more advanced signal processing algorithms have been used. With suitable signal processing algorithms, the inspection results can be significantly improved in size and depth identification, subsurface corrosion detection, emissivity variation reduction, and corrosion dimension quantification. Therefore, more advanced signal processing algorithms are needed to further improve the sensitivity and quantification ability of NDT system.

5.3. Combination of SHM and NDT. Structural integrity monitoring (SHM) systems for steel under coating can provide continuous real-time data directly to a central

control/monitoring room often located at a significant distance, unlike NDT (often scheduled inspection); thus, it can help identify corroded areas and plan condition-based structural repairs [77]. In [78], structural integrity monitoring of corroded steel marine structures with the use of NDE and SHM approach was presented. First of all, NDT systems for manufacturing quality control and for on-site inspection have been presented. However, the authors thought that the use of NDT was not simple for industry practices on the maintenance of marine steel structures. For that purpose, SHM techniques have to be considered.

5.4. Intelligent Inspection System. The efficiency of corrosion detection system can be improved by exploring an intelligent inspection system with artificial intelligence. As various types of corrosion can be acquired during material measurement, the treatment for different types of corrosion is different. Take uniform corrosion for example, this is characterized by the entirety of the surface area considered corroding at the same (or a similar) rate [79]. Uniform corrosion (or general corrosion) is relatively easily measured and predicted. This type of corrosion causes the loss of metal thickness and weight. This corrosion is the most typical corrosion during manufacturing which needs to identify to improve the manufacturing quality of the coating. Therefore, it is important to classify the corrosion type with an intelligent inspection system. As computers are becoming more and more powerful, artificial intelligence methods can be used to reduce inspection time and improve the reliability. For example, neural networks, artificial neural networks, the classic computer vision techniques, and the deep learning approach have been used for corrosion detection [80].

5.5. Implementation of Unmanned Systems. For a large sample under test, the measurement system needs to be placed in a mobile robot or a vehicle. Pixel physical size calibration has been combined with an unmanned aerial vehicle (UAV) to solve the difficulty of manually detecting the surface of the quay crane [81]. The inspection time for a large material can be significantly reduced. The whole inspection can be performed autonomously. With autonomous robots, corrosion detection can be arranged horizontally and vertically. The safety and efficiency of the NDT system can be significantly improved. However, lightweight equipment and advanced detection algorithms are also required in order to provide the automatic inspection ability.

6. Conclusion

The challenges posed by the development of corrosion under coating are inaccessibility (very difficult or impossible to reach) and lift-off effects caused by the variation of the coating layer. Lift-off effects can cause errors in the detection and measurement of corrosion. Moreover, thick coating layers result in a large lift-off, which leads to a reduction in sensitivity. In addition, the challenges associated with the characterization of corroded metal require an understanding of microstructural and physical changes prior to the initiation and growth of corrosion. In most cases of corrosion, changes in the intrinsic material properties are dominant in the early stages. Physical damages, such as defects, occur when corrosion has exceeded a critical limit. Furthermore, their concealed nature results in the accumulation of such changes for long periods of time, leading to the critical limit being exceeded, and potentially catastrophic failures will become more likely.

An extensive literature survey in this paper has been followed by a discussion of NDT techniques using ultrasonic, acoustic, electromagnetic, radiographic, thermographic methods for the detection of corrosion. It has been shown that the majority of techniques are limited when it comes to the online in situ monitoring of corrosion under coating, primarily due to the thick coating layer. Solutions to overcome this problem typically involve either applying much higher output power using bulky, expensive equipment or using inspection holes in the coating layer to send signals along the length of an insulated structure. Compare with traditional NDT methods, such as UT, EC, and PEC, microwave NDT provides a wider range of advantages, including noncontact nature and high sensitivity and resolution. However, the cost of microwave NDT systems is relatively high and therefore limits their applications. Therefore, RFID-based methods are made fast and affordable to extend microwave NDT applications in corrosion detection, which requires new approaches to obtaining many details of RFID architectures. To address challenges from corrosion under coating, the following issues is needed to investigate: new principles and methods, signal processing algorithms, combination of SHM and NDT, intelligent inspection system, and implementation of unmanned systems.

Abbreviations

AE: Acoustic emission
EC: Eddy current
ECPT: Eddy current pulsed thermography
EM: Electromagnetic
EMATs: Electromagnetic acoustic transducers
GMR: Giant magneto resistance
IR: Infrared
MFL: Magnetic flux leakage
MNDT: Microwave NDT
MWT: Microwave thermography
NDT: Nondestructive testing
PEC: Pulsed eddy current
THz: Terahertz
UT: Ultrasonic testing.

Data Availability

The data supporting this systematic review are from previously reported studies and datasets, which have been cited.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62071123 and 61601125), Natural Science Foundation of Fujian Province of China (Grant Nos. 2018J01787 and 2020J01312), the General Program of Natural Science Foundation of Hunan Province (2018JJ2458), and the Postdoctoral Science Foundation of China (2018M630898) and was also supported by the 2019 Fujian Provincial Marine Economic Development Subsidy Fund Project (FJHJF-L-2019-7), the Program for New Century Excellent Talents in Fujian Province University, and the cultivation plan of Outstanding Young Scientific Research Talents in Colleges and Universities of Fujian Province.

References

[1] Z. Ahmad, *Principles of Corrosion Engineering and Corrosion Control*, Elsevier, 2006.

[2] R. W. Revie and H. H. Uhlig, *Corrosion and Corrosion Control: An Introduction to Corrosion Science and Engineering: Fourth Edition*, John Wiley & Sons, 2008.

[3] T. Druet, B. Chapuis, M. Jules, G. Laffont, and E. Moulin, "Passive SHM system for corrosion detection by guided wave tomography," in *Sensors, Algorithms and Applications for Structural Health Monitoring*, Springer, 2018.

[4] C. J. Lissenden, I. Jovanovic, A. T. Motta et al., "Remote detection of stress corrosion cracking: Surface composition and crack detection," in *AIP Conference Proceedings*, Melville, NY, 2018.

[5] Q. Zhang and R. Xin, "The defect-length effect in corrosion detection with magnetic method for bridge cables," *Frontiers of Structural and Civil Engineering*, vol. 12, no. 4, pp. 662–671, 2018.

[6] M. Hattori, A. Nishikata, and T. Tsuru, "EIS study on degradation of polymer-coated steel under ultraviolet radiation," *Corrosion Science*, vol. 52, no. 6, pp. 2080–2087, 2010.

[7] V. F. Lvovich, *Impedance Spectroscopy: Applications to Electrochemical and Dielectric Phenomena*, John Wiley & Sons, 2012.

[8] Y. He, G. Tian, H. Zhang, M. Alamin, A. Simm, and P. Jackson, "Steel corrosion characterization using pulsed eddy current systems," *Sensors Journal, IEEE*, vol. 12, no. 6, pp. 2113–2120, 2012.

[9] F. Honarvar and A. J. U. Varvani-Farahani, "A Review of ultrasonic testing applications in additive manufacturing: Defect evaluation material characterization, and process control," *Ultrasonics*, vol. 108, p. 106227, 2020.

[10] V. Marcantonio, D. Monarca, A. Colantoni, and M. Cecchini, "Ultrasonic waves for materials evaluation in fatigue, thermal and corrosion damage: a review," *Mechanical Systems and Signal Processing*, vol. 120, pp. 32–42, 2019.

[11] D. Si, B. Gao, W. Guo, Y. Yan, G. Tian, and Y. Yin, "Variational mode decomposition linked wavelet method for EMAT denoise with large lift-off effect," *NDT & E International*, vol. 107, article 102149, 2019.

[12] M. Le, J. Kim, S. Kim, and J. Lee, "Nondestructive testing of pitting corrosion cracks in rivet of multilayer structures," *International Journal of Precision Engineering and Manufacturing*, vol. 17, no. 11, pp. 1433–1442, 2016.

[13] A. Yassin, M. S. U. Rahman, and M. Abou-Khousa, "Imaging of near-surface defects using microwaves and ultrasonic phased array techniques," *Journal of Nondestructive Evaluation*, vol. 37, no. 4, p. 71, 2018.

[14] T. T. To and T. N. Dang, "Researching on measurement strategies of fuel tank corrosion using phased array technology," in *Applied Mechanics and Materials*, pp. 499–507, Trans Tech Publications Ltd, 2019.

[15] T. C. Truong and J.-R. Lee, "Thickness reconstruction of nuclear power plant pipes with flow-accelerated corrosion damage using laser ultrasonic wavenumber imaging," *Structural Health Monitoring*, vol. 17, no. 2, pp. 255–265, 2018.

[16] H. Liu, L. Zhang, H. F. Liu et al., "High-frequency ultrasonic methods for determining corrosion layer thickness of hollow metallic components," *Ultrasonics*, vol. 89, pp. 166–172, 2018.

[17] Z. Zhang, X. Wu, and J. Tan, "In-situ monitoring of stress corrosion cracking of 304 stainless steel in high-temperature water by analyzing acoustic emission waveform," *Corrosion Science*, vol. 146, pp. 90–98, 2019.

[18] K. Wu and J.-W. Byeon, "Morphological estimation of pitting corrosion on vertically positioned 304 stainless steel using acoustic-emission duration parameter," *Corrosion Science*, vol. 148, pp. 331–337, 2019.

[19] A. Zaki, H. Chai, D. Aggelis, and N. Alver, "Non-destructive evaluation for corrosion monitoring in concrete: a review and capability of acoustic emission technique," *Sensors*, vol. 15, no. 8, pp. 19069–19101, 2015.

[20] L. Xie, B. Gao, G. Y. Tian, J. Tan, B. Feng, and Y. Yin, "Coupling pulse eddy current sensor for deeper defects NDT," *Sensors and Actuators A: Physical*, vol. 293, pp. 189–199, 2019.

[21] A. Raude, M. Bouchard, and M. Sirois, *Stress Corrosion Cracking Direct Assessment of Carbon Steel Pipeline Using Advanced Eddy Current Array Technology*, CORROSION 2018, Phoenix, Arizona, 2018.

[22] B. Yan, Y. Li, S. Ren, I. M. Zainal Abidin, Z. Chen, and Y. Wang, "Recognition and evaluation of corrosion profile via pulse-modulation eddy current inspection in conjunction with improved canny algorithm," *NDT & E International*, vol. 106, pp. 18–28, 2019.

[23] M. Grosso, C. J. Pacheco, M. P. Arenas et al., "Eddy current and inspection of coatings for storage tanks," *Journal of Materials Research and Technology*, vol. 7, no. 3, pp. 356–360, 2018.

[24] B. Rao and B. Raj, "NDE methods for monitoring corrosion and corrosion-assisted cracking," in *Non-Destructive Evaluation of Corrosion and Corrosion-assisted Cracking*, John Wiley & Sons, Inc., 2019.

[25] J. Bailey, N. Long, and A. Hunze, "Eddy current testing with giant magnetoresistance (GMR) sensors and a pipe-encircling excitation for evaluation of corrosion under insulation," *Sensors*, vol. 17, no. 10, p. 2229, 2017.

[26] D. Rifai, A. Abdalla, K. Ali, and R. Razali, "Giant magnetoresistance sensors: a review on structures and non-destructive eddy current testing applications," *Sensors*, vol. 16, no. 3, p. 298, 2016.

[27] Y. Qu, H. Zhang, R. Zhao, L. Liao, and Y. Zhou, "Research on the method of predicting corrosion width of cables based on the spontaneous magnetic flux leakage," *Materials*, vol. 12, no. 13, p. 2154, 2019.

[28] T. Azizzadeh and M. S. Safizadeh, "Design and manufacturing of the magnetic flux leakage inspection system for detection of pitting corrosion in gas pipelines," *Iranian Journal of Manufacturing Engineering*, vol. 5, no. 2, pp. 43–49, 2018.

[29] R. Xia, J. Zhou, H. Zhang, L. Liao, R. Zhao, and Z. Zhang, "Quantitative study on corrosion of steel strands based on self-magnetic flux leakage," *Sensors*, vol. 18, no. 5, p. 1396, 2018.

[30] Y. Ege and M. Coramik, "A new measurement system using magnetic flux leakage method in pipeline inspection," *Measurement*, vol. 123, pp. 163–174, 2018.

[31] H. Zhang, L. Xu, R. Wu, and A. Simm, "Sweep frequency microwave NDT for subsurface defect detection in GFRP," *Insight-Non-Destructive Testing and Condition Monitoring*, vol. 60, no. 3, pp. 123–129, 2018.

[32] R. Zoughi, *3D Microwave Camera for Concrete Delamination and Steel Corrosion Detection*, Missouri S & T, 2018.

[33] S. Kharkovsky and R. Zoughi, "Microwave and millimeter wave nondestructive testing and evaluation - overview and recent advances," *IEEE Instrumentation & Measurement Magazine*, vol. 10, no. 2, pp. 26–38, 2007.

[34] M. Adhvaryu, P. N. Patel, and C. D. Modhera, "Apertured EBG-based microwave patch antenna for characterization of corrosion in steel rebar of civil structures," *Sensing and Imaging*, vol. 20, no. 1, p. 34, 2019.

[35] H. Zhang, B. Gao, G. Y. Tian, W. L. Woo, and L. Bai, "Metal defects sizing and detection under thick coating using microwave NDT," *NDT & E International*, vol. 60, pp. 52–61, 2013.

[36] A. Wahab, M. M. A. Aziz, A. R. M. Sam, K. Y. You, A. Q. Bhatti, and K. A. Kassim, "Review on microwave nondestructive testing techniques and its applications in concrete technology," *Construction and Building Materials*, vol. 209, pp. 135–146, 2019.

[37] M. Dvorsky, S. Barker, M. T. A. Ghasr, and R. Zoughi, *Microwave Imaging for Corroded Rebars and Delamination in Concrete Structures*, Missouri S & T, 2018.

[38] S. Mukherjee, X. Shi, L. Udpa, S. Udpa, Y. Deng, and P. Chahal, "Design of a split-ring resonator sensor for near-field microwave imaging," *IEEE Sensors Journal*, vol. 18, no. 17, pp. 7066–7076, 2018.

[39] N. N. Qaddoumi, A. H. El-Hag, and Y. Saker, "Outdoor insulators testing using artificial neural network-based near-field microwave technique," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 2, pp. 260–266, 2014.

[40] R. E. Jones, F. Simonetti, M. J. S. Lowe, and I. P. Bradley, "Use of microwaves for the detection of water as a cause of corrosion under insulation," *Journal of Nondestructive Evaluation*, vol. 31, no. 1, pp. 65–76, 2012.

[41] A. Mazzinghi, A. Freni, and L. Capineri, "A microwave nondestructive testing method for controlling polymeric coating of metal layers in industrial products," *NDT & E International*, vol. 102, pp. 207–217, 2019.

[42] D. Xiao and Z. Yunhua, "A novel compressive sensing algorithm for SAR imaging," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 2, pp. 708–720, 2014.

[43] M. D. Navagato and R. M. Narayanan, "Microwave imaging of multilayered structures using ultrawideband noise signals," *NDT & E International*, vol. 104, pp. 19–33, 2019.

[44] H. Zhang, Y. He, B. Gao, G. Y. Tian, L. Xu, and R. Wu, "Evaluation of atmospheric corrosion on coated steel using -band sweep frequency microwave imaging," *IEEE Sensors Journal*, vol. 16, no. 9, pp. 3025–3033, 2016.

[45] S. Zhong, "Progress in terahertz nondestructive testing: a review," *Frontiers of Mechanical Engineering*, vol. 14, no. 3, pp. 273–281, 2019.

[46] R. F. Anastasi and E. I. Madaras, "In terahertz NDE for under paint corrosion detection and evaluation," in *AIP Conference Proceedings*, pp. 515–522, Melville, NY, 2006.

[47] W. Tu, S. Zhong, A. Incecik, and X. Fu, "Defect feature extraction of marine protective coatings by terahertz pulsed imaging," *Ocean Engineering*, vol. 155, pp. 382–391, 2018.

[48] C.-W. You, C. Lu, T.-Y. Wang et al., "Method for defect contour extraction in terahertz non-destructive testing conducted with a raster-scan THz imaging system," *Applied Optics*, vol. 57, no. 17, pp. 4884–4889, 2018.

[49] B. Cao, M. Wang, X. Li, M. Fan, and G. Tian, "Noncontact thickness measurement of multilayer coatings on metallic substrate using pulsed terahertz technology," *IEEE Sensors Journal*, vol. 20, no. 6, pp. 3162–3171, 2020.

[50] Y. He, B. Gao, A. Sophian, and R. Yang, "Chapter 5 - Active Thermography and Eddy Current Excited Thermography," in *Transient Electromagnetic-Thermal Nondestructive Testing*, Y. He and Butterworth-Heinemann, Eds., pp. 93–121, 2017.

[51] S. Sfarra, C. Ibarra-Castanedo, P. Avdelidis et al., "A comparative investigation for the nondestructive testing of honeycomb structures by holographic interferometry and infrared thermography," *Journal of Physics: Conference Series*, vol. 214, 2010.

[52] C. Maierhofer, R. Arndt, M. Röllig et al., "Application of impulse-thermography for non-destructive assessment of concrete structures," *Cement and Concrete Composites*, vol. 28, no. 4, pp. 393–401, 2006.

[53] W. Zhu, X. Cai, L. Yang, J. Xia, Y. Zhou, and Z. Pi, "The evolution of pores in thermal barrier coatings under volcanic ash corrosion using X-ray computed tomography," *Surface and Coatings Technology*, vol. 357, pp. 372–378, 2019.

[54] M. Margret, M. Menaka, V. Subramanian, R. Baskaran, and B. Venkatraman, "Non-destructive inspection of hidden corrosion through Compton backscattering technique," *Radiation Physics and Chemistry*, vol. 152, pp. 158–164, 2018.

[55] R. Kant, P. S. Chauhan, G. Bhatt, and S. Bhattacharya, "Corrosion monitoring and control in aircraft: a review," in *Sensors for Automotive and Aerospace Applications*, S. Bhattacharya, A. K. Agarwal, O. Prakash, and S. Singh, Eds., pp. 39–53, Singapore, Springer Singapore, 2019.

[56] T. J. Stannard, J. J. Williams, S. S. Singh, A. S. Sundaram Singaravelu, X. Xiao, and N. Chawla, "3D time-resolved observations of corrosion and corrosion-fatigue crack initiation and growth in peak-aged Al 7075 using synchrotron X-ray tomography," *Corrosion Science*, vol. 138, pp. 340–352, 2018.

[57] B. C. Barlow, A. Situm, B. Guo, X. Guo, A. P. Grosvenor, and I. J. Burgess, "X-ray microprobe characterization of corrosion at the buried polymer-steel interface," *Corrosion Science*, vol. 144, pp. 198–206, 2018.

[58] J. Wilson, G. Y. Tian, I. Z. Abidin, S. Yang, and D. Almond, "Modelling and evaluation of eddy current stimulated thermography," *Nondestructive Testing and Evaluation*, vol. 25, no. 3, pp. 205–218, 2010.

[59] L. Bai, B. Gao, G. Y. Tian, W. L. Woo, and Y. Cheng, "Spatial and time patterns extraction of eddy current pulsed thermography using blind source separation," *IEEE Sensors Journal*, vol. 13, no. 6, pp. 2094–2101, 2013.

[60] A. Yin, B. Gao, G. Yun Tian, W. L. Woo, and K. Li, "Physical interpretation and separation of eddy current pulsed thermography," *Journal of Applied Physics*, vol. 113, no. 6, p. 064101, 2013.

[61] Y. He, B. Gao, A. Sophian, and R. Yang, "Chapter 12 - Through Coating Imaging of Early Marine Corrosion Using ECPPT," in *Transient Electromagnetic-Thermal Nondestructive Testing*, Y. He, B. Gao, A. Sophian, and R. Yang, Eds., pp. 241–255, Butterworth-Heinemann, 2017.

[62] Y. He, G. Y. Tian, M. Pan, D. Chen, and H. Zhang, "An investigation into eddy current pulsed thermography for detection of corrosion blister," *Corrosion Science*, vol. 78, pp. 1–6, 2014.

[63] R. Yang, Y. He, H. Zhang, and S. Huang, "Through coating imaging and nondestructive visualization evaluation of early marine corrosion using electromagnetic induction thermography," *Ocean Engineering*, vol. 147, pp. 277–288, 2018.

[64] H. Zhang, R. Yang, Y. He, A. Foudazi, L. Cheng, and G. Tian, "A review of microwave thermography nondestructive testing and evaluation," *Sensors*, vol. 17, no. 5, p. 1123, 2017.

[65] A. Foudazi, M. T. Ghasr, and K. M. Donnell, "Characterization of corroded reinforced steel bars by active microwave thermography," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 9, pp. 2583–2585, 2015.

[66] D. Pieper, K. M. Donnell, M. T. Ghasr, and E. C. Kinzel, "Integration of microwave and thermographic NDT methods for corrosion detection," *AIP Conference Proceedings*, vol. 1581, no. 1, pp. 1560–1567, 2014.

[67] S. A. Keo, F. Brachelet, F. Breaban, and D. Defer, "Steel detection in reinforced concrete wall by microwave infrared thermography," *NDT & E International*, vol. 62, pp. 172–177, 2014.

[68] E. Amin, J. Saha, and N. Karmakar, "Smart sensing materials for low-cost chipless RFID sensor," *IEEE Sensors Journal*, vol. 14, no. 7, pp. 2198–2207, 2014.

[69] O. O. Rakibet, C. V. Rumens, J. C. Batchelor, and S. J. Holder, "Epidermal passive RFID strain sensor for assisted technologies," *IEEE Antennas and Wireless Propagation Letters*, vol. 13, pp. 814–817, 2014.

[70] S. Soodmand, A. Zhao, and G. Y. Tian, "UHF RFID system for wirelessly detection of corrosion based on resonance frequency shift in forward interrogation power," *IET Microwaves, Antennas & Propagation*, vol. 12, pp. 1877–1884, 2018.

[71] A. Zhao, G. Y. Tian, and J. Zhang, "IQ signal based RFID sensors for defect detection and characterisation," *Sensors and Actuators A: Physical*, vol. 269, pp. 14–21, 2018.

[72] V. P. Plessky and L. M. Reindl, "Review on SAW RFID tags," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 57, no. 3, pp. 654–668, 2010.

[73] S. Shevchenko, A. Kukaev, M. Khivrich, and D. Lukyanov, "Surface-acoustic-wave sensor design for acceleration measurement," *Sensors*, vol. 18, no. 7, p. 2301, 2018.

[74] J. Zhang, G. Tian, A. Marindra, A. Sunny, and A. Zhao, "A review of passive RFID tag antenna-based sensors and systems for structural health monitoring applications," *Sensors*, vol. 17, no. 2, p. 265, 2017.

[75] A. Zhao, J. Zhang, and G. Y. Tian, "Miniaturization of UHF RFID tag antenna sensors for corrosion characterization," *IEEE Sensors Journal*, vol. 17, no. 23, pp. 7908–7916, 2017.

[76] R. Scapaticci, I. Catapano, and L. Crocco, "Wavelet-based adaptive multiresolution inversion for quantitative microwave imaging of breast tissues," *IEEE Transactions on Antennas and Propagation*, vol. 60, no. 8, pp. 3717–3726, 2012.

[77] H. Zhang, Y. du, J. Tang, G. Kang, and H. J. S. Miao, "Circumferential SH wave piezoelectric transducer system for monitoring corrosion-like defect in large-diameter pipes," *Sensors*, vol. 20, no. 2, p. 460, 2020.

[78] J. Hua, X. Cao, Y. Yi, and J. Lin, "Time-frequency damage index of Broadband Lamb wave for corrosion inspection," *Journal of Sound and Vibration*, vol. 464, article 114985, 2020.

[79] M. G. Fontana, *Corrosion Engineering*, Tata McGraw-Hill, 2005.

[80] L. Petricca, T. Moss, G. Figueroa, and S. Broen, "Corrosion detection using A.I : a comparison of standard computer vision techniques and deep learning model," in *Proceedings of the Sixth International Conference on Computer Science, Engineering and Information Technology*, p. 99, Limerick City, Ireland, 2016.

[81] J. Liu, Y. Liu, and Y. Ke, "Detection and analysis of a quay crane surface based on the images captured by a UAV," *Remote Sensing Letters*, vol. 11, no. 1, pp. 76–85, 2020.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5548614, 17 pages
https://doi.org/10.1155/2021/5548614

Research Article

A Cyclic Consistency Motion Style Transfer Method Combined with Kinematic Constraints

Huaijun Wang,[1,2] Dandan Du,[1] Junhuai Li,[1,2] Wenchao Ji,[1] and Lei Yu[1,2]

[1]*School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China*
[2]*Shaanxi Key Laboratory for Network Computing and Security Technology, Xi'an 710048, China*

Correspondence should be addressed to Junhuai Li; lijunhuai@xaut.edu.cn

Received 18 February 2021; Accepted 24 May 2021; Published 30 June 2021

Academic Editor: Bin Gao

Motion capture technology plays an important role in the production field of film and television, animation, etc. In order to reduce the cost of data acquisition and improve the reuse rate of motion capture data and the effect of movement style migration, the synthesis technology of motion capture data in human movement has become a research hotspot in this field. In this paper, kinematic constraints (KC) and cyclic consistency (CC) network are employed to study the methods of kinematic style migration. Firstly, cycle-consistent adversarial network (CCycleGAN) is constructed, and the motion style migration network based on convolutional self-encoder is used as a generator to establish the cyclic consistent constraint between the generated motion and the content motion, so as to improve the action consistency between the generated motion and the content motion and eliminate the lag phenomenon of the generated motion. Then, kinematic constraints are introduced to normalize the movement generation, so as to solve the problems such as jitter and sliding step in the movement style migration results. Experimental results show that the generated motion of the cyclic consistent style transfer method with kinematic constraints is more similar to the style of style motion, which improves the effect of motion style transfer.

1. Introduction

Motion capture technology is based on the principles of computer graphics, recording the human body motion process through motion capture devices [1]. When the motion capture system is performing motion capture, it can track the motion trajectory of the moving object in the three-dimensional space and obtain the motion information of the moving object in the three-dimensional space through calculation processing. It has high precision, high quality, and complete motion information when representing human movements. The combination of motion capture data and computer animation technology can realistically restore actions. In recent years, it has been widely used in movies, games, medical treatment, sports, and other fields [2, 3]. In the field of film production, the use of motion capture technology has been particularly successful. Many animated films that use motion capture technology have achieved good box office results. In "Alita: Battle Angel," the use of motion capture technology to capture actors' actions and expressions is processed by computer animation technology, making it difficult for viewers to distinguish the boundary between reality and animation. In the field of game production, the application of motion capture technology makes the characters in the game more realistic. High-precision motion capture data ensures the fluency of fighting actions and brings better game experience to players [4]. In the medical field, Noitom's "Dr. Joint" [5] uses motion capture technology to address postoperative rehabilitation problems of knee patients and helps rehabilitation training by recording the patient's activity and gait data. In addition, in motion training, the motion capture system can capture the detailed sports situation of the athletes, so as to better analyze the problems of the athletes, and make corresponding adjustments to achieve better training goals.

With the widespread application of motion capture technology in film, animation, and other production fields [3], research institutions such as Carnegie Mellon University, the University of Edinburgh, and the University of Bonn have established huge human motion capture databases.

Due to differences in collection objects and collection sites, it is necessary to recollect motion capture data for the same type of action, resulting in low reusability of motion capture data and increasing the cost in practical applications. Data-driven motion synthesis is a key technology to realize the reuse of human motion data. Through the study of human motion synthesis methods such as motion style transfer, motion retargeting, and motion blending, based on the existing motion capture data, the motion data that meets the needs of users is synthesized [6]. The collection work that originally required repeated collection actions or even replacement of collection objects can now be reduced by human motion synthesis technology, saving a lot of manpower and material resources, and improving the production efficiency of movies, animations, etc. The editing and synthesis methods of motion capture data have high research and practical application value.

At present, deep learning provides great convenience for motion style transfer, without the need for complex data preprocessing. However, there are still two problems with the motion style transfer method based on deep learning: first, because the motion capture data is time series data, the pooling process of neural network reduces the temporal correlation of motion data when extracting motion features, resulting in the difference of motion of generated motion and content motion at the same time, and the phenomenon of generated motion lags relative to content motion; second, the reconstruction of motion features results in the missing of some motion data frames, which leads to some problems such as jitter and sliding step in the generated movement after the motion style migration. In this paper, by determining the training target of kinematic constraint loss function and combining kinematic constraint with cyclic consistent confrontation generation network, the problems of animation jitter and sliding step in the process of style transfer are solved and make the style of the generated motion and the style motion closer.

The paper contribution: by combining the cyclic consistent style transfer method with kinematic constraints (KC), the motion style transfer network based on the convolutional autoencoder is used as the generator, and the cyclic consistent generation adversarial network (CCycleGAN) is constructed to establish cyclical consistency constraints between generated motion and content motion to further improve the consistency of generated motion and content motion; introduce kinematic constraints, standardize generated motion, solve problems such as jitter and sliding in the result of style transfer, and improve motion style transfer effect.

2. Related Research

With the development and application of motion capture technology, human motion synthesis technology based on motion capture data has attracted more and more attention from researchers at home and abroad and has made considerable progress.

2.1. Motion Blending. Early motion data is mainly composed of high-level motion parameters such as joint angles and joint coordinates. Therefore, technologies in the field of image and signal processing are used in the design, modification, and adaptation of motion data. Human motion data is treated as a time series signal for editing or fusion. Troje [7] proposed a motion synthesis framework that encodes motion patterns and uses linear methods for motion analysis and motion synthesis. Shapiro et al. [8] proposed an interactive motion data editing method, which uses independent components to analyze the motion style in the motion data and reedit the motion data to change the motion style. Wang and Bodenheimer [9] used the linear mixing method to determine the transformation point on the motion sequence by calculating the optimal weight of the basic cost metric.

Since it is difficult to directly synthesize more complex or obviously different motion styles with the method of signal processing on motion data, in order to solve the problem of poor synthesis of complex motion, some scholars establish kinematic constraints during motion generation to achieve smooth processing of generated motion [10]. With the improvement of animation effect requirements, in order to deal with more complex motion data, nonlinear processing methods [11] are applied to motion capture data with complex structures.

2.2. Methods Based on Statistics and Learning. Some scholars use statistics and learning methods to analyze the motion data, extract representative motion features and motion patterns in the motion data, and change the motion mode by adding constraints to generate new motions while retaining the existing motion characteristics.

Matthew and Hertzmann [12] learn the motion pattern of each motion style from a set of motion data sequences containing multiple motion styles. Each motion sequence can have a different choreography, and each choreography element has a different style. Through learning it can identify the general arrangement elements in the sequence and use interpolation to synthesize new motion data according to the action choreography elements. Grochow et al. [13] proposed an inverse kinematic system based on a human pose learning model. Given a set of kinematic constraints, the poses that are most likely to meet these constraints can be generated. The system uses different motion data for learning, generates the probability distribution of the motion sequence pose, determines the probability of a motion pose in the motion pose space through the objective function, and matches the pose to generate a new motion.

2.3. Motion Graph. In 2002, Kovar et al. [14] first proposed the concept of "motion graph," through the relationship between different motion data constructed, search for the optimal path in the constructed motion graph, and synthesize a new motion sequence. Arikan and Forsyth [15] proposed a framework for synthesizing motion by editing motion capture data. They regard motion synthesis as a combination problem and combine them by randomly searching the hierarchical structure of motion graphs. Since motion graph can only combine and edit motion capture data to

meet user needs, Min and Chai [16] enrich motion capture datasets by mixing the same type of motion data or combining sketches. The construction of a motion graph requires more on the quantity and type of motion capture data in order to be able to express the changes of the entire motion, and the new motion generated finally depends too much on the existing motion dataset. Parametric motion synthesis is to add the human body's footing, speed, acceleration, and other parameters to the synthesis model, control the synthesis process, and improve the problems of animation jitter and foot sliding [17].

2.4. The Deep Learning Approach. At present, applying deep learning to human motion capture data has become the main method of motion style transfer. Deep learning is used to synthesize new data, and the framework based on deep learning automatically learns features from the dataset. Taylor et al. [18] applied restricted Boltzmann machines to synthesize animation. On this basis, Mittelman et al. [19] proposed a structured constrained Boltzmann machine to improve the animation reconstruction. Subsequently, Fragkiadaki et al. [20] used an autoencoder (AE) recursive decoder network, which is a recurrent neural network that combines deep learning with time dynamics and produces smooth interpolated motion while reducing slipping. To further improve the animation effect, Du et al. [21] used multisource large-scale motion datasets to construct a hierarchical recurrent neural network and synthesized smooth and natural motion animation. In the motion editing method proposed by Holden et al. [22], a single-layer convolutional autoencoder is used for feature extraction, which also shows a better ability to express motion data, which promotes the autoencoder (AE) using in motion synthesis. In motion style transfer, Zan [23] establishes a self-encoding network structure with three convolutional layers and establishes style constraints in the feature space to realize motion style transfer. A novel data-driven framework is present for motion style transfer [24], which supports style extraction from videos and learns from an unpaired collection of motions with style labels. In this paper, Yu et al. propose that style translation is an effective way [25] to transform adult motion capture data to the style of child motion. Our method is based on CycleGAN.

The deep learning method based on autoencoder (AE) improves the effect of motion data synthesis or motion style transfer. However, encoding the motion data will cause a certain amount of data loss, which leads to jitter and slipping in the result of motion style transfer. In this paper, the method of movement style transfer is studied by combining kinematic constraints to improve the reuse rate of motion capture data.

3. The Whole Process of Cycle-Consistent Motion Style Transfer Combined with Kinematic Constraints

Style motion and content motion have more similar motion features, so the generated motion can maintain a high level of consistency with the content motion. However, the collection of style motion is difficult and the types of actions are relatively few. There are relatively few content motion and style motion with similar motion content. When performing motion style transfer, using content motion with similar actions for motion style transfer can improve the transfer effect; when the content of the content motion and the style motion is quite different, the generated motion can remain similar to the content motion at the same time. However, the generated movement has obvious motion lag and causes the movement direction to deviate. Therefore, it is necessary to improve the similarity between the generated motion and the content motion while maintaining a high style similarity between the generated motion and the style motion. The transfer of motion style mainly involves problems such as difficulty in extracting motion features, poor reconstruction of motion effects, and establishment of motion style constraints. The overall process of motion style transfer of motion feature extraction and motion reconstruction network is shown in Figure 1.

It can be seen from Figure 1 that in order to transfer a specific style motion to content motion, it is first necessary to extract the motion features from the input motion data and then reconstruct the motion from the motion features to make the reconstructed motion data consistent with the input motion data. In order to realize the transfer of motion style, the style of the reconstruction motion to establish reasonable constraint, guarantee the reconstruction motion in reserves the content motion, content motion has the style of motion style and outputs the generated motion of the motion style transfer. Therefore, the motion style transfer network has the same structure as the motion feature extraction and motion reconstruction network, and the parameters are shared. The motion style constraints are established in the hidden layer feature space of the network, the motion characteristics are adjusted, and the motion style transfer is realized through motion reconstruction.

Aiming at the data loss caused by the use of autoencoder to encode the motion data in the process of motion style transfer, a cyclic consistent (CC) style transfer method combined with kinematic constraints (KC) proposed in this paper mainly includes two steps: (1) construction a cyclic consistent generated adversarial network; (2) combined with kinematic constraints to establish a cyclic consistent style transfer model.

4. The Construction of a Cyclic Consistent Generated Adversarial Network

4.1. Theoretical Basis. In the field of image processing, it is possible to use the cyclic consistent generated adversarial network to convert two image sample domains (nonpaired image domains) with large differences in style and improve the effect of nonpaired image style transfer [26]. Figure 2 shows the cycle-consistent generated adversarial network model. The cyclic consistent generated adversarial network is used for image-to-image mapping learning, and the learning method uses unpaired images for style transfer.

First, there are two unpaired image sample spaces X and Y with different contents. The goal of generating an adversarial network is to learn the mapping from X to Y. This

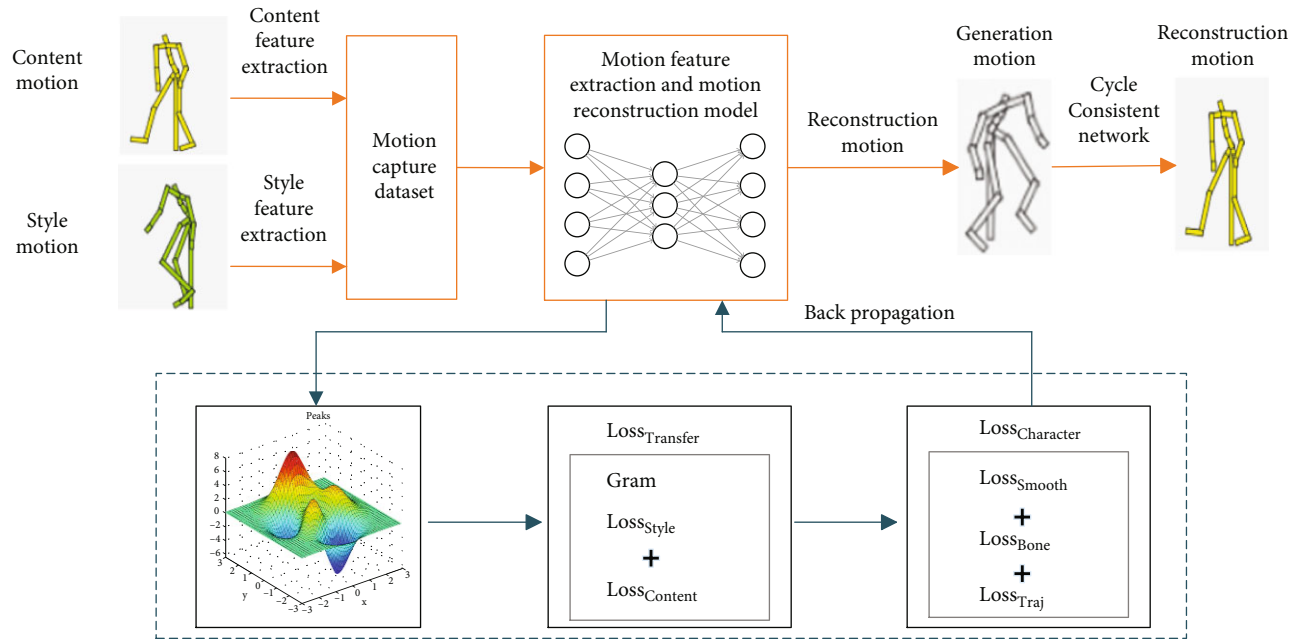


Figure 1: Motion style transfer process of motion capture data.

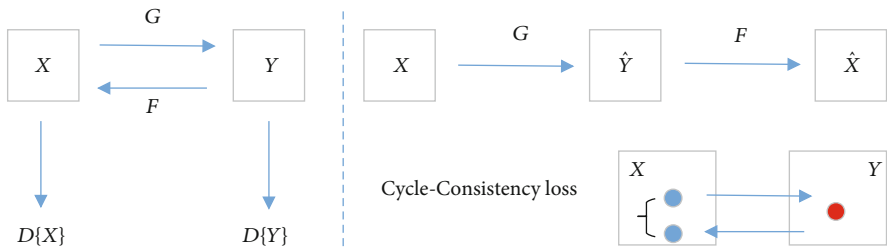


Figure 2: Cycle-consistent generative adversarial networks.

mapping is G, which corresponds to the generator in the generation adversarial network.

$$\hat{Y} = G(X, Y). \tag{1}$$

Among them, the generator G can convert the picture in the sample space X into a fake picture $\hat{Y}$ similar to the image sample space Y, and it is hoped that the style of the generated image $\hat{Y}$ and Y is as similar as possible.

For the generated picture $\hat{Y}$, the discriminator $D\{Y\}$ is used to determine whether it is a real Y picture, thereby forming a generated adversarial network.

$$\text{Loss}_{\text{GAN}}(G, D, A, B) = \log D(Y) + \log\left(1 - D(G(X))\right). \tag{2}$$

Using only this one loss will cause the mapping G to map all the images in the sample space X to the same image in the Y space, invalidating the loss. Therefore, by introducing the mapping F, $\hat{Y}$ can be transformed into a picture $\hat{X}$ similar to the sample space X.

$$\hat{X} = F\left(\hat{Y}, X\right). \tag{3}$$

And then establish the connection between $\hat{X}$ and X, forming a circular consistency constraint.

$$\text{Loss}_{\text{Cyc}}(G, F, X, Y) = ||F(G(X)) - X||_2^2. \tag{4}$$

The cyclic consistency constraint is applied to the image style transfer, and the information of the content image during the transfer can be retained as much as possible, so that the generated image after the transfer is more complete and natural. Applying it to motion style transfer can reconstruct the generated motion and establish the connection between the generated motion and the content motion through the cyclic consistency constraint, so that the generated motion retains more motion content and improves the consistency of the generated motion and the content motion and improves the effect of motion style transfer.

4.2. Establishment of Cyclic Consistency Constraint. In the motion style transfer, two sample spaces are defined: content motion C and style motion S. The cyclic consistency constraint (CC) is applied to the motion style transfer to establish a cyclic consistency style transfer model, as shown in Figure 3.

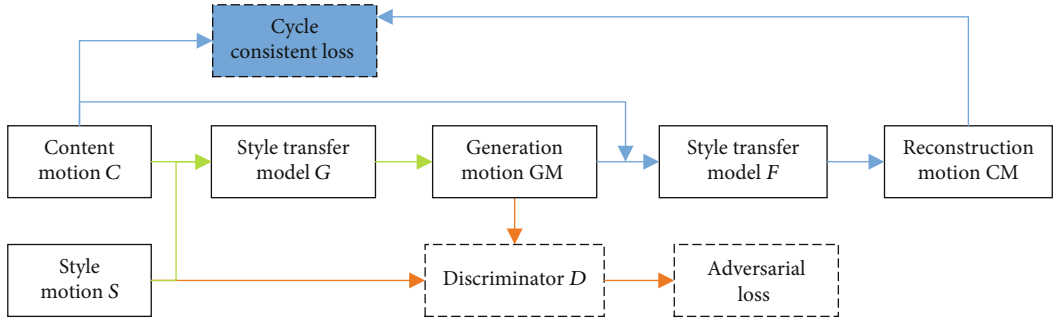


Figure 3: Cyclic-consistent style transfer model.

In the motion style transfer model, generators G and F are all motion style transfer models based on motion capture data. The content motion C and the style motion S are used as inputs of the generator G to perform motion style transfer to obtain the generated motion GM.

$$\mathrm{GM} = G(C, S). \tag{5}$$

The discriminator D judges the style difference between the generated motion GM and the real style motion S through equation (6), so that the motion style of the generated motion is close to the motion style of the input S.

$$\mathrm{Loss}_{\mathrm{Style}} = ||\mathrm{Gram}(H_s) - \mathrm{Gram}(R(H_s))||_2^2. \tag{6}$$

The motion feature H_s of the input style motion S and the motion feature H_{GM} of the generated motion GM are obtained through the discriminator D to obtain the confrontation loss:

$$\mathrm{Loss}_{\mathrm{GAN}} = \log D(H_s) + \log\left(1 - D(H_{\mathrm{GM}})\right). \tag{7}$$

Generative adversarial networks generally measure the generation effect through the log loss function [27].

The use of the discriminator for training will emphasize the features of the motion style, making it difficult for the generator to retain the motion content and structure of the content motion, and it is necessary to add cyclic consistency constraints to encourage the content of the motion to be retained in the alignment process. The generated motion GM is transferred to the generation network F, the motion style of the content motion C is transferred to the generated motion GM, the motion style of the style motion S in the generated motion GM is removed, and the reconstructed motion CM of the generated motion GM is obtained.

$$\mathrm{CM} = F(\mathrm{GM}, C). \tag{8}$$

The paradigm $L2$ is used to establish the consistent loss of the reconstructed motion CM and the content motion C, thereby effectively achieving cyclic consistency, so that the generated motion GM has more motion features of the content motion during reconstruction:

$$\mathrm{Loss}_{\mathrm{cyc}} = ||\mathrm{CM} - C||_2^2. \tag{9}$$

The cyclic consistency constraint makes the generated motion after the style transfer reconstitutes the original input content motion. By establishing the cyclic consistency constraint, the transferred motion style in the generated motion is removed to form a cycle. In this way, let the network learn the process of motion style transfer and then remove, so that the generated motion has more content motion features.

5. Cyclic Consistent Style Transfer Method Combined with Kinematic Constraints

The cyclic consistency constraint can establish the connection between the content motion and the generated motion and enhance the consistency of the generated motion and the content motion. However, due to the complex structure of the motion capture data, the inheritance relationship between the bone joint points makes the motion data highly correlated, and the data of the autoencoder is lossy, resulting in a gap between the motion generated after the encoding and decoding operation and the content motion. The resulting motion data frame is unreasonable, the action content is incomplete, and problems such as jitter and sliding footsteps occur.

At present, the method of solving motion jitter and foot slippage in motion synthesis is to add constraints to the motion synthesis results and standardize the motion data. By establishing kinematic constraints (KC), dynamic constraints and spatiotemporal constraints, and other methods [28], constraints are added to the generated motion to obtain complete and smooth and natural motion data. Lee and Shin [29] first used inverse kinematics to establish kinematic constraints for each frame of motion data and used multilevel spline curve interpolation to achieve smooth complete motion. Tak and Ko [30] added dynamic constraints on the basis of the previous kinematic constraints and transformed the spatiotemporal optimization problem into a constraint state estimation problem. Choi and Ko [31] based on inverse kinematics to calculate the joint angle change from the position of the extremity to realize the editing of motion data. Gleicher [32] realized motion synthesis based on spatiotemporal constraints but did not consider the kinematics and dynamic constraints of the generated motion and lacked the reality of motion. Zhang et al. [33] propose a motion retargeting method based on spatiotemporal constraints, which imposes spatiotemporal constraints on joint positions to avoid unreasonable motion. The kinematic constraints

established by Grochow et al. [13] use end effectors to clarify the position that the extremity needs to reach. Zhou et al. [34] construct a variety of kinematic constraints to edit motion data to realize motion retargeting.

Among kinematic constraints (KC), dynamic constraints, and spatiotemporal constraints, the motion synthesis method based on spatiotemporal constraints is computationally expensive and time-consuming. Dynamic constraints require fine-grained parameter control of the motion frame. Usually, dynamic parameters such as speed and acceleration are used to directly modify the motion features. It is necessary to rely on experience to achieve parameter control, and the generated results are uncertain. Kinematic constraints are further constraints on the consistency between the generated motion and the content motion and are targeted constraints. By determining the kinematic constraints to generate the motion, it provides a more reasonable generated motion for the cyclic consistent constraint and improves generation of the consistency of motion and content motion. Kinematic constraints have a large range of optional constraints and strong applicability. In this paper, three common constraints, such as smooth constraint, bone length constraint, and trajectory constraint, are selected for smoothing processing in character animation. Kinematic constraint loss function training objective is determined to combine kinematic constraint with cyclic consistent resistance generation network to solve the problems of jitter and sliding. The style transfer model is shown in Figure 4.

The kinematic constraints in the style transfer model shown in Figure 4 mainly include three aspects:

(1) Motion smoothing constraint

Villegas et al. [35] found that the data frames of continuous motion are highly dependent on the previous and subsequent data frames when performing motion retargeting, that is, the motion of each frame in the motion data is slightly changed compared with the motion of the previous frame, which can be generated by generating motion and content motion. The speed changes of the front and back data frames are used to the motion smoothly constrain and solve the problem of generating motion sliding. The smoothing constraint is defined as follows:

$$\text{Loss}_{\text{Smooth}}(H) = \sum_{j} \| v_j - v_{j-1} \|^2 - \| v_j{}' - v_{j-1}{}' \|^2. \tag{10}$$

Among them, $v_j{}'$ is the motion speed of the joint point j in the three-dimensional space coordinate system in the content motion, and v_j is the motion speed of the joint point j in the three-dimensional space coordinate system that generates the motion.

(2) Bone length constraint

The motion capture data collected by the same motion capture device has the same bone hierarchy, but the length of the bones is not the same. The generated motion data needs to be consistent with the bone data of the content motion. When Villegas et al. [35] use motion features to reconstruct motion data, it uses bone length constraints to ensure that the bones that generate motion will not be deformed and avoid the jitter of generated motion. This paper uses the three-dimensional space coordinates of the joint points as input data and imposes bone length constraints between adjacent joint points to maintain the stiffness of the body, so that the movement body that generates the motion will not cause movement dislocation due to deformation. The loss function of the bone length constraint is defined as follows:

$$\text{Loss}_{\text{Bone}}(H) = \sum_{i} \sum_{b} \left| \| p_{b1}^{i} - p_{b2}^{i} \| - l_b \right|^2. \tag{11}$$

Among them, i represents the number of motion frames, b represents the number of human bones, and p_{b1}^{i} and p_{b2}^{i} are the two end joint points that generate a segment of bone in motion. l_b is the length of bone b.

(3) Motion trajectory constraint

The motion style transfer hopes that the generated motion follows the motion trajectory of the content motion, so that the motion postures of the generated motion and the content motion are synchronized. Therefore, the motion needs to be precisely restricted to a certain trajectory. Holden et al. [22] edit and generate the motion data of the given trajectory through the high-level motion parameter of the given trajectory and generate the trajectory route of the motion through trajectory constraints. The motion trajectory constraint loss function is defined as follows:

$$\text{Loss}_{\text{Traj}}(H) = \| w - w' \|^2 + \| v_r - v_r{}' \|^2. \tag{12}$$

Among them, w is the axis angular velocity of the generated motion around the y axis, w' is the axis angular velocity of the content motion, v_r is the motion speed of the generated motion root joint point, and $v_r{}'$ is the motion speed of the content motion root joint point.

Therefore, the loss of the three kinematic constraints combined with smoothing constraint, bone length constraint, and trajectory constraint is defined as follows:

$$\text{Loss}_{\text{KC}} = \arg \min_{H} \text{Loss}_{\text{Smooth}}(H) + \text{Loss}_{\text{Bone}}(H) + \text{Loss}_{\text{Traj}}(H). \tag{13}$$

Through training, kinematic constraint loss is minimized to obtain kinematic constraint motion features of the hidden layer, which can constrain the joint to the desired position while maintaining the stiffness of each bone. The kinematic constraints are adjusted to generate the motion backpropagation to the feature H of the hidden layer, until the hidden layer feature that can minimize the kinematic constraint value is obtained to realize the kinematic constraint. Therefore, the overall transfer loss of the circular consistent style transfer model combined with kinematic constraints is as follows:

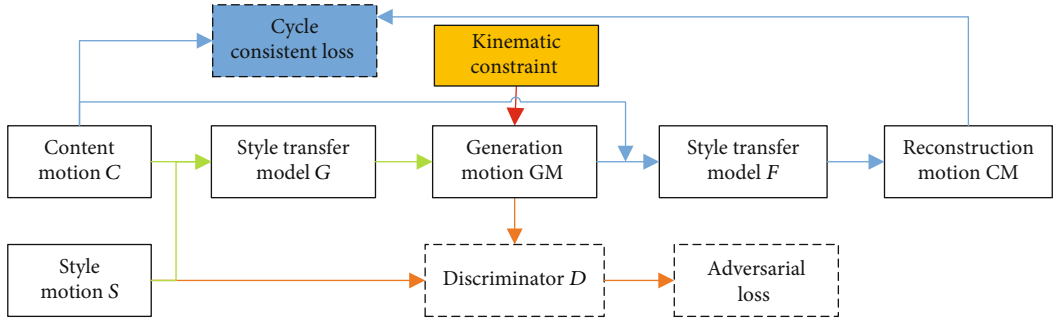


Figure 4: Cyclic-consistent style transfer model with kinematic constraints.

$$\text{Loss}_{\text{cycGAN}} = \text{Loss}_{\text{KC}} + \min_{G}\max_{D}\text{Loss}_{\text{GAN}} + \text{Loss}_{\text{cyc}}. \quad (14)$$

Among them, L_{Kine} is the kinematic constraint loss, Loss_{GAN} is the adversarial loss, and Loss_{cyc} is the cyclic consistency loss.

The specific network structure of the cyclic consistent style transfer model combined with kinematic constraints is shown in Table 1. The generated network takes the content motion and the style motion as the input of the generator to get the generated motion, and the discriminant network judges the style difference between the generated motion and the real style motion, so that the generated motion style is close to the input motion style. The two-generation networks have the same network structure as the motion style transfer network based on convolutional autoencoding, and the network parameters are shared. The discriminant network structure and motion feature extraction are the same as the coding network structure of the motion reconstruction model, sharing network parameters.

6. Experiment and Analysis

6.1. Data Processing and Model Training. In this paper, the CMU motion capture dataset [36] was used, with 2600 sets of BVH motion data of about 3 million frames as training data. The input motion capture data dimension is 73, consisting of 21 bone joints, plus the initial coordinate value of the bone root joint to form the data dimension. In order to process input data with large dimension, the number of hidden units in the neural network is generally more than twice of the data dimension, that is, the number of hidden units should be greater than 146, so the number of hidden units is set as 256. The collection frequency of the motion capture dataset in this paper is 120 frames per second. A human body motion will last for about 1 second to 2 seconds. In order to maintain a certain degree of integrity and continuity of the motion data of each batch, all motion data will be split once every two seconds and retain the action content of the second one after the previous motion, that is, the data is split by a 50% overlap window of 240 frames, and the motion data segment with less than 240 frames is filled with the last frame of the current data segment. Therefore, the size of the filter in the network is $5 * 5 * 1$, corresponding to about half a second of motion data, which is a reasonable sequence length for most motion [11]. Due to the large training dataset, in order to improve the training speed, gradient descent uses Adam to update the parameters W_0 and b_0 and sets the learning rate alpha = 0.01 [37]. Use all datasets for 100 complete training, that is, epoch = 100, and use all the data to update the network parameters in each backpropagation, that is, batchsize = 1. The initialization of $W_0 \in R^{m*d*w_0}$ selects a small random value, and the initialization of $b_0 \in R^m$ is 0.

Although the Euler angle representation method of BVH motion capture files is intuitive and convenient in animation playback, the Euler angle rotation component performs poorly in characterizing the spatiotemporal characteristics of motion. Therefore, this paper first transforms the BVH Euler angle data into three-dimensional space coordinates of joint points.

At the same time, in order to make the trained network have better stability, the motion capture data is normalized when constructing the dataset used for autoencoder training [22]. All spatial coordinates in BVH dataset are normalized as follows:

$$X = \frac{X_{\text{input}} - X_{\text{mean}}}{X_{\text{std}}}, \quad (15)$$

where X_{input} is the input motion capture dataset, X_{mean} is the average value of the input motion data, X_{std} is the standard deviation of the input motion data, and X is the standard motion data after processing.

This paper selects three styles of old people, zombies, and orangutans and transfers them to ordinary people's walking and running, respectively, and analyzes the effect of the cyclic consistent style transfer method combined with kinematic constraints. That is, the three kinds of motions of old people, zombies, and orangutans are style motions, and ordinary people's walking is the content motion. The style motion and content motion are used as the input data of the network to transfer the motion style to obtain the generated motion.

6.2. Experimental Analysis Process. This paper is mainly from the following six aspects of the experiment.

(1) Analysis on the results of motion style transfer

The migration results of ordinary people's walking and the three styles of motion are shown in Table 2. The three

Table 1: Cyclic-consistent style transfer networks with kinematic constraints.

	Layer	Shape	Param
Generated network	Input	(none, 256, 73)	
		Encoder network	
	DropoutLayer		Dropout = 0.25
	Conv1DLayer	(none, 73, 240)	
	Pool1DLayer	(none, 256, 240)	
		Decoder network	
	Depool1DLayer	(none, 256, 240)	
	DropoutLayer		Dropout = 0.25
	Conv1DLayer	(none, 256, 240)	
	Output	(none, 256, 73)	
		Adversarial network	
Discriminant network	DropoutLayer		Dropout = 0.25
	Conv1DLayer	(none, 73, 240)	
	Pool1DLayer	(none, 256, 240)	

styles of motion are transferred to ordinary people's walking movement through the cyclic consistent style transfer model combined with kinematic constraints. Compared with the style transfer results without constraints, the motion poses of the three style motion transfer results are closer to those of the content motion, and the style transfer effect is good.

It can be seen by performing the style transfer of the old people with or without constraints on the same posture that the difference in the effect of the constrained style transfer compared with the unconstrained style transfer is mainly reflected in the processing of complex motions, especially in the last column of the turning action; the constrained style transfer can maintain high consistency between generated motion and content motion. In contrast, unconstrained style transfer, although the motion posture is close to the content motion, is affected by the style motion, and there is room for improvement in the relative positions of joints and human body orientation; the obvious difference in the results of zombie style motion with and without constraint style transfer is the orientation of the generated motion and the content motion posture. In this group of forward and turn motions, the generative motion of unconstrained style transfer is relatively lagging. When the content motion turns, the generated motion is still going straight. Obviously, the style transfer effect with kinematic constraints is better for the posture constraints at the same time; the style transfer result of the orangutan style motion is with or without constraints, although the generative motion is with unconstrained style transfer. The posture and orientation are close to the content movement, but the relative positions of the feet are different compared to the content motion. In the constrained style transfer result, the relative positions of the feet of the generated motion are clearer and the footing point is more clear.

(2) Analysis on the effect of cyclic consistent constraint and kinematic constraint

In order to verify the effect of cyclic consistent constraint and kinematic constraint, the motion style transfer result (unconstrained transfer result) based on convolutional autoencoder is compared with the cyclic consistent style transfer result (constrained transfer result) combined with kinematic constraint.

The results of motion style transfer on the same level are mainly analyzed by observing the posture and footsteps of the human body. From the migration results of unconstrained migration and constrained migration in Table 3, it can be seen that the constrained generated motion and content motion in the style transfer of old people walking are more consistent in posture; the footsteps of generated motion in the style transfer of zombie walking. The horizontal contact is normal, which effectively solves the problem of ground penetration; the generated motion in the style transfer of the orangutan walking can move according to the position of the content motion.

(3) The trajectory of motion style transfer results

In order to further illustrate the effect of motion style transfer, the motion trajectory diagram is visualized. Figure 5 is a trajectory diagram of ordinary people walking and three styles of motion.

The following moves the walking movement of ordinary people to the three styles of old people, zombie, and orangutan. The results of the migration trajectory are shown in Table 4.

Compared with unconstrained style migration, the trajectory diagram of constrained style migration is closer to the content motion and the trajectory is more complete. Compared to the content motion, generated with constraint style migration movement trajectory distance is shorter, and generated trajectory diagram is small; this is mainly due to the old style, zombie style speed slowly, leading to generated movement of the whole movement distance is short,

Table 2: Comparison of unconstrained and constrained transfer results between ordinary people walking and the three styles of motion.

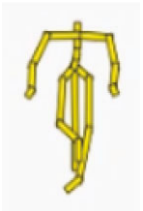
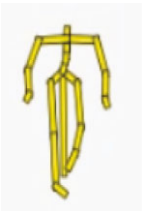
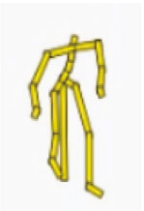
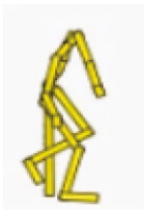
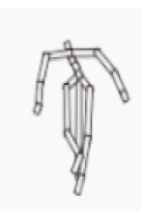
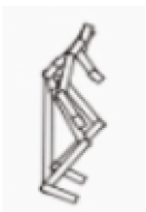
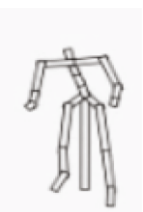
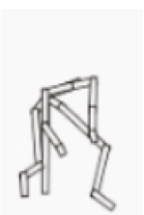
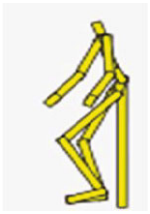
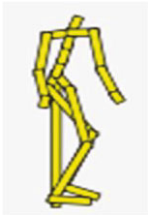
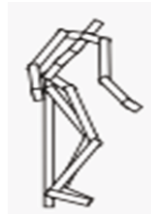
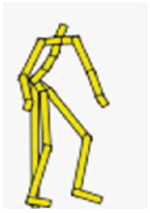
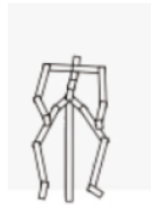
Motion type		Motion I	Motion II	Motion III	Motion IV
Content motion: ordinary people's walking					
Walking style of the old people	Unconstrained				
	Constrained				
Walking style of zombies	Unconstrained				
	Constrained				
Walking style of orangutans	Unconstrained				
	Constrained				

Table 3: Results of unconstrained and constrained movement style migration.

Behavior	Content motion	Unconstrained migration	Constrained
Old people—walking migration		The generated motion poses are similar, but the orientation is obviously different	
		Although the generated motion can maintain a posture similar to the content motion, the motion is lagging	
Zombies—walking migration		Motion lag	
		The overall position of the generated motion at the same level is lower than that of the content motion, and the right foot penetrates the horizontal plane	
Orangutans—walking migration		The problem of large distance between footholds is prone to occur, leading to slipping	
			

but from the whole, adding constraints for generated movement can effectively promote the migration effect.

Due to the characteristics of the orangutan style movement with large strides and fast speed, the movement distance of unconstrained generated movement is significantly larger, so the movement range exceeds the trajectory collection range. However, since the movement speed is faster than that of the old man and the zombie style, the constrained generated movement trajectory is closer to the content movement trajectory, and the migration effect is good. In addition, the constrained generated motion is significantly larger in distance between the feet and the stride length compared

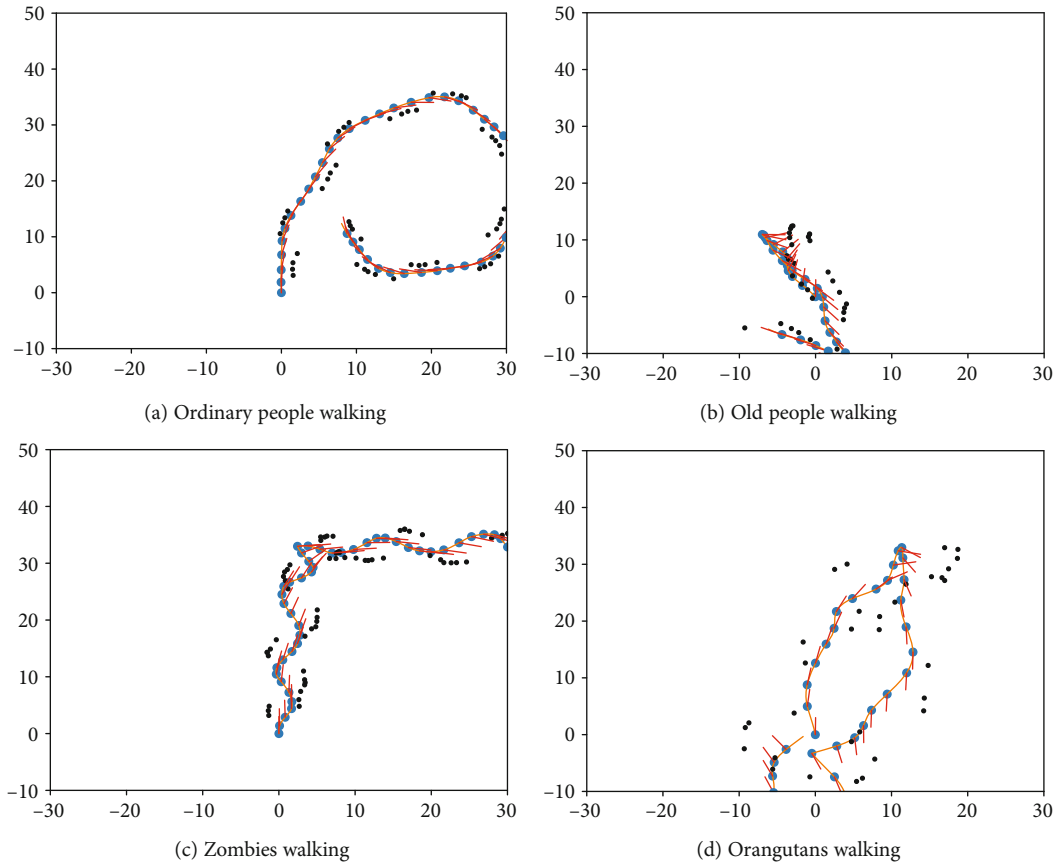


Figure 5: Content motion and style motion trajectories.

with the content motion, which is also in line with the characteristics of the large stride length of the orangutan style, indicating that the style feature retention effect is good.

According to the above analysis, unconstrained style transfer is prone to problems such as jitter and slippage caused by unclear foothold position, foot end penetration, and motion lag. The constrained style transfer can effectively solve the constraints on joint positions and motion trajectories, thereby further improving the effect of style transfer.

(4) Analysis of the style transfer results of unpaired motion data

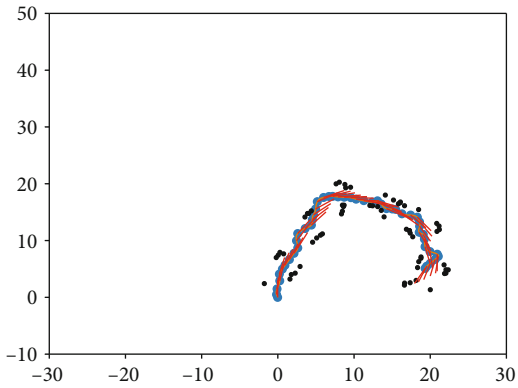
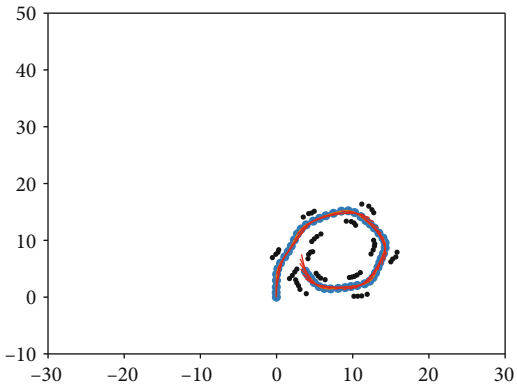
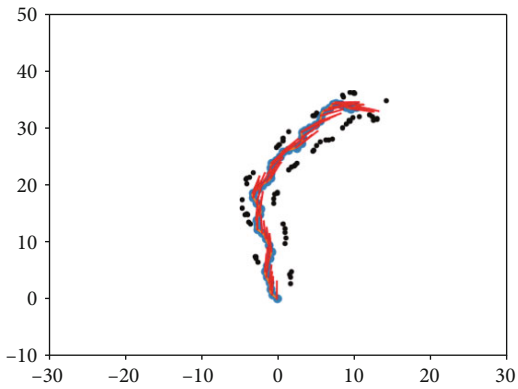
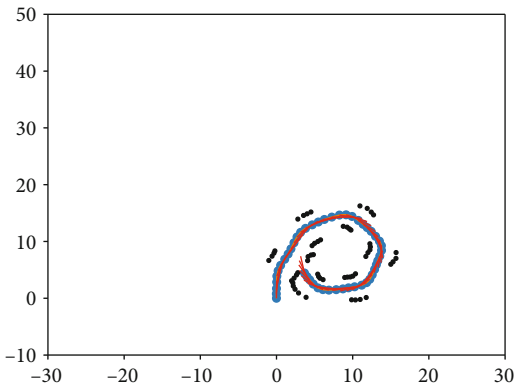
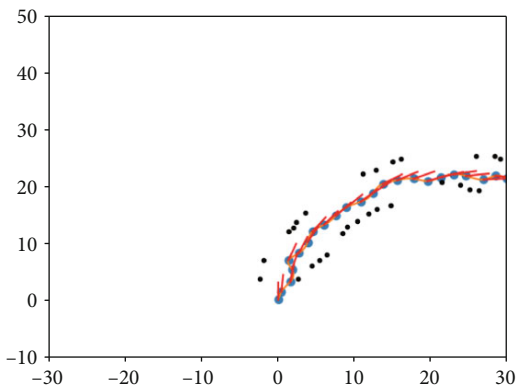
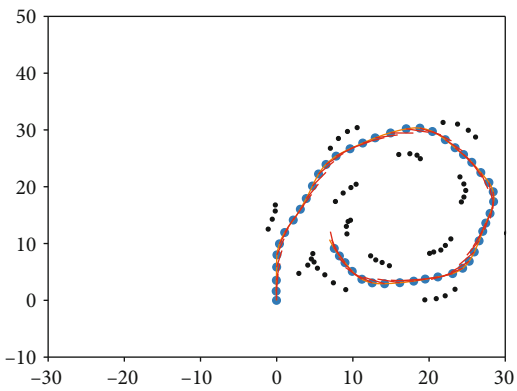
In order to verify the style transfer effect of the unpaired input motion data, this paper takes the running motion that is different from the content motion of the three styles as the content motion. Due to the cyclic consistency constraint, the training is carried out by reducing the difference between the generated motion and the content motion. Therefore, the constrained generated motion should be closer to the posture of the content motion than the unconstrained generated motion. The following is to judge the consistency of the motion posture of the generated motion and the content motion based on the footsteps in the trajectory diagram and transfer the three styles of motion to running. The trajectory diagram of the input motion data is shown in Figure 6. It shows the running motion of ordinary people. The walking motion of old people, the walking motion of zombies, and the walking motion of orangutans are the same as in Figure 5.

The following three styles of motions are transferred to ordinary people's running motions through a circular consistent style transfer model combined with kinematic constraints. The migration results are shown in Table 5.

The unconstrained style transfer results of the old people's style motions are poor. Compared with the content motion trajectory graph, the footsteps and the ground are in intensive contact, indicating that the motion content of the generated motion and the content motion is quite different. The generated motion trajectory of the constrained style transfer is closer to the trajectory diagram of the content motion, indicating that the cyclic consistency constraint still has a better migration effect on the style motion and the content motion that have different action content; zombie, orangutan style motion, and running motion are quite different, so there is a big difference between the unconstrained generated motion trajectory graph and the content motion trajectory graph, and it is impossible to determine the starting position of the movement and the similar trajectory paragraph. Constrained generated motion trajectory graphs still maintain a good migration effect, and the trajectories of generated motion and content motion are highly similar.

According to the results of the migration of the three styles of movement to the running movement, compared with the content movement, the generated movement of

Table 4: Results of moving movement migration.

Style motion	Unconstrained	Constrained
Old people—walking migration		
Zombies—walking migration		
Orangutans—walking migration		

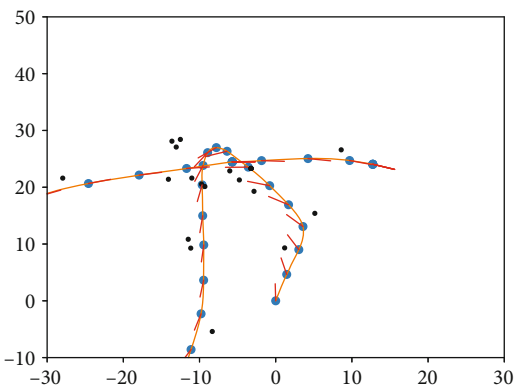


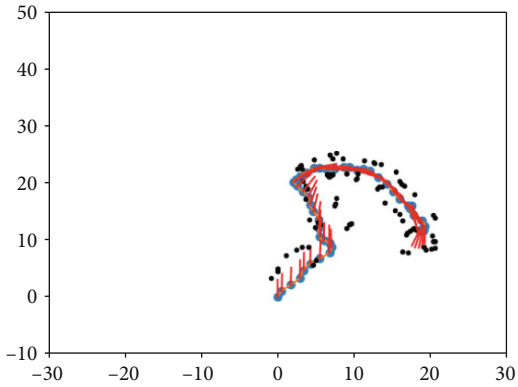
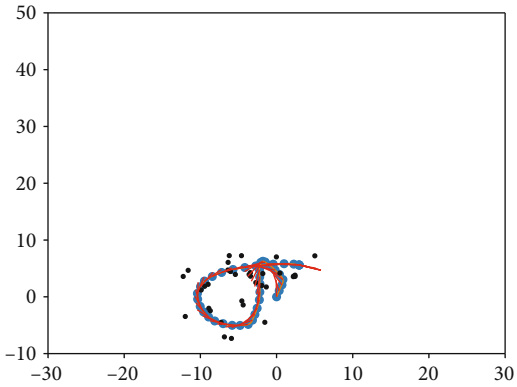
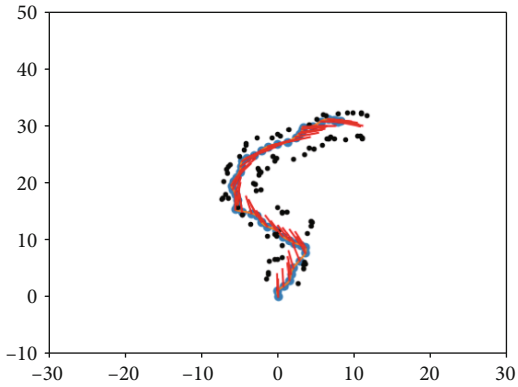
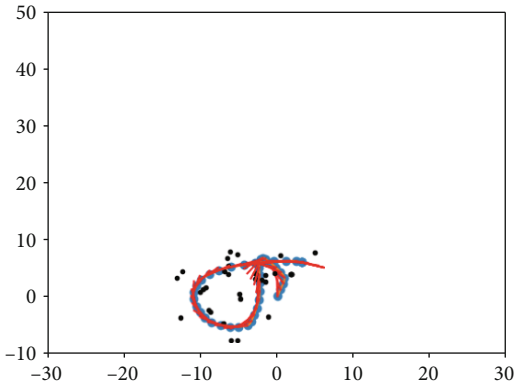
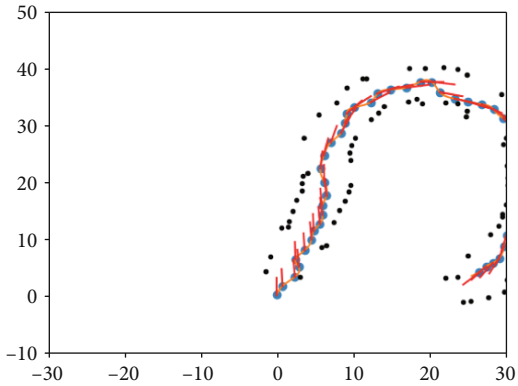
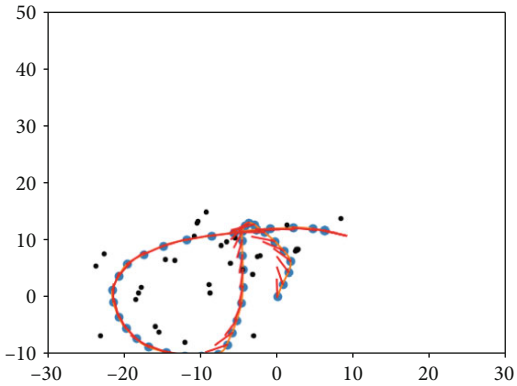
Figure 6: The ordinary human being running trajectory.

the unconstrained style transfer has lower similarity between the movement posture and the movement trajectory, and it is difficult to judge similar trajectory paragraphs. The cyclic consistent style transfer method combined with kinematic constraints has a good overall transfer effect. The generated motion retains style characteristics while ensuring a high degree of similarity of motion posture and motion trajectory.

(5) Motion style transfer training loss

Figures 7–9 show the changes in the loss values of the three styles of motion transfer to walking and running through the migration model. The solid line represents the

Table 5: Style migration results of unpaired motion data.

Trajectory of content motion	Unconstrained	Constrained
Old people—walking migration		
Zombies—walking migration		
Orangutans—walking migration		

change of the migration loss value of the walking content motion, and the dotted line represents the change of the migration loss value of the running content motion.

In Figures 7–9, loss value with the increase of the number of iterations quickly converge, and the top 50 iteration training effect is relatively obvious; since the number of iterations is 100 times, loss value change is leveling off, and three style movement of migration loss value can drop to a lower level from three style movement migration loss value variation that can be seen; as the movement style complexity increases, the initial loss value is more and more big, and in the old man and the zombie migration style, loss value is changing, due to differences in content and style motion increases, leading to loss of migration to the running value slightly higher than the migrated to loss value of the walking motion. Since the action content of orangutan style was close to that of running, there was no significant change in the loss value of moving to the two content motions.

(6) Evaluation of similarity of movement style transfer

Through equations (16) and (17), we calculated the style similarity of the generated movement from three styles of

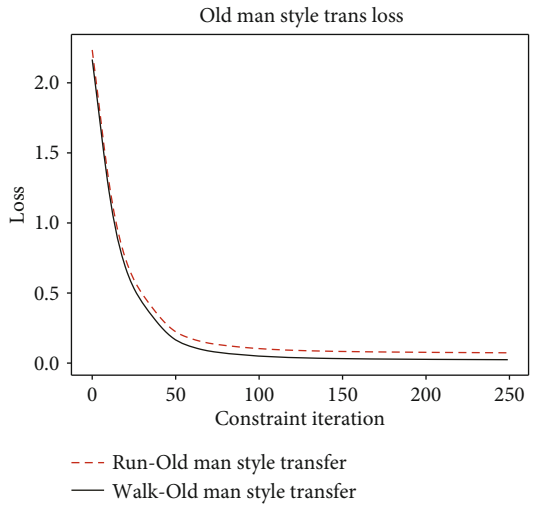


Figure 7: Changes of old man-run style translation loss.

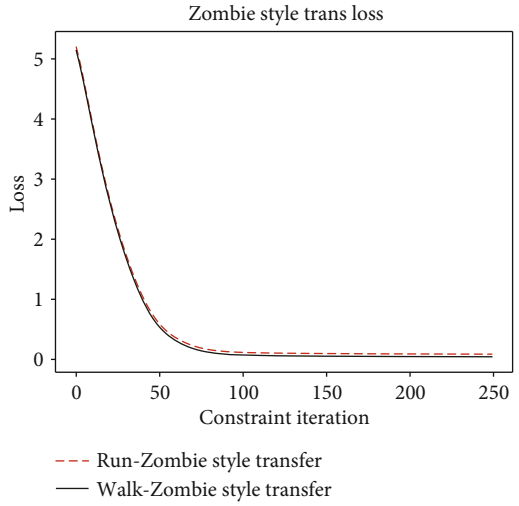


Figure 8: Changes of zombies-run style translation loss.

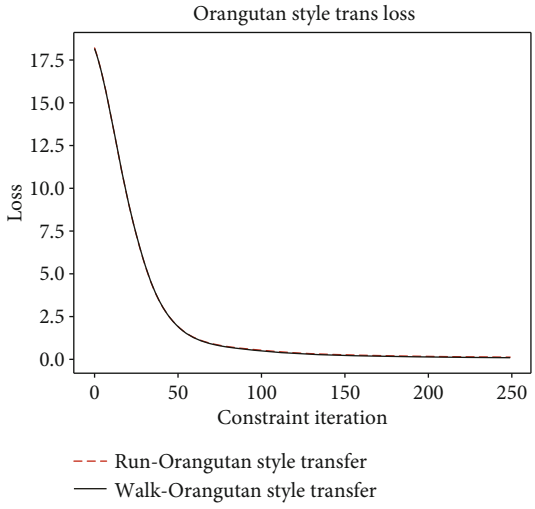


Figure 9: Changes of orangutan-run style translation loss.

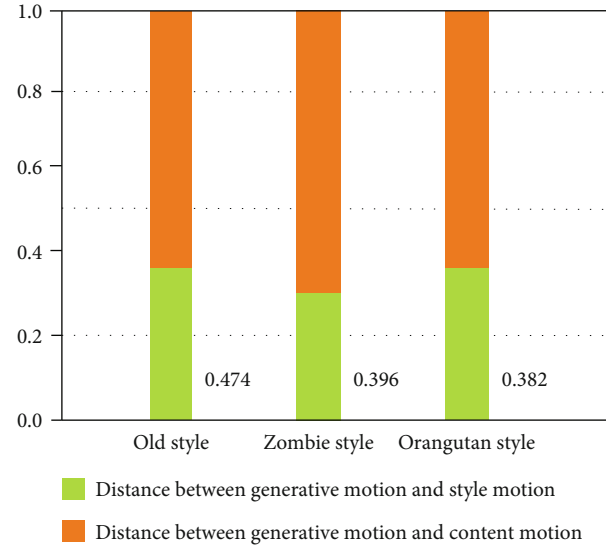


Figure 10: Style translation similarity.

motions to running.

$$D_S = \frac{D_{\text{style}}}{D_{\text{style}} + D_{\text{content}}}, \tag{16}$$

$$D_C = \frac{D_{\text{content}}}{D_{\text{style}} + D_{\text{content}}}, \tag{17}$$

where D_S is the style similarity between the generated movement and the style movement, and D_C is the style similarity between the generated movement and the content movement.

From the calculation results of style similarity in Figure 10, it can be seen that the style similarity value of the generated motion and the style motion of the circular consistent style transfer method combined with the kinematic constraint is lower than the style similarity value of the generated motion and the content motion, indicating the generated motion similar to the motion styles of the three styles of motions. In order to ensure the migration effect, the kinematic constraint and the cyclic consistency constraint are added, so that the generated motion and the content motion are more similar in the action content. In addition, the action content of running and the three styles of motion is quite different, while the action content of the walking and the three styles of motion is small, which leads to an increase in the style similarity value of the generated motion and the content motion.

In order to compare the effect of style transfer, this paper compares with Zan [23], Hu [38], Guo et al. [39], and Holden et al. [22]. Guo et al. [39] proposed a style transfer method combined with inverse kinematic constraints. First, the motion sequence is aligned through dynamic time warping, and then, the motion sequence is edited by establishing inverse kinematic constraints to realize the motion style transfer. Holden et al. stack a feedforward neural network on a single-layer convolutional autoencoder and edit the

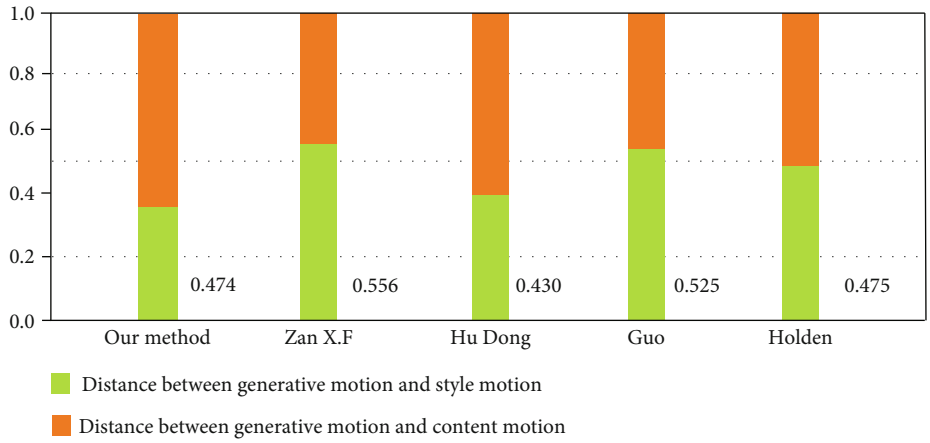


Figure 11: Style translation similarity comparison.

motion sequence through high-level parameters to generate the target motion sequence. Through the above five methods, the elderly style motions are transferred to running motions, and then, the style similarity is compared. From the calculation results of style similarity in Figure 11, it can be seen that the generated movement and content movement of Zan [23] and Guo et al. [39] are more similar, and the effect of style transfer is not ideal. The method of this paper and Hu [38] generates relatively small values of similarity between the movement and the style movement, and the effect of style transfer is better.

7. Conclusion

Aiming at the problem of generated motion postures lagging behind content motion at the same time and generated motion jitter sliders in the motion style migration method based on convolutional autoencoder, a cyclic consistent style migration method combining kinematic constraints is proposed. By constructing a cyclic consistent generation adversarial network, the motion style transfer network based on the convolutional autoencoder is used as a generator to establish a cyclic consistency constraint between the generated motion and the content motion, which improves the consistency of the generated motion and the content motion, and eliminates generated motion lagging. Kinematic constraints are introduced to standardize the generation of motion, which solves the problems of jitter and sliding in the results of motion style transfer and improves the effect of motion style transfer.

In the cyclic consistent style transfer model combined with kinematic constraints, physical factors are not considered, and physical constraints on generated motion are lacking. For example, when constraining the position of joint points, it did not consider whether the matching between footing point and motion speed is reasonable after the motion style transfer, the change of human muscles, gravity, and other factors. Reasonable physical constraints are also a way to improve the effect of generated motion, which is planned to be the content of subsequent research.

Data Availability

The data used to support the findings of this study have been deposited in the CMU, Carnegie-Mellon Mocap Database repository (http://mocap.cs.cmu.edu/).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

H.J.W and J.H.L performed the conceptualization and methodology; D.D.D and W.C.J contributed to the software and validation; W.C.J helped in the original draft preparation; D.D.D and L.Y wrote, reviewed, and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This work was funded by the National Key R&D Program of China (No. 2017YFB1402103), Natural Science Foundation of China (No. 61971347), Scientific Research Program of Shaanxi Province (2016KTZDNY01-06 and 2018HJCG-05), Shaanxi Water Conservancy Technology Project (2020slkj-17), and Project of Xi'an Science and Technology Planning Foundation (2020KJRC0093).

References

[1] R. Arai and K. Murakami, "Hierarchical human motion recognition by using motion capture system," in *2018 International Workshop on Advanced Image Technology (IWAIT)*, pp. 1–4, Chiang Mai, Thailand, 7-10 January 2018.

[2] I. Gajniyarov, I. Mikhailov, I. Starodubtsev et al., "The motion capture as behavior analyzing method of spontaneous motor activity in human infants," in *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, pp. 681–684, Novosibirsk, Russia, 2019.

[3] S. Sharma, S. Verma, M. Kumar, and L. Sharma, "Use of motion capture in 3D animation: motion capture systems,

challenges, and recent trends," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 289–294, Faridabad, India, 2019.

[4] Z. Yang, M. H. Rafiei, A. Hall et al., "A novel methodology for extracting and evaluating therapeutic movements in game-based motion capture rehabilitation systems," *Journal of Medical Systems*, vol. 42, no. 12, p. 255, 2018.

[5] L. Noitom, "EB/OL," 2020, https://www.noitom.com.cn/.

[6] K. Aberman, R. Wu, D. Lischinski, B. Chen, and D. Cohen-Or, "Learning character-agnostic motion for motion retargeting in 2D," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–14, 2019.

[7] N. F. Troje, "Decomposing biological motion: a framework for analysis and synthesis of human gait patterns," *Journal of Vision*, vol. 2, no. 5, pp. 371–387, 2002.

[8] A. Shapiro, Y. Cao, and P. Faloutsos, "Style components," in *Proceedings of Graphics Interface 2006*, pp. 33–39, Québec, Canada, 2006.

[9] J. Wang and Bodenheimer, "Synthesis and evaluation of linear motion transitions," *ACM Transactions on Graphics*, vol. 27, no. 1, pp. 1–15, 2008.

[10] T. Mukai and S. Kuriyama, *Geostatistical motion interpolation*, vol. 24, no. 3, 2005, ACM SIGGRAPH 2005 Papers, New York, NY, USA, 2005, Association for Computing Machinery.

[11] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 283–298, 2008.

[12] B. Matthew and A. Hertzmann, "Style machines," in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 183–192, USA, 2000.

[13] K. Grochow, S. L. Martin, A. Hertzmann, and Z. Popovic, *Style-based inverse kinematics*, ACM SIGGRAPH 2004 Papers, New York, NY, USA, 2004.

[14] L. Kovar, M. Gleicher, and F. Pighin, "Motion graphs," in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH'02*, pp. 473–482, New York, NY, USA, 2002.

[15] O. Arikan and D. Forsyth, "Interactive motion generation from examples," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques - SIGGRAPH '02 (2002)*, vol. 21no. 3, pp. 483–490, New York, NY, USA, 2002.

[16] J. Min and J. Chai, "Motion graphs++: a compact generative model for semantic motion analysis and synthesis," in *ACM Transactions on Graphics*, vol. 31, no. 6pp. 1–12, Association for Computing Machinery, New York, NY, USA, 2012.

[17] Y. Lee, K. Wampler, G. Bernstein, J. Popović, and Z. Popović, *Motion fields for interactive character locomotion*, vol. 29, no. 6, 2010, ACM Transactions on Graphics (TOG), New York, NY, USA, 2010, Association for Computing Machinery.

[18] G. W. Taylor, G. E. Hinton, and S. T. Roweis, "Two distributed-state models for generating high-dimensional time series," *Journal of Machine Learning Research*, vol. 12, pp. 1025–1068, 2011.

[19] R. Mittelman, B. Kuipers, S. Savarese, and H. Lee, "Structured recurrent temporal restricted Boltzmann machines," *Proceedings of Machine Learning Research*, P. X. Eric and J. Tony, Eds., , , pp. 3620–3628, PMLR, Bejing, China, 2014.

[20] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proceedings of the IEEE International Conference on Computer Vision. 2015*, pp. 4346–4354, Santiago, Chile, 2015.

[21] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118, Boston, MA, 2015.

[22] D. Holden, J. Saito, and T. Komura, "A deep learning framework for character motion synthesis and editing," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 1–11, 2016.

[23] X. F. Zan, *Research on Editing and Reuse Technology of Human Motion Capture Data*, Beijing Jiaotong University, Beijing, China, 2019, A master's degree.

[24] K. Aberman, Y. Weng, D. Lischinski, D. Cohen-Or, and B. Chen, "Unpaired motion style transfer from video to animation," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, 2020.

[25] Z. D. Yu, A. Andreas, S. Ariel, M. Moshe, and J. Eakta, "Adult 2child: motion style transfer using cycle GANs," in *{MIG} '20: Motion, Interaction and Games*, J. G. Stephen, S. Shinjiro, K. Ioannis, and B. Z. Victor, Eds., vol. 13, pp. 1–11, Virtual Event, SC, USA, 2020.

[26] S. J. Shin, S. C. You, H. Jeon et al., "Style transfer strategy for developing a generalizable deep learning application in digital pathology," *Computer Methods and Programs in Biomedicine*, vol. 198, article 105815, 2021.

[27] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, Venice, Italy, 2017.

[28] Z. Yan, Z. Du, and D. Wu, "Ball interaction model for characteristics simulation of soft tissue," *Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao/Journal of Computer-Aided Design and Computer Graphics*, vol. 26, no. 8, pp. 1346–1353, 2014.

[29] J. Lee and S. Y. Shin, "A hierarchical approach to interactive motion editing for human-like figures," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 39–48, ACM Press/Addison-Wesley Publishing Co.,, USA, 1999.

[30] S. Tak and H. S. Ko, "A physically-based motion retargeting filter," *ACM Transactions on Graphics*, vol. 24, no. 1, pp. 98–117, 2005.

[31] K. Choi and H. Ko, "Online motion retargetting," *The Journal of Visualization and Computer Animation*, vol. 11, no. 5, pp. 223–235, 2000.

[32] M. Gleicher, "Retargetting motion to new characters," in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH*, pp. 33–42, Budmerice Castle, Slovakia, 1998.

[33] Y. Zhang, L. Ye, J. Wang, and Q. Zhang, "Motion retargeting based on terminal effector constraints," in *Proceedings of 2015 IEEE Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2015*, pp. 513–517, Chongqing, China, 2016.

[34] Y. Zhou, S. J. Li, H. S. Zhu, and X. P. Liu, "An all-purpose bidirectional recurrent autoencoder for retargeting of motion data represented by joint position," *Journal of Computer-Aided Design & Computer Graphics*, vol. 32, no. 2, pp. 315–324+333, 2020.

[35] R. Villegas, J. Yang, D. Ceylan, and H. Lee, "Neural kinematic networks for unsupervised motion retargetting," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8639–8648, Salt Lake City, UT, 2018.

[36] J. Lee, J. Chai, P. S. A. Reitsma, J. K. Hodgins, and N. S. Pollard, "Interactive control of avatars animated with human motion data," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 491–500, 2002.

[37] J. Lee and K. H. Lee, "Precomputing avatar behavior from human motion data," *Graphical Models*, vol. 68, no. 2, pp. 158–174, 2006.

[38] D. Hu, *Character Motion Synthesis and Style Transfer Based on Deep Learning and Spatio-Temporal Constraint*, Huaqiao University, FuJian, China, 2019, A master's degree.

[39] X. Guo, S. Xu, W. Che, and X. Zhang, "Automatic motion generation based on path editing from motion capture data," in *Transactions on Edutainment IV*, pp. 91–104, Springer, Berlin, Heidelberg, 2010.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5563698, 18 pages
https://doi.org/10.1155/2021/5563698

Research Article

Weak-Light Image Enhancement Method Based on Adaptive Local Gamma Transform and Color Compensation

Wencheng Wang,[1] **Xiaohui Yuan**,[2] **Zhenxue Chen**,[3] **XiaoJin Wu**,[1] **and Zairui Gao**[1]

[1]*Weifang University, College of Information and Control Engineering, Weifang 261061, China*
[2]*University of North Texas, College of Engineering, Denton, TX 76207, USA*
[3]*Shandong University, College of Control Science and Engineering, Jinan 250061, China*

Correspondence should be addressed to Wencheng Wang; wwcwfu@126.com

Received 27 January 2021; Revised 6 February 2021; Accepted 10 March 2021; Published 25 June 2021

Academic Editor: Bin Gao

In weak-light environments, images suffer from low contrast and the loss of details. Traditional image enhancement models are usually failure to avoid the issue of overenhancement. In this paper, a simple and novel correction method is proposed based on an adaptive local gamma transformation and color compensation, which is inspired by the illumination reflection model. Our proposed method converts the source image into YUV color space, and the Y component is estimated with a fast guided filter. The local gamma transform function is used to improve the brightness of the image by adaptively adjusting the parameters. Finally, the dynamic range of the image is optimized by a color compensation mechanism and a linear stretching strategy. By comparing with the state-of-the-art algorithms, it is demonstrated that the proposed method adaptively reduces the influence of uneven illumination to avoid overenhancement and improve the visual effect of low-light images.

1. Introduction

The computer vision system has been widely used in a variety of fields such as industrial production, video surveillance, intelligent transportation, and remote sensing [1] and plays a more and important role in human's life. Nevertheless, during image acquisition, many uncontrollable factors will lead to various defects in the acquired images. Especially under poor and complex light conditions, such as low light, uneven light, backlight, and hazy conditions, the weak reflection of light from the object's surface causes color distortion and noise amplification in the images, which seriously affects the image quality [2]. As shown in Figure 1, the top row includes uneven-light images, in which uneven illumination can cause some areas of an image to be overexposed while others are underexposed, affecting not only human visual perception but also the accuracy of image segmentation and object recognition, sometimes resulting in the failure of a machine vision system. Therefore, it is of great importance to enhance the contrast and observability of images collected from poor lighting conditions [3–5].

Weak-light image enhancement has become a focus of research in the image processing field, and its interdisciplinary characteristics have attracted considerable attention from researchers worldwide. For example, in a facial recognition system, Oloyede et al. [6] applied a new evaluation function in conjunction with metaheuristic-based optimization algorithms to automatically select the best-enhanced face image. To enhance underwater images, Hou et al. [7] presented a novel underwater color image enhancement approach based on hue preservation by combining the HSI and HSV color models. Fu and Cao [8] combine the merits of deep learning and conventional image enhancement technology to improve the quality of underwater image. To improve the contrast of retinal fundus images, Soomro et al. [9] used independent component analysis (ICA) for image enhancement to effectively achieve quick and accurate segmentation of the eye vessels. Kallel et al. [10] proposed a new enhancement algo-

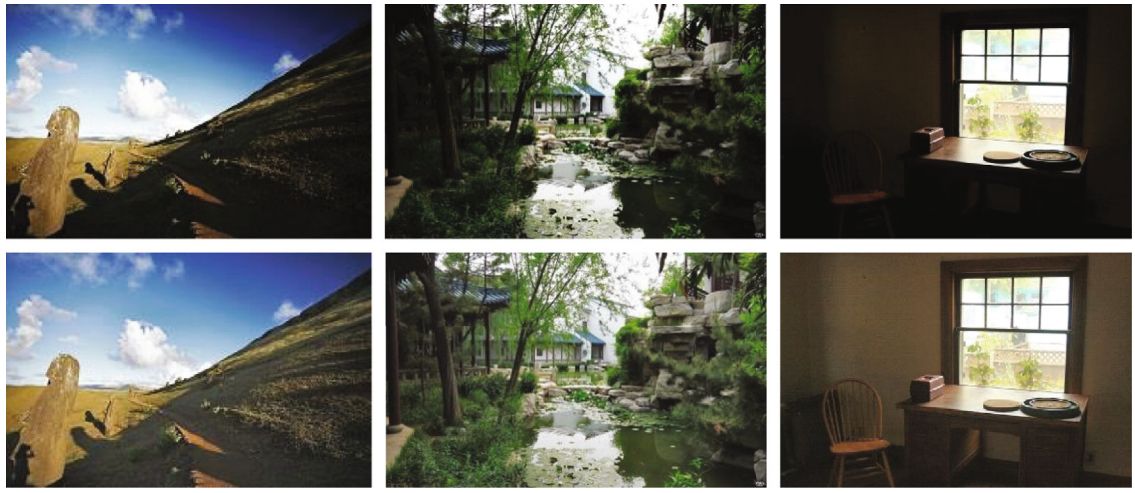


Figure 1: Examples of images acquired under uneven illumination (top row: original images; bottom row: enhanced images).

rithm dedicated to computed tomography (CT) scans based on the discrete wavelet transform with singular value decomposition (DWT–SVD) followed by adaptive gamma correction (AGC), which consistently produced good contrast enhancement with excellent brightness and edge detail conservation.

However, those conventional methods are limited in adaptivity and tend to overenhance some local areas in the case of uneven illumination. They are also difficult to strike the balance between computational complexity and visual effect. Therefore, in this research work, a self-adaptive enhancement method is proposed for processing images from uneven light environments, which is inspired by the illumination-reflection model. Figure 1 shows samples processed with our method. This method can effectively enhance the visual effect of an image, revealing more details in dark areas while preserving the overall detail information, thus providing a valuable reference for the study of the correction of images acquired under uneven lighting conditions. The contributions are as follows:

(1) It is a simple and effective image enhancer based on a novel local gamma transformation and illumination reflection model. This method can effectively enhance the visual effect of an image, revealing more details in dark areas while preserving the overall detail information

(2) The method has a color compensation mechanism; it is suitable for the processing of color images captured with monitoring system

(3) The proposed method can adjust the parameters according to the light distribution and adaptively reduce the influence of uneven illumination on the image, thus providing a valuable reference for the study of the correction of images acquired under uneven lighting conditions

(4) Our method can produce the satisfied results with less computational complexity

The rest of this paper is organized as follows. In Section 2, some related works are summarized. Section 3 introduces the flowchart of the proposed method. In Section 4, the comparisons of the experimental results are presented. Finally, the research work is concluded in Section 5.

2. Related Works

Traditional image enhancement methods of weak-light images include histogram equalization (HE) and grayscale transformation (GT) [11, 12], which usually obtains the correction parameters based on the cumulative probability distribution of gray values. For example, Huang et al. [13] proposed a gamma correction algorithm that adaptively obtains gamma correction parameters based on the cumulative probability distribution. Later, Liu et al. proposed a low-light image enhancement method based on the optimal hyperbolic tangent function [14]. In Ref. [15], a block-iterative histogram method was used to enhance the contrast of an image while processing each different part of the image with partially overlapped subblock histogram equalization (POSHE) using a moving template. Subsequently, Chen and Ramli proposed the minimum mean brightness error bihistogram equalization (MMBEBHE) [16] algorithm to minimize the error between the brightness mean values of the output image and the original image. Celik and Tjahjadi [17] proposed the contextual and variational contrast enhancement (CVC) algorithm, which performs nonlinear data mapping using context information and a 2D gray histogram to achieve the contrast improvement. These methods are simple in their computation rules and low in computational complexity but are prone to various processing issues, such as color loss and noise amplification. Huang et al. propose an effective image enhancement strategy named as contrast limited dynamic quadri-histogram

equalization (CLDQHE) which includes three steps can yield pleasing results with the preservation of brightness and structures [18].

Based on the computational theory of color constancy, Jobson et al. proposed the single-scale retinex (SSR) algorithm, which was further developed into various multiple-scale retinex (MSR) algorithms, such as the multiscale retinex with color restoration (MSRCR) algorithm [19, 20] and multiscale retinex with chromaticity preservation (MSRCP) [21]. Later, Fu et al. presented a weighted variational model to estimate illumination from a weak-light image, which can not only extract the reflection information accurately but also suppress the amplification of noises [22]. Wang et al. [23] introduced an NPE (naturalness preserved enhancement) algorithm with a bright-pass filter to preserve the image naturalness by integrating the neighborhood information of pixels, thereby improving the image contrast while avoiding excessive local enhancement. In 2011, Dong et al. [24] inverted low-light images to generate images similar to those acquired on foggy days and then used a defogging algorithm to enhance the contrast of the original images. Later, Zhang et al. proposed the framework of real-time enhancer by combining the dehazing methods with bilateral filter techniques, in which the DCP (dark channel prior) model is used for parameter optimization and a joint bilateral filter is used to reduce the noise interferences [25]. Park et al. introduced the bright channel prior (BCP) and combined the BCP estimation with retinex theory to realize the weak-light image enhancement [26] and achieved good results. Such methods effectively enhance the details in the dark areas of an image, but they also incur high computational complexity and tend to produce halo effects in dark areas.

In the past decade, machine-learning-based techniques have been widely adopted to improve the contrast of weak-light images [27]. For example, Lore et al. [28] adopted SSDA (stacked sparse denoising autoencoder) method to develop an image enhancer based on the simulation of a low-light environment, in which a machine-learning algorithm was used for training a self-encoder to adjust the brightness adaptively for several low-illumination image signals. Shen et al. [29] analyzed the performance of the MSR algorithm from the perspective of CNNs and designed an MSR network with a CNN architecture for enhancing low-light images. Tao et al. proposed an LLCNN (low-light convolutional neural network) model for image enchantment based on a deep learning technique, in which the enhanced images can finally be generated from multilevel feature graphs after learning on the low-light image database [30]. Park et al. introduced the retinex theory into the deep learning framework and proposed a double self-encoding network [31], in which a convolutional autoencoder and a stacked autoencoder are used to achieve brightness enhancement and noise suppression. Inspired by image-fusion-based methods [32] developed a single-image enhancer by combing the image-fusion-based technique to train an end-to-end CNN model, which is based on building a multiexposure image dataset with different contrast-scale images. Methods of this kind offer a good image enhancement effect, but their computational models often require an excessively long time or too many expensive resources for training.

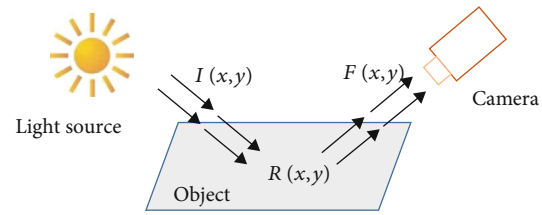


Figure 2: The illumination-reflection model.

3. Framework of Proposed Method

According to the basic principle of imaging, an image is produced by the light that is reflected or emitted from the surface of an object in a scene and reaches the camera. Generally, it often is regarded as a two-dimensional function $F(x, y)$, where the value of this function is the brightness of the pixel at coordinates (x, y) in the image, and $F(x, y)$ is the composition of the illumination component $(I(x, y))$ that enters the scene and the reflection component $(R(x, y))$from the object surface. The mathematical expression of this illumination-reflection model is as follows:

$$F(x, y) = I(x, y)R(x, y). \quad (1)$$

The spatial relations of this model are illustrated in Figure 2.

It is shown that the intensity of incident light mainly relies on the light source, and its distribution function $(I(x, y))$ shows little spatial variation. The spectrum of $I(x, y)$ mainly concentrated in the low-frequency region to reflect the lighting environment during the imaging process, while that of the reflection component $R(x, y)$ is mainly concentrated over a wide range in the high-frequency band, corresponding to the image details that reflects the natural attributes of the target. If the illumination in a scene is even, then the illumination component is uniformly distributed in the space, and the acquired image is considered to have natural lighting and high visual quality; however, if the illumination in an imaged scene is uneven, then areas with excessively strong illumination will be overexposed, while those with insufficient illumination will be underexposed, causing various visual questions for the human eyes. If we can find a way to estimate the reflection component, i.e., separating $I(x, y)$ from $F(x, y)$, then, we can eliminate the effects of light on imaging, thus helping to achieve the goal of image enhancement [33].

Inspired by the above model and theory, we propose the framework of image enhancer based on adaptive local gamma transform and color compensation in this paper. The proposed method eliminates the associations among color components by modifying color space; thus, the goal of image enhancement is achieved by processing the color components in a different space. First of all, the source color image is transformed to the YUV space, where the brightness

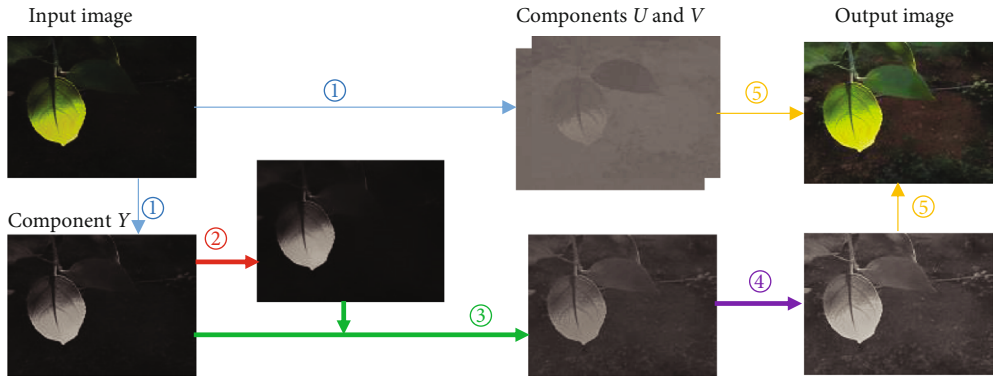
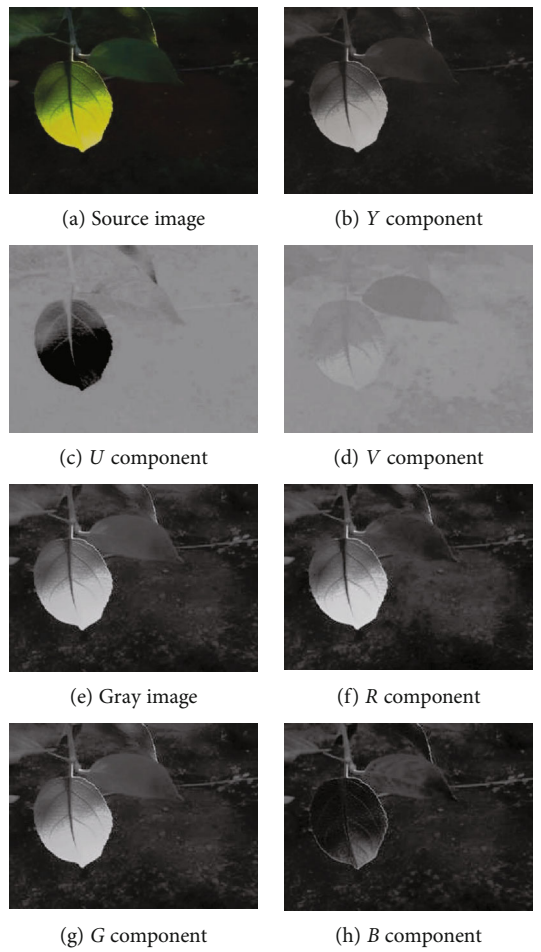


Figure 3: The framework of the proposed method: ① RGB space to YUV space; ② illumination component estimation; ③ local gamma transformation; ④ grayscale stretching; ⑤ saturation enhancement.



(a) Source image (b) *Y* component

(c) *U* component (d) *V* component

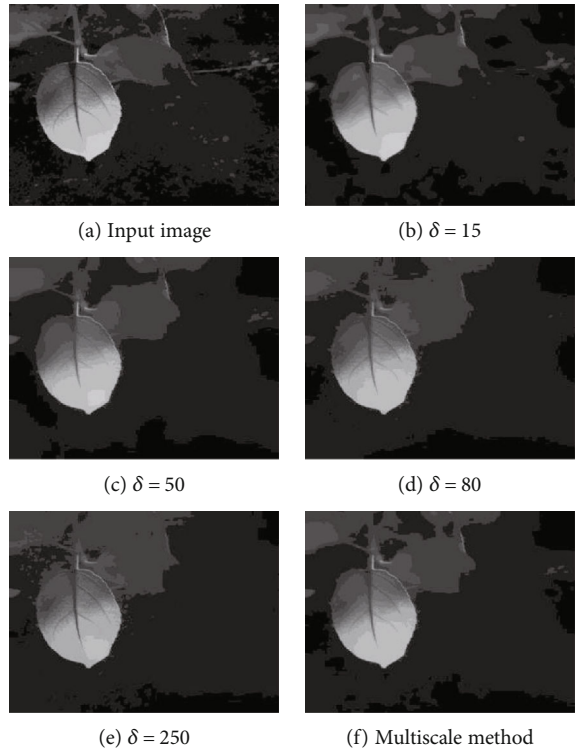
(e) Gray image (f) *R* component

(g) *G* component (h) *B* component

Figure 4: Example of RGB space to YUV space.

part of the scene is estimated from the *Y* component using a fast guided filtering function, and then, local gamma transform enhancement is performed on the image through adaptive adjustment according to the gray distribution of the brightness component. Finally, the contrast of the image is adjusted via grayscale linear stretching, and a color compensation strategy is applied to the RGB image. The whole flowchart is shown in Figure 3.

3.1. Color Space Conversion. As known from the neural mechanism of the visual perception system, the human eyes are more sensitive to luminance than to color; thus, the enhancement of luminance is the key to the proposed algorithm for the correction of unevenness in illumination. For color images, the chrominance information and brightness information cannot be effectively distinguished in the RGB (red, green, blue) color space; consequently, applying a direct



(a) Input image (b) $\delta = 15$

(c) $\delta = 50$ (d) $\delta = 80$

(e) $\delta = 250$ (f) Multiscale method

Figure 5: Single-scale and multiscale illumination component extraction.

correction to the three channels in the RGB color space not only leads to color distortion but also increases the computation load. By contrast, in the YUV color space, each color corresponds to two chrominance components (U and V) and one brightness component (Y); hence, the separation of brightness and chrominance makes it possible to alter illumination intensity without affecting the color. Therefore, in this study, we propose a YUV-space grayscale mapping based chrominance-luminance recombination algorithm in which the luminance component Y is processed while leaving the chrominance components U and V unchanged for enhancement. The relation of RGB color space and YUV color space is [33]:

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.147 & -0.289 & 0.436 \\ 0.615 & -0.515 & -0.100 \end{bmatrix} \times \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \quad (2)$$

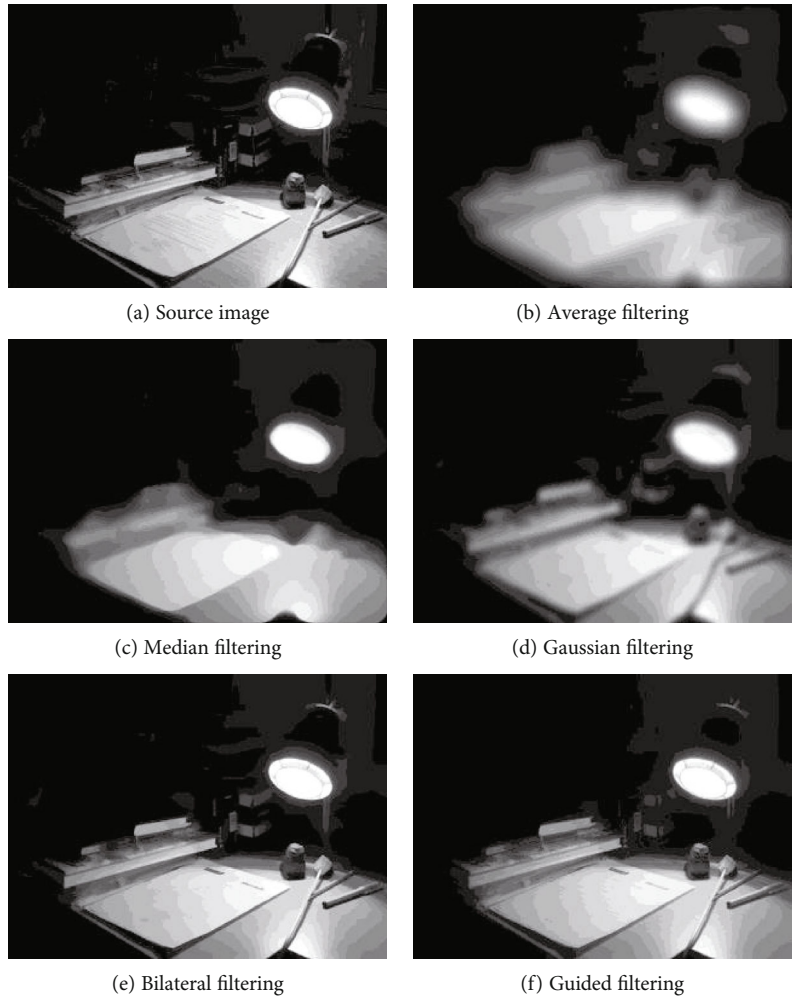
After the conversion to YUV space, the images corresponding to each component are as shown in Figure 4. In Figure 4, (a) is the source color image, (e) is the corresponding gray image of (a), (b–d) are the three components of Y, U, and V, respectively, and (f–h) are the components of R, G, and B, respectively.

3.2. Estimation of the Illumination Component. To effectively reduce the effect of uneven illumination on image quality, the accurate extraction of the lighting information from a scene is particularly important. Currently, the main methods for extracting the illumination component include average filter, bilateral filter, and Gaussian filter. The average filtering method smooths images by calculating the mean value of each pixel with its neighbors. It is fast but can be strongly influenced by neighboring pixels. The Gaussian filtering method is poor at retaining edges, causing the extracted illumination component to have fuzzy edges and thus to perform poorly in the retention of detailed information. The bilateral filtering algorithm shows better edge preservation characteristics but has a very high computational complexity, which limits its use in practical engineering applications. The guided filtering algorithm is a guided image-based local linear transformation that obtains the low-frequency information from the image while retaining the edge information and has low computational complexity. It is the fastest available edge-retaining filtering algorithm and was therefore used in this study to extract the illumination component [34, 35].

Let the images for inputting, outputting, and guiding are denoted by p, q, and I, respectively. Then, for any given pixel s, its guided filtering process is:

$$q_j = a_s I_j + b_s, \forall_j \in \omega_s, \quad (3)$$

where j is the pixel index and a_s and b_s are the linear transformation factors. The minimum reconstruction difference between p and q is calculated as follows:



(a) Source image (b) Average filtering

(c) Median filtering (d) Gaussian filtering

(e) Bilateral filtering (f) Guided filtering

Figure 6: Extraction of the illumination component with different methods.

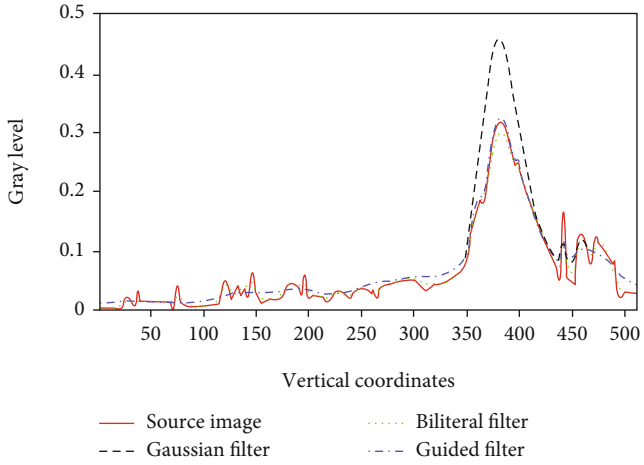


Figure 7: A one-dimensional plot of the illumination components extracted with different algorithms.

$$a_s = \frac{1/|\omega| \sum I_j p_j - \mu_s \bar{p}_s}{\sigma_s^2 + \xi}, \tag{4}$$

$$b_s = \bar{p}_s - a_s \mu_s, \tag{5}$$

where σ_s^2 and μ_s are the variance and the mean value of the guided image I within the window ω_s, respectively; ξ is a parameter that controls the degree of smoothness of the filter; $|\omega|$ is the pixel number of ω_s; and $\bar{p}_s$ is the mean value of the input image p. Thus, the output of the filter will be:

$$q_j = \frac{1}{|\omega|} \sum_{s:j \in \omega_s} (a_s I_j + b_s), \tag{6}$$

$$q_j = \bar{a}_j I_j + \bar{b}_j, \tag{7}$$

where $\bar{b}_j$ and $\bar{a}_j$ are the mean values of b and a, respectively, within the neighborhood window $\omega_s$centered on pixel j.

Therefore, the guided filtering process can be seen as the convolution of the guided filtering function and the original

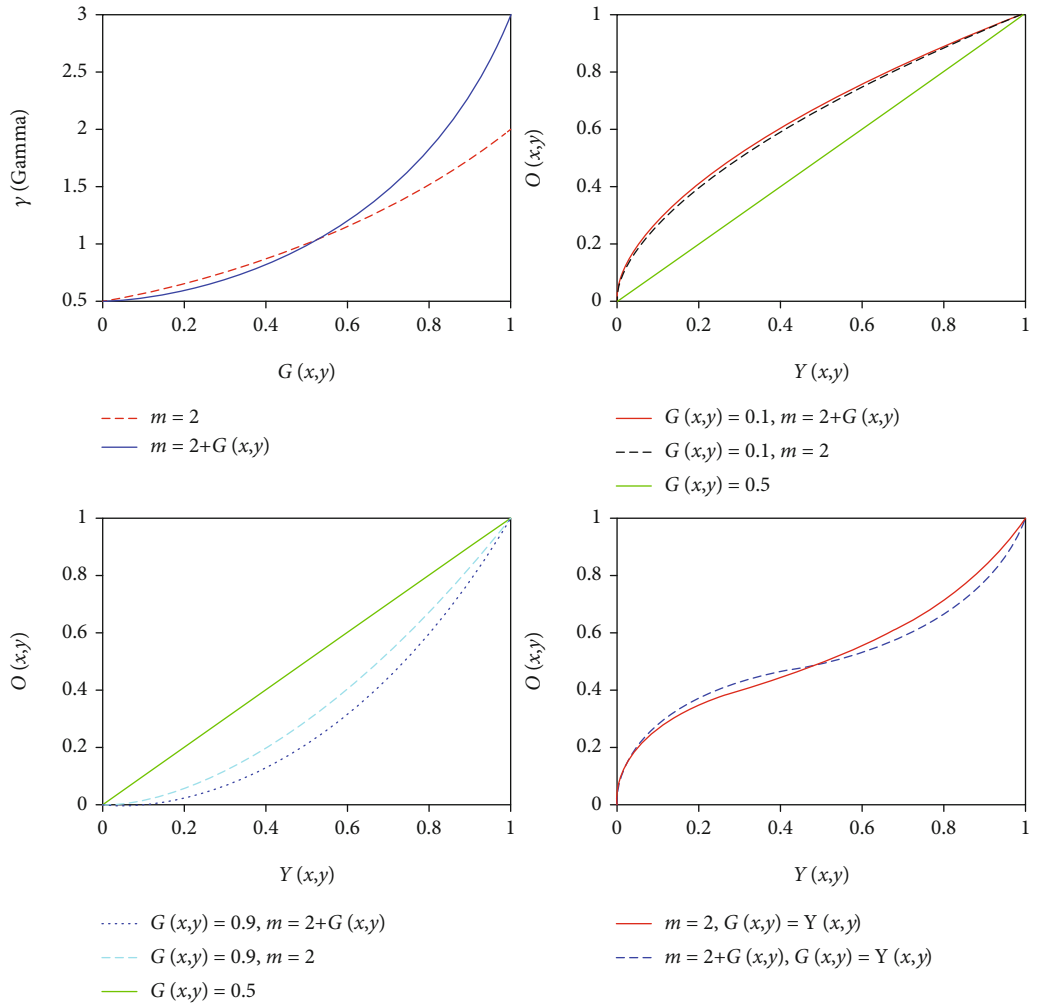


Figure 8: Curve changes with illumination.

image, which gives rise to the following estimate of the illumination component:

$$\widehat{I}(x,y) = Y(x,y) * G^F(x,y), \tag{8}$$

where $G^F(x,y)$ is the guided filter, $Y(x,y)$ is the input image, and $\widehat{I}(x,y)$ is the output image which denotes the estimation of the luminance component.

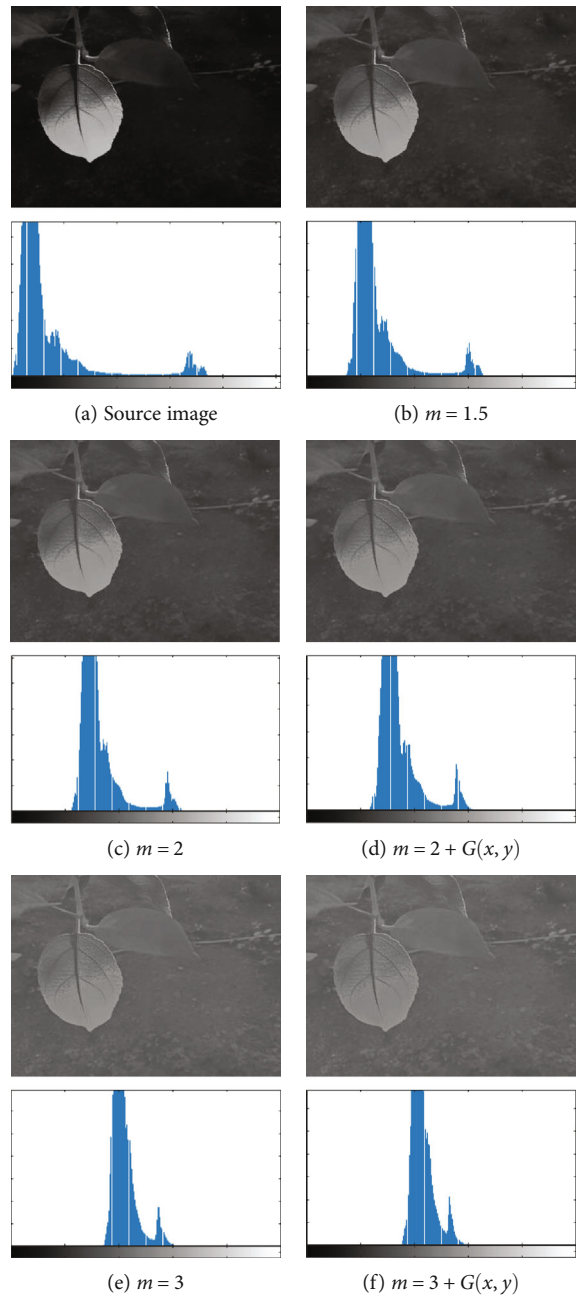
To consider both the local and global characteristics of the estimated luminance values, we introduced the multiscale guided filtering, which will extract the illumination components of the scene using filtering windows of different scales and weights, which ultimately gives rise to the following estimate of the illumination component:

$$G(x,y) = \widehat{I}(x,y) = \sum_{t=1}^{N} \lambda_t \left[Y(x,y) * G_t^F(x,y) \right], \tag{9}$$

where λ_i is the weights for the illumination component extracted at the tth scale and N is the number of scales used; $\widehat{I}(x,y)$ is the value of the weighted combination of the illumination components at point (x,y) that is extracted using the guided filtering function with windows of various scales. In Figure 5, the values of the illumination components extracted using three different scales are shown. In Figure 5(f), the result of the fusion of these three different scales (15, 80, and 250) is shown, where the weight of the illumination component extracted at each scale was set to 1/3.

From Figure 5, it can be seen that the method based on multiscale guided filtering can well extract the illumination component from the source image, which describes the variation of illumination while get rid of the details, to meet the requirements of practical application. However, this method requires multiple filtering operations to be performed on the image. Based on the comprehensive consideration of both computational complexity and performance, we propose a calculation method with an adaptive window, in which the window size c is 1/4 of the smaller dimension of the image, as follows:

$$c = \mathrm{Int}\left[\frac{\min(w,h)}{4}\right], \tag{10}$$



(a) Source image (b) $m = 1.5$

(c) $m = 2$ (d) $m = 2 + G(x, y)$

(e) $m = 3$ (f) $m = 3 + G(x, y)$

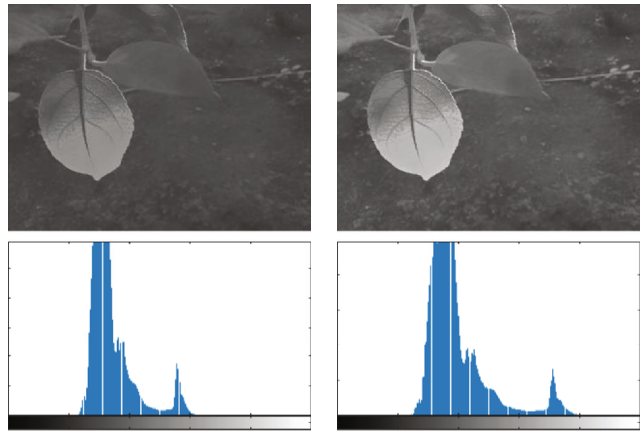
Figure 9: Enhancement results with different m.

where Int[] is a function to extract an integer that down to the nearest number and h and w are the height and width of the image, respectively. To demonstrate the advantages of guided filtering, the effects of various filtering methods (including Gaussian filter, average filter, median filter, and bilateral filter) are compared in Figure 6.

It shows that guided filter, bilateral filter, and Gaussian filter all yield good descriptions of the illumination variations in the scene, consistent with the distribution of the illumination component. To further compare the edge-retaining characteristics of the guided filter, bilateral filter, and Gaussian filter, we consider the pixels on Line 110 in Figures 6(a) and 6(d)–6(f) as examples. In Figure 7, we present a one-dimensional brightness diagram generated from the gray-scale values acquired at the corresponding positions in these images.

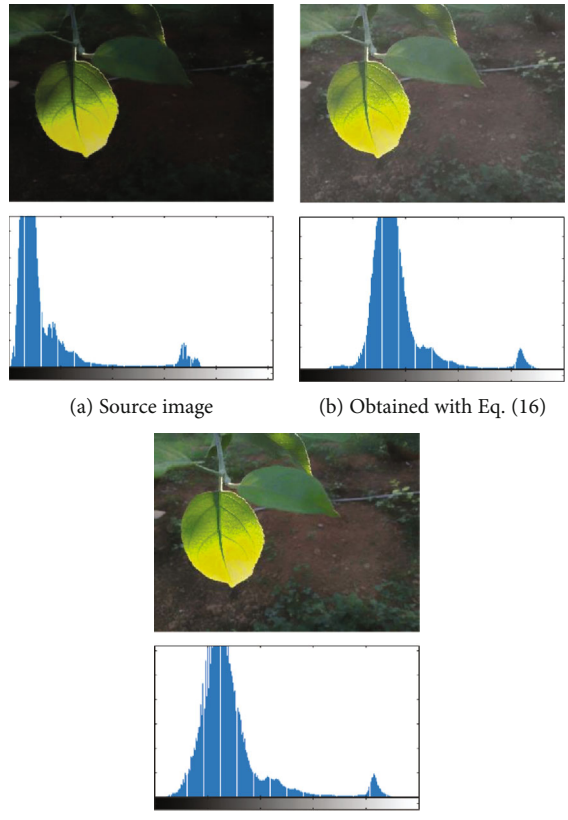
Figure 7 shows that Gaussian filtering results in larger deviations in sharp edge regions compared with the edges in the original image, while the fast guided filtering algorithm performs the best approximating the brightness distribution of the original image, especially in the edge areas, while maintaining low computational complexity and a high speed.

3.3. Local Gamma Transform. To adaptively increase the brightness of low-illumination areas while decreasing the brightness of high-illumination areas based on the gray



(a) Before stretching (b) After stretching

Figure 10: Images before and after gray stretching.



(a) Source image (b) Obtained with Eq. (16)

(c) Obtained with Eq. (17)

Figure 11: Color saturation enhancement.

distribution, we attempt to improve and expand the conventional method of gamma correction, which has the following standard form:

$$O(x, y) = [Y(x, y)]^{\gamma}, \tag{11}$$

where $O(x, y)$ is the corrected brightness with a range of [0 1], $Y(x, y)$ is the source image to be enhanced, and γ is a control parameter. When γ is less than 1 but greater than 0, the overall brightness increases, and when γ is greater than 1, the overall brightness decreases.

For uniformly overexposed or underexposed images, this algorithm can produce satisfactory results through the adjustment of the parameter γ, but when both overexposed and underexposed areas are present in the same image, it is difficult for the algorithm to achieve satisfactory effectiveness

(1) Input weak-light image F
(2) Transform source image into YUV space using Eq. (2) and separate the luminance and the chrominance components V, Y, and U
(3) Extract the illumination component from Y using Eq. (8) to obtain image G
(4) Obtain the enhanced brightness image O using Eq. (13)
(5) Perform linear stretching on image O using Eqs. (14) and (15) to generate image Y'
(6) Make color compensation to enhance the saturation of image using Eq. (17)
(7) Output the enhanced image J

Algorithm 1

when the same parameter is used across the entire image. Therefore, we introduce an algorithm that allows γ to vary with the local information of the image, as follows:

$$\gamma = m^{(2\times G(x,y)-1)}, \tag{12}$$

where $G(x, y)$ is the illumination extracted from $Y(x, y)$ and m is the base of the exponential function. In general, areas with low illumination need more aggressive correction, so a small γ value should be adopted, i.e., a greater value of m should be adopted in Eq. (12); for images with excessively high contrast, a γ value greater than 1 should be adopted, i.e., the value of m should be low, to suppress the illumination intensity. In Ref. [36], a piecewise function is formulated based on whether the mean value of the input image is greater than 0.5. However, for images with both overexposed and underexposed areas, the mean value can be very close to 0.5, and if such images are processed using this algorithm, it is possible that no notable change may result, meaning that the actual needs of image correction will not be met. Therefore, we present a formulation of the gamma correction parameter γ that varies with the illumination component of the scene and propose an adaptive brightness adjustment function based on this local gamma transformation, which adaptively adjusts the control parameters according to the illumination distribution of the input image, as follows:

$$O(x, y) = [Y(x, y)]^{(2+G(x,y))^{[2\times G(x,y)-1]}}. \tag{13}$$

According to Eq. (12), when the base values are set to 2 and $2 + G(x, y)$, the changes in the output γ with the input $G(x, y)$ are as shown in Figure 8.

According to Eq. (13), when the base is set to 2, $2 + G(x, y)$, or $3 + G(x, y)$, the changes of correction effect are as shown in Figure 9. Figure 9 shows that as m increases, low pixel values are enhanced, and high pixel values are suppressed. This compresses the image's dynamic range and leads to an overall enhancement in the image brightness, at the cost of reduced contrast.

3.4. Grayscale Linear Stretching. To mitigate the problem of image gray value concentration, we use a grayscale stretching function to improve the image. A simple linear pointwise operation is performed to expand the histogram of the image to include the entire grayscale range. The rationale for this action is to improve the dynamic grayscale range for image processing.

Let $O(x, y)$ denote the input image, whose minimum grayscale value $L_{\min}$ and maximum grayscale value $L_{\max}$ are defined as follows:

$$L_{\min} = \min\,[O(x, y)], \quad L_{\max} = \max\,[O(x, y)]. \tag{14}$$

By linearly mapping the dynamic range from $[A, B]$ to $[A, 1]$, then the output image $Y'(x, y)$ will be:

$$Y'(x, y) = \frac{(1 - L_{\min})}{L_{\max} - L_{\min}} O(x, y) + \frac{(L_{\max} - 1)L_{\min}}{L_{\max} - L_{\min}}. \tag{15}$$

As shown in Figure 10, processing with the proposed algorithm expands the dynamic range of the image, facilitating the identification of details in overexposed and underexposed areas of the image.

3.5. Color Compensation. Using the following formulas, we convert the image back from the YUV color space into the RGB color space using the enhanced component Y' while leaving the U and V components unchanged:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} 1.000 & 0.000 & 1.140 \\ 1.000 & -0.395 & -0.581 \\ 1.000 & 2.032 & 0.001 \end{bmatrix} \times \begin{bmatrix} Y' \\ U \\ V \end{bmatrix}, \tag{16}$$

where Y', U, and V are the brightness component and the two chrominance signal components, respectively, in YUV space.

However, after the conversion to RGB space using the above method, the image may show a decrease in color saturation. To ensure that the color saturation of the output image is consistent with that of the input image, we adopt the following expressions:

$$\begin{cases} R' = \varepsilon \times \left[\left(\dfrac{Y'}{Y}\right) \times (R + Y) + R - Y\right], \\ G' = \varepsilon \times \left[\left(\dfrac{Y'}{Y}\right) \times (G + Y) + G - Y\right], \\ B' = \varepsilon \times \left[\left(\dfrac{Y'}{Y}\right) \times (B + Y) + B - Y\right], \end{cases} \tag{17}$$

(a) Night (b) Bridge (c) Castle (d) Town

(e) Girl (f) Street (g) Pine (h) Dawn

Figure 12: Samples of image enhancement with our method.

where ε is set to an empirical value of 0.5; R, G, and B denote the red, green, and blue components in the RGB space, respectively; and Y and Y' are the brightness components in YUV space before and after enhancement, respectively.

The images obtained using Eqs. (16) and (17) and their corresponding grayscale histograms are shown in Figure 11. The image obtained using Eq. (17) has better color saturation and higher contrast than that obtained using Eq. (16).

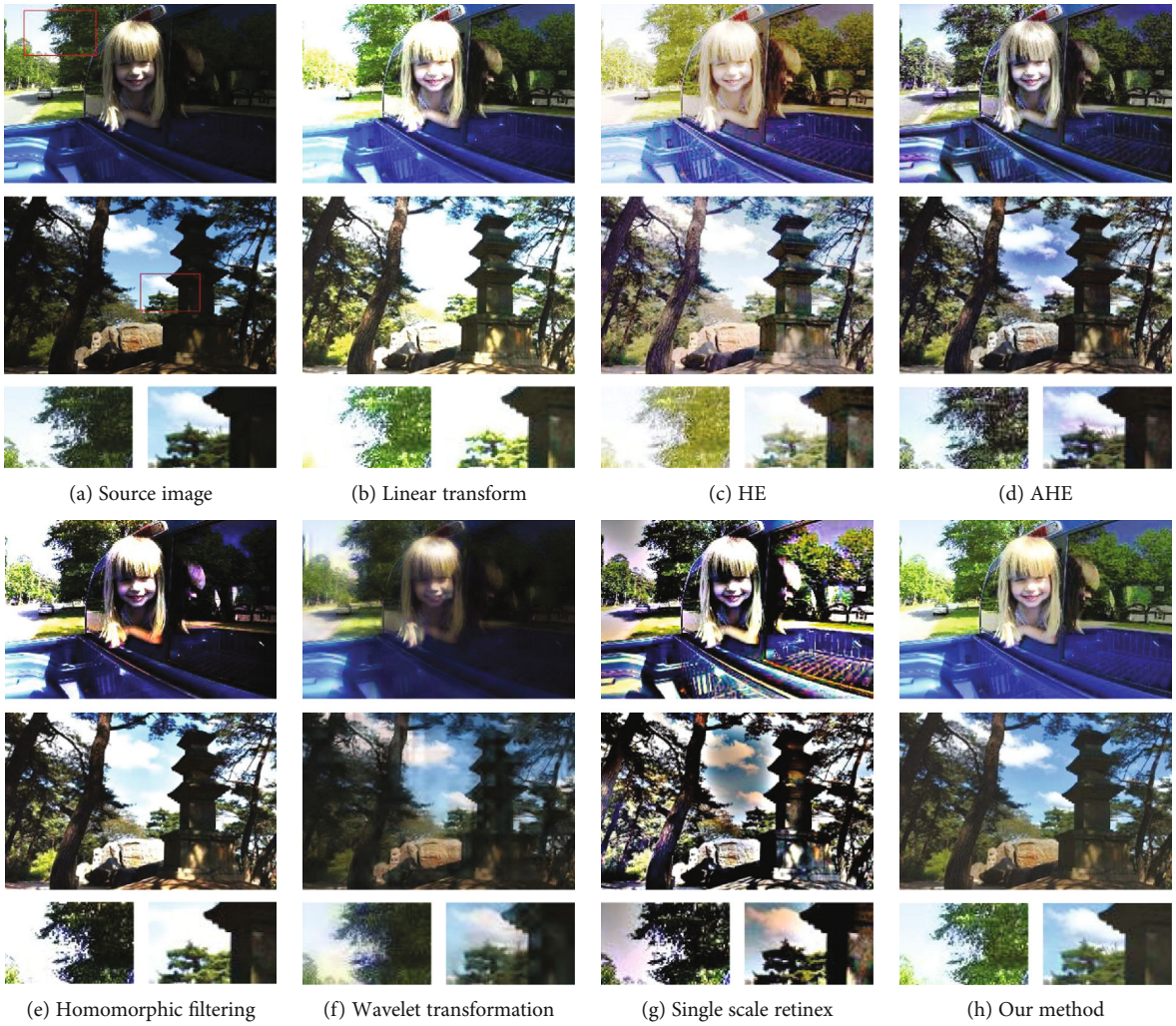
The specific steps of implementation of the above-described adaptive enhancement method for low-light images acquired under uneven illumination are summarized as follows:

4. Experiments and Analysis

To test the performance of our method, we used an experimental platform consisting of a computer (with an Intel(R) Core (TM) i7-6700 and 16 GB of RAM) and the simulation software MATLAB. The images used for testing included an urban streetscape, some natural scenery, and an indoor scene and have the common features of a large dynamic range and uneven illumination. Some of the experimental results are shown in Figure 12 for the images "Night," "Bridge," "Castle," "Town," "Girl" [37], "Street," "Pine," and "Dawn." As shown in Figure 12, after processing with the proposed algorithm, the areas with low illumination are enhanced, and those with high illumination are suppressed. The enhanced images are natural in color and clear in detail, indicating that the proposed method can adaptively mitigate the impact of uneven scene illumination on image quality. Next, we will compare the processing results of the proposed algorithm with those of various mainstream algorithms in terms of both a subjective visual assessment and an objective quantitative analysis.

4.1. Subjective Evaluation

4.1.1. Comparison with Traditional Enhancement Methods. In Figure 13, the results of the proposed method and other conventional image enhancement methods are shown. Figure 13(a) shows the original images [37], and Figures 13(b)–13(h) show the experimental results of a linear transformation (LT), histogram equalization (HE), adaptive histogram equalization (AHE), homomorphic filtering (HF), the wavelet transform (WT), the Retinex method, and the proposed method, respectively. The corresponding amplification effects in the areas demarcated by boxes in Figure 13(a) are shown in rows 3 and 6. The results indicate



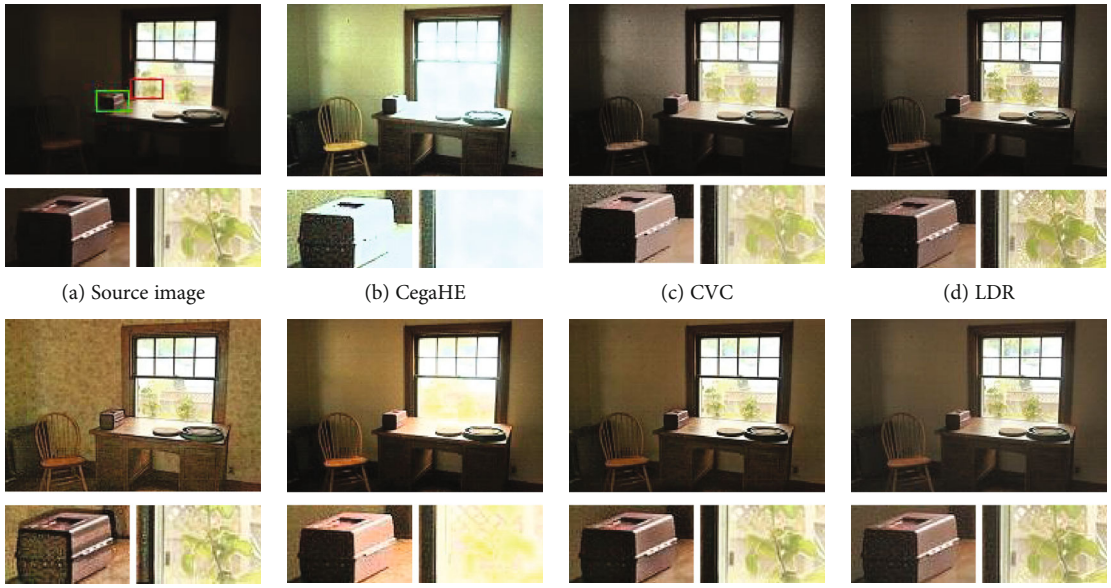
(a) Source image (b) Linear transform (c) HE (d) AHE

(e) Homomorphic filtering (f) Wavelet transformation (g) Single scale retinex (h) Our method

Figure 13: Comparison of the proposed algorithm with several conventional algorithms.

that the images processed using the various methods show changes of varying degrees relative to the original image. For example, Figures 13(c) and 13(g) are significantly enhanced in terms of contrast, showing greater detail but a shift in hue. In addition, the severe "halo" noise in Figure 13(g) results in poor visual quality. Figures 13(e) and 13(f) show no overall hue shift but exhibit inadequate improvement in details and are fuzzy. Figures 13(b) and 13(d) show good overall effects, but excessive enhancement is evident in bright regions due to the linear transformation method, whereas AHE causes the color to be significantly darkened. In contrast, the proposed method yields remarkable improvements in both color and contrast, achieving a better visual effect than the other methods.

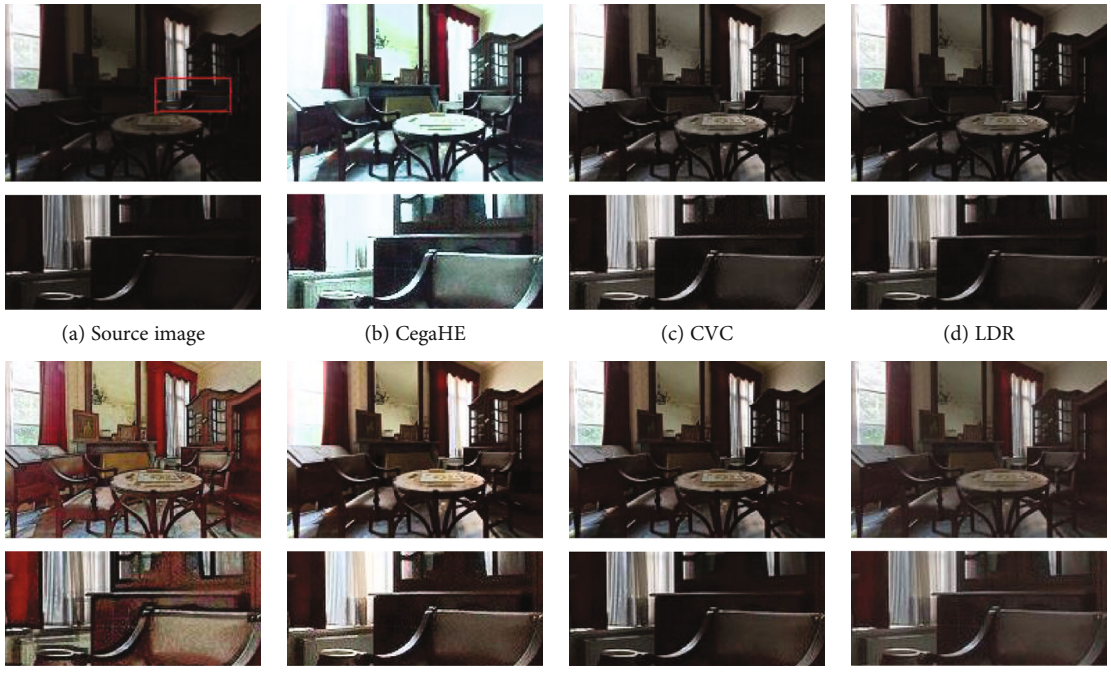
4.1.2. Comparison with State-of-the-Art Methods. We further compared the enhancement effect of the proposed method with those of some state-of-the-art methods using "Window" and "Furniture" as test images. The results are shown in Figures 14 and 15. In each of these figures, (a) shows the original image and enlarged views of the areas demarcated by the boxes, and (b–h) show the results obtained using CegaHE [38], CVC [16], the linear dynamic range (LDR) technique [39], DCP [24], MSRCP [21], SRIE [22] and the proposed algorithm, respectively, along with the corresponding enlarged areas. The results show that compared with the original image, the overall visibility and contrast of the enhanced images obtained using the various enhancement methods are greatly improved, achieving good enhancement effectiveness. However, the CegaHE method results in a severe hue shift. The CVC and LDR methods achieve only a slight enhancement while amplifying the noise in the dark regions, while the CVC method is additionally unable to restore color to low-light pixels. The MSRCP and DCP methods enhance the overall image brightness, but the MSRCP method results in overenhancement, while the DCP method shows a significant overenhancement effect in edge regions. Relative to the other methods, the SRIE method and the proposed method both strike a balance between color information and brightness information, thereby achieving good enhancement effects. However, the SRIE method is unable to achieve uniform results for images with alternating bright and dark regions, resulting in inferior overall performance compared to the proposed method. With regard to



(a) Source image (b) CegaHE (c) CVC (d) LDR

(e) DCP (f) MSRCP (g) SRIE (h) Proposed method

Figure 14: Experimental results on the "Window" image.



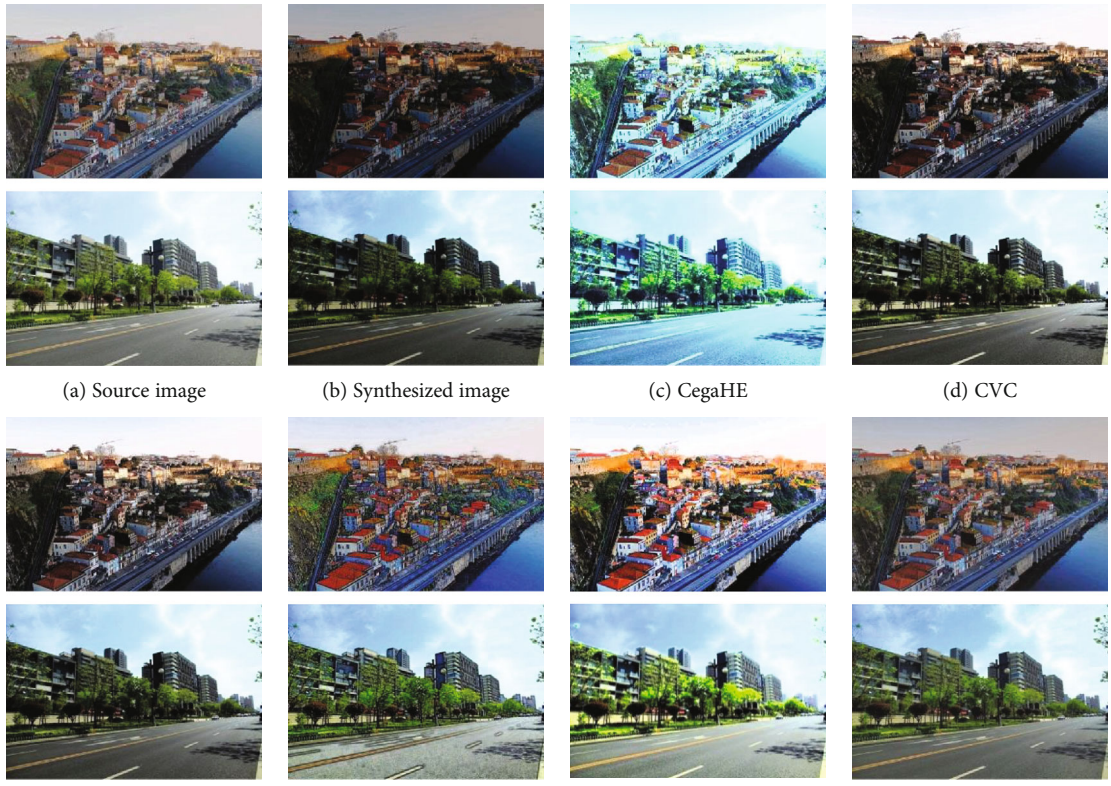
(a) Source image (b) CegaHE (c) CVC (d) LDR

(e) DCP (f) MSRCP (g) SRIE (h) Proposed method

Figure 15: Experimental results on the "Furniture" image.

local details, in the areas of the images demarcated by boxes, the DCP method results in overenhancement and consequent noise at the edges. The CVC and LDR methods lead to underenhancement, the CegaHE and MSRCP methods lead to local overenhancement, and the SRIE method generates shadows in some local areas. By contrast, the proposed method shows no excessive amplification of the noise in dark areas in the enhanced image while significantly enhancing the areas that need highlighting without overenhancement, thereby achieving superior sharpness, contrast, and image color.

To further compare the processing effects of the different algorithms, we also tested the algorithms on artificially synthesized images, as shown in Figure 16. In this figure, (a) shows two images acquired under proper lighting, and (b) shows corresponding low-light images that have been synthesized through gamma transformation (with a γ value of 2). (c–h) show the image enhancement results obtained using



(a) Source image (b) Synthesized image (c) CegaHE (d) CVC

(e) LDR (f) DCP (g) MSRCP (h) Proposed method

Figure 16: Experimental results on synthesized images.

Table 1: The assessment results on the images in Figure 16.

	CegaHE	CVC	LDR	DCP	MSRCP	Our method
MSE	1087.535	2462.565	864.13	718.64	1088.675	*329.58*
PSNR	17.8	14.585	18.78	19.635	17.77	*23.13*
SSIM	0.85	0.825	0.895	0.835	0.87	*0.97*

the different methods. The results indicate that the proposed method can adaptively enhance the brightness of low-light areas while suppressing that of high-illuminance areas, and the enhancement effects are consistent with those observed on the actual images presented above.

4.2. Objective Evaluation. Because different methods focus on different aspects of an image, a subjective evaluation is likely to be biased [40]. Therefore, we adopt several objective evaluation criteria to further examine the processing effects of different methods. We adopt the mean squared error (MSE), the peak signal-to-noise ratio (PSNR), and the structural similarity index measure (SSIM) as objective evaluation metrics for comparison and evaluation [41]. The objective evaluation data corresponding to Figure 16 are shown in Table 1, where the best results are italicized.

To conduct a more general test, we subjected a number of synthesized images to processing with various methods, including CegaHE [38], CVC [16], LDR [39], DCP [24], EFF [42], MSRCP [21], SRIE [22], and the proposed algorithm. Some of the experimental results are shown in Figure 17, where Figure 17(a) shows the original images, Figure 17(b) shows the artificial quality-reduced images, and Figure 17(c) shows the results obtained after the enhancement of the images in Figure 17(b). The objective evaluation metrics achieved by the various methods based on these images are shown in Table 2, in which the values indicating the best performance are italicized.

Tables 1 and 2 indicate that the enhanced images generated using the proposed method most closely match the original images in terms of both gray value distribution and structure. The proposed method greatly outperforms the other methods in terms of its comprehensive effect, generating the best results. These results show that the proposed algorithm can mitigate the influence of uneven illumination on images and achieve effective correction for images of diverse scenes acquired under uneven lighting.

4.3. Computational Complexity. To compare the computational complexity of the above methods, we tested the methods on images of different sizes in the MATLAB experimental environment and report the average run time calculated from 20 operations on images of the same size. The results presented in Table 3 show that the SRIE method has the lowest computational efficiency when processing a single image, requiring 242.22 seconds to process an image with 2048×1536 pixels, while CVC, MSRCP,

(a) Source images

(b) Artificial processed images

(c) Output images with proposed method

Figure 17: Comparison with state-of-the-art methods.

Table 2: The assessment results on the images in Figure 17.

		CegaHE	CVC	LDR	DCP	EFF	MSRCP	SRIE	Our method
	MSE	839.31	4554.18	732.72	873.17	1292.50	1131.82	1135.16	*368.12*
Group 1	PSNR	18.89	11.55	19.48	18.72	17.02	17.59	17.58	*22.47*
	SSIM	0.82	0.74	0.85	0.83	0.86	0.82	0.80	*0.95*
	MSE	1212.11	1389.11	695.13	976.86	427.98	1167.74	1184.34	*683.90*
Group 2	PSNR	17.30	16.70	19.71	18.23	21.82	17.46	17.40	*19.78*
	SSIM	0.90	0.93	0.93	0.90	*0.96*	0.88	0.92	*0.96*
	MSE	712.05	3084.25	487.98	453.46	1172.70	712.31	588.00	*142.47*
Group 3	PSNR	19.61	13.24	21.25	21.57	17.44	19.60	20.44	*26.59*
	SSIM	0.80	0.77	0.86	0.85	0.87	0.86	0.88	*0.95*
	MSE	914.60	1518.88	1193.66	*672.35*	685.98	992.04	1105.18	748.61
Group 4	PSNR	18.52	16.32	17.36	*19.85*	19.77	18.17	17.70	19.39
	SSIM	0.89	0.88	0.89	0.91	0.93	0.88	0.90	*0.95*
	MSE	1016.39	3160.82	1106.58	492.06	651.73	1061.99	655.86	*478.82*
Groupe5	PSNR	18.06	13.13	17.69	21.21	19.99	17.87	19.96	*21.33*
	SSIM	0.81	0.78	0.83	0.82	0.90	0.83	0.89	*0.96*

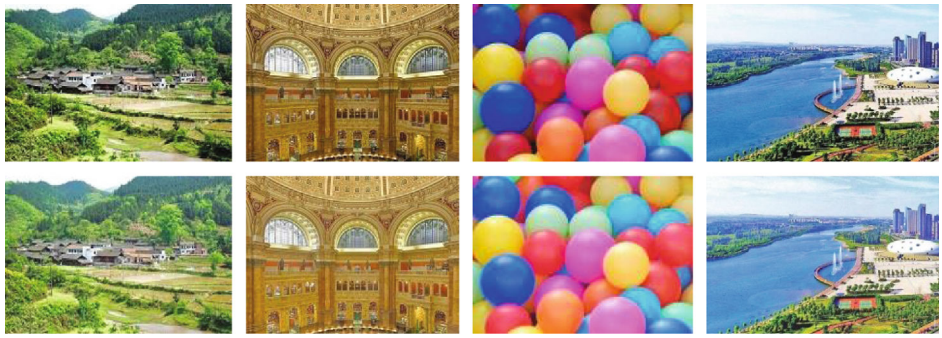
Table 3: Experimental results of computational complexity (unit: seconds).

	600×400	800×600	1024×768	1600×1200	2048×1536
CVC	0.27	0.40	0.60	1.27	2.33
MSRCP	*0.17*	0.40	0.88	3.29	7.41
DCP	0.33	0.60	1.06	2.42	3.89
SRIE	7.08	13.61	22.38	101.91	242.02
EFF	0.42	0.62	0.94	1.86	3.03
Proposed method	0.19	*0.32*	*0.51*	*1.09*	*2.14*

DCP, EFF, and the proposed method all require similar seconds to process the same image. With the size of the image increasing, the processing time of the MSRCP method increases more rapidly, while that of the other methods increases linearly. The proposed method requires the least run time and thus has the lowest time complexity.

4.4. Adaptivity of Our Method. We also tested the methods on images acquired under extremely low illumination as well as

(a) Processed on some images with extreme low light



(b) Processed on some images with adequate light

Figure 18: Examples of adaptive image enhancement under different illumination conditions.

images obtained in normal light conditions; the experimental results are shown in Figure 18. In the top panel of this figure, the first row contains the original images acquired under extremely low illumination, and the second row shows the corresponding enhancement results.

The results show that for the enhancement of images acquired under extremely low illumination, which has presented great challenges in the field of image processing, although the enhancement effect of the proposed method is unsatisfactory, no blocky effect is not present in the restored images; in this sense, they are consistent with human visual perception. In the bottom panel of the figure, the first and second rows show images acquired under normal illumination and the corresponding enhancement results obtained using the proposed method, respectively, and the results showed that for images acquired under normal illumination conditions, the processing results of the proposed method are identical to the original images, indicating that the proposed method can adaptively adjust its parameters for different scenes and thus shows good robustness and adaptability.

5. Conclusion

In this paper, we propose a color image correction method based on local gamma transformation and color compensation. In which the illumination-reflection model is adopted to address the problems of local overenhancement due to uneven illumination in low-light images and the lack of adaptability of the parameter settings encountered in previous methods. First, we convert the original RGB color image into the YUV color space and extract the illumination distribution of the scene from the Y component using a guided filtering function. Then, we perform illuminance enhancement based on an adaptive local gamma transformation and expansion of the dynamic range. Finally, we enhance the color saturation of the image. Comparisons between the proposed method and other conventional algorithms indicate that the proposed algorithm can not only effectively improve the visual effect of the processed image but also reveal more detailed information in dark regions. Because the proposed algorithm uses the distribution characteristics of the illumination component of the scene to dynamically adjust the parameters of the gamma function, it can effectively improve the visual quality of an image, allowing better identification of details in both overexposed and underexposed areas of the image.

Data Availability

Some or all data, models, or code generated or used during the study are available from the corresponding author by request (Wencheng Wang).

Conflicts of Interest

The authors declare that they have no conflicts of interest related to this work.

Acknowledgments

This research was funded by the Shandong Provincial Natural Science Foundation (No. ZR2019FM059), the Science and

Technology Plan for Youth Innovation of Shandong Universities (No. 2019KJN012), and the National Natural Science Foundation of China (No. 61403283).

References

[1] Z. Huang, Y. Zhang, Q. Li et al., "Joint analysis and weighted synthesis sparsity priors for simultaneous denoising and destriping optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 6958–6982, 2020.

[2] W. Wang, X. Yuan, X. Wu, and Y. Liu, "Fast image dehazing method based on linear transformation," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1142–1155, 2017.

[3] Z. Zhu, H. Wei, G. Hu, Y. Li, G. Qi, and N. Mazur, "A novel fast single image dehazing algorithm based on artificial multiexposure image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, article 5001523, pp. 1–23, 2021.

[4] W. Wang, F. Chang, T. Ji, and X. Wu, "A fast single-image dehazing method based on a physical model and gray projection," *IEEE Access*, vol. 6, pp. 5641–5653, 2018.

[5] M. Zheng, G. Qi, Z. Zhu, Y. Li, H. Wei, and Y. Liu, "Image dehazing by an artificial image fusion method based on adaptive structure decomposition," *IEEE Sensors Journal*, vol. 20, no. 14, pp. 8062–8072, 2020.

[6] M. Oloyede, G. Hancke, H. Myburgh, and A. Onumanyi, "A new evaluation function for face image enhancement in unconstrained environments using metaheuristic algorithms," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, 2019.

[7] G. Hou, Z. Pan, B. Huang, G. Wang, and X. Luan, "Hue preserving-based approach for underwater colour image enhancement," *IET Image Processing*, vol. 12, no. 2, pp. 292–298, 2017.

[8] X. Fu and X. Cao, "Underwater image enhancement with global-local networks and compressed- histogram equalization," *Signal Processing: Image Communication*, vol. 86, article 115892, 2020.

[9] T. Soomro, T. Khan, M. Khan, J. Gao, M. Paul, and L. Zheng, "Impact of ICA-based image enhancement technique on retinal blood vessels segmentation," *IEEE Access*, vol. 6, no. 6, pp. 3524–3538, 2018.

[10] F. Kallel, M. Sahnoun, A. Ben Hamida, and K. Chtourou, "CT scan contrast enhancement using singular value decomposition and adaptive gamma correction," *Signal Image & Video Processing*, vol. 12, no. 5, article 1232, pp. 905–913, 2018.

[11] Z. Huang, T. Zhang, Q. Li, and H. Fang, "Adaptive gamma correction based on cumulative histogram for enhancing near-infrared images," *Infrared Physics and Technology*, vol. 79, pp. 205–215, 2016.

[12] W. Wang, X. Wu, X. Yuan, and Z. Gao, "An experiment-based review of low-light image enhancement methods," *IEEE Access*, vol. 8, pp. 87884–87917, 2020.

[13] S. Huang, F. Cheng, and Y. Chiu, "Efficient contrast enhancement using adaptive gamma correction with weighting distribution," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 1032–1041, 2013.

[14] S. Chi Liu, S. Liu, H. Wu et al., "Enhancement of low illumination images based on an optimal hyperbolic tangent profile," *Computers & Electrical Engineering*, vol. 70, no. 8, pp. 538–550, 2018.

[15] J. Kim, L. Kim, and S. Hwang, "An advanced contrast enhancement using partially overlapped sub-block histogram equalization," *IEEE Transactions Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 475–484, 2001.

[16] Soong-der Chen and A. Ramli, "Minimum mean brightness error bi-histogram equalization in contrast enhancement," *IEEE Transactions on Consumer Electronics*, vol. 49, no. 4, pp. 1310–1319, 2003.

[17] T. Celik and T. Tjahjadi, "Contextual and variational contrast enhancement," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3431–3441, 2011.

[18] Z. Huang, Z. Wang, J. Zhang, Q. Li, and Y. Shi, "Image enhancement with the preservation of brightness and structures by employing contrast limited dynamic quadri-histogram equalization," *Optik*, vol. 226, article 165877, 2021.

[19] D. Jobson, Z. Rahman, and G. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Transactions on Image Processing*, vol. 6, no. 7, pp. 965–976, 1997.

[20] Z. Rahman, D. Jobson, and G. Woodell, "Retinex processing for automatic image enhancement," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 100–110, 2004.

[21] A. Petro, C. Sbert, and J. Morel, "Multiscale retinex," *Image Processing on Line*, vol. 4, no. 4, pp. 71–88, 2014.

[22] X. Fu, D. Zeng, Y. Huang, X. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2782–2790, Las Vegas, NV, USA, 2016.

[23] S. Wang, J. Zheng, H. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3538–3548, 2013.

[24] X. Dong, G. Wang, Y. Pang et al., "Fast efficient algorithm for enhancement of low lighting video," in *2011 IEEE International Conference on Multimedia and Expo*, pp. 1–6, Barcelona, Spain, 2011.

[25] L. Zhang, P. Shen, X. Peng et al., "Simultaneous enhancement and noise reduction of a single low-light image," *IET Image Processing*, vol. 10, no. 11, pp. 840–847, 2016.

[26] S. Park, B. Moon, S. Ko, S. Yu, and J. Paik, "Low-light image restoration using bright channel prior-based variational retinex model," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, 2017.

[27] Z. Huang, Y. Zhang, Q. Li et al., "Unidirectional variation and deep CNN denoiser priors for simultaneously destriping and denoising optical remote sensing images," *International Journal of Remote Sensing*, vol. 40, no. 15, pp. 5737–5748, 2019.

[28] K. Lore, A. Akintayo, and S. Sarkar, "LLNet: a deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[29] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, "MSR-net: low-light image enhancement using deep convolutional network," 2017, https://arxiv.org/abs/1711.02488.

[30] L. Tao, C. Zhu, G. Xiang, Y. Li, H. Jia, and X. Xie, "LLCNN: a convolutional neural network for low-light image enhancement," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4, St. Petersburg, FL, USA, 2017.

[31] S. Park, S. Yu, M. Kim, K. Park, and J. Paik, "Dual autoencoder network for Retinex-based low-light image enhancement," *IEEE Access*, vol. 6, no. 3, pp. 22084–22093, 2018.

[32] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.

[33] W. Wang, Z. Chen, X. Yuan, and X. Wu, "Adaptive image enhancement method for correcting low-illumination images," *Information Sciences*, vol. 496, pp. 25–41, 2019.

[34] W. Wang and X. Yuan, "Recent advances in image dehazing," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 410–436, 2017.

[35] K. He, J. Sun, and X. Tang, "Guided image filtering," *Institute of Electrical and Electronics Engineers transactions on pattern analysis and machine intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[36] R. Schettini, F. Gasparini, S. Corchs, F. Marini, A. Capra, and A. Castorina, "Contrast image correction method," *Journal of Electronic Imaging*, vol. 19, no. 2, article 023005, 2010.

[37] T. Pu and S. Wang, "Perceptually motivated enhancement method for non-uniformly illuminated images," *IET Computer Vision*, vol. 12, no. 4, pp. 424–433, 2017.

[38] C. Chiu and C. Ting, "Contrast enhancement algorithm based on gap adjustment for histogram equalization," *Sensors*, vol. 16, no. 6, p. 936, 2016.

[39] C. Lee, C. Lee, and C. Kim, "Contrast enhancement based on layered difference representation of 2D histograms," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5372–5384, 2013.

[40] W. Wang, X. Yuan, X. Wu, and Y. Liu, "Dehazing for images with large sky region," *Neurocomputing*, vol. 238, pp. 365–376, 2017.

[41] W. Wang, Z. Chen, X. Yuan, and F. Guan, "An adaptive weak light image enhancement method," in *The 12th International Conference on Signal Processing Systems*, vol. 11719, Shanghai, China, November 2020.

[42] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, "A new image contrast enhancement algorithm using exposure fusion framework," in *International Conference on Computer Analysis of Images and Patterns*, vol. 10425, pp. 36–46, Ystad, Sweden, August 2017.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5525602, 12 pages
https://doi.org/10.1155/2021/5525602

Research Article

A Biologically Inspired Algorithm for Low Energy Clustering Problem in Body Area Network

Mengying Xu and Jie Zhou

College of Information Science and Technology, Shihezi University, Shihezi, China

Correspondence should be addressed to Jie Zhou; jiezhou@shzu.edu.cn

Received 24 January 2021; Revised 11 March 2021; Accepted 15 April 2021; Published 26 April 2021

Academic Editor: Bin Gao

The growing application of body area networks (BANs) in different fields makes the low energy clustering a paramount issue. A clustering optimization algorithm in BANs is a fundamental scheme to guarantee that the essential collected data can be forwarded in a reliable path and improve the lifetime of BANs. Low energy clustering is a technique, which provides a method that shows how to reduce network communication costs in BANs. A careful low energy clustering scheme is one of the most critical means in the research of BANs, which has attracted considerable attention, comprising monitoring capability constraints. However, the classical clustering method leads to high cost when constraints such as large overall energy consumption are undertaken. Hence, a binary immune hybrid artificial bee colony algorithm (BIHABCA), a randomized swarm intelligent scheme applied in BANs, motivated by immune theory and hybrid scheme is introduced. Furthermore, we designed the formulation that considers both distances between two nodes and the length of bits. Finally, we have compared the energy cost optimized by BIHABCA with a shuffled frog leaping algorithm, ant colony optimization, and simulated annealing in the simulation with different quantity of nodes in terms of energy cost. Results show that the energy cost of the network optimized by the proposed BIHABCA method decreased compared to those by the other three methods which mean that the proposed BIHABCA finds the global optima and reduces the energy cost of transmitting and receiving data in BANs.

1. Introduction

As a branch of wireless sensor networks (WSN), a special network in medical applications called body area network (BAN) is an important network in biomedical and many other fields, which plays an important role in telemedicine, special population monitoring, and community medical services [1, 2]. It is a special type of sensor network, which has brought tremendous changes to human society. It is used for health care, individual health recovery, and sports, even in social public health to collect the data of electrocardiograph signal, blood pressure, blood sugar, temperature, etc. [3, 4]. BAN is one of the most important networks in telemedicine, special population monitoring (such as infants and the elderly), and community health care and has broad application prospects [5].

The typical wireless BAN consists of three parts: medical sensor nodes (the nodes or devices are generally placed on the human body), sink nodes, and network management nodes [6]. BAN is generally constructed with a distributed strategy. Medical sensor nodes and portable mobile devices are employed to collect health data of the human body. The specific process is as follows: sensor nodes collect physiological data by monitoring the human body, and the collected information is transferred to the sink node. The sink node communicates with network management nodes using the Internet, satellite, and other communication methods, which can control and manage the BAN. Medical sensor nodes are evenly distributed in different areas of the human body, including the head, limbs, and trunk. In addition, in order to transmit the collected vital data effectively, the transmission distance, transmission rate, and residual energy of each node must also be fully considered [7, 8]. A BAN consists of distributed autonomous sensor nodes, and each sensing unit consists of sensing boards, processor, short-range radio transmitter unit, and battery. Each node is a small and

compact device to sense the healthy data and send to the base station [9–11].

Clustering optimization is an important sensor deployment issue in many industrial, consumer, and environmental monitoring applications. The storage capacity of BANs is generally small, which limits the network's lifetime. In order to save communication energy and further prolong the lifetime of BANs, an effective clustering optimized method is designed to ensure the reasonable allocation of cluster heads on each path and improve transmission speed [12]. However, there are many defects and shortcomings in the research of low energy clustering. Firstly, the random selection of cluster heads leads to uneven distribution of cluster and cluster heads. Secondly, in the cluster head selection, the remaining energy of nodes, the amount of neighboring nodes, and the amount of times that the cluster head has been used are not considered, which aggravates the burden of cluster head and makes the communication energy and lifetime uneven. Thirdly, cluster heads near sink nodes consume more energy and are prone to premature death [13–15].

The artificial bee colony algorithm (ABCA) is a swarm intelligent method, which imitates the honey gathering behavior of bees. It is similar to many heuristic algorithms and belongs to an intelligent optimization algorithm. The optimization performance of ABCA is superior because few parameters should be adjusted and the complexity of the algorithm is low. Therefore, it is suitable for solving NP-hard problems, such as a clustering algorithm. However, many other heuristic algorithms, such as ant colony optimization, genetic algorithm, and evolutionary algorithm, have many parameters to adjust and are easy to fall into evolutionary stagnation. Therefore, they are not the optimal methods to optimize the clustering algorithm.

In this paper, in order to optimize the position of cluster heads and energy cost of BAN, we propose a binary immune hybrid artificial bee colony algorithm (BIHABCA) in the low energy clustering problem for reducing network communication costs. To solve the problem, the distance between nodes and the number of data are considered to model the low energy clustering problem and introduce the fitness function of calculating energy cost. In BIHABCA, the immune operator and hybrid mechanism are considered into the artificial bee colony algorithm (ABCA) to develop global search capability, and the fitness function for low energy clustering is given to calculate the energy cost of the network in each generation.

In the simulation, we compare the BIHABCA with the shuffled frog leaping algorithm (SFLA), ant colony optimization (ACO), and simulated annealing (SA) in BANs with different quantity of cluster heads on the human body. The BIHABCA has demonstrated to have good potential in the optimization scheme of cluster heads to reduce communication energy. It also avoids local optima and improves the quality of solutions. The main contributions are given as follows:

(1) Firstly, the BIHABCA method can successfully minimize the network energy consumption in BANs. After iterations, the energy cost optimized by the BIHABCA is 13.00%, 21.38%, and 27.38% less than SFLA, ACO, and SA, respectively, with 10% cluster heads when the number of nodes in BAN is 100. Furthermore, when the number of nodes increases, the similar conclusions can be deduced by comparing with other three algorithms. The clustering method based on BIHABCA requires less communication energy consumption of nodes in BANs, which can successfully strengthen the energy utilization efficiency

(2) Secondly, the BIHABCA combined with immune and hybrid operators has better performance with no premature convergence. When the cluster head nodes account for 10% and 20% of total sensor nodes, respectively, in BANs. The BIHABCA has higher convergence rate than the SFLA, ACO, and SA. After iteration, the fitness optimized by BIHABCA converges to the optimal value compared with those by the other three algorithms

(3) Finally, total energy cost of a system depends on energy costs on transmission and reception of all node. Therefore, with the increase of the amount of sensors in BANs, the demand for data transferred increases and the energy cost also improves correspondingly

2. Related Work

In the initial position management, the medical sensors are arranged arbitrarily in the BANs and different detection areas have different densities. Due to the small sensing radius of the sensor, the energy of the node is not only used for sensing data but also for transmitting data. If the node approaches the sink sensor, more energy is needed for transmitting the data. When the battery is exhausted, the data cannot be transmitted to the sink node, which will lead to the phenomenon of an energy hole. The research hotspot in this field is to optimize this problem by a heuristic swarm intelligent method for the practical situation.

In recent years, many researchers use different methods to solve problems of routing, clustering, and mobile sink for WSNs. In the WSN, a clustering strategy depending on a wolf pack algorithm using levy flight is definitely given to enhance the overall efficiency of system in [16]. Through simulation evaluation, a network's lifetime is steadily raised and the energy efficiency is even better balanced.

In [17], the authors give a hexagon beehive model. Nodes are allocated throughout a hexagon arbitrarily. Therefore, the experiment results of the suggested model indicate the improvement in the residual energy between sensors, minimizing over-all energy cost and finally strengthening the life span of the system.

In order to reduce the risk of premature sensor death in the system, in [18], a cluster head assortment strategy with ACO-based MDC is given. In the experiment, the suggested approach can easily improve the sensor network lifespan considerably; nevertheless, the computing complex is too high.

In [19], the authors study a survey on clustering methods or algorithms reported in WSNs. They notice the clustering algorithms have certainly few overall efficiency in lessening system communication cost. It is also appropriate for limited equipment constraints of WSNs.

There is an effective clustering strategy with a hybrid anomaly monitoring technique for misdirection as well as black hole. An experiment was performed to lessen the energy consumption of the system in [20]. The outcomes illustrate that the suggested approach is important in security application.

In [21], the change mechanism of sensing radius is proposed to improve the network lifetime. It can be regarded as a linear programming problem. In [22], the lifetime is developed by increasing the number of sensors. Although the method improves the lifetime, it increases the energy consumption of WSN.

In BANs, an effective clustering method that considers large scale nodes and computational time is proposed. Experiment data represent that the method improves communication efficiency and lifetimes of networks [23]. Some recent works on the clustering issue are mainly focused on the optimization algorithms which are given in the following in-depth description.

In the WSN, the cluster head selection is improved by a grey wolf optimizer that considers both average intracluster distance and residual energy to lessen energy consumption in [24]. The suggested clustering method improved by the grey wolf optimizer is applied to raise the network lifespan. The acquired results show that the overall performance and efficiency of the given method outperform other metaheuristics. It expands network lifetime successfully.

In [25], the authors study a gravitational search algorithm to optimize gradient clustering selection in the system. They study the length from the cluster heads to the gateway nodes as well as the remaining energy of the gateway nodes in the model. The suggested gravitational search algorithm is dependent on an evolutionary optimization. Experimental tests display the effectiveness and scalability of the offered strategy and demonstrate a great balance between exploration and search capabilities.

Particle swarm optimization is used to lessen the energy optimized dynamic clustering, to additionally choose the ideal cluster head in [26]. In the fitness computation, Manhattan distance is specially designed to compute the shortest route between the cluster heads and the base station. Energy consumption has been decreased by a suggested algorithm.

Optimal cluster head selection is explored employing an artificial bee colony metaheuristic in [27]. The provided approach uses an advanced population sampling strategy. The provided technique raises the global convergence and boosts energy effectiveness in a system.

In [28], a distributed clustering strategy based on ACO is given to improve communication efficiency. Furthermore, the allocation of cluster heads is selected using the Manhattan distance. Sensing data is transmitted between the two nodes. Experience results show that the energy of the network reduces efficiently. In [29], the SFLA is given in the clustering algorithm in the BAN to reduce total computational time. It can also effectively optimize communication costs after a limited number of iterations. Simulation results show that SFLA effectively extends the BAN's stable period and lifetime, and the energy consumption is balanced effectively. More heuristic optimization algorithms are introduced in recent years.

3. System Model

In this section, after determining how many clusters in the BAN should be divided into, the cluster heads in each cluster must be carefully selected. The distance between nodes, the amount of data, and other factors must be considered when the cluster heads are selected, to make the cluster stable and extend the effective working time of the BANs. Therefore, we build a system model of the low energy clustering problem. With the continuous work to sense data, sensor nodes will stop working because of the exhaustion of energy. Energy cost is used to evaluate the lifetime of the BANs, which is defined as the time from the beginning of the BAN to death.

BAN is divided into several clusters, which include cluster member nodes and cluster heads. The normal nodes send medical information to the cluster heads. After that, data is sent to the base station from sink nodes. Cluster effectively saves the total energy consumption of BAN when the amount of information is huge. It is significant to the sensor network for transferring data.

The energy cost of a node is mostly consumed in the phase of transmission and reception. The low energy clustering model is simplified, and the wireless communication module is only considered in this paper. The energy consumption required by a sensor to transmit data includes transmission energy and reception energy. Our goal is to calculate the communication cost consumed in the process of transmission energy and reception energy of all sensor nodes in BANs.

Energy consumption of transmitting data includes energy consumption of signal transmitting circuit and signal amplifying circuit. Reception energy is the consumption of receiving data in the signal receiving circuit.

Transmission energy is related not only to the length of bits but also to the distance of data transmission between nodes. The energy consumed by a node to send data with k bits to another node whose distance is d. Energy costs on transmission and reception of one node are given in

$$E_t(k, d) = E_{\text{elec}} \cdot k + k \cdot \varepsilon_{\text{amp}} \cdot d^3, \tag{1}$$

$$E_r(k) = E_{\text{elec}} \cdot k, \tag{2}$$

where E_t is the transmission energy, E_r is the reception energy, and E_{elec} represents the consumption parameter of transmitting or receiving per bit. Furthermore, $E_{\text{elec}} \cdot k$ in the above equations represents the energy consumption of transmitting or receiving k bits. ε_{amp} is the power amplification parameter. Transmission energy cost can be affected by the distance between two connected nodes.

$$E = E_t + E_r. \tag{3}$$

In (3), E is the total energy cost of a node, which includes transmission energy and reception energy.

$$E_{\text{sum}} = \sum_{m=1}^{M} E_m. \quad (4)$$

Suppose that there are M nodes in the coordinate area. E_{sum} is the total cost of M nodes in BANs.

In (5), the function is designed to find the minimum energy of whole medical nodes in BANs when the cluster head nodes are allocated.

$$f = \min\ (E_{\text{sum}}), \quad (5)$$

where f indicates the quality of the individuals after clustering. It also means the energy consumption of the allocation scheme of cluster heads in a round when communicating with other nodes. The individual with lower energy consumption in a round has better performance than other individuals.

The goal of building the low energy clustering model of BANs is to reduce the energy cost by allocating the cluster head nodes when data is transmitted and received in the network. In this way, the lifetime of BANs is improved effectively.

4. BIHABCA for Minimizing Energy Cost in BANs

4.1. Basic ABCA. The ABCA is a swarm intelligent method proposed by a Turkish scholar in 2005, which imitates the process of honey gathering of bees. Honeybees carry out different activities and perform the information sharing and selection between the colonies, to further obtain the best solution of the problem. It is similar to many heuristic algorithms and belongs to an intelligent optimization algorithm. Good results have been achieved in solving continuous combinatorial optimization problems [30].

Reference [31] proves that the convergence speed of the ABCA is superior to those of heuristic algorithms in solving multiobjective and multiextremum function problems, avoiding falling into local optimum as well as solving engineering problems including complex and multiextremum value. ABC is employed to cope with many projects involving traveling salesman problem, knapsack problem, engineering, software testing, neural networks, job scheduling, etc.

In the ABCA, one half of the bees is composed of employed bees, and the other half is composed of onlookers. The amount of employed bees is equal to that of onlookers. A food source represents a possible solution.

In the ABCA, each iteration consists of the three parts: employed bees are employed to collect the location of the honey source and calculate its solution; onlookers evaluate the fitness and use wheel roulette selection to calculate the possibility. Food source will be chosen with probability by calculating the value of fitness. Scout bees find a new individual to take the place of the abandoned one.

The optimization process of the ABCA is given as follows: firstly, initialize the location of food location and analyze the fitness of each individual. Employed bees generate a new food location nearby as well as assess the fitness of a new individual. Greedy selection is executed between the new solution and the old one [32].

After the employed bee has finished the search process, the onlookers evaluate the fitness as well as its location from all the bees in the area and select the honey location according to the probability of fitness. With the increase of the quality of solution of the honey source, the possibility of the honey source to be selected also increases. If a solution is abandoned, the scout bee randomly generates a honey source to replace the abandoned one.

4.1.1. Phase of Employed Bees. In the ABCA, the quantity of food locations is determined by the quantity of employed bees. In the BIHABCA, each bee generates a random location $v_{n,m}$ as follows:

$$v_{n,m} = x_{n,m} + \varphi(x_{n,m} - x_{k,m}) \quad n = 1, 2, \cdots, N\ (n \neq k), \quad (6)$$

where $x_{n,m}$ is the n_{th} food source in the population; $m = 1, 2, \cdots, M$; M is the dimension of the population; $x_{k,m}$ is another food source near $x_{n,m}$; φ is a random number subject to mean and distribution, $\varphi \in [-1, 1]$; and N represents the number of food locations.

If the solution is better than the earlier one, the employed bee will remember the better solution; otherwise, it will still remember the old solution.

4.1.2. Phase of Onlookers. In the ABCA, for each food location in the phase of employed bees, the probability of each individuals being selected in the whole population is calculated in

$$\text{Possibility}_n = \frac{\text{fitness}_n}{\sum_{n=1}^{N} \text{fitness}_n}, \quad (7)$$

where fitness_n represents the quality of the n_{th} honey location and Possibility_n is the possibility that the onlooker chooses the food location in the population. It indicates that the honey source with high possibility has better solution to the problem.

4.1.3. Phase of Scout Bees. When an individual optimized many times in the stage of employed bees and onlookers is not improved, it will be discarded. The new honey source will be randomly produced by scout bees after limit times, limit is the upper limit of the algorithm. The new food location is randomly selected depending on (6).

4.2. Binary Immune Hybrid Artificial Bee Colony Algorithm. In this paper, the BIHABCA is used to solve the low energy clustering problem in BANs. It is improved in encoding and updating of honey source of BIHABCA in BANs to decrease the total energy cost when monitoring human health data and improve the global exploration abilities of proposed algorithm. The energy clustering problem is

Table 1: Description of parameters.

Parameters	Description
Number of individuals	The number of individuals is equal to the sum of the employed bees and following bees. The number of employed bees is equal to the number of following bees
Number of iterations	Optimization times of algorithm
Number of food source	The number of food source is equal to the number of employed bees. A food source represents a solution of clustering problem
Limit	Number of times when there is search honey source
Number of groups	A population can be divided into several sub-memeplexes
Number of local iterations	Optimization times of local search
Weight of pheromone	The pheromones released by ant: it determines the selection of path
Weight of heuristic information	It determines the possibility that the ant will take the path before it chooses
Pheromone volatilization coefficient	The rate at which pheromones evaporate over time
Initial temperature	Maximum temperature in the first generation
Annealing temperature coefficient	The rate of temperature drop

regarded as antigen. A possible solution is an antibody. There are N antibodies in the population.

The steps of BIHABCA include solution encoding and initialization, fitness evaluation, stage of employed bees, stage of onlookers, stage of scout bees, and termination condition.

4.2.1. Solution Encoding and Initialization. BANs are divided into cluster head nodes and sensing nodes to collect heathy data. Binary coding is used to describe the clustering problem in BANs. In (8), P is a population that contains N antibodies and each antibody represents the mode of M sensors: $P = \{P_1, P_2, \cdots, P_N\}$ and $P_n = \{p_{n,1}, p_{n,2}, \cdots, p_{n,M}\}$. When $p_{n,m} = 1$, the sensor node is a cluster head node. On the contrary, when $p_{n,m} = 0$, the sensor node is a common node. For instance, $p_n = \{1, 0, 0, 0, 1, 0, 1, 0, 0, 0\}$ represents that there are 10 nodes in the n_{th} antibody, where the 1st, 5th, and 7th nodes are cluster heads and the rest of the individuals represent ordinary nodes. In (9), the maximum amount of cluster heads in an antibody is T.

In the BIHABCA, the initial generation is 0 and the population is optimized for $\text{GEN}_{\max}$ times.

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,M-1} & p_{1,M} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,M-1} & p_{2,M} \\ \cdots & \cdots & p_{n,m} & \cdots & \cdots \\ p_{N-1,1} & p_{N-1,2} & \cdots & p_{N-1,M-1} & p_{N-1,M} \\ p_{N,1} & p_{N,2} & \cdots & p_{N,M-1} & p_{N,M} \end{bmatrix}, \quad (8)$$

$$T = \sum_{m=1}^{M} p_{n,m}. \quad (9)$$

4.2.2. Fitness Evaluation. Evaluate the energy cost of each food source in (4) and (5), and record the minimum energy cost and its food source. The fitness function of an individual can be calculated in (5).

Table 2: Parameter setting of BIHABCA.

Number of individuals	Number of food source	Limit	Number of iterations
80	40	20	100

Table 3: Parameter setting of SFLA.

Number of individuals	Number of groups	Number of local iterations	Number of iterations
80	4	10	100

Table 4: Parameter setting of ACO.

Number of individuals	Weight of pheromone	Weight of heuristic	Number of individuals	Weight of pheromone
80	2	2	0.98	100

Table 5: Parameter setting of SA.

Number of individuals	Initial temperature	Annealing temperature coefficient	Number of iterations
80	300	0.85	100

4.2.3. Phase of Employed Bees. In the BIHABCA, a honey source represents a feasible solution in the low energy clustering problem in (6) that is randomly generated by employed bees [33, 34]. In addition, the quantity of honey locations is equal to the quantity of employed bees. Employed bees and onlookers are fifty percent of the total amount of bees, respectively. In the low energy clustering problem, the

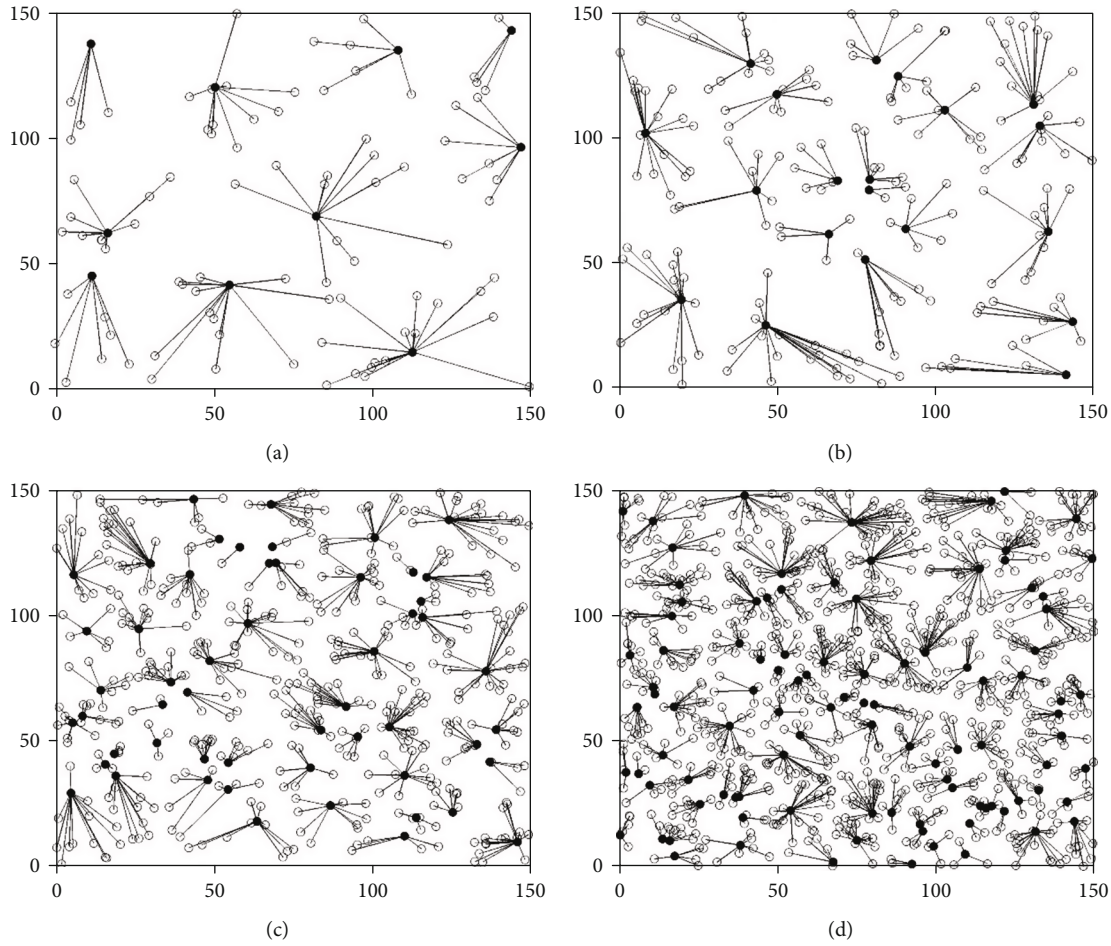


Figure 1: The changes in the number of sensor nodes and cluster head nodes: (a) 100 nodes; (b) 200 nodes; (c) 500 nodes; (d) 1000 nodes.

binary updating method is designed to calculate the position of food source. A sigmoid function is given to transform food source as follows:

$$\mathrm{Sig}(x_{n,m}) = \frac{1}{1 + \exp(-x_{n,m})},$$
$$p_{n,m} = \begin{cases} 1\ r, < \mathrm{sig}(x_{n,m}), \\ 0\ r, \geq \mathrm{sig}(x_{n,m}). \end{cases} \quad (10)$$

4.2.4. Phase of Onlookers. The onlookers choose the honey source depending on the quality of solution. The probability of the onlookers choosing a honey source based on the energy cost of all nodes in BANs can be calculated in

$$\mathrm{Possibility}_n = \frac{1/f_n}{\sum_{n=1}^{N} 1/f_n}, \quad (11)$$

where $\mathrm{Possibility}_n$ is the possibility of the n_{th} individual being selected by onlookers and f_n is the energy cost of transmitting data and receiving data in BANs. However, individuals with low energy costs are more likely to be selected. Therefore, fitness is converted to reciprocal to calculate the possibility. In a low energy clustering problem, the food locations with lower communication costs are more likely to be chosen by the onlookers.

4.2.5. Phase of Scout Bees. When an individual is not updated within limit times, it should be discarded by employed bees. The scout bees will generate a new food location. In the BIHABCA, a new food location can be randomly generated instead of the abandoned one in

$$p_{n,m} = \begin{cases} 1\ r, \geq 0.5, \\ 0\ r, < 0.5. \end{cases} \quad (12)$$

4.2.6. Termination Condition. In this work, stop searching when the iteration counter is equal to $\mathrm{GEN}_{\max}$, and output the best clustering scheme with the lowest energy cost.

4.2.7. Basic Steps of BIHABCA. The basic steps of the BIHABCA are described as follows:

Step 1. Population and its parameters are initialized. Set the quantity of food sources N, limit times in the phase of scout bees, and maximum number of iterations $\mathrm{GEN}_{\max}$

Step 2. Binary code is designed to describe the food source in low energy clustering problem. Initial antibodies are randomly created

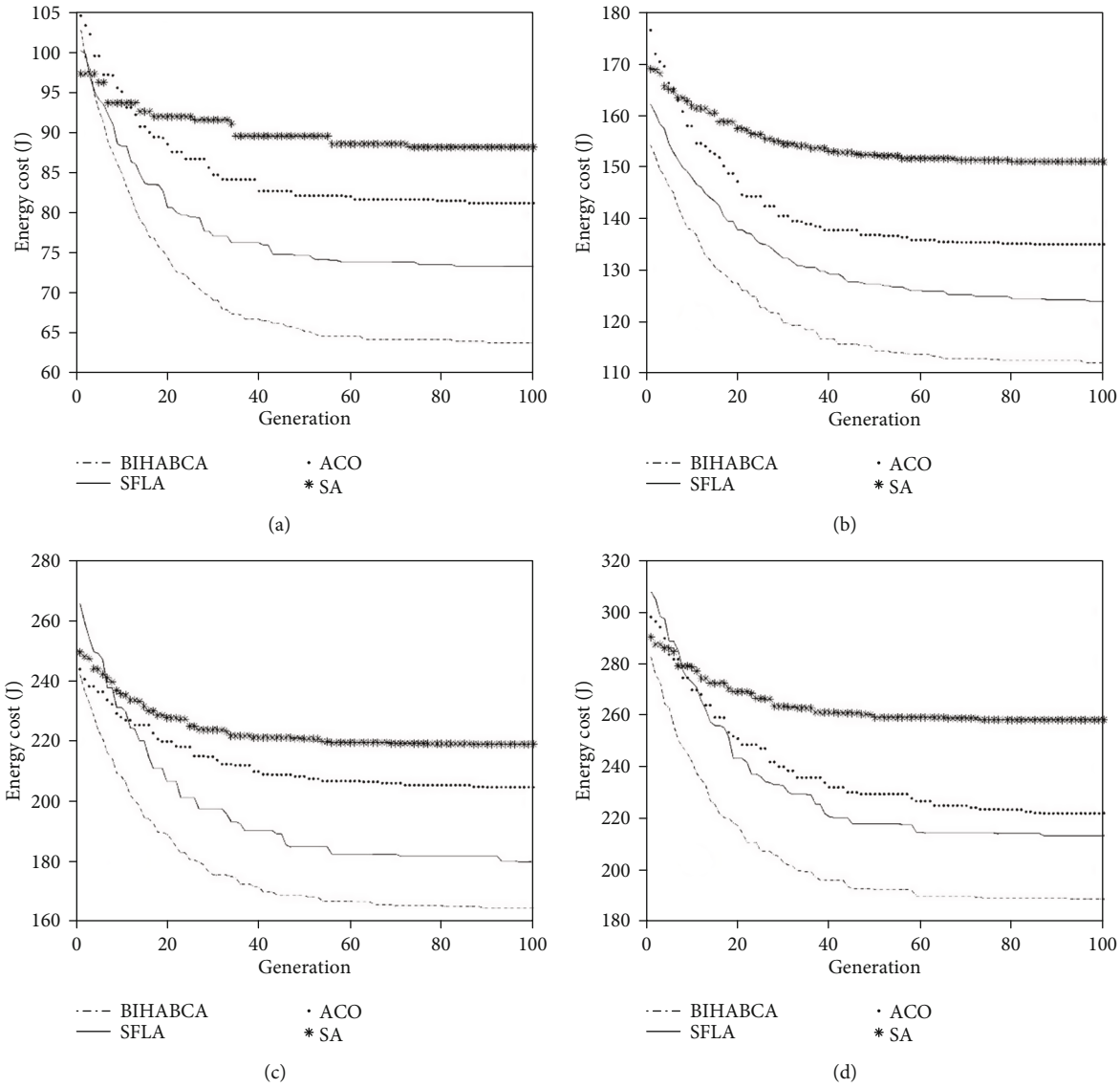


Figure 2: The changes in energy cost with 10% cluster head nodes: (a) 100 nodes; (b) 200 nodes; (c) 300 nodes; (d) 400 nodes.

Step 3. Calculate the affinity (total transmission energy and reception energy of M nodes in BAN), and find the optimal solution. Individuals with lower network energy consumption have better affinity

Step 4. Add one iteration, $t = t + 1$

Step 5. Add one bee, $n = n + 1$

Step 6. Use employed bees to update the food source in (6) and convert to binary code in (7)

Step 7. The onlookers are selected according to the probability Possibility_n to produce a new solution

Step 8. If food source is not updated after limit times by employed bees, it will be abandoned. Scout bees will randomly produce a new food source in equation (12)

Step 9. The several antibodies with the lowest energy cost in each generation are replaced to update some highest antibodies in population to the next generation

Step 10. If $\text{GEN}_{\max}$ is met, output the best individual and its energy cost. Otherwise, turn to step 4

4.2.8. Computational Complexity Analysis. Computational complexity of the proposed method is considered in this part to demonstrate the performance. The distance and energy cost need to be calculated from each sensor to other sensors. Thus, $O(M^2)$ complex multiplications are needed in energy calculation in the system model, where M is the number of sensors. For the BIHABCA, each population contains N individuals and each individual represents the working mode of M sensors. The optimization is executed G generation times. Thus, as for the worst situation, the computational complexity in the BIHABCA for minimizing energy cost in BANs is $O(M^2) + O(GNM)$.

5. Simulation Results and Analysis

In this section, the energy cost optimized by the BIHABCA method and SFLA, ACO, and SA with different quantities of medical sensors are simulated using MATLAB R2104a.

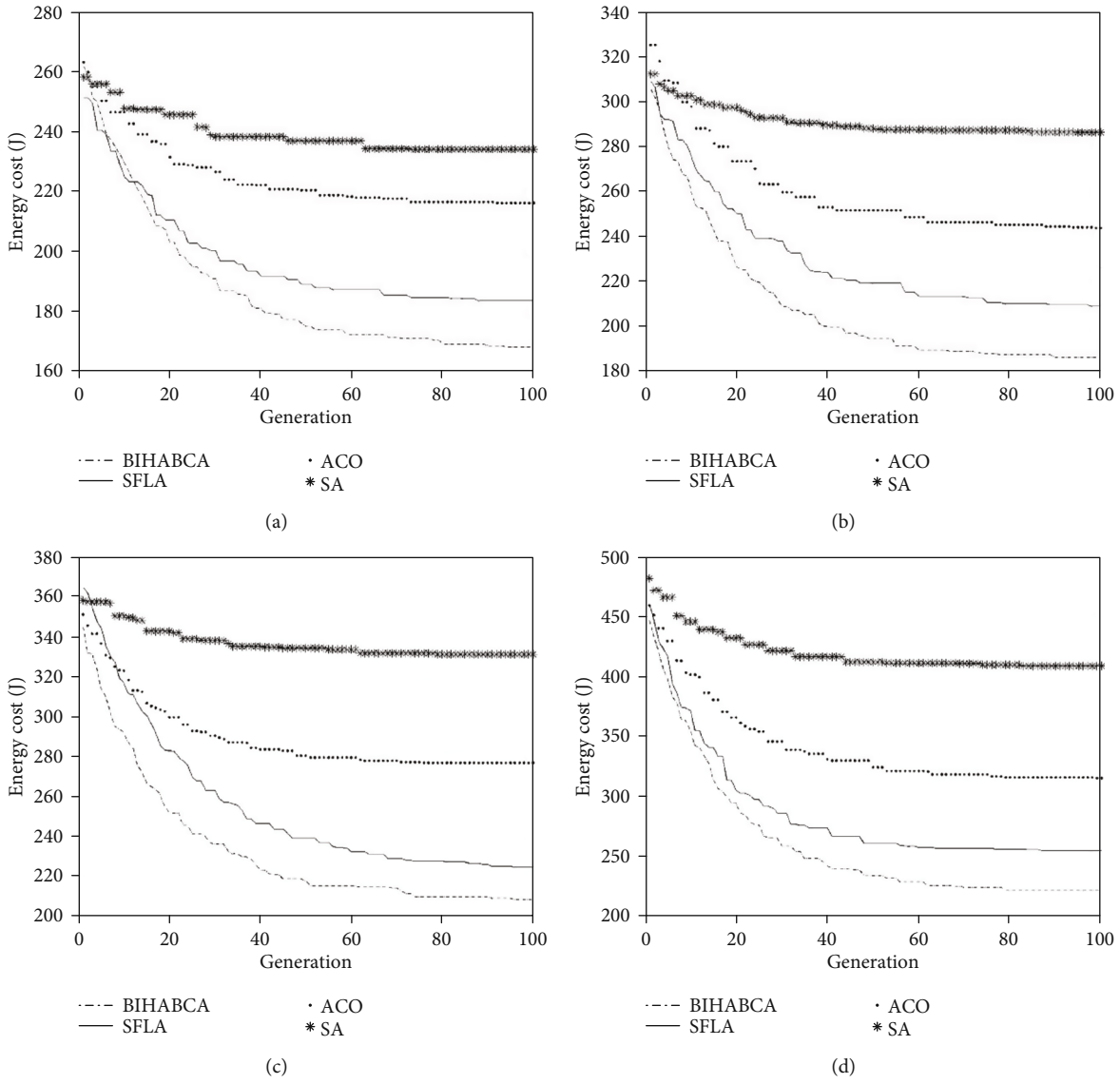


Figure 3: The changes in energy cost with 20% cluster head nodes: (a) 100 nodes; (b) 200 nodes; (c) 300 nodes; (d) 400 nodes.

The fitness of algorithms can be calculated in part III. As indicated earlier, the objective function in section III is to calculate the effectiveness of schemes in the low energy clustering problem. We assume that the monitoring area is the hospital. Patients are equipped with sensor nodes to monitor blood pressure, blood sugar, temperature, etc. In the simulation, all sensors are initially considered uniformly distributed in the monitoring area (150 m × 150 m). Optimization times of the BIHABCA, SFLA, ACO, and SA are 100, respectively. Monte Carlo simulation is employed in the experiments. The final result in each generation is the average value of 100 experiments; the advantage of proposed algorithm is statistically significant in the experiment.

Some critical parameters in BANs are given as follows: $E_{\text{elec}} = 50\,\text{nJ/bit}$, $\varepsilon_{\text{amp}} = 100\,\text{pJ/bit/m}^2$, and $k = 1\,\text{Mbps}$. The total energy cost of receiving data and transmitting data optimized by the BIHABCA will be compared with the SFLA, ACO, and SA.

Table 6: Energy costs optimized by the BIHABCA, SFLA, ACO, and SA for 20% cluster head nodes (J).

Number of nodes	BIHABCA	SFLA	ACO	SA
100	167.32	182.65	215.80	233.54
200	184.67	207.63	242.65	286.02
300	207.37	222.50	275.82	330.62
400	220.20	253.52	312.32	407.87

As for the BIHABCA, there are few parameters that have to be adjusted. However, the parameters in BIHABCA are sensitive, because they have to be repeated and tested a number of times in the process of adjusting parameters in the experiments. If the parameters of the BIHABCA are not suitable for the clustering problem, the performance of the algorithm will be reduced and energy cost cannot be optimized effectively.

Each controlling parameter has a range of empirical values and they need to be adjusted in the range of empirical values until the optimal energy cost is reached.

Table 1 gives the description of parameters.

In order to ensure that the comparison between the comparative methods is fair, the same values of parameters of the BIHABCA, SFLA, ACO, and SA are given in the compare experiment, which include the consumption parameter of transmitting or receiving per bit E_{elec}, transmitting or receiving k bits, and power amplification parameter ε_{amp}. Furthermore, the parameters in the four optimization algorithms are the same, which include experiment times, iteration times, and number of individuals

In Table 2, we set the parameters of the BIHABCA, which are tested many times to get the optimal solution. In Tables 3–5, the parameters of the SFLA, ACO, and SA are given, respectively.

In Figure 1, the examples of clustering schemes optimized by BIHABCA are given from Figures 1(a)–1(d) when the numbers of nodes are 100, 200, 500, and 1000, respectively. In the following figures, the sensing area is a square (150 m × 150 m). Hollow circles mean normal nodes, and filled circles are cluster heads.

The total communication costs of all medical sensors receiving data and transmitting data optimized by BIHABCA, SFLA, ACO and SA, respectively, when with 10% cluster heads are shown from Figures 2(a)–2(d). The figures show the energy consumption of the BANs of the BIHABCA, SFLA, ACO, and SA on the low energy clustering problem when the quantities of sensor nodes are 100, 200, 300, and 400, respectively. After 100 runs, BIHABCA yielded much lower energy costs compared with the SFLA, ACO, and SA as shown in Figure 2.

In Figure 2(a), the energy optimized by the BIHABCA is 63.71 J when the number of nodes in the BAN is 100. Furthermore, the energy cost of receiving and transmitting data optimized by SFLA is 73.23 J with the same number of nodes in the BAN; the energy optimized by the ACO is 81.04 J, and the energy optimized by SA is the 87.73 J with 100 nodes in the BAN. Compared with the traditional SFLA, ACO, and SA method, the BIHABCA method we designed can, respectively, reduce total transmission energy and reception energy by 13.00%, 21.38%, and 27.38%. Furthermore, the proposed optimized BIHABCA is capable of providing a global optimum solution with a faster convergent speed.

Similar results can be obtained in Figures 2(b)–2(d). In Figure 2(b), the energy costs of receiving and transmitting data optimized by the BIHABCA, SFLA, ACO, and SA are 111.64 J, 123.72 J, 134.82 J, and 150.60 J, respectively, when the number of medical nodes is 200. In Figure 2(c), the energy costs of receiving and transmitting data optimized by the BIHABCA, SFLA, ACO, and SA are 164.13 J, 179.26 J, 203.92 J, and 218.52 J, respectively, when the number of medical nodes is 300. In Figure 2(d), the energy costs of receiving and transmitting data optimized by the BIHABCA, SFLA, ACO, and SA are 187.56 J, 210.87 J, 220.54 J, and 257.33 J, respectively, with 400 medical nodes. Simulations indicate that BIHABCA can reduce the total energy cost to further extend the lifetime and show the effectiveness of the proposed strategy.

Table 7: Percentage of energy cost reduction optimized by the BIHABCA than the other three algorithms with 10% cluster heads.

Number of nodes	SFLA	ACO	SA
100	13.00%	21.38%	27.38%
200	9.76%	16.58%	25.87%
300	8.44%	19.51%	24.89%
400	11.05%	14.95%	27.11%

Table 8: Percentage of energy cost reduction optimized by the BIHABCA than the other three algorithms with 20% cluster head nodes.

Number of nodes	SFLA	ACO	SA
100	8.39%	22.47%	28.35%
200	11.06%	23.81%	35.43%
300	6.80%	24.82%	37.28%
400	13.14%	29.72%	46.01%

Table 9: Runtime (s).

Percentage of cluster head	Number of sensors	BIHABCA	SFLA	ACO	SA
10%	100	211.06	223.15	247.62	266.12
	200	357.21	367.44	393.68	451.67
	300	521.12	567.90	637.56	623.34
	400	652.81	676.49	745.34	787.24
20%	100	207.32	222.84	230.56	236.13
	200	364.22	374.97	387.47	413.56
	300	547.81	578.92	640.02	594.23
	400	621.69	581.37	674.90	701.22

As for a very deep analysis from different parameters of different metaheuristics, Figure 2 shows the energy cost in the BAN when the percentage of cluster head is 10% and the number of nodes is given from 100 to 400, respectively, based on the BIHABCA, SFLA, ACO, and SA. In the BIHABCA, an immune operator can accelerate the search of honey source location and overcome the shortcomings of random search of the ABCA. It even improves the evolution speed. The improved BIHABCA has a great improvement in search accuracy and stability. As for the SA, the energy cost of the clustering method can be seen from the simulation results. It reduces slowly with the increase of the generation times and falls easily into stagnation state in the later running period of the algorithm. The energy cost optimized by ACO is relatively stable, and the energy cost of BAN is difficult to be reduced. The pheromone volatilization coefficient is a significant factor that effects the energy cost of BAN. A pheromone volatilization coefficient will lead to the ACO falling into the local optimal solution, which affects the global search ability of the algorithm. The frogs in SFLA adopt different jumping way according to steps, and the number of iterations of the subgroup affects the local search ability of the SFLA and reduces the search performance. It can be seen from

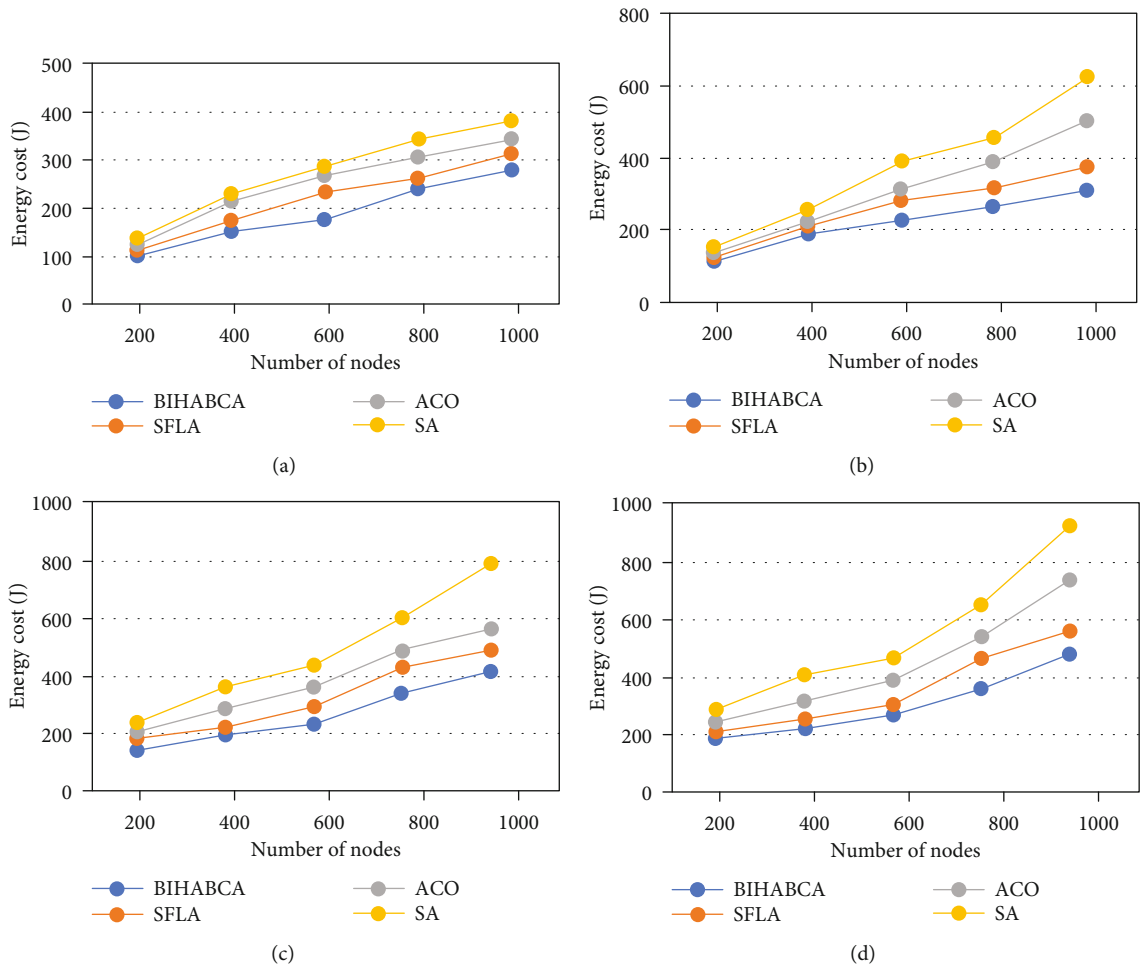


Figure 4: The changes in energy cost with different number of cluster head nodes: (a) 5% cluster heads; (b) 10% cluster heads; (c) 15% cluster heads; (d) 20% cluster heads.

the figure that under the condition of different numbers of nodes, the energy cost is raised because of the increased number of nodes.

In Figure 3, the convergence and energy cost of the BIHABCA, SFLA, ACO, and SA are given when the cluster heads are 20% in BANs. In the initial phase of iteration, the convergence speed of the BIHABCA is faster because of the immune and hybrid operator. In the later stage, the convergent speed is slow. However, the BIHABCA performs better in energy efficiency than the SFLA, ACO, and SA. Furthermore, the increase of nodes will lead to the increase of energy consumption. As the amount of cluster heads increases, the amount of transmission data increases. Therefore, the energy consumption of transmission data increases. In addition, the SFLA, ACO, and SA show a slower convergence rate than the BIHABCA. The BIHABCA combined with immune and hybrid operators has better performance with no premature convergence.

As shown in Figure 3, the cluster head nodes account for 20% of total sensor nodes in BANs. The figures show energy costs optimized by the BIHABCA, SFLA, ACO, and SA with different numbers of nodes in BANs. It can be seen from Figure 3 that with the increase of cluster heads, the increment of the communication energy cost optimized by the BIHABCA, SFLA, ACO, and SA is continuously increased. The main reason is that the transmission energy and reception energy of total nodes in the communication process increase exponentially with the increase of head sensors. Furthermore, the SFLA, ACO, and SA are easy to fall into premature convergence. The BIHABCA can accelerate the search of honey source location and overcome the shortcomings of random search of the ABCA. At the same time, the immune operator helps the population evolutes toward the optimal search space. It avoids the stagnation of evolution and improves the convergence speed of the algorithm. In conclusion, the performance of the BIHABCA is better than those of the other three algorithms.

In Table 6, the best results of the BIHABCA, SFLA, ACO, and SA after 100 iterations are shown. After 100 iterations, the results optimized by the four methods are described in Table 5. The BIHABCA indicates the ideal effectiveness to decrease the energy cost with 100 medical nodes. The SFLA, ACO, and SA provide suboptimal results with 200, 300, and 400 nodes, respectively.

Tables 7 and 8 list the percentage of energy cost reduction optimized by BIHABCA than those by the other

three algorithms with 10% cluster heads and 20% cluster heads, respectively. Experiment results show that the BIHABCA method can reduce the network energy consumption in BANs.

Table 9 shows the comparison of the runtime of the BIHABCA, SFLA, ACO, and SA after 100 iterations when the number of sensors increases and the percentage of cluster heads increases. Each runtime is given with 100 experimental tests. Table 9 shows that runtime is independent of the number of cluster heads. Furthermore, with the increase of sensor number, the computational complexity of algorithms raises, too. Therefore, the runtime of the algorithms also increases.

Figure 4 shows the energy cost optimized by the BIHABCA, SFLA, ACO, and SA, respectively, with different numbers of cluster heads. As shown in the following figure, with the increase of the amount of sensors in BANs, the demand for data transferred increases and the energy cost also improves correspondingly.

It can be concluded from Figure 4 that the clustering method based on BIHABCA requires less communication energy consumption of nodes in BANs, which can effectively improve the energy utilization efficiency.

6. Conclusion

Hence, a binary immune hybrid artificial bee colony algorithm (BIHABCA) in low energy clustering optimization in BANs in this paper is introduced to optimize the total energy cost. We first describe the clustering problem intimately and introduce the energy cost formulation. Then, immune and hybrid operators are designed into the ABCA to improve the effectiveness of the system. Extensive tests are carried out to validate the efficiency gain in terms of the energy efficiency when compared with the SFLA, ACO, and SA. After iterations, the energy cost optimized by the BIHABCA is 13.00%, 21.38%, and 27.38% less than those by the SFLA, ACO, and SA, respectively, with 10% cluster heads when the number of nodes in BAN is 100. The increment of the communication energy cost optimized by the BIHABCA, SFLA, ACO, and SA is continuously increased mainly because the transmission energy and reception energy of total nodes in the communication process raise with the increase of the cluster heads. Experiment results show that BIHABCA has minimized communication costs in BANs when compared with the SFLA, ACO, and SA approach with a large quantity of nodes in BANs.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper was funded by the Corps Innovative Talents Plan (grant number 2020CB001), the project of Youth and Middle-Aged Scientific and Technological Innovation Leading Talents Program of the Corps (grant number 2018CB006), the China Postdoctoral Science Foundation (grant number 220531, the Funding Project for High Level Talents Research in Shihezi University (grant number RCZK2018C38), the Project of Shihezi University (grant number ZZZC201915B), and the Postgraduate Education Innovation Program of the Autonomous Region.

References

[1] M. Ambigavathi and D. Sridharan, "A review of channel access techniques in wireless body area network," in *2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)*, Tindivanam, India, 2017.

[2] R. Punj and R. Kumar, "Technological aspects of WBANs for health monitoring: a comprehensive review," *Wireless Networks*, vol. 25, no. 3, pp. 1125–1157, 2019.

[3] R. Kaur, R. Pasricha, and B. Kaur, "A study of wireless body area networks and its routing protocols for healthcare environment," *Recent Advances in Electrical & Electronic Engineering*, vol. 13, no. 2, pp. 136–152, 2020.

[4] M. A. Panhwar, S. Jatoi, and K. A. Memon, "Wireless body area networks: architecture, standards, challenges, and applications," *International Journal of Computer Science and Network Security*, vol. 19, pp. 173–178, 2019.

[5] K. Hasan, K. Biswas, K. Ahmed, N. S. Nafi, and M. S. Islam, "A comprehensive review of wireless body area network," *Journal of Network and Computer Applications*, vol. 143, pp. 178–198, 2019.

[6] J. E. Bower, K. R. Kuhlman, M. D. Haydon, C. C. Boyle, and A. Radin, "Cultivating a healthy neuro-immune network: a health psychology approach," *Social and Personality Psychology Compass*, vol. 13, no. 9, article e12498, 2019.

[7] D. Vera, N. Costa, L. Roda-Sanchez, T. Olivares, A. Fernández-Caballero, and A. Pereira, "Body area networks in healthcare: a brief state of the art," *Applied Sciences-Basel*, vol. 9, no. 16, p. 3248, 2019.

[8] F. Wu, X. Li, A. K. Sangaiah et al., "A lightweight and robust two-factor authentication scheme for personalized healthcare systems using wireless medical sensor networks," *Future Generation Computer Systems*, vol. 82, pp. 727–737, 2018.

[9] R. Kadel, N. Islam, K. Ahmed, and S. J. Halder, "Opportunities and challenges for error correction scheme for wireless body area network-a survey," *Journal of Sensor and Actuator Networks*, vol. 8, no. 1, p. 1, 2019.

[10] T. Rashid, S. Kumar, A. Verma, P. R. Gautam, and A. Kumar, "Co-REERP: cooperative reliable and energy efficient routing protocol for intra body sensor network (Intra-WBSN)," *Wireless Personal Communications*, vol. 114, no. 2, pp. 927–948, 2020.

[11] M. Mohammadhosseini, S. Najafzadeh, and E. Mahdipour, "Reduce energy consumption in sensors using a smartphone, smartwatch, and the use of SFLA algorithms (REC-SSS)," *Journal of Supercomputing*, vol. 77, no. 1, pp. 909–935, 2021.

[12] D. Wohwe Sambo, B. O. Yenke, A. Förster, and P. Dayang, "Optimized clustering algorithms for large wireless sensor networks: a review," *Sensors*, vol. 19, no. 2, p. 322, 2019.

[13] R. Yarinezhad and S. N. Hashemi, "Exact and approximate algorithms for clustering problem in wireless sensor networks," *IET Communications*, vol. 14, no. 4, pp. 580–587, 2020.

[14] B. Jang, "An effective clustering protocol for wireless sensor networks," *Basic & Clinical Pharmacology & Toxicology*, vol. 125, pp. 199-200, 2019.

[15] D. Maheswari, S. Sudha, and M. Meenalochani, "Fuzzy based adaptive clustering to improve the lifetime of wireless sensor network," *China Communications*, vol. 16, no. 12, pp. 56–71, 2019.

[16] Y. U. Xiu-wu, Y. U. Hao, L. Yong, and X. Ren-rong, "A clustering routing algorithm based on wolf pack algorithm for heterogeneous wireless sensor networks," *Computer Networks*, vol. 167, article 106994, 2020.

[17] S. M. Amini, A. Karimi, and S. R. Shehnepoor, "Improving lifetime of wireless sensor network based on sinks mobility and clustering routing," *Wireless Personal Communications*, vol. 109, no. 3, pp. 2011–2024, 2019.

[18] M. Krishnan, S. W. Yun, and Y. M. Jung, "Dynamic clustering approach with ACO-based mobile sink for data collection in WSNs," *Wireless Networks*, vol. 25, no. 8, pp. 4859–4871, 2019.

[19] A. Ranganathan and B. Rangasamy, "Analysis of energy-efficient clustering algorithms for wireless sensor network (WSN)," *Journal of Testing and Evaluation*, vol. 47, no. 6, pp. 20180487–20183877, 2019.

[20] B. Ahmad, W. Jian, Z. A. Ali, S. Tanvir, and M. S. A. Khan, "Hybrid anomaly detection by using clustering for wireless sensor network," *Wireless Personal Communications*, vol. 106, no. 4, pp. 1841–1853, 2019.

[21] D. Adhikary and D. K. Mallick, "Energy-aware on-demand fuzzy-unequal clustering protocol for wireless sensor networks," *Journal of Engineering Science and Technology*, vol. 14, pp. 1200–1219, 2019.

[22] Z. Chen, H. Lin, L. Wang, and B. Zhao, "Interference-free clustering protocol for large-scale and dense wireless sensor networks," *KSII Transactions on Internet and Information Systems*, vol. 13, no. 3, pp. 1238–1259, 2019.

[23] S. Khriji, D. El Houssaini, I. Kammoun, K. Besbes, and O. Kanoun, "Energy-efficient routing algorithm based on localization and clustering techniques for agricultural applications," *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 3, pp. 56–66, 2019.

[24] D. Agrawal, M. H. Wasim Qureshi, P. Pincha et al., "GWO-C: grey wolf optimizer-based clustering scheme for WSNs," *International Journal of Communication Systems*, vol. 33, no. 8, article e4344, 2020.

[25] B. Mamalis and M. Perlitis, "Energy balanced two-level clustering for large-scale wireless sensor networks based on the gravitational search algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, pp. 32–42, 2019.

[26] J. K. C and R. Venkataraman, "EODC: an energy optimized dynamic clustering protocol for wireless sensor network using PSO approach," *International Journal Of Computers Communications & Control*, vol. 14, no. 2, pp. 183–198, 2019.

[27] P. S. Mann and S. Singh, "Improved artificial bee colony metaheuristic for energy-efficient clustering in wireless sensor networks," *Artificial Intelligence Review*, vol. 51, no. 3, pp. 329–354, 2019.

[28] K. Guleria and A. K. Verma, "Meta-heuristic ant colony optimization based unequal clustering for wireless sensor network," *Wireless Personal Communications*, vol. 105, no. 3, pp. 891–911, 2019.

[29] S. Anandamurugan and T. Abirami, "Antipredator adaptation shuffled frog leap algorithm to improve network life time in wireless sensor network," *Wireless Personal Communications*, vol. 94, no. 4, pp. 2031–2042, 2017.

[30] H. Wang, W. J. Wang, S. Y. Xiao, Z. Cui, M. Xu, and X. Zhou, "Improving artificial bee colony algorithm using a new neighborhood selection mechanism," *Information Sciences*, vol. 527, pp. 227–240, 2020.

[31] A. Saxena, S. Shekhawat, A. Sharma, H. Sharma, and R. Kumar, "Chaotic step length artificial bee colony algorithms for protein structure prediction," *Journal of Interdisciplinary Mathematics*, vol. 23, no. 2, pp. 617–629, 2020.

[32] Z. Muhammad, N. Saxena, I. M. Qureshi, and C. W. Ahn, "Hybrid artificial bee colony algorithm for an energy efficient internet of things based on wireless sensor network," *IETE Technical Review*, vol. 34, sup1, pp. 39–51, 2017.

[33] C. Ozturk, E. Hancer, and D. Karaboga, "A novel binary artificial bee colony algorithm based on genetic operators," *Information Sciences*, vol. 297, pp. 154–170, 2015.

[34] C. J. Santana Jr., M. Macedo, H. Siqueira, A. Gokhale, and C. J. A. Bastos-Filho, "A novel binary artificial bee colony algorithm," *Future Generation Computer Systems*, vol. 98, pp. 180–196, 2019.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5579252, 14 pages
https://doi.org/10.1155/2021/5579252

Research Article

A Novel QoS Routing Energy Consumption Optimization Method Based on Clone Adaptive Whale Optimization Algorithm in IWSNs

Jing Xiao, Yang Liu, Hu Qin, Chaoqun Li, and Jie Zhou

College of Information Science and Technology, Shihezi University, Shihezi 832000, China

Correspondence should be addressed to Jie Zhou; jiezhou@shzu.edu.cn

Received 26 February 2021; Revised 29 March 2021; Accepted 7 April 2021; Published 23 April 2021

Academic Editor: Bin Gao

Routing requests in industrial wireless sensor networks (IWSNs) are always restricted by QoS. Therefore, finding a high-quality routing path is a key problem. In this paper, a clone adaptive whale optimization algorithm (CAWOA) is designed for reducing the routing energy consumption of IWSNs with QoS constraints, and a novel clone operator is proposed. More importantly, CAWOA innovatively adopts a discrete binary-based routing coding method, which provides strong support for optimal routing schemes. In addition, a novel routing model of IWSNs combined with QoS constraints has been designed, which involves comprehensive consideration of bandwidth, delay, delay jitter, and packet loss rate. Subsequently, in a series of simulations, the proposed algorithm is compared with other heuristic-based routing algorithms, namely, whale optimization algorithm (WOA), simulated annealing (SA), particle swarm optimization (PSO), and genetic algorithm (GA). The simulation results suggest that the CAWOA-based routing algorithm outperforms other methods in terms of routing energy consumption, convergence speed, and optimization ability. Compared with GA, SA, PSO, and WOA under the conditions that the number of nodes is 120, the maximum delay is 120 ms, the maximum delay jitter is 25 ms, the maximum bandwidth is 9 Mbps, and the packet loss rate is 0.02; the energy consumption of CAWOA-based routing is reduced by 12%, 17%, 19%, and 7%, respectively.

1. Introduction

With the improvement in productivity and the popularization of industrial automation, industrial wireless sensor networks (IWSNs) have become an important tool for monitoring the production environment [1, 2]. In addition, IWSNs adopt the concept of the Industrial Internet of Things (IIOT); therefore, IWSNs have the characteristics of flexible, mobile, and large scale [3–5]. There is a large number of wireless sensor nodes in IWSNs, and these nodes are connected to the gateway through the Network Manager (NM), thereby transmitting information to the plant automation network. In IWSNs, there are requirements for low latency, high reliability, and real time. To meet these requirements, reasonable planning and careful design of IWSNs are necessary, and the characteristics of sensor devices and the properties of IWSNs make these tasks more complex and challenging.

To reduce the transmission distance of data and improve the quality of service (QoS) of IWSN applications, routing optimization is a commonly used method. Routing refers to the data transmission path from the source node to the destination node. Since there are usually many choices for the routing path, the optimal routing is a key issue in IWSNs. The so-called optimal routing represents the path with the least energy consumption of routing under the given evaluation criteria, generally QoS restrictions. Nowadays, there are many literatures on the reasonable planning of IWSN routing, and the effect of routing schemes obtained by different methods and different evaluation criteria is not the same [6].

Hizal and Zengin [7] proposed that it is necessary to do research on a novel QoS routing protocol, which not only considers the shortest path and the lowest cost but also studies the coverage, location, energy level, and mobility of the sensor. Therefore, routing algorithms in IWSNs also need to consider QoS constraints. More specifically, Indumathi

and Vaithianathan [8] started from the energy consumption and delay in QoS constraints and proposed a scheme that can reduce energy consumption and delay for multiconstrained QoS multicast routing. Vafaei et al. [9] proposed a QoS routing protocol that can increase the average data packet transmission rate in a vehicle ad hoc network. In addition, due to the rapid development of machine learning in recent years, Sun et al. [10] proposed a QoS routing path selection algorithm combined with machine learning, which can improve link congestion.

The selection of the optimal routing path under QoS constraints is an NP-hard problem [11]. In this case, heuristic swarm intelligence algorithms may be more applicable for solving the optimal routing problem in IWSNs [12]. Compared with conventional algorithms, the most significant advantage of swarm intelligence algorithms is that they are not limited to the complexity of the problem to be solved. They can obtain a set of potential optimal solutions and use their unique mechanisms to continuously optimize the solution during the iterative process. Therefore, in this paper, we propose a novel clone adaptive whale optimization algorithm (CAWOA), establish a novel routing model in IWSNs with QoS constraints, and prove the effectiveness of the proposed algorithm in reducing routing energy consumption through a series of simulation experiments under different conditions.

The main goal of this paper is to find a routing path with the lowest energy consumption in IWSNs under multiple QoS constraints. In general, the main contributions of this paper can be listed as follows:

(1) A novel clone adaptive whale optimization algorithm (CAWOA) is proposed, which combines the advantages of clonal expansion and adaptive operator

(2) CAWOA has made significant innovations to make it applicable for solving the discrete binary-based routing energy optimization problem in IWSNs with QoS constraints

(3) A new cloning operator is proposed, which can perform hierarchical cloning of the population, thus effectively avoiding the situation of local optimum

(4) A novel routing model of IWSNs that comprehensively considers network bandwidth, delay, delay jitter, and packet loss rate is established, and a fitness function for evaluating routing energy consumption is designed

(5) CAWOA is compared with the genetic algorithm, particle swarm optimization, simulated annealing, and whale optimization algorithm in routing energy consumption, convergence speed, and optimization ability

The structure of this paper is organized as follows. The related works are discussed in Section 2. Then, Section 3 shows the IWSN QoS routing model and the fitness function. In Section 4, the process of CAWOA is introduced. Subsequently, there are simulation experiments and discussions of the results in Section 5. Finally, the conclusion is given in Section 6.

2. Related Work

In recent years, the problem of wireless sensor QoS routing has attracted more and more people's attention. In different application scenarios, the QoS constraints that need to be considered are not exactly the same, so the routing energy consumption generated is also different.

Nayyar and Singh [13] conducted a comprehensive review of 31 WSN simulators for the convenience of researchers in WSN simulation. Zhang et al. [14] proposed a routing algorithm combining quantum genetics and heuristic Q learning strategy to adapt to rapid changes in the network structure. In addition, to further reduce the routing energy consumption, some energy-saving routing protocols have been designed. Mostafaei et al. [15] modeled the QoS routing problem as a multiconstrained optimal path problem and proposed an algorithm based on distributed learning automata to save it. Zhang et al. [16] proposed a clustering method based on compressed sensing based on the clustering structure of wireless sensor networks, which can reduce the total energy consumption of the network within and between clusters. Mostafaei and Obaidat [17] proposed an algorithm based on irregular cell learning automata, which can reduce energy consumption while ensuring the safety of WSN. However, in IWSNs with QoS constraints, it is also necessary to consider factors such as delay and delay jitter to meet actual production requirements.

Many researchers use heuristic algorithms to solve the optimal routing problem. Varshney et al. [18] proposed a lightning-based lion optimization routing algorithm for minimizing the energy consumption in IWSNs. The algorithm determines the most suitable sensor placement method by considering the throughput, lifetime, delay, and coverage area of the sensor. In [19], Xu et al. proposed a routing protocol based on the genetic algorithm. They observe that the minimal sensor node coverage set can be found by avoiding redundant coverage of sensor nodes, thereby reducing the energy consumption of wireless sensor networks. Kirsan et al. proposed a multihop LEACH routing protocol based on simulated annealing to reduce network energy consumption [20]. However, these methods are neglected to consider the convergence speed of the algorithm when optimizing the energy consumption of sensor routing. Thus, some improved algorithms with faster convergence speed have been proposed. Kavitha and Velusamy [21] proposed a routing algorithm that combines genetic algorithm and simulated annealing. Bilandi et al. [22] proposed a hybrid routing algorithm based on simulated annealing and particle swarm optimization (PSO) to optimize energy consumption. Similar to [21, 22], Mohanakrishnan and Ramakrishnan [23] designed a routing protocol combining the genetic algorithm and whale optimization algorithm (WOA). However, these methods have the possibility of falling into a local optimum. Nayyar and Singh [24] proposed an energy-saving routing protocol that optimizes the real-time performance of WSN. The protocol is based on ant colony optimization and can provide the best solution in terms of throughput and packet delivery. Duan et al. [25] proposed a hybrid IWSN solution based on a task-oriented model and used heuristic modeling

methods to design collaborative routing algorithms, thereby reducing the energy consumption of wireless nodes and data communication delays. Jin et al. [26] proposed a convergent broadcast scheduling algorithm for IWSNs with multiple radio interfaces and, based on this algorithm, proposed a fast heuristic algorithm to minimize routing under the time constraints of industrial production. To solve the load balancing and fault tolerance in QoS routing, Moussa and El Alaoui [27] proposed a routing protocol based on ant colony optimization and unequal clustering, which improves the convergence speed of the algorithm and avoids local optimization. However, the parameters used in these routing algorithms are all fixed and lack the ability to dynamically adjust the algorithm parameters along with the running process.

With the purpose of minimizing the energy consumption of IWSN routing under QoS constraints, we propose a novel IWSN routing model. The model takes into account the impact of QoS constraints that are very important in IWSNs on routing energy consumption, including delay, bandwidth, delay jitter, and packet loss rate. In addition, a novel clone adaptive whale optimization algorithm (CAWOA) is proposed to solve the routing path with the lowest energy consumption. Compared with other algorithms for routing optimization, CAWOA has faster convergence speed, higher solution quality, and stronger ability to jump out of the local optimum. Furthermore, the important parameters in CAWOA can be dynamically adjusted along with the running process of the algorithm, thereby enhancing its ability to search for the solution space.

3. System Model

3.1. Problem Description. IWSNs include sensor nodes, sink nodes, gateway nodes, and base stations. Normally, the sensor nodes send data to the sink nodes, the sink nodes receive the data and transmit it to the gateway nodes after preliminary processing, and then, the gateway nodes transmit the data to the base station so that the technical staff can plan the next step according to the data content. Since the principle of routing from sensor nodes to sink nodes is the same as that of routing from sink nodes to a gateway node, the optimal routing of IWSNs can be defined as the path with the lowest routing energy consumption under QoS constraints. An industrial wireless sensor network with 6 sensor nodes, 8 sink nodes, and one gateway node is shown in Figure 1.

In IWSNs, the main energy consumption of sensor nodes comes from data transmission, which is the so-called routing energy consumption. Routing refers to the data transmission path from the specified source node to the destination node. As shown in Figure 2, there are many paths from node A to node D, such as $\{A \to E \to D\}$, $\{A \to D\}$, and $\{A \to C \to D\}$. In the routing energy consumption optimization problem of IWSNs with QoS constraints, it is not that the less the number of nodes passing by, the lower the energy consumption of the routing, but the overall energy consumption of a route is obtained through a combination of multiple factors.

Subsequently, to solve the routing energy consumption optimization problem of IWSNs with QoS constraints, the mathematical model is given in Sections 3.2 and 3.3.

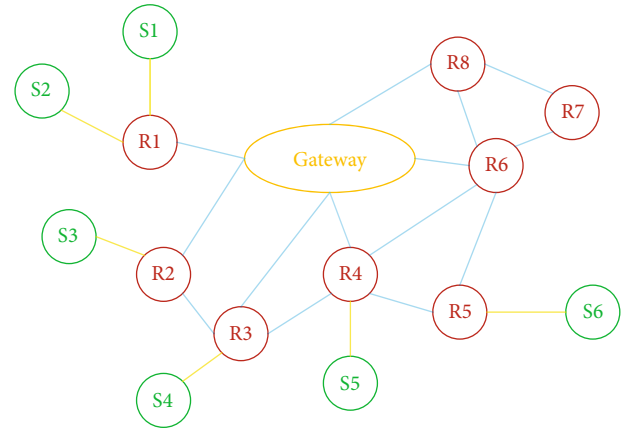


Figure 1: Composition of an industrial wireless sensor network.

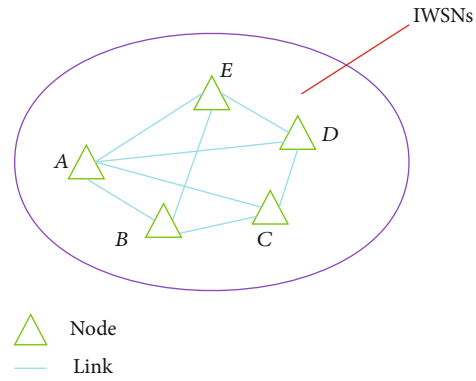


Figure 2: Route map in IWSNs.

3.2. Sensing Model. In the IWSN sensing model, since the sensing capabilities of sensors are limited, nodes can only transmit data to other nodes within the sensing range. The specific perception method is shown in

$$s_{a,b} \begin{cases} 1, & \text{if node } b \text{ is within the sensing range of node } a, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $S_{a,b}$ is the perception ability of node a to node b. If $S_{a,b} = 1$, it means that a can send data to b; otherwise, it cannot.

3.3. QoS Routing Model. The IWSN routing model with QoS constraints can be represented by an undirected weight graph $G = <V, E>$, where E is the collections of node links in the undirected weight graph and V represents the set of all nodes. Each link in G represents a direct path between two adjacent nodes. QoS routing is defined as finding an optimal path that satisfies QoS constraints and minimization of energy consumption. In actual factory production, five evaluation parameters are usually used as QoS constraints in IWSNs. They are delay, delay jitter, bandwidth, cost, and packet loss rate.

Table 1: QoS function definition.

Function name	Abbreviation	Definition
Delay	DL	Average duration of data transmission
Delay jitter	DLJ	Fluctuations of the transmission time
Bandwidth	BW	The amount of data per unit time
Packet loss rate	PLR	Data loss during transmission
Cost	C	The overhead of data transmission

Delay represents the average time it takes for data packets to be transmitted in IWSNs. Delay jitter denotes the fluctuation of data packet transmission time. Bandwidth refers to the amount of data that the sensor can transmit per unit time. Packet loss rate is the loss or damage of data packets during transmission. The above factors will affect the routing transmission quality of IWSNs. What is more, the specific QoS function definition is shown in Table 1.

Given a source node i and destination node h in the IWSNs, a path P that satisfies equations (2), (3), (4), and (5) and has the lowest transmission cost is a QoS routing request that can be accepted:

$$\mathrm{BW}(i) \geq B_{\mathrm{lim}}, \tag{2}$$

$$\sum_{i \in EP} \mathrm{DL}(i) + \sum_{h \in VP} \mathrm{DL}(h) \leq \mathrm{DL}_{\mathrm{lim}}, \tag{3}$$

$$\sum_{i \in EP} \mathrm{DLJ}(i) + \sum_{h \in VP} \mathrm{DLJ}(h) \leq \mathrm{DLJ}_{\mathrm{lim}}, \tag{4}$$

$$\mathrm{PLR}(h) \leq \mathrm{PLR}_{\mathrm{lim}}, \tag{5}$$

where B_{lim}, $\mathrm{DL}_{\mathrm{lim}}$, $\mathrm{DLJ}_{\mathrm{lim}}$, and $\mathrm{PLR}_{\mathrm{lim}}$, respectively, represent the constraint values for bandwidth, delay, delay jitter, and packet loss rate in QoS routing. EP is the collection of links on the routing request P, and VP is the set of nodes on the route request P. In the IWSN QoS routing model, we first define the total number of sensor nodes in G as $n = |V|$, the source node s belongs to V, and the destination node $d \in |V - |s||$. Then, with the purpose of measuring the advantage of a route, we have defined 5 measurement functions in Table 1, which are delay function, delay jitter function, bandwidth function, packet loss rate function, and cost function. These five functions are represented by equations (6), (7), (8), (9), and (10), respectively,

$$\mathrm{DL}(r(s,d)) = \sum_{e \in r(s,d)} \mathrm{DL}(e) + \sum_{n \in r(s,d)} \mathrm{DL}(n), \tag{6}$$

$$\mathrm{BW}(r(s,d)) = \min\{\mathrm{BW}(e)\}, \quad e \in r(s,d), \tag{7}$$

$$\mathrm{DLJ}(r(s,d)) = \sum_{e \in r(s,d)} \mathrm{DLJ}(e) + \sum_{n \in r(s,d)} \mathrm{DLJ}(n), \tag{8}$$

$$\mathrm{PLR}(r(s,d)) = 1 - \prod_{n \in r(s,d)} (1 - \mathrm{PLR}(n)), \tag{9}$$

$$C(r(s,d)) = \sum_{e \in r(s,d)} C(e), \tag{10}$$

where $r(s,d)$ represents all paths between the node s and the node d that meet the QoS constraints. Then, the routing energy consumption optimization problem of IWSNs is to find the path with the least energy consumption that satisfies equations (2), (3), (4), and (5) at the same time.

Specifically, $C(e)$ can be represented by

$$C(e) = C_a + C_b. \tag{11}$$

In (11), $C(e)$ is the total energy consumption between two adjacent nodes, which is composed of C_a and C_b. C_a represents the energy consumption of data transmission, and C_b denotes the energy consumption of receiving information between two nodes.

Assuming that the distance between two nodes is dist and the amount of information transmitted is x bits, the energy consumption C_a of the transmitted information can be expressed as

$$C_a(x, d) = E_e \cdot x + \eta_{\mathrm{amp}} \cdot x \cdot \mathrm{dist}^3. \tag{12}$$

In (12), E_e is the energy parameter. The power amplification parameter η_{amp} used for multipath fading determines the energy of the amplifier. dist and x are the distance between two nodes and the number of bits, respectively. In addition, the power consumption C_b for receiving information is shown as

$$C_b(x) = E_e \cdot x. \tag{13}$$

With the purpose of evaluating the energy consumption of routing, a fitness function is designed as shown in

$$\mathrm{fitness} = \min\left\{\frac{C(A) + \mathrm{DL}(A) * 1 + \mathrm{DLJ}(A) + \mathrm{PLR}(A) * \mathrm{PLC}}{r * \mathrm{BW}(A)}\right\}. \tag{14}$$

In (14), $A = r(v_s, v_d)$ represents all routing paths that meet the QoS constraints from the node s to the node d in IWSNs. $C(A)$ is the energy consumption between two nodes, $\mathrm{DL}(A)$ is the delay between two nodes, $\mathrm{DLJ}(A)$ is the delay jitter, $\mathrm{PLR}(A)$ is the packet loss rate, PLC is the cost of packet loss, and $\mathrm{BW}(A)$ is the network bandwidth. r is the bandwidth factor. However, if a route fails to meet QoS constraints, which include delay, delay jitter, bandwidth, packet loss rate, and cost, then the route request will be discarded.

The design reason for the fitness function (14) is that an optimal route should consider not only the energy

consumption of data transmission but also other factors that affect the overall routing. These factors include data transmission delay, delay jitter, packet loss rate, and bandwidth. The increase in delay will cause the degradation of routing quality, and the delay is not fixed, so the delay jitter must be considered. In addition, there is no guarantee that every transmission is successful during data transmission; therefore, the packet loss rate is used to deal with this situation. After the loss of packet, there is a corresponding packet loss cost, and PLC is used to indicate the packet loss cost. It also considers the bandwidth of data transmission. Usually, the routing performance is proportional to the bandwidth. Since the bandwidth will fluctuate due to the influence of the actual environment, the bandwidth factor r is introduced for regulation. In general, a qualified route in IWSNs has the necessity of taking QoS factors into consideration.

4. CAWOA-Based Routing Algorithm for Minimizing Energy Consumption in IWSNs with QoS Constraints

To optimize the routing energy consumption of IWSNs with QoS constraints, a novel clone adaptive whale optimization algorithm (CAWOA) is proposed. The whale optimization algorithm (WOA) has low computational complexity. In the early stage of the algorithm, WOA performs a global search, while in the later stage of the algorithm, it performs a local search, which can effectively obtain the routing path that meets the QoS constraints. Compared with other heuristic algorithms, WOA's local search ability is stronger. Its major disadvantage is that it is easy to fall into the local optimum. However, the addition of the clone operator can effectively avoid the emergence of local optimal conditions. Furthermore, WOA has a faster convergence speed; this advantage can make it have higher practicability. As a result, CAWOA is inspired by the traditional WOA but has made significant improvements in convergence speed and optimization capabilities. By using CAWOA, the optimal routing path with the least energy consumption can be found; therefore, the network lifetime of IWSNs can be effectively extended for saving factory costs. In addition, different from the traditional WOA, the significant improvement of the CAWOA is the addition of the clone operator and the adaptive operator.

The process of CAWOA includes population coding and initialization, calculating fitness and finding the leading whale, adaptive encircling predation, bubble-net attacking, random search for prey, cloning operation, and termination operation.

4.1. Population Coding and Initialization. The first step of applying CAWOA to the routing energy consumption optimization problem of IWSNs is to determine the encoding method. It is difficult to achieve the expected goal using conventional decimal coding in the routing problem, because the data does not need to pass through all nodes during the routing process, and the optimal routing is the path that has minimal energy consumption under the QoS constraints. Therefore, binary encoding is a desirable coding method, which has the characteristics of simple and easy encoding and decoding. Under the problem of binary encoding, 1 means passing through the node, and 0 means not passing through the node. Assuming there are 5 sensor nodes in IWSNs, the binary code of the individual whale can be expressed as

$$\text{whale} = [1, 0, 1, 1, 1]. \tag{15}$$

In (15), since the source node and the destination node must be passed, the first and last bits are both 1. In addition, the third and fourth bits are also 1, indicating that the nodes through which the data passes during this routing process are the first, third, fourth, and fifth nodes. However, equation (15) cannot indicate the order of access between nodes. For example, we do not know whether to pass through the third node or the fourth node first, so the application of equation (15) needs to be combined with

$$\text{order} = \text{randperm}(N), \tag{16}$$

where N is the quantity of sensors and $\text{randperm}(N)$ is a function to scramble the number. If $N = 5$, $\text{randperm}(N)$ can be $\{3, 4, 2, 5, 1\}$. Since the starting point number is always 1, the destination point number is always N, so the logical routing order is $\{1, 3, 4, 2, 5\}$. Combining equation (15), it can be concluded that the actual route is $\{1, 4, 2, 5\}$, so that the optimal solution search for the problem can be realized in the entire solution space.

After determining the coding method of the individual whale, the encoding of the whale population can be expressed as

$$\text{pop} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,M-1} & w_{1,M} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,M-1} & w_{2,M} \\ \vdots & & w_{n,m} & & \vdots \\ w_{N-1,1} & w_{N-1,2} & \cdots & w_{N-1,M-1} & w_{N-1,M} \\ w_{N,1} & w_{N,2} & \cdots & w_{N,M-1} & w_{N,M} \end{bmatrix} (w_{n,m} \in \{0, 1\}). \tag{17}$$

In (17), N is the number of whales and M is the number of sensors. In the initialization process, the first column and the last column are all set to 1, and the remaining positions are initialized randomly.

4.2. Calculating Fitness and Finding the Leading Whale. In the routing energy consumption optimization problem with QoS constraints, the fitness value of the whale represents the energy consumption of a route. Before other operations of CAWOA, it is necessary to calculate the fitness of each individual to find the position of the leading whale. The fitness value of each individual is obtained according to formula (14). Different from traditional WOA, CAWOA uses binary coding when encoding the population, so there is no need to judge whether the position is out of bounds, which reduces the computational complexity of the algorithm to a certain extent. What is more, the whale with the lowest

fitness is the leading whale, and its position will affect the activities of other whales.

4.3. Adaptive Encircling Predation. In CAWOA, the predation behavior of whales symbolizes the process of finding the optimal solution in the QoS routing energy consumption optimization problem. In the problem, the individual whale finds the position of the prey first and then surrounds the prey. The prey can be regarded as the leading whale, which means that other whales update their positions toward the position of the leader whale for carrying out the predation operation. Therefore, the first step in encircling predation is to calculate the distance between the individual whale and the leading whale, which is derived from

$$D = |C \cdot W^*(\text{gen}) - W(\text{gen})|, \tag{18}$$

where gen is the current iteration, $W^*(\text{gen})$ represents the position of the leading whale in the gen_{th} generation, and $W^*(\text{gen})$ should be updated during each iteration if a better solution appears. $W(\text{gen})$ denotes the position of the individual whale in the gen_{th} generation, and the constant C is the coefficient vector, which is calculated adaptively by

$$C = \begin{cases} C_1 - \dfrac{(C_1 - C_2)\left(f_c - f_{\text{avg}}\right)}{f_{\max} - f_{\text{avg}}}, & f_c \geq f_{\text{avg}}, \\ C_1, & f_c < f_{\text{avg}}, \end{cases} \tag{19}$$

where f_c is the fitness value of the current whale, f_{avg} is the average fitness value of the population, $f_{\max}$ is the fitness value of the leading whale, and C_1 and C_2 are two constants.

Then, the position of the leading whale affects the update of the position of the individual whale, and its formula is shown in

$$W(\text{gen} + 1) = W(\text{gen}) - A \cdot D, \tag{20}$$

where A is another coefficient vector, which is calculated by

$$A = \begin{cases} A_1 - \dfrac{(A_1 - A_2)\left(f_c - f_{\text{avg}}\right)}{f_{\max} - f_{\text{avg}}}, & f_c \geq f_{\text{avg}}, \\ A_1, & f_c < f_{\text{avg}}, \end{cases} \tag{21}$$

where f_c, $f_{\max}$, and f_{avg} are the same as those in equation (19) and A_1 and A_2 are two constants.

The addition of adaptive operators allows CAWOA to dynamically adjust the parameters according to the fitness value when the whale is preying, which speeds up the convergence speed of the algorithm.

4.4. Bubble-Net Attacking. In the problem of IWSN routing energy consumption optimization under QoS constraints, the bubble-net attacking behavior of whales helps to find a better solution. There are two strategies for simulating the bubble-net attacking: one is the shrinking encircling mechanism, and the other is the spiral updating position.

In CAWOA, the shrinking encircling means that the position update of the whale is performed according to equation (20). The coefficient vector A can be adaptively adjusted according to the fitness of the population. If the value of A is between $[-1, 1]$, the updated position of the current whale can be any value between its own position and the position of the leading whale.

The spiral updating position means that whales swim to the surface with a spiral posture and spit out varying size bubbles for preying on shrimp and fish. In this stage, the distance between the whale and the leading whale is first calculated, and the calculation formula is shown in

$$D' = |W^*(\text{gen}) - W(\text{gen})|. \tag{22}$$

Then, the individual whale updates its position, as shown in

$$W(\text{gen} + 1) = D' \cdot e^{bl} \cdot \cos(2\pi l) + W^*(\text{gen}), \tag{23}$$

where b is the spiral constant, l is a random number in $[-1, 1]$, and $W^*(\text{gen})$ denotes the position of the leading whale in the gen_{th} generation. Assuming that the probability of performing the two bubble-net attacking behaviors of the whale during the predation process is 0.5, the update formula of the whale position can be summarized as

$$W(\text{gen} + 1) = \begin{cases} W^*(gen) - A \cdot D, & r < 0.5, \\ D' \cdot e^{bl} \cdot \cos(2\pi l) + W^*(gen), & r \geq 0.5, \end{cases} \tag{24}$$

where r is a random number and $r \in [0, 1]$.

4.5. Random Search for Prey. With the purpose of avoiding falling into a local optimum in solving the IWSN routing energy optimization problem under QoS constraints, the position of the whale in CAWOA cannot be updated only by the position of the leading whale, and sometimes, it must be updated with the position of the partner. Specifically, CAWOA at this stage is to conduct a random search for prey and obtain the next position of the whale, and this operation is carried out under the influence of the coefficient vector A. If $|A| > 1$, CAWOA conducts the random search behavior, thereby increasing the global search ability of the algorithm, as shown in

$$D = |C \cdot W_r - W(\text{gen})|, \tag{25}$$

$$W(\text{gen} + 1) = W_r - A \cdot D, \tag{26}$$

where W_r can be a random position in the whale population.

4.6. Cloning Operation. In CAWOA, the cloning operation is inspired by the concept of cloning in biology, and its purpose is to improve the local search capability and convergence speed of the algorithm. Normally, the optimal solution of the IWSN routing problem with QoS constraints is related

```
Input: The population sorted according to the energy consumption value from small to large is set to pop1, the number of individuals is pop_num,
Output: The cloned population is set to pop2
First perform multi-level cloning operations
for p ∈ [1, pop_num]
    if p<pop_num×0.2
        Assign the individual of pop1(1) to pop2(p)
    else if p <pop_num×0.5
        Assign the individual of pop1(2) to pop2(p)
    else if p <pop_num×0.7
        Assign the individual of pop1(3) to pop2(p)
    else
        Assign the individual of pop1(4) to pop2(p)
    end if
end for
Then perform high-probability mutation operations
for p ∈ [1, pop_num]
    r is a random number and r∈ [0,1]
    if r<0.3
        Perform mutation operation on pop2(p)
    end if
end for
Return pop_2
```

Algorithm 1: Clonal expansion and high-probability mutation.

to the optimal individual in the current iteration process; however, traditional WOA performs many unnecessary operations. Therefore, the cloning operation can effectively improve the performance of the algorithm by expanding the number of individuals with high fitness. The cloning operation in CAWOA is divided into two parts: one is clonal expansion, and the other is high-probability mutation. The purpose of high-probability mutation operation is to reduce the possibility of falling into a local optimum, and for the same purpose, clonal expansion generally uses the way of multilevel cloning. The pseudocode of the cloning operation is shown in Algorithm 1.

4.7. Termination Operation. If CAWOA reaches the specified number of iterations, the algorithm stops looping and outputs the result; otherwise, it returns to Section 4.2.

4.8. Steps of CAWOA. The specific steps can be expressed as follows.

Step 1. Determine the number of sensors and the population size, randomly generate the location of sensor nodes in the monitoring area, and randomly generate whale locations. Initialize the parameters of the CAWOA and set the initial iteration gen = 1.

Step 2. Calculate the fitness of individual whales according to equation (14) and compare them, select the leading whale, and define it as W^*.

Step 3. Entering the main loop of the algorithm, if $r < 0.5$ and $|A| < 1$, then the whale individual updates its position according to $W(\text{gen}+1) = W(\text{gen}) - A \cdot D$; otherwise, it updates its position with $W(\text{gen}+1) = W_r - A \cdot D$. If $r > 0.5$, the individual whale updates its position according to $W(\text{gen}+1) = D' \cdot e^{bl} \cdot \cos(2\pi l) + W^*(\text{gen})$.

Step 4. Update coefficient vector C and A according to equations (19) and (21).

Step 5. Calculate and sort the fitness value of the whale population according to equation (14), perform cloning operation according to Algorithm 1, and then perform mutation operations on the cloned population with high probability.

Step 6. Calculate the fitness value according to equation (14), find the globally optimal individual and set it as W^*, and set the cloned population as the initial population for the next iteration.

Step 7. gen = gen + 1, if the maximum number of iterations of the algorithm is reached, the algorithm ends; otherwise, go to Step 3 to continue the iteration.

Step 8. Output the optimal solution W^*.

The flow chart of CAWOA can be expressed in Figure 3.

5. Results and Discussion

We made the following assumptions in the simulation:

(1) Delay, bandwidth, delay jitter, and packet loss exist in the link and affect routing

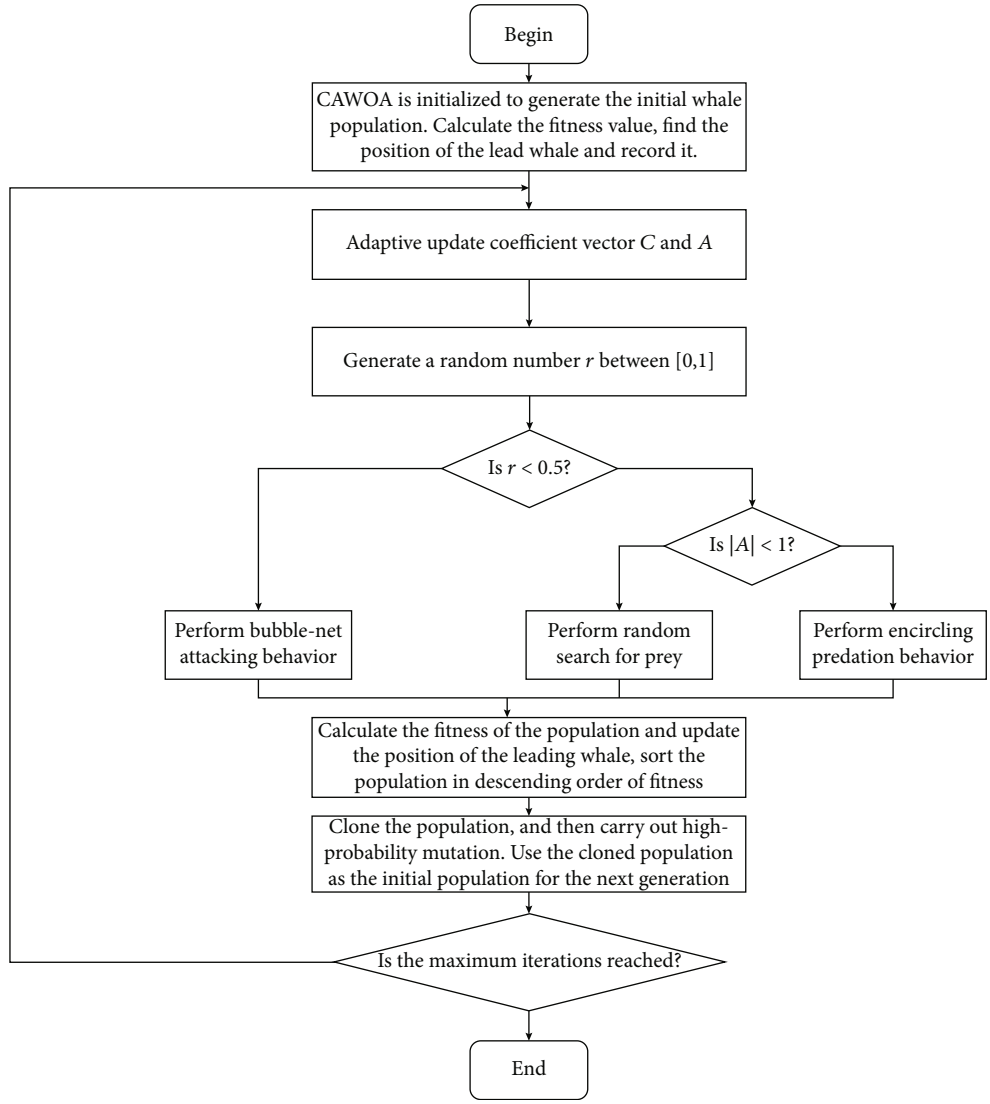


Figure 3: Steps of CAWOA.

(2) The node is stimulated by physical signals before sending data

(3) There are sending energy and receiving energy. The sending node generates sending energy by sending data, and the receiving node generates receiving energy by receiving data

To prove the effectiveness of the proposed algorithm in reducing the energy consumption of IWSN routing with QoS constraints, a series of simulations is carried out, and CAWOA is compared with the whale optimization algorithm, particle swarm optimization, genetic algorithm, and simulated annealing. The simulation experiment includes the trend of the algorithm for reducing the routing energy consumption, the speed of algorithm convergence, the percentage of energy consumption obtained after optimization, and the trend graph of energy consumption in large-scale IWSNs. Different types of simulations are carried out under different numbers of sensors, which can better reflect the practicality of the algorithm. In addition, all simulations are performed on a computer equipped with R7 4800H 2.9 GHz CPU, and the fitness function used in the algorithms is according to formula (14).

For the IWSN routing energy optimization problem with QoS constraints, the unified definition of public parameters helps to compare algorithms in a relatively fair situation. Therefore, the population size of the algorithms is set to 40, the number of iterations is set to 100, the sensors in IWSNs are distributed in a square area with a side length of 400, and the sensor coordinates are generated randomly. Specifically, CAWOA uses multilevel cloning, with the mutation probability set to 0.2, C_1 is 2, C_2 is 0.5, $A_1 = 2$, and $A_2 = -2$. The genetic algorithm uses a two-point crossover method with a crossover probability of 0.7, and the probability of mutation is set to 0.05. The initial temperature of simulated annealing is set to 200, and an exponential annealing method with an annealing coefficient of 0.96 is adopted. The individual learning factor and social learning factor of the particle

Table 2: Simulation parameters in Figure 4.

	Sensors	Delay (max)	Delay jitter (max)	Bandwidth (max)	Packet loss rate (max)
Figure 4(a)	30	50 ms	10 ms	6 Mbps	0.01
Figure 4(b)	40	60 ms	15 ms	6 Mbps	0.01
Figure 4(c)	50	70 ms	20 ms	7 Mbps	0.01
Figure 4(d)	60	80 ms	25 ms	7 Mbps	0.01

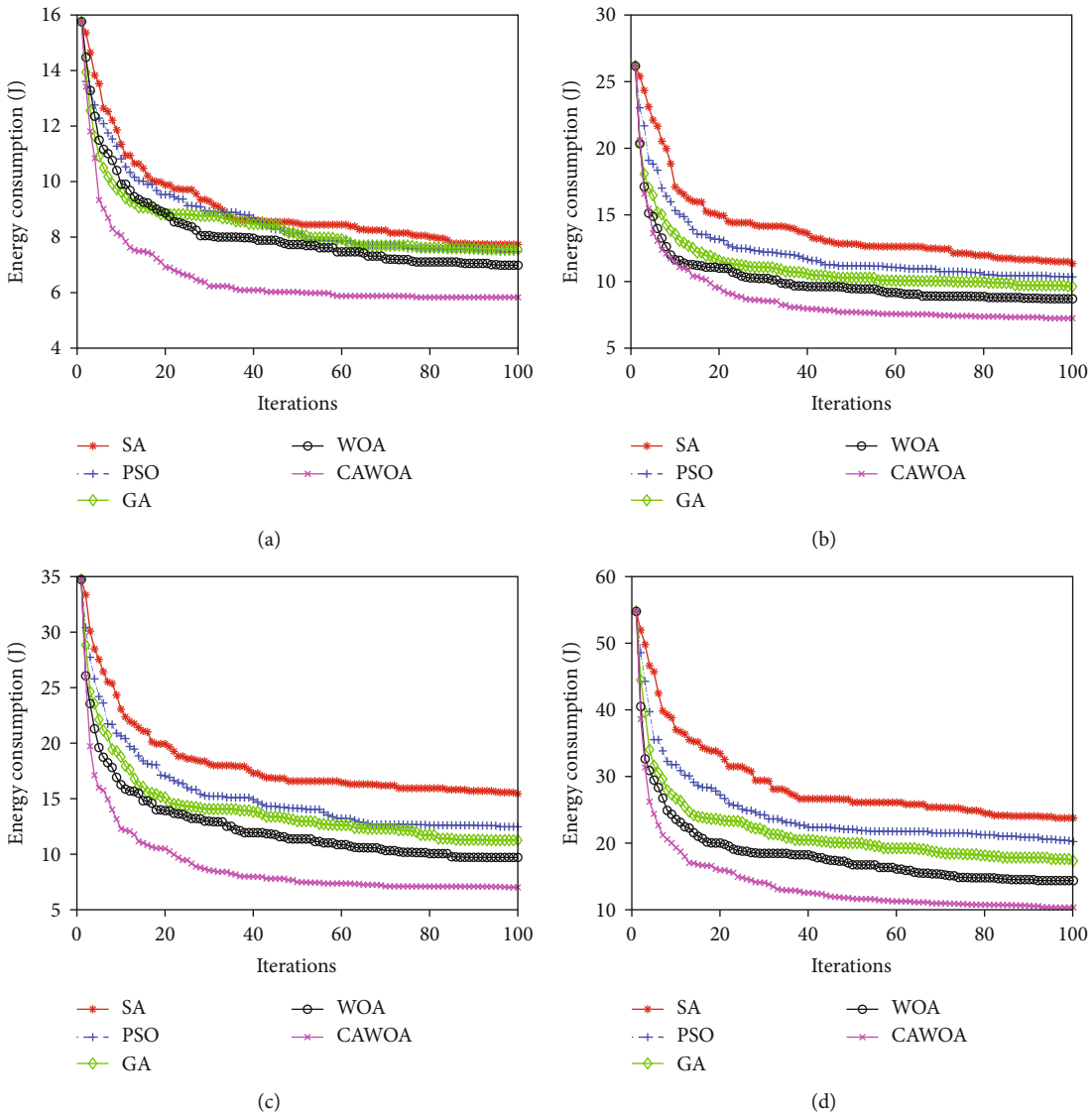


Figure 4: Energy consumption comparing five algorithms: (a) 30 sensors; (b) 40 sensors; (c) 50 sensors; (d) 60 sensors.

swarm optimization algorithm are both 2, the weighting factor ω_{max} is 0.9, and ω_{min} is 0.4. In the whale optimization algorithm, if the position is positive, set it to 1; otherwise, it is 0. Table 2 shows the simulation parameters in Figure 4.

In Figures 4(a)–4(d), the graphs of the optimization of routing energy consumption by five algorithms are shown. It can be found in Figure 4(a) that the convergence speeds of GA, SA, and PSO are slow, while CAWOA and WOA converge faster than them. However, WOA is stuck in premature convergence, and the routing energy consumption 6.98 J obtained by WOA is not the optimal solution. In contrast, CAWOA obtains the optimal routing solution with an energy consumption of 5.83 J while maintaining a faster convergence speed. In Figure 4(b), CAWOA has achieved lower routing energy consumption than other algorithms at the beginning of the iteration, and this trend has been maintained until the termination of the algorithm. In Figure 4(c), the performance shown by SA is the most unsatisfactory,

Table 3: Simulation parameters in Figure 5.

	Sensors	Delay (max)	Delay jitter (max)	Bandwidth (max)	Packet loss rate (max)
Figure 5(a)	40	80 ms	10 ms	6 Mbps	0.01
Figure 5(b)	60	90 ms	20 ms	7 Mbps	0.01
Figure 5(c)	80	100 ms	30 ms	8 Mbps	0.01
Figure 5(d)	100	110 ms	40 ms	9 Mbps	0.01

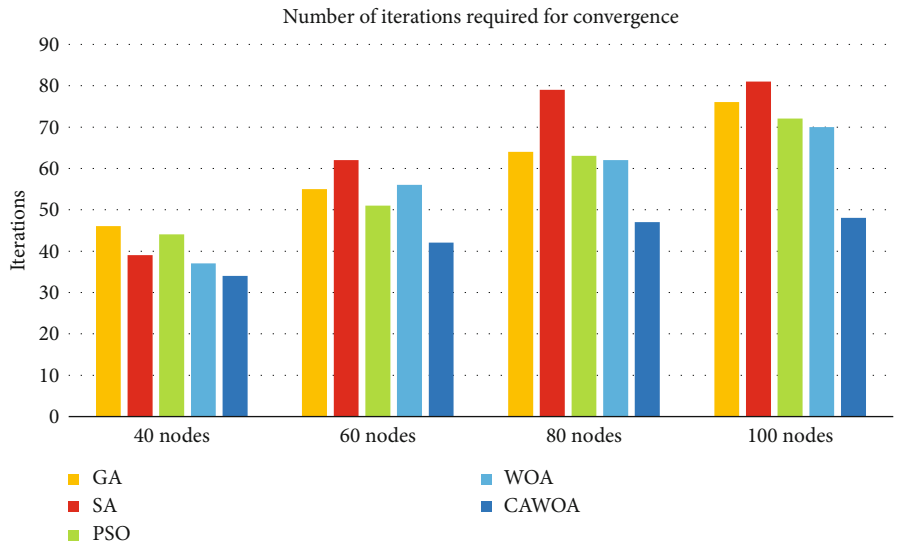


Figure 5: Convergence iteration comparison.

Table 4: Simulation parameters in Figure 6.

	Sensors	Delay (max)	Delay jitter (max)	Bandwidth (max)	Packet loss rate (max)
Figure 6(a)	60	60 ms	10 ms	6 Mbps	0.01
Figure 6(b)	80	80 ms	15 ms	7 Mbps	0.01
Figure 6(c)	100	100 ms	20 ms	8 Mbps	0.02
Figure 6(d)	120	120 ms	25 ms	9 Mbps	0.02

followed by PSO, GA, and WOA which have roughly the same performance when the number of iterations reaches 70. The 7.04 J obtained through CAWOA optimization is the optimal value among the five algorithms. What is more, in Figure 4(d) with 60 sensor nodes, the energy consumption value obtained by the CAWOA-based routing algorithm is 10.39 J, while the results obtained by GA, SA, PSO, and WOA are 17.39 J, 23.82 J, 20.28 J, and 14.42 J, respectively; the percentages of CAWOA's solution better than GA, SA, PSO, and WOA are 40.25%, 56.39%, 48.77%, and 27.98%, respectively. Table 3 shows the simulation parameters in Figure 5.

Figure 5 shows the changes in convergence iterations required by the five algorithms as the number of sensor nodes in IWSNs increases. In the case of 40 sensor nodes, the convergence iterations required by the routing algorithm based on GA, SA, PSO, and WOA are 46, 39, 44, and 37, respectively, while the CAWOA-based routing algorithm only needs 34 iterations, so convergence speed of CAWOA is faster than that of GA, SA, PSO, and WOA. When the number of sensor nodes increases to 60, 80, and 100, the iterations required for the CAWOA-based routing algorithm to achieve convergence are 42, 47, and 48, respectively. It can be seen from Figure 5 that the convergence iterations of CAWOA are less than that of GA, SA, PSO, and WOA, which can prove that CAWOA has a good convergence performance. Table 4 shows the simulation parameters in Figure 6.

In Figures 6(a)–6(d), in order to prove the quality of the solution obtained by the CAWOA-based routing algorithm in IWSNs with QoS constraints, pie charts of the energy consumption results obtained by the five algorithms are shown. Figure 6(a) suggests that the energy consumption of routing based on CAWOA accounts for 12% and is the lowest among the five algorithms, while the energy consumption based on GA, SA, PSO, and WOA accounts for 20%, 28%, 23%, and 17%, respectively. More importantly, with the increase in

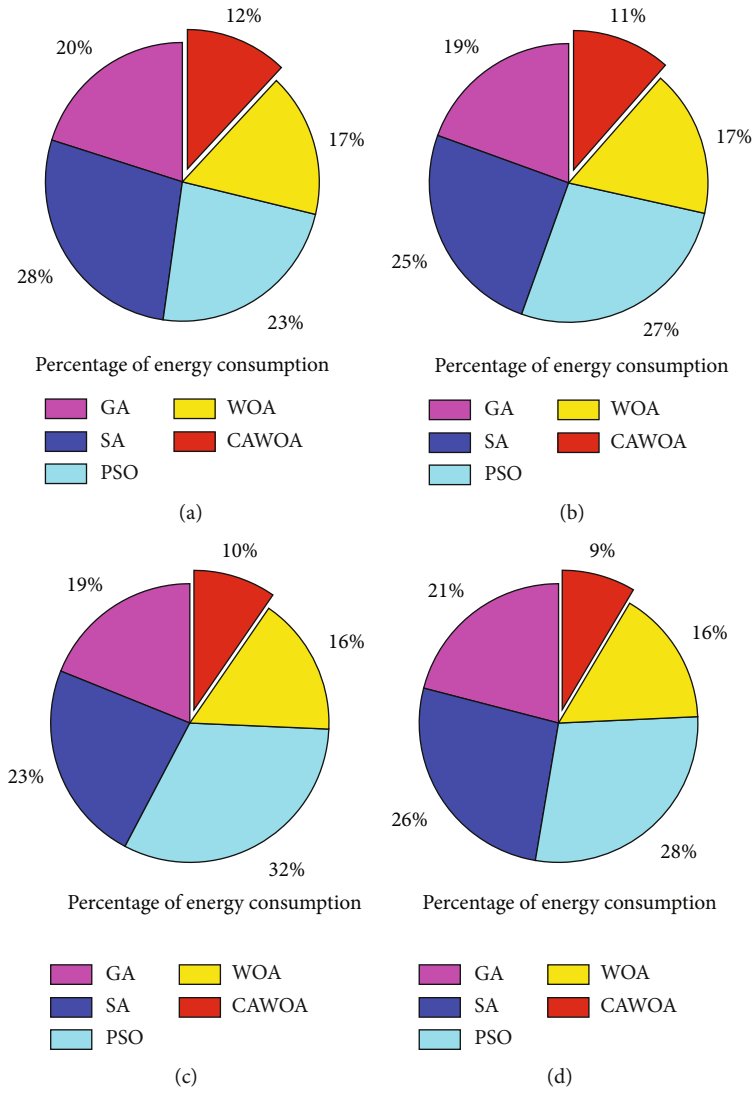


Figure 6: Comparison of energy consumption optimized by algorithms: (a) 60 sensors; (b) 80 sensors; (c) 100 sensors; (d) 120 sensors.

Table 5: Simulation parameters in Figure 7.

	Sensors	Delay (max)	Delay jitter (max)	Bandwidth (max)	Packet loss rate (max)
Figure 7(a)	50	70 ms	10 ms	6 Mbps	0.02
Figure 7(b)	100	80 ms	15 ms	7 Mbps	0.02
Figure 7(c)	150	90 ms	20 ms	8 Mbps	0.02
Figure 7(d)	200	100 ms	25 ms	9 Mbps	0.02

the number of sensor nodes, the energy consumption obtained by the CAWOA-based routing algorithm has been maintained at a low level. The energy consumption ratio is 11%, 10%, and 9% in Figures 6(b)–6(d), respectively. Especially in Figure 6(d), where there are 120 sensor nodes in its simulation conditions, the energy consumption of the CAWOA-based routing algorithm is much lower than that based on GA, SA, PSO, and WOA. Their proportions are 21%, 26%, 28%, and 16%, respectively. In general, the CAWOA-based routing algorithm has the capability of obtaining excellent routing schemes under certain QoS constraints in IWSNs. Table 5 shows the simulation parameters in Figure 7.

In the real industrial production environment, there is often a large number of sensor nodes in IWSNs. Therefore, Figures 7(a)–7(d) show the routing performance based on different algorithms in large-scale IWSNs with QoS constraints. Firstly, in Figure 7(a) with 50 nodes, it is obvious that the energy consumption of the CAWOA-based routing algorithm is lower than that of CA, SA, PSO, and WOA. In

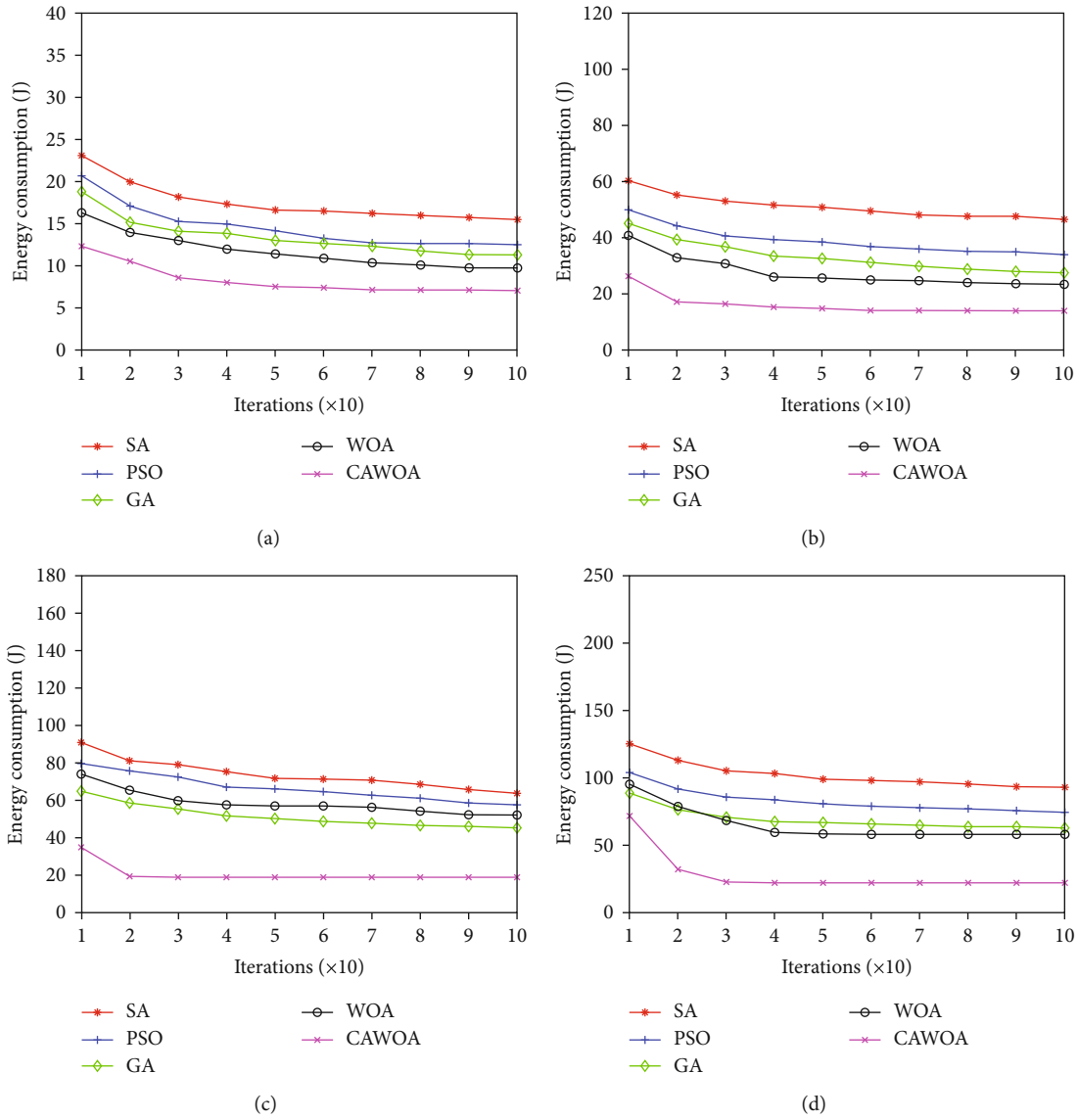


Figure 7: Line charts of energy consumption comparison of algorithms in large-scale IWSNs: (a) 50 sensors; (b) 100 sensors; (c) 150 sensors; (d) 200 sensors.

the first 10 iterations, the energy consumption of CAWOA-based routing is reduced to 12.31 J, while the energy consumption based on CA, SA, PSO, and WOA only reached 18.79 J, 23.08 J, 20.67 J, and 16.29 J. Then, the number of sensor nodes in Figures 7(b)–7(d) is 100, 150, and 200, respectively, and the energy consumption of the CAWOA-based routing algorithm is always lower than that of GA, SA, PSO, and WOA. The results clearly demonstrate that the CAWOA-based routing algorithm solves the problem of IWSN routing optimization with a large number of nodes by its slow growth of routing energy consumption, which is beneficial for saving the energy of the IWSNs. Table 6 shows the further comparison of routing energy consumption in large-scale IWSNs with QoS constraints. Table 7 shows the simulation parameters in Table 6.

According to the results in Table 6, the energy consumption of the CAWOA-based routing algorithm is always the lowest among the five algorithms. With the continuous expansion of IWSNs, the energy consumption based on GA, SA, PSO, and WOA has risen sharply, while the energy consumption of routing based on CAWOA has always been maintained at a low value. When the number of nodes is 100, 200, 300, 400, and 500, the energy consumption values

Table 6: Comparison of optimization results of algorithms.

Number of sensors	Routing energy consumption (J)				
	GA	SA	PSO	WOA	CAWOA
100	27.48	33.95	46.51	23.37	13.94
200	62.91	93.12	74.51	58.14	22.18
300	101.46	115.67	131.80	100.94	35.94
400	144.34	161.05	185.89	134.33	58.78
500	213.56	236.23	270.84	191.21	92.89

Table 7: Simulation parameters in Table 6.

Sensors	Delay (max)	Delay jitter (max)	Bandwidth (max)	Packet loss rate (max)
100	60 ms	10 ms	6 Mbps	0.01
200	70 ms	15 ms	7 Mbps	0.01
300	80 ms	20 ms	8 Mbps	0.01
400	90 ms	25 ms	9 Mbps	0.02
500	100 ms	30 ms	10 Mbps	0.02

of CAWOA-based routing are 13.94 J, 22.18 J, 35.94 J, 58.78 J, and 92.89 J, respectively. Therefore, the algorithm CAWOA proposed in this paper can effectively reduce energy consumption in large-scale IWSNs with QoS constraints.

6. Conclusions

The purpose of this paper is to reduce the routing energy consumption of IWSNs under the QoS constraints; therefore, a novel clone adaptive whale optimization algorithm (CAWOA) is designed. The significant innovation of CAWOA is the adoption of the latest adaptive technology and cloning operation. What is more, the routing method of IWSNs is optimized and the energy consumption of the network is reduced. In addition, under the premise of considering the QoS constraints in IWSNs, we designed a novel IWSN routing model that takes into account the influence of network bandwidth, delay, delay jitter, and packet loss rate on sensor routing energy consumption. Subsequently, CAWOA is compared with the genetic algorithm, whale optimization algorithm, particle swarm optimization, and simulated annealing for proving its optimization of IWSN routing energy consumption with QoS constraints. The simulation results show that the proposed routing algorithm based on CAWOA is better than other algorithms in terms of routing energy consumption, convergence speed, and optimization capability. In addition, CAWOA is especially suitable for large-scale IWSNs. As the number of sensor nodes increases, the effect of CAWOA-based routing algorithms in reducing energy consumption becomes more obvious. Therefore, it can be concluded that the application of CAWOA can effectively reduce the routing energy consumption in IWSNs.

Future research should consider more complex IWSNs, including but not limited to heterogeneous IWSNs and mobile IWSNs. In addition, in more complex situations, machine learning technology can be used as a powerful tool, such as reinforcement learning, to further improve the reliability of IWSNs and reduce routing energy consumption.

Data Availability

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Disclosure

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This paper was funded by the Corps innovative talents plan, grant number 2020CB001; project of Youth and Middleaged Scientific and Techno-logical Innovation Leading Talents Program of the Corps, grant number 2018CB006; China Postdoctoral Science Foundation, grant number 220531; Funding Project for High Level Talents Research in Shihezi University, grant number RCZK2018C38; and Project of Shihezi University, grant number ZZZC201915B.

References

[1] E. Shayo, P. Mafole, and A. Mwambela, "A survey on time division multiple access scheduling algorithms for industrial networks," *SN Applied Sciences*, vol. 2, no. 12, p. 10, 2020.

[2] N. H. Nguyen and M. K. Kim, "Link quality estimation from burstiness distribution metric in industrial wireless sensor networks," *Energies*, vol. 13, no. 23, p. 12, 2020.

[3] J. Sengupta, S. Ruj, and S. D. Bit, "A secure fog-based architecture for industrial Internet of things and industry 4.0," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2316–2324, 2021.

[4] M. M. Hassan, S. Huda, S. Sharmeen, J. Abawajy, and G. Fortino, "An adaptive trust boundary protection for IIoT networks using deep-learning feature-extraction-based semi-supervised model," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2860–2870, 2021.

[5] Z. W. Ren, M. Mukherjee, J. Lloret, and P. Venu, "Multiple kernel driven clustering with locally consistent and selfish graph in industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2956–2963, 2021.

[6] D. G. Zhang, J. X. Gao, X. H. Liu, T. Zhang, and D. X. Zhao, "Novel approach of distributed & adaptive trust metrics for MANET," *Wireless Networks*, vol. 25, no. 6, pp. 3587–3603, 2019.

[7] S. Hizal and A. Zengin, "lResearch on quality of service based routing protocols for mobile ad hoc networks," *Journal of ICT Research and Applications*, vol. 14, no. 2, pp. 185–205, 2020.

[8] G. Indumathi and V. Vaithianathan, "Optimal relay and channel selection schemes for multiconstrained QoS multicast routing in cognitive radio ad hoc networks," *International Journal of Communication Systems*, vol. 34, article e4674, 2021.

[9] M. Vafaei, A. Khademzadeh, and M. A. Pourmina, "A new QoS adaptive multi-path routing for video streaming in urban VANETs integrating ant colony optimization algorithm and fuzzy logic," *Wireless Personal Communications*, p. 34, 2021.

[10] W. F. Sun, Z. Wang, and G. H. Zhang, "A QoS-guaranteed intelligent routing mechanism in software-defined networks," *Computer Networks*, vol. 185, p. 12, 2021.

[11] E. Kharati and M. Khalily-Dermany, "Determination of the multicast optimal route for mobile sinks in a specified deadline using network coding and tabu search algorithm in wireless sensor networks," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, p. 13, 2020.

[12] M. K. Shahzad, S. M. R. Islam, M. Hossain, M. Abdullah-Al-Wadud, A. Alamri, and M. Hussain, "GAFOR: genetic algorithm based fuzzy optimized re-clustering in wireless sensor networks," *Mathematics*, vol. 9, no. 1, p. 18, 2021.

[13] A. Nayyar and R. Singh, "A comprehensive review of simulation tools for Wireless Sensor Networks (WSNs)," *Journal of Wireless Networking and Communications*, vol. 5, no. 1, pp. 19–47, 2015.

[14] D. G. Zhang, T. Zhang, Y. Dong, X. H. Liu, Y. Y. Cui, and D. X. Zhao, "Novel optimized link state routing protocol based on quantum genetic strategy for mobile learning," *Journal of Network and Computer Applications*, vol. 122, pp. 37–49, 2018.

[15] H. Mostafaei, "Energy-efficient algorithm for reliable routing of wireless sensor networks," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 7, pp. 5567–5575, 2019.

[16] D. G. Zhang, T. Zhang, J. Zhang, Y. Dong, and X. D. Zhang, "A kind of effective data aggregating method based on compressive sensing for wireless sensor network," *EURASIP Journal on Wireless Communications and Networking*, 2018.

[17] H. Mostafaei and M. S. Obaidat, "Learning automaton-based self-protection algorithm for wireless sensor networks," *Iet Networks*, vol. 7, no. 5, pp. 353–361, 2018.

[18] S. Varshney, C. Kumar, and A. Swaroop, "Lightning-based lion optimization algorithm for monitoring the pipelines using linear wireless sensor network," *Wireless Personal Communications*, vol. 117, no. 3, pp. 2475–2494, 2021.

[19] Y. Xu, W. G. Jiao, and M. Q. Tian, "Energy-efficient connected-coverage scheme in wireless sensor networks," *Sensors*, vol. 20, no. 21, p. 19, 2020.

[20] Aidil Saputra Kirsan, U. H. al Rasyid, Iwan Syarif, and Dian Neipa Purnamasari, "Energy efficiency optimization for intermediate node selection using MhSA-LEACH: multi-hop simulated annealing in wireless sensor network," *Emitter*, vol. 8, no. 1, pp. 1–18, 2020.

[21] A. Kavitha and R. L. Velusamy, "Simulated annealing and genetic algorithm-based hybrid approach for energy-aware clustered routing in large-range multi-sink wireless sensor networks," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 35, no. 2, pp. 96–116, 2020.

[22] N. Bilandi, H. K. Verma, and R. Dhir, "hPSO-SA: hybrid particle swarm optimization-simulated annealing algorithm for relay node selection in wireless body area networks," *Applied Intelligence*, vol. 51, no. 3, pp. 1410–1438, 2021.

[23] U. Mohanakrishnan and B. Ramakrishnan, "MCTRP: an energy efficient tree routing protocol for vehicular ad hoc network using genetic whale optimization algorithm," *Wireless Personal Communications*, vol. 110, no. 1, pp. 185–206, 2020.

[24] A. Nayyar and R. Singh, "IEEMARP- a novel energy efficient multipath routing protocol based on ant colony optimization (ACO) for dynamic sensor networks," *Multimedia Tools and Applications*, vol. 79, no. 47-48, pp. 35221–35252, 2020.

[25] Y. Duan, Y. Luo, W. F. Li, P. Pace, G. Aloi, and G. Fortino, "A collaborative task-oriented scheduling driven routing approach for industrial IoT based on mobile devices," *Ad Hoc Networks*, vol. 81, pp. 86–99, 2018.

[26] X. Jin, H. T. Xu, C. Q. Xia, J. T. Wang, and P. Zeng, "Convergecast scheduling and cost optimization for industrial wireless sensor networks with multiple radio interfaces," *Wireless Networks*, vol. 24, no. 8, pp. 3205–3219, 2018.

[27] N. Moussa and A. E. El Alaoui, "An energy-efficient cluster-based routing protocol using unequal clustering and improved ACO techniques for WSNs," *Peer-to-Peer Networking and Applications*, p. 14, 2021.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5550850, 18 pages
https://doi.org/10.1155/2021/5550850

Research Article

End-Effector Pose Estimation in Complex Environments Using Complementary Enhancement and Adaptive Fusion of Multisensor

Mingrui Luo,[1,2] **En Li**,[2] **Rui Guo**,[3] **Jiaxin Liu**,[4] **and Zize Liang**[2]

[1]*The School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China*
[2]*The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[3]*State Grid Shandong Electric Power Company, Jinan, China*
[4]*State Grid Liaoning Electric Power Company Limited, Shenyang, China*

Correspondence should be addressed to En Li; en.li@ia.ac.cn

Received 25 February 2021; Revised 20 March 2021; Accepted 25 March 2021; Published 16 April 2021

Academic Editor: Bin Gao

Redundant manipulators are suitable for working in narrow and complex environments due to their flexibility. However, a large number of joints and long slender links make it hard to obtain the accurate end-effector pose of the redundant manipulator directly through the encoders. In this paper, a pose estimation method is proposed with the fusion of vision sensors, inertial sensors, and encoders. Firstly, according to the complementary characteristics of each measurement unit in the sensors, the original data is corrected and enhanced. Furthermore, an improved Kalman filter (KF) algorithm is adopted for data fusion by establishing the nonlinear motion prediction of the end-effector and the synchronization update model of the multirate sensors. Finally, the radial basis function (RBF) neural network is used to adaptively adjust the fusion parameters. It is verified in experiments that the proposed method achieves better performances on estimation error and update frequency than the original extended Kalman filter (EKF) and unscented Kalman filter (UKF) algorithm, especially in complex environments.

1. Introduction

In special manufacturing, troubleshooting, and other fields, there will inevitably be a closed and narrow operating environment, which is high risk and inefficient for manual operation. Therefore, the research on robots that can be used in this complex environment is significant and challenging, especially since the traditional manipulators are unable to do the job due to their bulky structure.

Kinematic redundancy is one of the most important properties that determine whether the robots can accomplish the task in the highly constrained environments mentioned above [1]. Manipulators with kinematic redundancy are called redundant manipulators. Redundant manipulators can be implemented in different ways, such as the wheeled mobile manipulator used for part handling [2, 3], the hyper-redundant manipulator used for engine manufacturing [4] or spacecraft maintenance [5], and the flexible bionic manipulator inspired by the octopus claw [6] or the elephant trunk [7]. Although redundant manipulators have excellent kinematic performance, they also face the challenge of control accuracy due to the accuracy of end-effector pose estimation.

Traditional manipulators generally use joint encoders as sensor inputs and use kinematics models to calculate the pose of the end-effector as information feedback for motion control. Since the accuracy of the joint encoder can be very high, the pose calculated by the kinematics model has low noise. However, redundant manipulators usually have a large number of joints and long slender links, which means the pose will be affected by the transmission gap, mechanical vibration, elastic deformation, and other factors that deviate from the calculated value. As a result, the manipulators are in an uncertain state of

kinematics and dynamics, so complex control models need to be established to ensure control accuracy [8–10].

Complex control models may introduce system uncertainty, so many researchers try to solve the problem of unreliable encoders in redundant manipulators by studying sensor processing methods. Analyzing the source of the measurement error and establishing a compensation model can effectively improve the sensor accuracy. For example, in a cable-driven redundant manipulator, the accuracy of the encoder-based pose estimation can be improved by correcting the cable hole location error, cable length error [11], and nonnegligible cable mass [12].

In addition to error correction for a single type of sensor, multisensor fusion is also the main research direction. The most commonly used sensors fused with encoders including angle sensors, vision sensors, and inertial sensors. For one thing, adding angle sensors at joints and fusing with motor encoders can correct transmission errors and improve pose estimation accuracy [13]. For another, adding one or more eye-to-hand cameras to the environment and performing the visual measurement on redundant manipulators can correct deformation errors [14] and is also an effective method for flexible robot pose estimation [15]. Moreover, the Kalman filter and its improved methods are also used to fuse motor encoders and inertial sensors, which have good dynamic response performance and can correct vibration errors [16]. Because the sensor and the fusion algorithm are lightweight, it is especially suitable for small bionic robots [17].

The sensor processing methods mentioned above can indeed solve the problem of end-effector pose estimation, but there are also many unsatisfactory shortcomings, especially when applied in complex environments. Therefore, the estimator constructed in this paper uses the compensation model and multisensor fusion at the same time, but in different processing stages. In a dynamic environment, although the preestablished compensation model is prone to failure and cannot accurately reflect the relationship between the sensor and the estimated state, it can effectively correct the wrong measurement [18]. Therefore, this paper does not directly use the compensation model for state estimation but as the preprocessing part of the state estimation to enhance the original data of the sensor. In terms of multisensor fusion, although visual-inertial fusion has become an effective method of mobile robot navigation [19], it is not common in end-effector pose estimation of redundant robots, and it is mostly the fusion of eye-to-hand camera and kinematic sensors [20]. In a narrow environment, it is difficult to apply an eye-to-hand camera, so the vision sensor needs to be arranged at the end of the manipulator, which means that the vision sensor is more susceptible to the influence of obstacles. Therefore, based on the Kalman filter algorithm, this paper establishes a nonlinear motion prediction of the end-effector and a synchronization update model of the multirate sensors that combines vision sensors, inertial sensors, and encoders. Moreover, since the neural network can improve the robustness of pose estimation [21], this paper introduces the radial basis function neural network to adjust parameters adaptively. Through the above-mentioned improved fusion strategy, the pose estimation of the end-effector can be more adaptable to complex environments and can better overcome the adverse effects of sensor noise and different sampling rates of multisensor fusion.

The rest of this paper is organized as follows: Section 2 introduces the overall method. Section 3 introduces the complementary enhancement of the original data. Section 4 introduces the fusion strategy including the nonlinear prediction, the synchronization update, and the adaptive adjustment. Section 5 introduces the experimental results and analysis, and Section 6 summarizes the paper.

2. The Overall Method of the Pose Estimation

Figure 1 shows the small redundant robot studied in this paper, which can enter the narrow space to accomplish the specified tasks.

Due to the miniaturized mechanical structure, when the end-effector is equipped with a load, the connecting rod between the joints will inevitably be bent and deformed. This deformation error will accumulate at the end-effector; as a result, the pose estimation based only on the joint encoders produces deviation. Therefore, structured-light RGB-D camera, magnetic angular rate, and gravity (MARG) sensor are added, together with the joint encoders.

In this paper, the fusion pose estimation method proposed mainly includes two parts: preenhancement and adaptive fusion. The system block diagram is shown in Figure 2.

The RGB-D camera and the MARG sensor can compensate for the error caused by a single encoder measurement to a certain extent, but at the same time, it is easy to reduce the robustness of the system due to the interference of the external environment. Considering that the RGB-D camera and the MARG sensor themselves are integrated sensors composed of multiple sensing units, each sensing unit has complementary characteristics, which can be used to enhance the original data of the sensors so that the fusion of multiple sensors can be completed based on robust data. According to the above theory, preenhancement is added before multisensor fusion.

The preenhancement part includes the image enhancement of the RGB-D camera and the quaternion extended Kalman filter (QEKF) fusion attitude calculation of the MARG sensor. In the image enhancement part, the color image is used to repair the depth image, which can make the measurement of the RGB-D camera more accurate and reliable in a narrow environment (Section 3.1). In the QEKF part, the magnetometer, accelerometer, and gyroscope, which are all parts of the MARG sensor, are fused to correct the drift error of each independent sensing unit (Section 3.2).

Through the preenhancement part, the robustness of each independent sensor is improved, but the following problems still need to be overcome in the multisensor fusion process: (a) nonlinear state model, (b) the sampling rate of different sensors differs greatly, and (c) sensor abnormal handling in extreme environments.

In the adaptive fusion part, the RBF and adaptive UKF (RBF-AUKF) method is modified to solve the above problems, which mainly includes pose prediction, synchronization update, and adaptive adjustment. In the pose prediction part, by establishing the pose prediction model of the end-effector

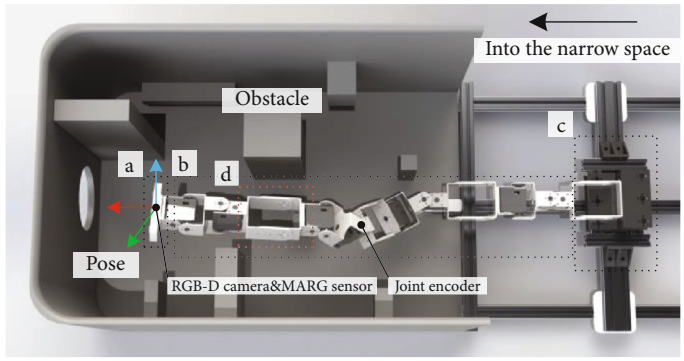


Figure 1: The small redundant robot (a) end-effector, (b) redundant manipulator, (c) base, and (d) connecting rod.

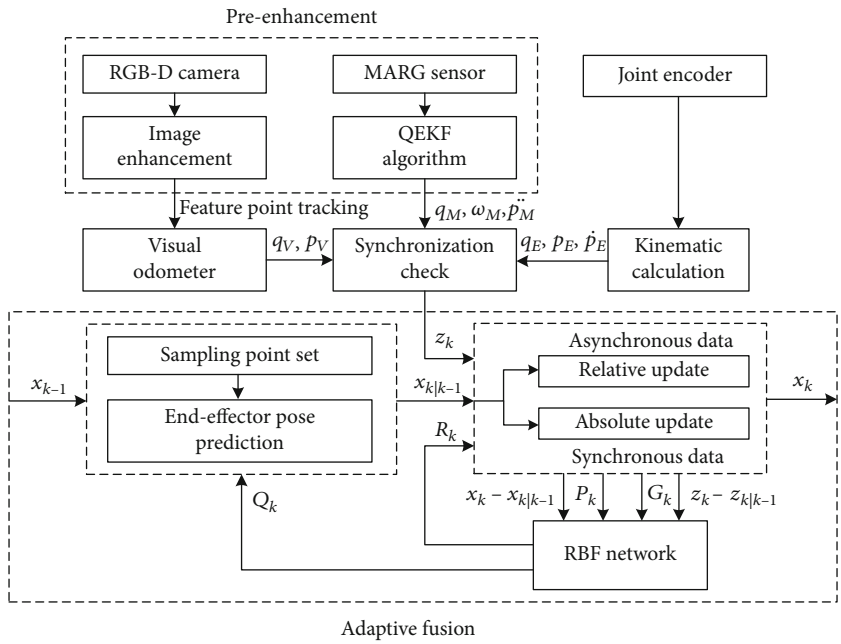


Figure 2: Block diagram of the pose estimation method.

and the nonlinear processing method, the pose state $x_{k|k-1}$ at the next moment can be predicted based on the pose state x_{k-1} at the previous moment (Section 4.1). In the synchronization update part, a synchronization check is performed on the preenhanced quaternion attitude q_V and position p_V obtained by the visual odometer; quaternion attitude q_M, angular velocity ω_M, and acceleration $\ddot{p}_M$ obtained by the QEKF algorithm; and quaternion attitude q_E, position p_E, and velocity $\dot{p}_E$ obtained by the kinematic calculation. According to the check result, the control variable is relatively updated when the sensor data z_k is asynchronous, while the state variable is absolutely updated when the sensor data z_k is synchronous. After the prediction and update, the fusion pose estimation x_k at this moment is obtained (Section 4.2). In the adaptive adjustment part, this paper selects the prediction error $x_k - x_{k|k-1}$, estimation error covariance matrix P_k, Kalman gain G_k, and observation error $z_k - z_{k|k-1}$ after updating as features. Based on the RBF neural network, the process, and observation noise covariance matrix Q_k, R_k is adaptively adjusted to improve the robustness of fusion (Section 4.3).

3. Sensor Enhancement Based on Complementary Characteristics

3.1. Enhancement of the Depth Image. The structured-light RGB-D camera is composed of a color camera, an infrared camera, and an infrared transmitter and obtains the depth value through the speckle feature of the infrared structured light projected on the surface of the object [22]. In a closed and narrow environment, infrared light is easily reflected and absorbed by objects multiple times, so it is easier to produce invalid depth values locally, which in turn introduces additional noise for pose estimation.

Since the invalid depth value occurs locally, the invalid point can be repaired by a valid point that is in a similar region to the invalid point. Since the color camera is not affected by

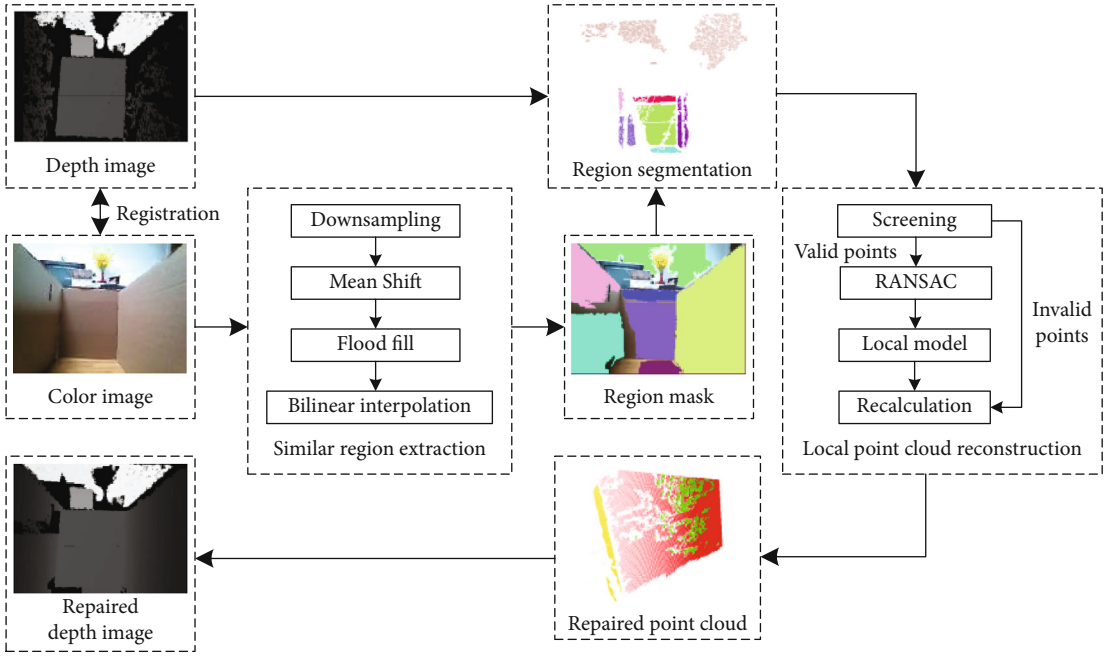


Figure 3: Depth image enhancement in a narrow environment.

the narrow environment, it can be used to determine whether the valid point and the invalid point belong to the same type of patch gathered by the point cloud, that is, the similar region. Therefore, the enhanced method proposed in this paper includes two steps, namely, similar region extraction and local point cloud reconstruction, as shown in Figure 3. The former extracts similar regions in the depth image through the texture features of the registered color image. The latter uses valid points in the similar region to construct a local model of the region and then recalculates invalid points in the region with the model to complete the enhancement repair process.

In the similar region extraction step, the Mean Shift algorithm is used to extract the color texture information of the color image, and the Flood Fill algorithm is used to extract the corresponding color-connected domains to form a similar region mask. Finally, the mask is applied to the depth image to obtain the region segmentation of the similar point cloud. Since the Mean Shift algorithm needs to perform multiple iterative calculations for each pixel, it will affect the real-time performance of the system in practical applications. In order to speed up the calculation process, this paper downsamples the input color image to a lower resolution, then forms the similar region mask on the low-resolution image, and finally performs bilinear interpolation on the mask to restore the original scale.

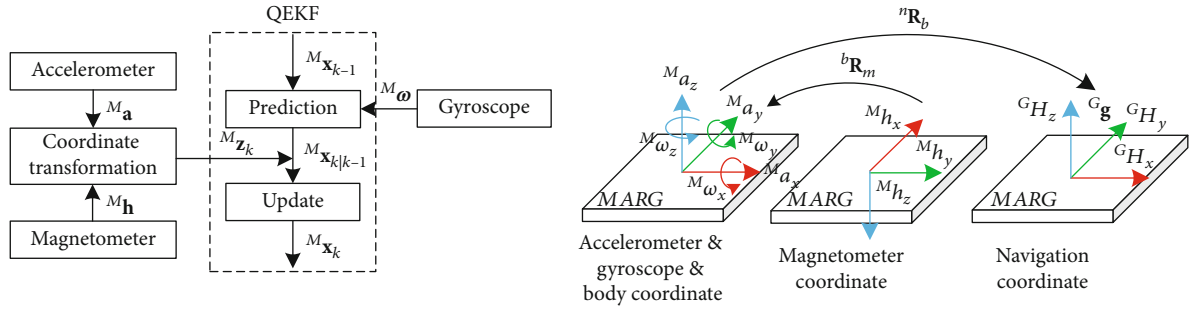
In the local point cloud reconstruction step, by fitting the patch model in the similar region and recalculating the depth value of the invalid point on the patch, the invalid point is repaired. The least-squares method is the basic method to fit the patch model, but in practical applications, the region segmented by the color image has deviation, and not all valid points can be used to fit the patch model. In order to form a more reliable local model, the RANSAC algorithm is used to filter the points used to fit the model.

Although the invalid point does not have the correct depth value, it contains the coordinate information (u, v) of the pixel plane. Combining with the local model $\hat{z} = ax + by + c$ that has been fitted, the spatial coordinate $(x, y, \hat{z})$ of the invalid point can be recalculated, as shown in equation (1). Finally, the invalid point in the depth image is repaired.

$$\begin{cases} x = (u - c_x)\dfrac{\widehat{z}}{f_x}, \\ y = (v - c_y)\dfrac{\widehat{z}}{f_y}, \\ \widehat{z} = ax + by + c, \end{cases} \tag{1}$$

where c_x, c_y is the optical center offset of the camera and f_x, f_y is the focal length of the camera.

3.2. Enhancement of the Attitude Calculation. The MARG sensor is a microelectromechanical system (MEMS) composed of a three-axis accelerometer, a three-axis gyroscope, and a three-axis magnetometer. However, due to the impact of sensor manufacturing accuracy and operating environment interference, the raw measurement data of the MARG sensor is generally noisy. From the analysis of its working principle, the gyroscope has good dynamic response characteristics, but it is easy to produce cumulative errors in the attitude calculation, and the accelerometer and magnetometer are just the opposite. Therefore, the QEKF algorithm is used to fuse the three measurement units of the MARG sensor so that the output attitude angle data can be enhanced, as shown in Figure 4.



(a) Block diagram of attitude calculation

(b) The coordinate system of MARG sensor

Figure 4: Attitude calculation enhancement.

There are three coordinate systems in this method. Firstly, the acceleration vector ${}^{M}\mathbf{a} = [{}^{M}a_x \; {}^{M}a_y \; {}^{M}a_z]^T$ measured by the accelerometer and the angular velocity vector ${}^{M}\boldsymbol{\omega} = [{}^{M}\omega_x \; {}^{M}\omega_y \; {}^{M}\omega_z]^T$ measured by the gyroscope are both in the body coordinate system. Secondly, the magnetic field intensity vector ${}^{M}\mathbf{h} = [{}^{M}h_x \; {}^{M}h_y \; {}^{M}h_z]^T$ measured by the magnetometer is in the coordinate system that can be transformed to the body coordinate system by the rotation matrix ${}^{b}\mathbf{R}_m$. Lastly, the geomagnetic field intensity vector ${}^{G}\mathbf{H} = [{}^{G}H_x \; {}^{G}H_y \; {}^{G}H_z]^T$ and the gravitational acceleration vector ${}^{G}g = [0 \; 0 \; -g]^T$ are in the navigation coordinate system that can be transformed from the body coordinate system by the rotation matrix ${}^{n}\mathbf{R}_b$.

Suppose the system state variables ${}^{M}\mathbf{x}_k$ at the k-th time are

$${}^{M}\mathbf{x}_k = \left[{}^{M}q_{0,k} \quad {}^{M}q_{1,k} \quad {}^{M}q_{2,k} \quad {}^{M}q_{3,k} \quad {}^{M}b_{x,k} \quad {}^{M}b_{y,k} \quad {}^{M}b_{z,k}\right]^T, \tag{2}$$

where ${}^{M}q_{0,k}, {}^{M}q_{1,k}, {}^{M}q_{2,k}, {}^{M}q_{3,k}$ is the quaternion value of the rotation transformation matrix ${}^{n}R_b$ at the k-th time and ${}^{M}b_{x,k}, {}^{M}b_{y,k}, {}^{M}b_{z,k}$ is the accumulated drift value of the gyroscope on the X, Y, and Z axes at the k-th time.

Suppose the system control variables ${}^{M}\mathbf{u}_k$ at the k-th time is

$${}^{M}\mathbf{u}_k = \left[{}^{M}\omega_{x,k} - {}^{M}b_{x,k-1} \quad {}^{M}\omega_{y,k} - {}^{M}b_{y,k-1} \quad {}^{M}\omega_{z,k} - {}^{M}b_{z,k-1}\right]^T. \tag{3}$$

Suppose the system observation variables ${}^{M}\mathbf{z}_k$ at the k-th time is

$${}^{M}\mathbf{z}_k = \left[{}^{M}\mathbf{a}_k \quad {}^{M}\mathbf{h}_k\right]^T. \tag{4}$$

According to the discrete-time model of the quaternion attitude differential equation [23], the following equation can be obtained:

$${}^{M}\mathbf{q}_k = \left[\mathbf{I}_{4\times4} + {}^{M}\boldsymbol{\Omega}_{k-1}\frac{T}{2}\right]{}^{M}\mathbf{q}_{k-1}, \tag{5}$$

$${}^{M}\boldsymbol{\Omega}_{k-1} = \begin{bmatrix} 0 & {}^{M}b_{x,k-1} - {}^{M}\omega_{x,k} & {}^{M}b_{y,k-1} - {}^{M}\omega_{y,k} & {}^{M}b_{z,k-1} - {}^{M}\omega_{z,k} \\ {}^{M}\omega_{x,k} - {}^{M}b_{x,k-1} & 0 & {}^{M}\omega_{z,k} - {}^{M}b_{z,k-1} & {}^{M}b_{y,k-1} - {}^{M}\omega_{y,k} \\ {}^{M}\omega_{y,k} - {}^{M}b_{y,k-1} & {}^{M}b_{z,k-1} - {}^{M}\omega_{z,k} & 0 & {}^{M}\omega_{x,k} - {}^{M}b_{x,k-1} \\ {}^{M}\omega_{z,k} - {}^{M}b_{z,k-1} & {}^{M}\omega_{y,k} - {}^{M}b_{y,k-1} & {}^{M}b_{x,k-1} - {}^{M}\omega_{x,k} & 0 \end{bmatrix}, \tag{6}$$

where T is the discrete-time interval, $\mathbf{I}_{i\times j}$ is the identity matrix with i rows and j columns, ${}^{M}\mathbf{q}_k = [{}^{M}q_{0,k} \; {}^{M}q_{1,k} \; {}^{M}q_{2,k} \; {}^{M}q_{3,k}]^T$ is the quaternion vector of ${}^{n}\mathbf{R}_b$ at the k-th time, and ${}^{M}\boldsymbol{\Omega}_k$ is the component of angular velocity in the body coordinate system.

Substituting equations (2) and (3) into equation (5), (6) can obtain the state prediction equation:

$${}^{M}f\left({}^{M}\mathbf{x}_{k-1}, {}^{M}\mathbf{u}_{k-1}\right) = \begin{bmatrix} \mathbf{I}_{4\times4} + {}^{M}\boldsymbol{\Omega}_{k-1}\frac{T}{2} & 0_{4\times3} \\ 0_{3\times4} & \mathbf{I}_{3\times3} \end{bmatrix} \begin{bmatrix} {}^{M}\mathbf{q}_k \\ {}^{M}b_{x,k-1} \\ {}^{M}b_{y,k-1} \\ {}^{M}b_{z,k-1} \end{bmatrix}, \tag{7}$$

where $0_{i\times j}$ is the zero matrix with i rows and j columns.

To transform the measured value of the sensor, the following equations can be obtained:

$${}^{M}\mathbf{a} = {}^{n}\mathbf{R}_b{}^{-1}{}^{G}\mathbf{g}, \tag{8}$$

$${}^{M}\mathbf{h} = {}^{n}\mathbf{R}_b{}^{-1}{}^{b}\mathbf{R}_m{}^{-1}{}^{G}\mathbf{H}. \tag{9}$$

Then, the system observation equation ${}^{M}h({}^{M}\hat{\mathbf{x}}_k)$ was obtained:

$$\begin{cases} {}^{M}\widehat{a}_{x,k} = 2g\left({}^{M}\widehat{q}_{1,k}{}^{M}\widehat{q}_{3,k} - {}^{M}\widehat{q}_{0,k}{}^{M}\widehat{q}_{2,k}\right), \\ {}^{M}\widehat{a}_{y,k} = 2g\left({}^{M}\widehat{q}_{2,k}{}^{M}\widehat{q}_{3,k} + {}^{M}\widehat{q}_{0,k}{}^{M}\widehat{q}_{1,k}\right), \\ {}^{M}\widehat{a}_{z,k} = g\left({}^{M}\widehat{q}_{0,k}^{2} - {}^{M}\widehat{q}_{1,k}^{2} - {}^{M}\widehat{q}_{2,k}^{2} + {}^{M}\widehat{q}_{3,k}^{2}\right), \\ {}^{M}\widehat{h}_{x,k} = {}^{G}H_x\left({}^{M}\widehat{q}_{0,k}^{2} - {}^{M}\widehat{q}_{1,k}^{2} + {}^{M}\widehat{q}_{2,k}^{2} - {}^{M}\widehat{q}_{3,k}^{2}\right) + 2{}^{G}H_y\left({}^{M}\widehat{q}_{1,k}{}^{M}\widehat{q}_{2,k} - {}^{M}\widehat{q}_{0,k}{}^{M}\widehat{q}_{3,k}\right) - 2{}^{G}H_z\left({}^{M}\widehat{q}_{2,k}{}^{M}\widehat{q}_{3,k} + {}^{M}\widehat{q}_{0,k}{}^{M}\widehat{q}_{1,k}\right), \\ {}^{M}\widehat{h}_{y,k} = 2{}^{G}H_x\left({}^{M}\widehat{q}_{1,k}{}^{M}\widehat{q}_{2,k} + {}^{M}\widehat{q}_{0,k}{}^{M}\widehat{q}_{3,k}\right) + {}^{G}H_y\left({}^{M}\widehat{q}_{0,k}^{2} + {}^{M}\widehat{q}_{1,k}^{2} - {}^{M}\widehat{q}_{2,k}^{2} - {}^{M}\widehat{q}_{3,k}^{2}\right) - 2{}^{G}H_z\left({}^{M}\widehat{q}_{1,k}{}^{M}\widehat{q}_{3,k} - {}^{M}\widehat{q}_{0,k}{}^{M}\widehat{q}_{2,k}\right), \\ {}^{M}\widehat{h}_{z,k} = -2{}^{G}H_x\left({}^{M}\widehat{q}_{2,k}{}^{M}\widehat{q}_{3,k} - {}^{M}\widehat{q}_{0,k}{}^{M}q_{1,k}\right) - 2{}^{G}H_y\left({}^{M}\widehat{q}_{1,k}{}^{M}\widehat{q}_{3,k} + {}^{M}\widehat{q}_{0,k}{}^{M}\widehat{q}_{2,k}\right) + {}^{G}H_z\left({}^{M}\widehat{q}_{0,k}^{2} - {}^{M}\widehat{q}_{1,k}^{2} - {}^{M}\widehat{q}_{2,k}^{2} + {}^{M}\widehat{q}_{3,k}^{2}\right), \end{cases} \tag{10}$$

where ${}^{M}a\widehat{a}_{x,k}$, ${}^{M}\widehat{a}_{y,k}$, ${}^{M}\widehat{a}_{z,k}$, ${}^{M}\widehat{h}_{x,k}$, ${}^{M}\widehat{h}_{y,k}$, ${}^{M}\widehat{h}_{z,k}$ is the theoretical measurement value of the sensors and ${}^{M}\widehat{q}_{0,k}$, ${}^{M}\widehat{q}_{1,k}$, ${}^{M}\widehat{q}_{2,k}$, ${}^{M}\widehat{q}_{3,k}$ is the quaternion attitude predicted by equation (7).

According to the EKF algorithm, calculate the enhanced quaternion attitude at the k-th time:

$$\begin{cases} {}^{M}\widehat{\mathbf{x}}_k = {}^{M}f\left({}^{M}\mathbf{x}_{k-1}, {}^{M}\mathbf{u}_{k-1}\right), \\ {}^{M}\widehat{\mathbf{P}}_k = {}^{M}\mathbf{F}_{k-1}{}^{M}\mathbf{P}_{k-1}{}^{M}\mathbf{F}_{k-1}{}^{T} + {}^{M}\mathbf{Q}_{k-1}, \\ {}^{M}\mathbf{G}_k = {}^{M}\widehat{\mathbf{P}}_k{}^{M}\mathbf{H}_k{}^{T}\left({}^{M}\mathbf{H}_k{}^{M}\widehat{\mathbf{P}}_k{}^{M}\mathbf{H}_k{}^{T} + {}^{M}\mathbf{R}_k\right)^{-1}, \\ {}^{M}\mathbf{x}_k = {}^{M}\widehat{\mathbf{x}}_k + {}^{M}\mathbf{G}_k\left[{}^{M}\mathbf{z}_k - {}^{M}h\left({}^{M}\widehat{\mathbf{x}}_k\right)\right], \\ {}^{M}\mathbf{P}_k = \left(\mathbf{I} - {}^{M}\mathbf{G}_k{}^{M}\mathbf{H}_k\right){}^{M}\widehat{\mathbf{P}}_k, \end{cases} \tag{11}$$

where ${}^{M}\widehat{\mathbf{x}}_k$ is the prediction of the system state, ${}^{M}\widehat{\mathbf{P}}_k$ is the prediction of the estimation error covariance matrix ${}^{M}\mathbf{P}_k$, ${}^{M}\mathbf{F}_{k-1} = (\partial {}^{M}f(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}))/(\partial \mathbf{x}_{k-1})|_{\mathbf{x}_{k-1}={}^{M}\mathbf{x}_{k-1}}$, ${}^{M}\mathbf{H}_k = (\partial {}^{M}h(\mathbf{x}_k))/(\partial \mathbf{x}_k)|_{\mathbf{x}_k={}^{M}\widehat{\mathbf{x}}_{k-1}}$, ${}^{M}\mathbf{G}_k$ is the Kalman gain, $\mathbf{I}$ is the identity matrix, and ${}^{M}\mathbf{Q}_{k-1}$ and ${}^{M}\mathbf{R}_k$ are the noise matrix.

To further improve the robustness of the system, the process noise matrix ${}^{M}\mathbf{Q}_{k-1}$ and the observation noise matrix ${}^{M}\mathbf{R}_k$ in equation (11) adopt an adaptive form.

Suppose that the angular velocity measurement value ${}^{M}\boldsymbol{\omega}$ obtained by the gyroscope obeys the Gaussian distribution with variance ${\sigma_g}^2$, and the drift value ${}^{M}\mathbf{b} = \begin{bmatrix} {}^{M}b_x & {}^{M}b_y & {}^{M}b_z \end{bmatrix}^{T}$ obeys the Gaussian distribution with variance ${\sigma_b}^2$, substituting equation (7) to obtain the following:

$${}^{M}\mathbf{Q}_{k-1} = \mathbf{I}_{7\times 7}\begin{bmatrix} \left({\sigma_g}^2 + {\sigma_b}^2\right)\left({}^{M}q_{1,k-1}{}^2 + {}^{M}q_{2,k-1}{}^2 + {}^{M}q_{3,k-1}{}^2\right)\dfrac{T^2}{4} & & & & & & \\ & \left({\sigma_g}^2 + {\sigma_b}^2\right)\left({}^{M}q_{0,k-1}{}^2 + {}^{M}q_{2,k-1}{}^2 + {}^{M}q_{3,k-1}{}^2\right)\dfrac{T^2}{4} & & & & & \\ & & \left({\sigma_g}^2 + {\sigma_b}^2\right)\left({}^{M}q_{0,k-1}{}^2 + {}^{M}q_{1,k-1}{}^2 + {}^{M}q_{3,k-1}{}^2\right)\dfrac{T^2}{4} & & & & \\ & & & \left({\sigma_g}^2 + {\sigma_b}^2\right)\left({}^{M}q_{0,k-1}{}^2 + {}^{M}q_{1,k-1}{}^2 + {}^{M}q_{2,k-1}{}^2\right)\dfrac{T^2}{4} & & & \\ & & & & {\sigma_b}^2 & & \\ & & & & & {\sigma_b}^2 & \\ & & & & & & {\sigma_b}^2 \end{bmatrix}. \tag{12}$$

Suppose that the acceleration measurement value ${}^{M}\mathbf{a}$ obtained by the accelerometer obeys the Gaussian distribution with variance ${\sigma_a}^2$, and the magnetic field intensity measurement value ${}^{M}\mathbf{h}$ obtained by the magnetometer obeys the Gaussian distribution with variance ${\sigma_m}^2$. Considering the influence of external acceleration and external magnetic field strength,

$${}^{M}\mathbf{R}_k = \begin{bmatrix} \left[k_a\left\|{}^{M}\mathbf{a}_k - {}^{G}\mathbf{g}_k\right\| + {\sigma_a}^2\right]\mathbf{I}_{3\times 3} & 0_{3\times 3} \\ 0_{3\times 3} & \left[k_m\left\|{}^{M}\mathbf{h}_k - {}^{G}\mathbf{H}_k\right\| + {\sigma_m}^2\right]\mathbf{I}_{3\times 3} \end{bmatrix}, \tag{13}$$

where k_a and k_m are correction coefficients.

4. Sensor Fusion Based on Modified RBF-AUKF Method

4.1. End-Effector Pose Prediction Model. The RGB-D camera and MARG sensor are both arranged at the end of the

manipulator, so the sensor data after preenhancement processing is directly in the end-effector coordinate system. The joint encoders are arranged at each joint of the manipulator, so it is necessary to transform the angle of each joint to the pose in the end-effector coordinate system through forward kinematics calculation.

Suppose the system state variables $\mathbf{x}_k$ at the k-th time is

$$\mathbf{x}_k = \begin{bmatrix} \mathbf{p}_k & \mathbf{q}_k \end{bmatrix}^T, \tag{14}$$

where $\mathbf{p}_k = \begin{bmatrix} p_{x,k} & p_{y,k} & p_{z,k} \end{bmatrix}^T$ is the position of the end-effector at the k-th time and $\mathbf{q}_k = \begin{bmatrix} q_{0,k} & q_{1,k} & q_{2,k} & q_{3,k} \end{bmatrix}^T$ is the quaternion attitude of the end-effector at the k-th time.

Suppose the system control variables $\mathbf{u}_k$ at the k-th time is

$$\mathbf{u}_k = \begin{bmatrix} \boldsymbol{\omega}_k & \dot{\mathbf{p}}_k & \ddot{\mathbf{p}}_k \end{bmatrix}^T, \tag{15}$$

where $\boldsymbol{\omega}_k = \begin{bmatrix} \omega_{x,k} & \omega_{y,k} & \omega_{z,k} \end{bmatrix}^T$ is the X, Y, and Z axis angular velocities of the end-effector, $\dot{\mathbf{p}}_k = \begin{bmatrix} \dot{p}_{x,k} & \dot{p}_{y,k} & \dot{p}_{z,k} \end{bmatrix}^T$ is the position velocity of the end-effector at the k-th time, and $\ddot{\mathbf{p}}_k = \begin{bmatrix} \ddot{p}_{x,k} & \ddot{p}_{y,k} & \ddot{p}_{z,k} \end{bmatrix}^T$ is the position acceleration of the end-effector at the k-th time.

According to the differential model of motion [24], the system state prediction equation $f(\mathbf{x}_{k-1}, \mathbf{u}_{k-1})$ can be obtained:

$$\begin{cases} \mathbf{p}_k = \mathbf{p}_{k-1} + {}^n\mathbf{R}_{b,k-1} \begin{bmatrix} \dot{p}_{x,k-1} T + \ddot{p}_{x,k-1} \dfrac{T^2}{2} \\ \dot{p}_{y,k-1} T + \ddot{p}_{y,k-1} \dfrac{T^2}{2} \\ \dot{p}_{z,k-1} T + \ddot{p}_{z,k-1} \dfrac{T^2}{2} \end{bmatrix}, \\ \mathbf{q}_k = \mathbf{q}_{k-1} + \dfrac{T}{2} \begin{bmatrix} 0 & -\omega_{x,k-1} & -\omega_{y,k-1} & -\omega_{z,k-1} \\ \omega_{x,k-1} & 0 & \omega_{z,k-1} & -\omega_{y,k-1} \\ \omega_{y,k-1} & -\omega_{z,k-1} & 0 & \omega_{x,k-1} \\ \omega_{z,k-1} & \omega_{y,k-1} & -\omega_{x,k-1} & 0 \end{bmatrix} \mathbf{q}_{k-1}, \end{cases} \tag{16}$$

where T is the discrete-time interval and ${}^n\mathbf{R}_{b,k-1}$ is the rotation matrix from the body coordinate system to the navigation coordinate system as shown in equation (17).

$${}^n\mathbf{R}_{b,k-1} = \begin{bmatrix} q_{0,k-1}^2 + q_{1,k-1}^2 - q_{2,k-1}^2 - q_{3,k-1}^2 & 2\left(q_{1,k-1}q_{2,k-1} - q_{0,k-1}q_{3,k-1}\right) & 2\left(q_{1,k-1}q_{3,k-1} + q_{0,k-1}q_{2,k-1}\right) \\ 2\left(q_{1,k-1}q_{2,k-1} + q_{0,k-1}q_{3,k-1}\right) & q_{0,k-1}^2 - q_{1,k-1}^2 + q_{2,k-1}^2 - q_{3,k-1}^2 & 2\left(q_{2,k-1}q_{3,k-1} - q_{0,k-1}q_{1,k-1}\right) \\ 2\left(q_{1,k-1}q_{3,k-1} - q_{0,k-1}q_{2,k-1}\right) & 2\left(q_{2,k-1}q_{3,k-1} + q_{0,k-1}q_{1,k-1}\right) & q_{0,k-1}^2 - q_{1,k-1}^2 - q_{2,k-1}^2 + q_{3,k-1}^2 \end{bmatrix}. \tag{17}$$

The state prediction equation mentioned in equation (16) has obvious nonlinear characteristics. Both EKF and UKF algorithms can solve the problem of nonlinear state estimation. However, when the calculation performance can meet the requirements, the estimation accuracy of the UKF algorithm is higher [25]. By sampling, the UKF algorithm approximates the probability density distribution of the nonlinear prediction equation, and the constructed sampling point set is as follows:

$$[\boldsymbol{\chi}_k]_i = \begin{cases} \mathbf{x}_k, & i = 0, \\ \mathbf{x}_k + \left[\sqrt{(N+\lambda)\mathbf{P}_{k-1}}\right]_i, & i = 1, \cdots, N, \\ \mathbf{x}_k - \left[\sqrt{(N+\lambda)\mathbf{P}_{k-1}}\right]_i, & i = N+1, \cdots, 2N, \end{cases} \tag{18}$$

where $[\boldsymbol{\chi}_k]_i$ is the i-th column of the sample point set matrix at the k-th time, which is a 7-dimensional column vector, $[\sqrt{(N+\lambda)\mathbf{P}_{k-1}}]_i$ is the i-th column of the matrix $\sqrt{(N+\lambda)\mathbf{P}_{k-1}}$, which is an N-dimensional column vector, λ is the scale parameter to adjust the approximation accuracy, and $\mathbf{P}_{k-1}$ is the estimation error covariance matrix.

4.2. Multirate Update Model. Due to the limitation of sensor processing capacity, the measurement update frequency of different sensors is not the same, or even quite different, which means that even if the sensors are triggered to collect at the same time, the measurement of each sensor cannot be received at the same time.

Although it is difficult for sensors to publish data synchronously all the time, it is possible to approximate the software synchronization of multiple sensors through timestamp alignment. The basic strategy is to set a sliding time window; when all sensor data falls within this window, it can be considered that the sensor data is synchronous. As is shown in Figure 5, d is the size of the sliding time window, which determines the accuracy of data synchronization. Its maximum size is the smallest update time interval among all sensors. The smaller the size, the higher the accuracy of the time, but the lower the frequency of synchronous data.

The UKF algorithm corrects the prediction model through sensor observations. Therefore, when the algorithm is used for multisensor fusion, the fusion accuracy is greatly affected by the time synchronization [26]. Because the

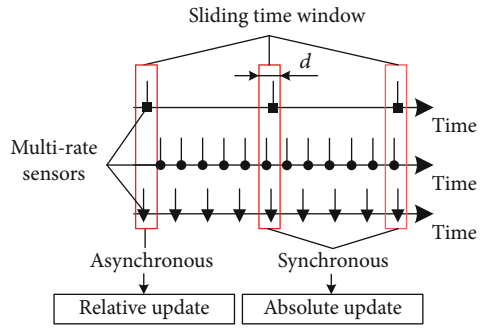


FIGURE 5: An update model that maintains an average update frequency.

update frequency of synchronous data is generally much lower than the average update frequency of data, the real-time performance of synchronization fusion estimation is poor. To solve this shortcoming, a modified synchronization fusion algorithm is proposed including two update models, namely, the absolute state update model and the relative state update model.

Define a matrix $\mathbf{H}_{m\times n}$ with m rows and n columns as a state update matrix, and its i-th row and j-th column satisfy the following:

$$[\mathbf{H}_{m\times n}]_{i,j} = \begin{cases} 1, \text{measurement } i \text{ is related to state } j, \\ 0, \text{otherwise.} \end{cases} \tag{19}$$

The absolute state update model updates the system state variables including the position and attitude. The number of rows and columns in the state update matrix $\mathbf{H}_{m\times n}$ satisfies the following:

$$m = \sum_{k=1}^{N_A} {}^{A}s_k, \quad n = 7, \tag{20}$$

where N_A is the total number of sensors synchronized with the system state variables and N_A must be equal to $N_{\max}$ which is the maximum number of sensors; otherwise $\mathbf{H}_{m\times n} = 0$. ${}^{A}s_k$ is the total number of the system state variables that the k-th sensor can observe.

The relative state update model updates the system control variables including velocity and acceleration. The number of rows and columns in the state update matrix $\mathbf{H}_{m\times n}$ satisfies the following:

$$m = \sum_{k=1}^{N_R} {}^{R}s_k, \quad n = 9, \tag{21}$$

where N_R is the total number of sensors synchronized with the system control variables and N_R must satisfy $2 \le N_R \le N_{\max}$; otherwise $\mathbf{H}_{m\times n} = 0$. ${}^{R}s_k$ is the total number of system control variables that the k-th sensor can observe.

When the multisensor data is not all synchronized, the relative state update is used to increase the update frequency locally, and when the data is all synchronized, the system state is globally corrected to improve the estimation accuracy. The final observation equation is as follows:

$$h(\mathbf{x}_{k|k-1}, \mathbf{u}_k) = \begin{bmatrix} \mathbf{H}_A & 0 \\ 0 & \mathbf{H}_R \end{bmatrix} \begin{bmatrix} \mathbf{x}_{k|k-1} \\ \mathbf{u}_k \end{bmatrix}, \tag{22}$$

where $\mathbf{H}_A$ is the absolute state update matrix determined by equation (20) and $\mathbf{H}_R$ is the relative state update matrix determined by equation (21).

According to the UKF algorithm, calculate the fusion pose estimation $\mathbf{x}_k$ of the end-effector:

$$\begin{cases} \boldsymbol{\chi}_{k|k-1} = f(\boldsymbol{\chi}_{k-1}, \mathbf{u}_{k-1}), \\ \mathbf{x}_{k|k-1} = \sum_{i=0}^{2n} W_i [\boldsymbol{\chi}_{k|k-1}]_i, \\ \mathbf{P}_{x,k|k-1} = \sum_{i=0}^{2n} W_i \left[[\boldsymbol{\chi}_{k|k-1}]_i - \mathbf{x}_{k|k-1} \right] \left[[\boldsymbol{\chi}_{k|k-1}]_i - \mathbf{x}_{k|k-1} \right]^T + \mathbf{Q}_k, \\ \boldsymbol{\gamma}_{k|k-1} = h(\boldsymbol{\chi}_{k|k-1}, \mathbf{u}_k), \\ \mathbf{z}_{k|k-1} = \sum_{i=0}^{2n} W_i [\boldsymbol{\gamma}_{k|k-1}]_i, \\ \mathbf{P}_{z,k|k-1} = \sum_{i=0}^{2n} W_i [[\boldsymbol{\gamma}_{k|k-1}]_i - \mathbf{z}_{k|k-1}] [[\boldsymbol{\gamma}_{k|k-1}]_i - \mathbf{z}_{k|k-1}]^T + \mathbf{R}_k, \\ \mathbf{P}_{xz,k|k-1} = \sum_{i=0}^{2n} W_i \left[[\boldsymbol{\chi}_{k|k-1}]_i - \mathbf{x}_{k|k-1} \right] [[\boldsymbol{\gamma}_{k|k-1}]_i - \mathbf{z}_{k|k-1}]^T, \\ \mathbf{G}_k = \mathbf{P}_{xz,k|k-1} {\mathbf{P}_{z,k|k-1}}^{-1}, \\ \mathbf{x}_k = \mathbf{x}_{k|k-1} + \mathbf{G}_k (\mathbf{z}_k - \mathbf{z}_{k|k-1}), \\ \mathbf{P}_k = \mathbf{P}_{x,k|k-1} - \mathbf{G}_k \mathbf{P}_{z,k|k-1} {\mathbf{G}_k}^T, \end{cases} \tag{23}$$

where $\boldsymbol{\chi}_{k|k-1}$ is the state prediction of sampling points determined by equation (16), $\mathbf{x}_{k|k-1}$ is the weighted mean of the state prediction, W_i is the weight value determined by equation (24), $\mathbf{P}_{x,k|k-1}$ is the weighted variance of the state prediction, $\boldsymbol{\gamma}_{k|k-1}$ is the state observation of sampling points determined by equation (22), $\mathbf{z}_{k|k-1}$ is the weighted mean of the state observation, $\mathbf{P}_{z,k|k-1}$ is the weighted variance of the state observation, $\mathbf{P}_{xz,k|k-1}$ is the weighted variance of the state prediction and observation, $\mathbf{z}_k$ is the sensor measurement value, $\mathbf{G}_k$ is the Kalman gain, $\mathbf{Q}_k$ and $\mathbf{R}_k$ are the noise matrix.

$$W_i = \begin{cases} \dfrac{N}{(N+\lambda)}, & i = 0N, \\ \dfrac{N}{2(N+\lambda)}, & i = 1, \cdots, 2N. \end{cases} \tag{24}$$

4.3. Adaptive Adjustment of the Noise Matrix. It can be seen from equation (23) that the process noise covariance matrix $\mathbf{Q}_k$ and the observation noise covariance matrix $\mathbf{R}_k$ can affect

the confidence level of the data during state prediction and update. In general applications, the variance of system state noise and sensor measurement noise is a fixed value that can be counted in advance. However, in a complex environment, the variance of noise will fluctuate to a certain extent. If the fixed value is still set in advance, it will affect the accuracy of the final fusion state estimation. Therefore, the system identification ability of the RBF network needs to be used to adaptively adjust the noise variance.

The RBF network includes a three-layer structure of input layer, hidden layer, and output layer, as shown in Figure 6. The prediction error $\mathbf{x}_k - \mathbf{x}_{k|k-1}$, estimation error covariance matrix $\mathbf{P}_k$, Kalman gain $\mathbf{G}_k$, and observation error $\mathbf{z}_k - \mathbf{z}_{k|k-1}$ shown in equation (23) can reflect the deviation and fluctuation during the prediction and update process in the UKF and can be used as features of the input layer.

Considering the similarity between the features, select the features according to the position state, attitude state, and the observations of each sensor:

$$\begin{cases} \mathbf{x}_k - \mathbf{x}_{k|k-1} = \begin{bmatrix} \mathbf{e}_{p,k} & \mathbf{e}_{q,k} \end{bmatrix}^T, \\ \mathbf{P}_k = \begin{bmatrix} \mathbf{P}_{pp,k} & \mathbf{P}_{pq,k} \\ \mathbf{P}_{qp,k} & \mathbf{P}_{qq,k} \end{bmatrix}, \\ \mathbf{G}_k = \begin{bmatrix} \mathbf{G}_{s1,k} & \mathbf{G}_{s2,k} & \mathbf{G}_{s3,k} \end{bmatrix}, \\ \mathbf{z}_k - \mathbf{z}_{k|k-1} = \begin{bmatrix} \mathbf{e}_{s1,k} & \mathbf{e}_{s2,k} & \mathbf{e}_{s3,k} \end{bmatrix}^T, \end{cases} \tag{25}$$

where $\mathbf{e}_{p,k}$, $\mathbf{e}_{q,k}$ are the position and attitude prediction error; $\mathbf{P}_{pp,k}$, $\mathbf{P}_{pq,k}$, $\mathbf{P}_{qp,k}$, $\mathbf{P}_{qq,k}$ are the estimation error covariance matrix of position, position-attitude, attitude-position, and attitude; $\mathbf{G}_{s1,k}$, $\mathbf{G}_{s2,k}$, $\mathbf{G}_{s3,k}$ are the Kalman gain of the three sensors; and $\mathbf{e}_{s1,k}$, $\mathbf{e}_{s2,k}$, $\mathbf{e}_{s3,k}$ are the observation error of the three sensors.

Extract the main information of the feature matrix in equation (25) to obtain the input layer feature vector $\mathbf{x}_{input}$, which contains 10 feature information:

$$\begin{aligned} \mathbf{x}_{\text{input}} = \big[& \left\|\mathbf{e}_{p,k}\right\| \quad \left\|\mathbf{e}_{q,k}\right\| \quad tr\left(\mathbf{P}_{pp,k}\right) \quad tr\left(\mathbf{P}_{qq,k}\right) \quad \left\|\mathbf{G}_{s1,k}\right\| \\ & \cdot \left\|\mathbf{G}_{s2,k}\right\| \left\|\mathbf{G}_{s3,k}\right\| \left\|\mathbf{e}_{s1,k}\right\| \left\|\mathbf{e}_{s2,k}\right\| \left\|\mathbf{e}_{s3,k}\right\| \big]^T. \end{aligned} \tag{26}$$

From input layer to hidden layer, radial basis function is used to activate neurons:

$$\phi_i\left(\mathbf{x}_{\text{input}}\right) = e^{\left\|\mathbf{x}_{\text{input}} - \mu_i\right\| / {\sigma_i}^2}, \tag{27}$$

where μ_i, σ_i are the center and width of the i-th hidden layer neuron.

The linear function is used from hidden layer to output layer:

$$\mathbf{y}_{\text{output},j} = \sum_i \omega_{ij} \phi_i\left(\mathbf{x}_{\text{input}}\right), \tag{28}$$

where ω_{ij} is the weight of the i-th hidden layer neuron to the j-th output layer neuron.

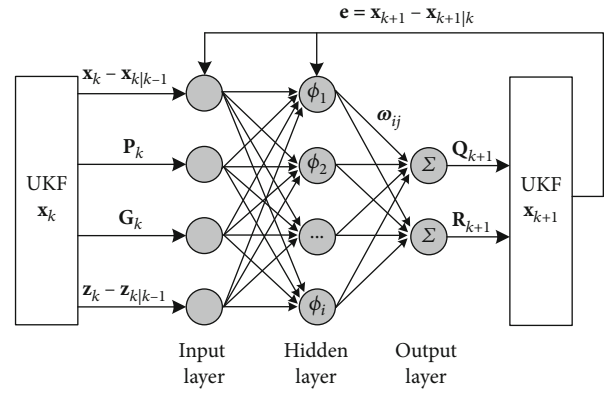


Figure 6: RBF network for adjusting noise matrix.

The output of the output layer is as follows:

$$\mathbf{y}_{\text{output}} = \begin{bmatrix} \sigma_p & \sigma_q & \sigma_{s1} & \sigma_{s2} & \sigma_{s3} \end{bmatrix}^T, \tag{29}$$

$$\begin{cases} \mathbf{Q}_{k+1} = \mathbf{I}\begin{bmatrix} \sigma_p & \sigma_q \end{bmatrix}^T, \\ \mathbf{R}_{k+1} = \mathbf{I}\begin{bmatrix} \sigma_{s1} & \sigma_{s2} & \sigma_{s3} \end{bmatrix}^T, \end{cases} \tag{30}$$

where σ_p and σ_q are the standard deviations of the process noise of the position and attitude and σ_{s1}, σ_{s2}, and σ_{s3} are the standard deviations of the observation noise of the three sensors.

The training process of the network is divided into two stages: offline and online. The offline training stage uses the data with known noise variance to obtain the initial neuron center and connection weights. At this stage, the output error of the network is the difference between the true noise variance and the network predicted noise variance. Since the noise variance in the offline stage is not the noise variance in the actual experiment, it is easy to produce overfitting in the offline training stage, which needs to be corrected through the online stage. The online training stage uses the updated prediction error $\mathbf{x}_{k+1} - \mathbf{x}_{k+1|k}$ at the next time as the output error of the network. The training goal is to minimize the prediction error:

$$\text{Loss} = \arg\min_{\mu_i, \omega_{ij}} \left\|\mathbf{x}_{k+1} - \mathbf{x}_{k+1|k}\right\|. \tag{31}$$

5. Experiments and Results

5.1. Experimental Device. The mechanical structure of the experimental device consisted of a 1-DOF motion platform and an 8-DOF redundant manipulator, as shown in Figure 7. The joint encoders were arranged at each rotary joint, and the RGB-D camera and the MARG sensor were arranged at the end of the manipulator. The RGB-D camera was connected to the industrial personal computer (IPC) through USB, and the MARG sensor and the joint encoders were connected to the sensor collector through the 485 bus and then connected to the IPC through USB.

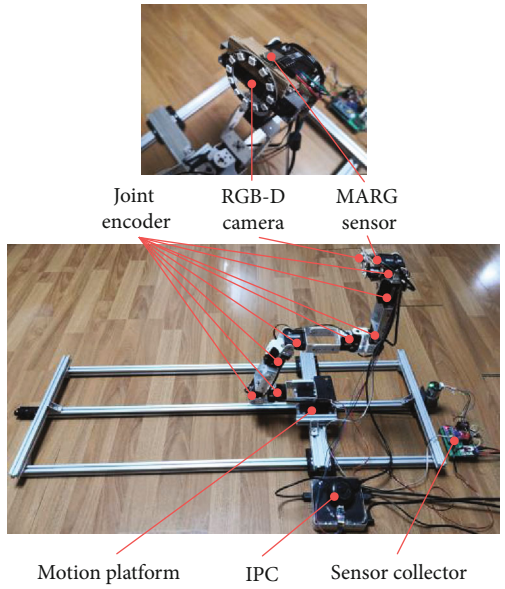


Figure 7: Experimental device.

5.2. Experiment of the Enhancement Part. In order to test the effect of the sensor complementary enhancement method proposed in this article, the experimental device was placed in a narrow artificially built environment, considering that the RGB-D camera focuses more on static repeatability and the MARG sensor focuses more on dynamic stability. Therefore, when testing the enhancement effect of the RGB-D camera, a periodically repeated expected motion trajectory was used, while the MARG sensor used a random expected trajectory. The experimental results are shown in Figures 8 and 9.

Comparing Figures 8(a) and 8(b), it can be seen that the point cloud after enhancement is closer to the real environment and has a less invalid blind area caused by the narrow environment. Figure 8(c) further demonstrates the impact of depth image enhancement on the accuracy of position estimation. It can be seen from the figure that the position trajectory obtained by using the original image information fluctuates more drastically and deviates from the reference trajectory to a greater degree. But the position trajectory obtained by using the enhanced image information is closer to the reference trajectory, and the overall fluctuation is smaller, which is more conducive to the fusion pose estimation with other sensors.

Comparing the desired and enhanced quaternion attitude in Figure 9, the overall calculation error of the QEKF algorithm proposed in this paper is within the expected range, which means acceleration information can compensate for the integral drift of angular velocity and geomagnetic field information can calibrate angle measurement errors. Through the enhancement of attitude calculation, it can provide more reliable attitude data for multisensor fusion.

5.3. Simulation of the Fusion Part. Because the statistical tests are very sensitive when the sample size is small, both simulation tests and actual experiments were carried out to complete the verification of the multisensor fusion pose estimation method. Simulation tests were mainly used to test the stability of the method under extreme conditions, and actual experiments were mainly used to test the fusion effect of the method in real environments. It was easy to adjust the sensor parameters in simulation environments, so a large number of samples can be generated to test the performance of the method. Although the experimental samples that were collected in the real environments were limited, these samples were more targeted and reliable after the screening of the experimental methods by the simulation tests. At the same time, to present the experimental results as objectively as possible, the root mean square (RMS) errors of multiple experimental results were used to evaluate the performance differences of different methods.

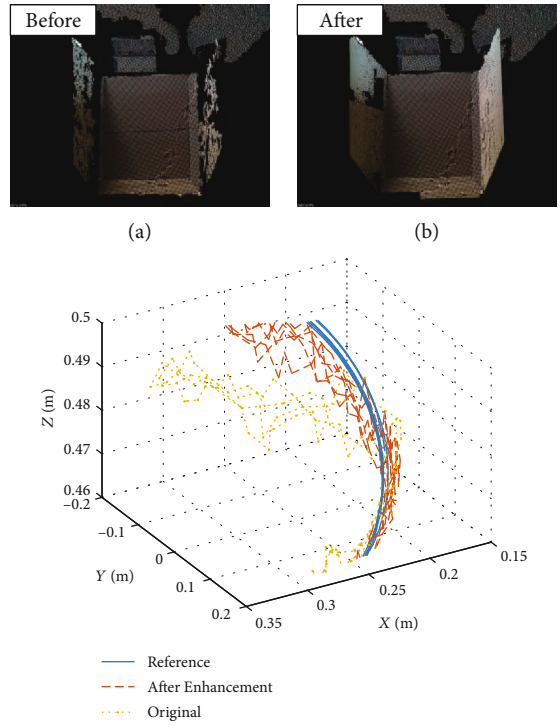


Figure 8: Results of depth image enhancement.

Since real sensors can only be simulated numerically under simulation conditions, the actual state and sensors were simplified during the simulation. Different from the real system, an observation state and two sensors were set up in the simulation environment. The observation noise of sensor 1 was smaller than that of sensor 2, and sensor 1 only produced a stable zero offset, while the zero offset of sensor 2 increased with time. Figure 10 and Table 1 show the effect of the synchronization update method on the fusion result.

Since the noise of sensor 2 is greater than that of sensor 1, the RMS error of sensor 2 is greater than that of sensor 1 when the state of a single sensor is estimated. In the simulation, the data update frequency f_1 of sensor 1 was different from the data update frequency f_2 of sensor 2. If the data synchronization of the sensor was not considered, the state update was performed when the data of any sensor inputted.

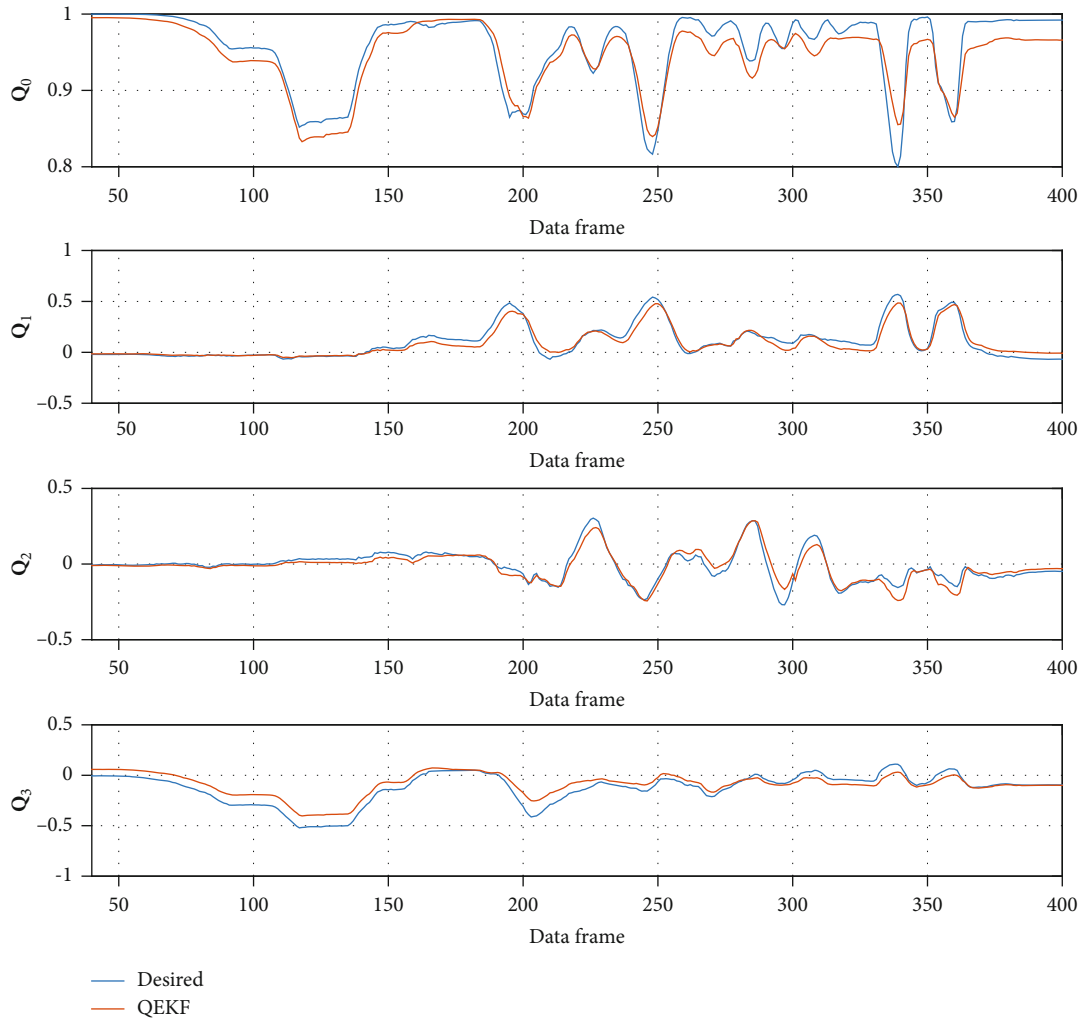


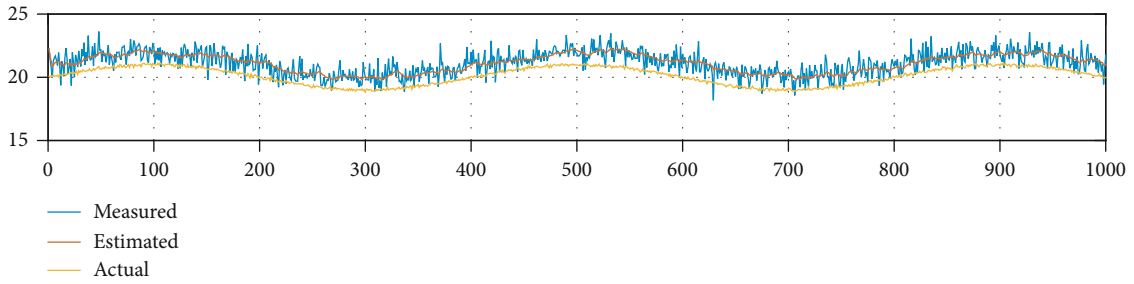
Figure 9: Results of attitude calculation enhancement.

From the experimental results shown in Figures 10(c) and 10(d), the RMS error when $f_1 > f_2$ is less than that when $f_1 < f_2$, indicating that the fusion result is affected by the sensor update frequency, and it is biased towards high-sampling-rate data sources. If the high-sampling-rate data is noisy, the fusion accuracy cannot be significantly improved. The above defects are improved in the modified synchronization update method. It can be seen from the experimental results shown in Figure 10(e) that the RMS error of the fusion data is significantly reduced, and there will be no sawtooth fluctuations in the asynchronous fusion. Therefore, the data synchronization of sensors is an important factor in multirate sensor fusion.

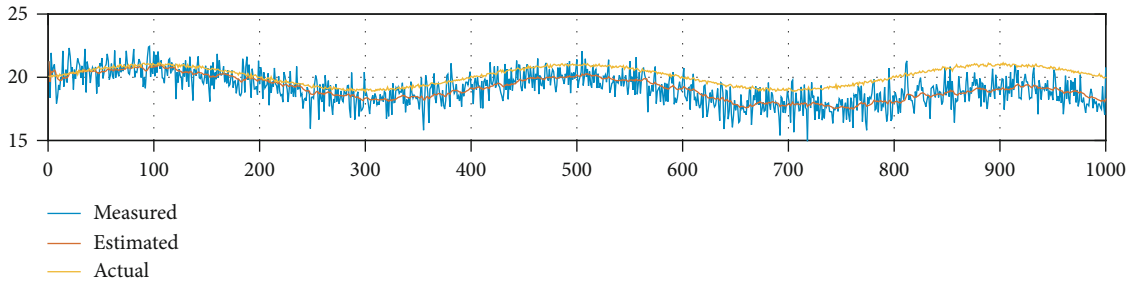
In the simulation environment, the process and observation noise conformed to the Gaussian distribution, but the variance of the distribution changed at every moment and satisfied a nonlinear model. The RBF network was designed to predict the variances of the process and observation noise and was trained when the filter was updated. In the simulation, the variances predicted by the network in the test set were compared with the preset variances, and the coefficient of determination (marked as R-square) was calculated. The closer the coefficient of determination is to 1, the higher the accuracy of the network. It can be seen from the results shown in Figure 11 that the network can predict the variances of the noise accurately and has no low coefficients of determination (0.71 for the process noise and 0.84 for the observation noise).

Although the coefficient of determination cannot be obtained when the true variance is unknown, the accuracy of the network can be indirectly reflected through the state estimation deviation. In other words, if the state estimation deviation can remain stable even in extreme cases, it means that the network's prediction of the parameters is accurate. In the simulation, the process noise and the observation noise of the sensor were designed to increase with time. Figure 12 and Table 2 show the fusion effects of different fusion algorithms in this case.

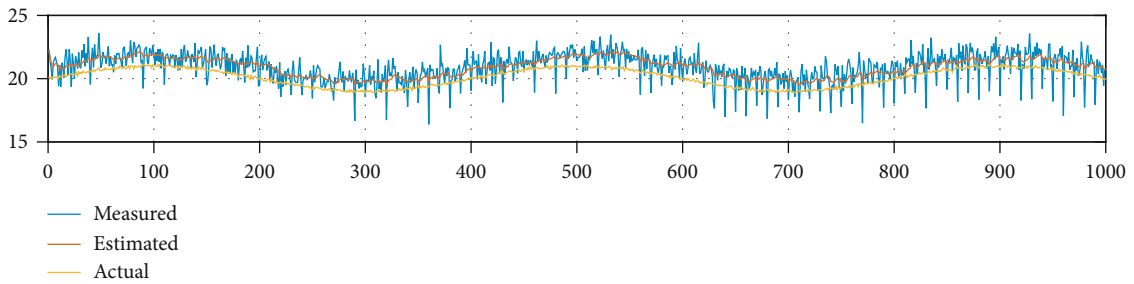
It can be seen from the experimental results that at the beginning, the system noise is small and the estimation error of each fusion method is not much different, but as time goes by, the system noise increases. The original EKF



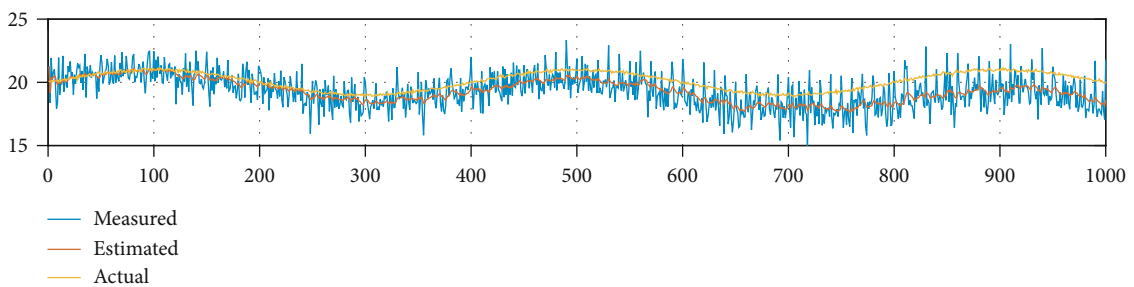
(a) Single sensor 1



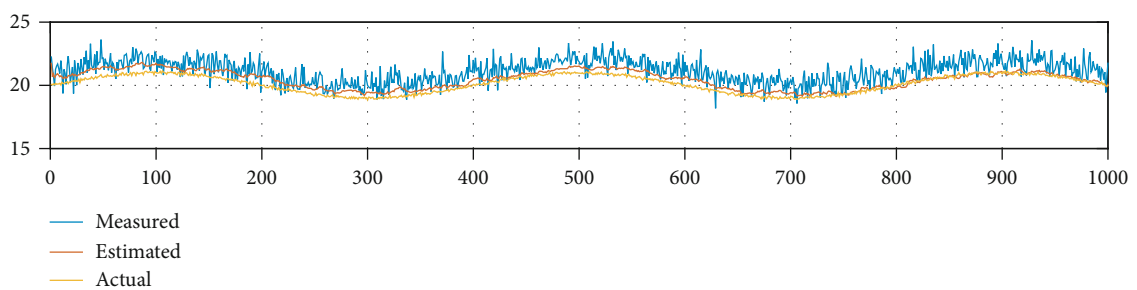
(b) Single sensor 2



(c) Sensor 1 and 2 update frequency: $f_1 > f_2$



(d) Sensor 1 and 2 update frequency: $f_1 < f_2$



(e) Sensor 1 and 2 synchronized

Figure 10: The impact of synchronization.

Table 1: RMS errors of different update methods.

	Update method				
	Sensor1 Single	Sensor2 Single	Sensors 1 and 2 $f_1 > f_2$	Sensors 1 and 2 $f_1 > f_2$	Sensors 1 and 2 Synchronized
RMS error	1.02	1.16	0.824	0.962	0.413

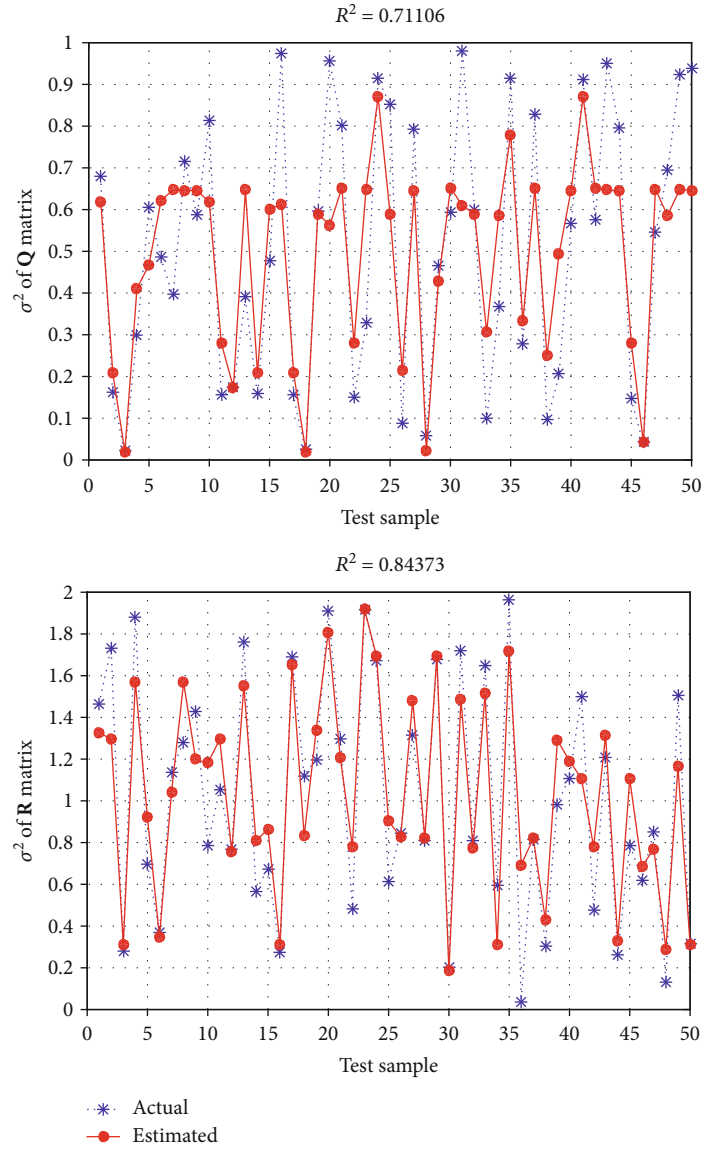


Figure 11: Prediction of the noise matrix.

and UKF algorithms cannot adaptively adjust the filter parameters, which are easily affected by noise and produce state estimation deviations. Through the RBF network proposed in this paper, the noise covariance matrix can be predicted, so that the EKF and UKF algorithms have adaptive adjustment capabilities. The modified RBF-AEKF and RBF-AUKF methods have smaller RMS errors than the original EKF and UKF methods, so they can be more robust in complex environments.

5.4. Experiment of the Fusion Part. Since the environment established in the simulation can only approximately reflect the actual situation, it is still necessary to conduct experiments in the real environment. In actual experiments, the end-effector of the manipulator periodically moved around the Z-axis of the base coordinate system, and the pose of the end-effector was set as the desired value. The position estimations p_x, p_y, p_z along the X, Y, and Z axes were obtained by fusion of the visual odometer based on the

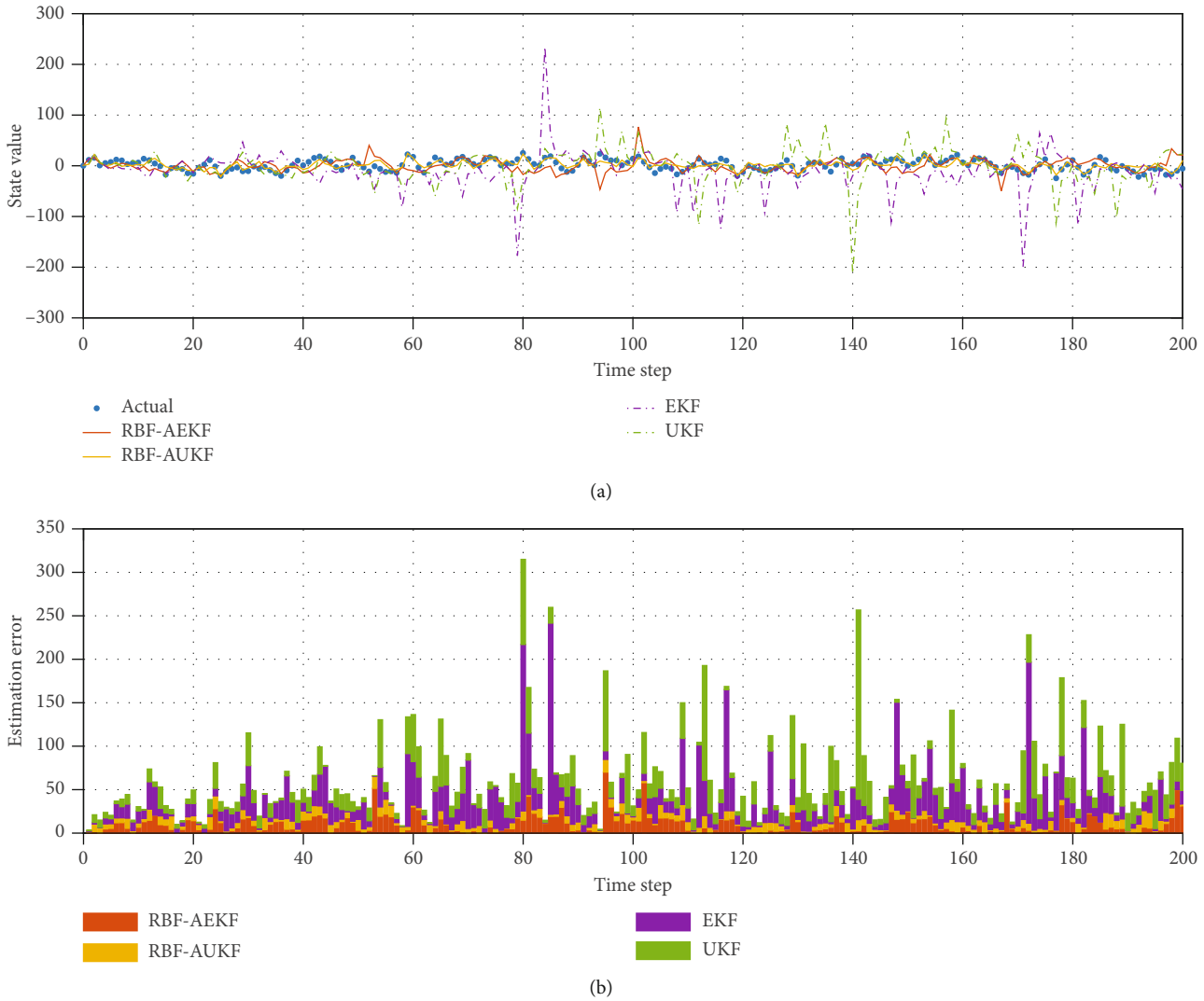


Figure 12: The fusion effects of different fusion algorithms.

Table 2: RMS errors of different fusion methods.

	Fusion method			
	EKF	UKF	RBF-AEKF	RBF-AUKF
RMS error	39.46	33.42	14.98	7.863

RGB-D camera and the kinematics calculation based on joint encoders. And the quaternion attitude estimations q_0, q_1, q_2, q_3 were obtained by fusion of the MARG sensor, visual odometer, and encoders. The experiment compares the original UKF method and the modified adaptive UKF method proposed in this paper, and the results are shown in Figures 13 and 14 and Table 3.

It can be seen from Table 3 that the RMS error of the pose estimation after multisensor fusion is smaller than that of a single sensor. Compared with that of the original method, the RMS error of the modified method proposed in this paper is significantly reduced in position estimation, but there is little difference in attitude estimation. The main reason is the modified method proposed in this paper is better than the traditional method due to its better adaptive effect when the noise fluctuation is large. In this experiment, the overall fluctuation of the sensor in attitude measurement is smaller than that of position measurement, so the modified method is close to the original method in attitude estimation.

It can be seen from Figure 13 that the visual odometer is prone to local sudden changes due to environmental interference and the encoder is prone to measurement offset. Compared with the original method, the modified method handles the measurement data of local sudden change better, so the overall RMS error is smaller. And because of the combination of absolute and relative updates, the overall update rate is maintained. In this experiment, the original method was updated only 258 times, while the modified method was updated 479 times, so the real-time response rate of state estimation was improved. It can be seen from Figure 14 that the rotation angle of the desired trajectory on the X and Y axes is small, and the rotation angle on the Z-axis is large. When the rotation angle is small, it can be seen that the sensor has obvious measurement drift, when the rotation

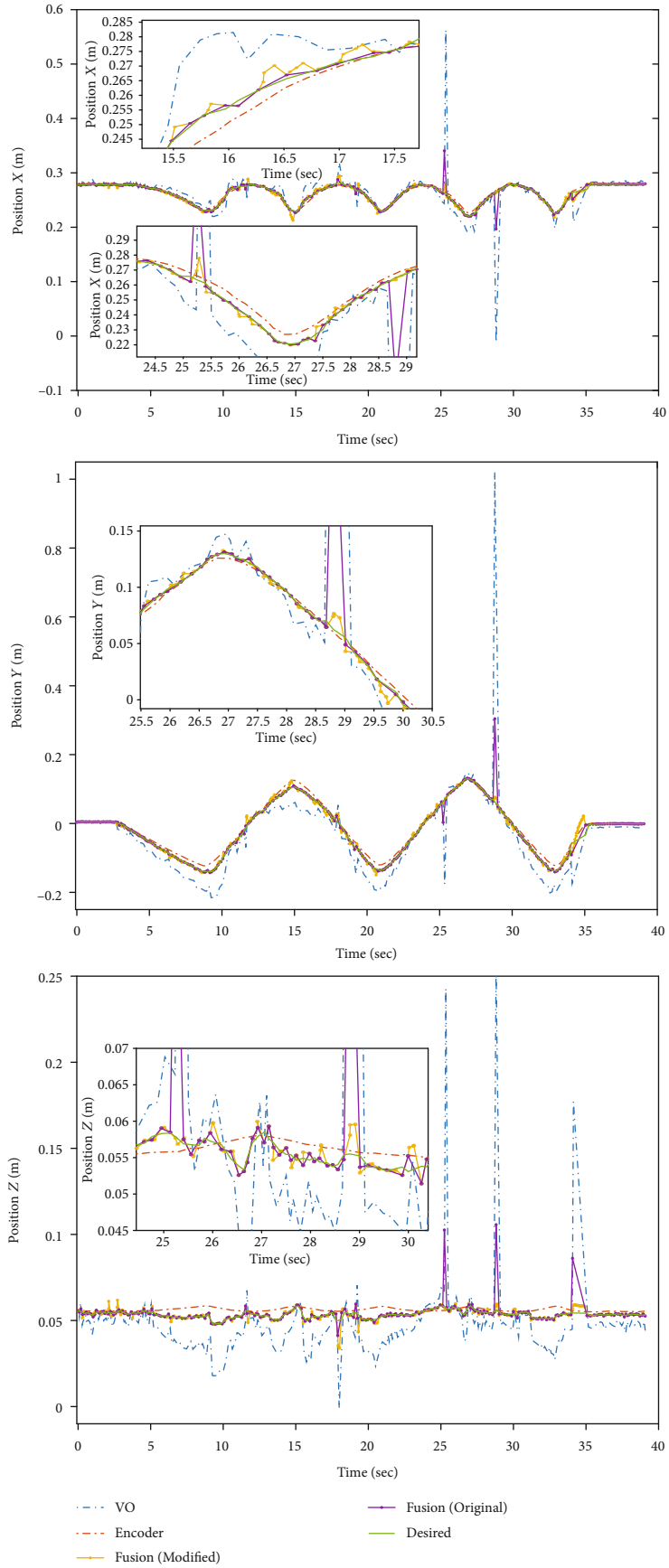


Figure 13: Results of position estimation.

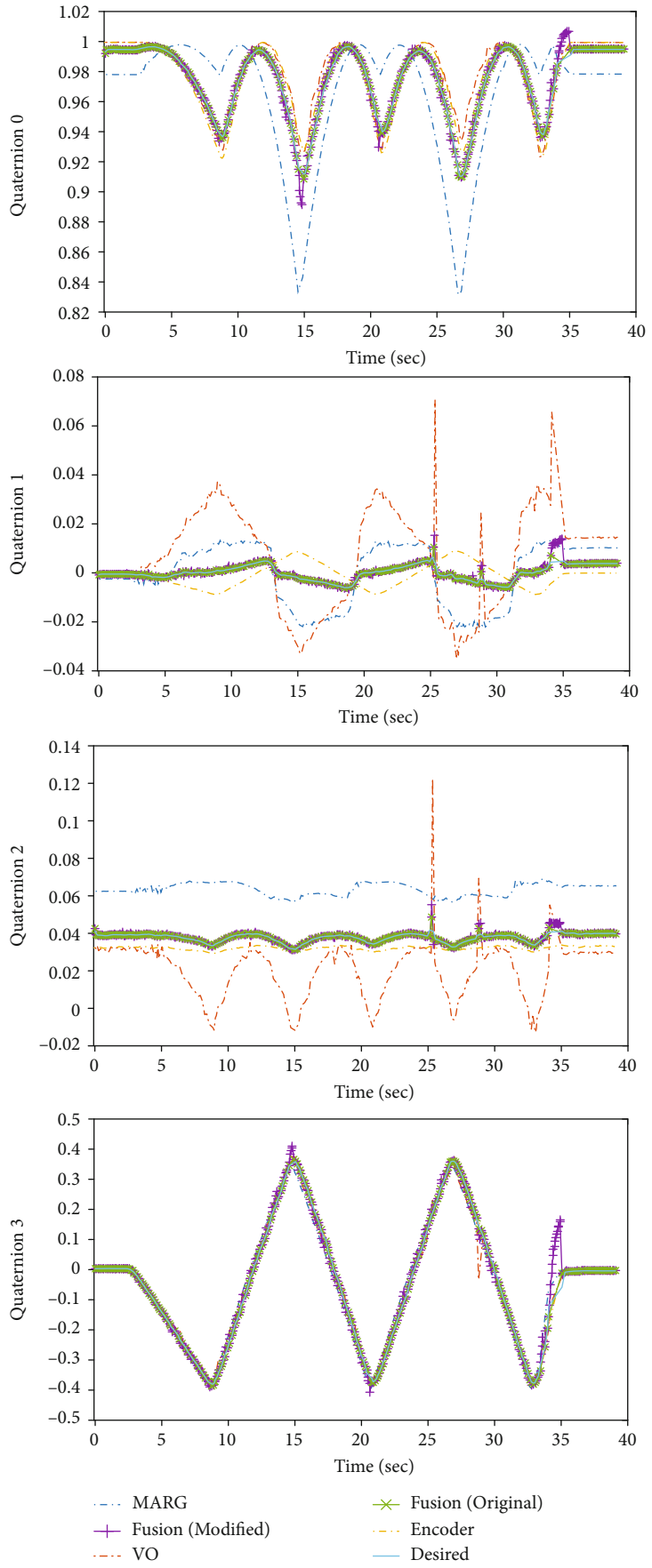


Figure 14: Results of attitude estimation.

Table 3: RMS errors of the original method and modified method.

RMS error	Estimated variable						
	p_x (cm)	p_y (cm)	p_z (cm)	q_0	q_1	q_2	q_3
MARG	—	—	—	0.319	0.101	0.260	0.273
VO	2.76	7.02	2.19	0.109	0.194	0.220	0.232
Encoder	0.303	1.08	0.360	0.0933	0.0624	0.0584	0.133
Fusion (modified)	0.175	0.364	0.108	0.00805	0.00564	0.00645	0.0482
Fusion (original)	0.663	1.59	0.475	0.00805	0.00564	0.00645	0.0482

angle is large, it is easy to produce local sudden changes. The modified method in this paper can compensate the measurement drift of each sensor, reduce the influence of local sudden changes on the state estimation, and thus can better follow the desired trajectory.

6. Conclusions

In order to improve the accuracy of the end-effector pose estimation, this paper proposes a modified method based on multisensor fusion, including preenhancement and adaptive fusion. The following conclusions can be drawn from the experimental results. Firstly, the depth image obtained by the RGB-D camera and the attitude obtained by the MARG sensor is enhanced, which can make the multisensor fusion process in a complex environment more stable. Secondly, the synchronization of sensors has a great influence on the results of multisensor fusion. The control variable is relatively updated when asynchronous, and the state variable is absolutely updated when synchronous, which can balance the update frequency and fusion accuracy. Thirdly, predicting the noise matrix through a neural network can make the algorithm adjust the fusion ratio of each sensor adaptively according to changes in the environment, which helps improve robustness. Finally, in the simulation environment, the RMS errors of the original UKF and EKF method are 4.25 times and 5.02 times that of the modified RBF-UKF method, respectively, and in the actual environment, the average pose estimation errors of the single sensor and the original UKF method are 8.57 times and 3.91 times that of the modified RBF-UKF method, and the update frequency of the original method is only 54% of the modified method. In summary, the method proposed in this paper can improve the accuracy of end-effector pose estimation and has high application value and practical engineering value.

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 that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1307400 and in part by the National Natural Science Foundation of China under Grant 61873267.

References

[1] G. S. Chirikjian and J. W. Burdick, "The kinematics of hyper-redundant robot locomotion," *IEEE Transactions on Robotics and Automation*, vol. 11, no. 6, pp. 781–793, 1995.

[2] M. H. Korayem and H. Ghariblu, "Maximum allowable load on wheeled mobile manipulators imposing redundancy constraints," *Robotics and Autonomous Systems*, vol. 44, no. 2, pp. 151–159, 2003.

[3] M. H. Korayem, V. Azimirad, A. Nikoobin, and Z. Boroujeni, "Maximum load-carrying capacity of autonomous mobile manipulator in an environment with obstacle considering tip over stability," *International Journal of Advanced Manufacturing Technology*, vol. 46, no. 5-8, pp. 811–829, 2010.

[4] X. Dong, D. Axinte, D. Palmer et al., "Development of a slender continuum robotic system for on-wing inspection/repair of gas turbine engines," *Robotics and Computer-Integrated Manufacturing*, vol. 44, pp. 218–229, 2017.

[5] J. Q. Peng, W. F. Xu, T. L. Liu, H. Yuan, and B. Liang, "End-effector pose and arm-shape synchronous planning methods of a hyper-redundant manipulator for spacecraft repairing," *Mechanism and Machine Theory*, vol. 155, p. 25, 2021.

[6] F. Renda, M. Giorelli, M. Calisti, M. Cianchetti, and C. Laschi, "Dynamic model of a multibending soft robot arm driven by cables," *IEEE Transactions on Robotics*, vol. 30, no. 5, pp. 1109–1122, 2014.

[7] M. Rolf and J. J. Steil, "Efficient exploratory learning of inverse kinematics on a bionic elephant trunk," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 6, pp. 1147–1160, 2014.

[8] C. C. Cheah, M. Hirano, S. Kawamura, and S. Arimoto, "Approximate Jacobian control for robots with uncertain kinematics and dynamics," *IEEE Transactions on Robotics and Automation*, vol. 19, no. 4, pp. 692–702, 2003.

[9] B. Xiao, S. Yin, and O. Kaynak, "Tracking control of robotic manipulators with uncertain kinematics and dynamics," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 10, pp. 6439–6449, 2016.

[10] M. H. Korayem, A. M. Shafei, F. Absalan, B. Kadkhodaei, and A. Azimi, "Kinematic and dynamic modeling of viscoelastic robotic manipulators using Timoshenko beam theory: theory and experiment," *International Journal of Advanced*

Manufacturing Technology, vol. 71, no. 5-8, pp. 1005–1018, 2014.

[11] C. T. Lin, W. J. Zhang, and H. Yuan, "Investigation on the tip positioning accuracy of cable-driven serpentine manipulators," *Applied Sciences-Basel*, vol. 10, no. 20, p. 17, 2020.

[12] N. Riehl, M. Gouttefarde, S. Krut, C. Baradat, and F. Pierrot, "Effects of non-negligible cable mass on the static behavior of large workspace cable-driven parallel mechanisms," in *IEEE International Conference on Robotics and Automation ICRA*, p. 2503, Kobe, Japan, 2009.

[13] A. Fortin-Cote, P. Cardou, and A. Campeau-Lecours, "Improving cable driven parallel robot accuracy through angular position sensors," in *2016 Ieee/Rsj International Conference on Intelligent Robots and Systems (Iros 2016)*, pp. 4350–4355, Daejeon, South Korea, 2016.

[14] T. Dallej, M. Gouttefarde, N. Andreff, P. E. Herve, and P. Martinet, "Modeling and vision-based control of large-dimension cable-driven parallel robots using a multiple-camera setup," *Mechatronics*, vol. 61, pp. 20–36, 2019.

[15] X. Ma, P. W.-Y. Chiu, and Z. Li, "Shape sensing of flexible manipulators with visual occlusion based on Bezier curve," *IEEE Sensors Journal*, vol. 18, no. 19, pp. 8133–8142, 2018.

[16] M. H. Korayem, M. Yousefzadeh, and S. Kian, "Precise end-effector pose estimation in spatial cable-driven parallel robots with elastic cables using a data fusion method," *Measurement*, vol. 130, pp. 177–190, 2018.

[17] Z. Bing, L. Cheng, A. Knoll, A. Thong, K. Huang, and F. Zhang, "Slope angle estimation based on multi-sensor fusion for a snake-like robot," in *2017 20th International Conference on Information Fusion (Fusion)*, Xian, People's Republic of China, 2017.

[18] Z. Gao and S. X. Ding, "Sensor fault reconstruction and sensor compensation for a class of nonlinear state-space systems via a descriptor system approach," *IET Control Theory and Applications*, vol. 1, no. 3, pp. 578–585, 2007.

[19] J. Kelly and G. S. Sukhatme, "Visual-Inertial sensor fusion: localization, mapping and sensor-to-sensor self-calibration," *International Journal of Robotics Research*, vol. 30, no. 1, pp. 56–79, 2011.

[20] S. A. Khalilpour, R. Khorrambakht, H. Damirchi, H. D. Taghirad, and P. Cardou, "Tip-trajectory tracking control of a deployable cable-driven robot via output redefinition," *Multibody System Dynamics*, vol. 4, pp. 1–28, 2020.

[21] A. Kuzdeuov, M. Rubagotti, and H. A. Varol, "Neural network augmented sensor fusion for pose estimation of tensegrity manipulators," *IEEE Sensors Journal*, vol. 20, no. 7, pp. 3655–3666, 2020.

[22] H. Sarbolandi, D. Lefloch, and A. Kolb, "Kinect range sensing: structured-light versus time-of-flight kinect," *Computer Vision and Image Understanding*, vol. 139, pp. 1–20, 2015.

[23] A. M. Sabatini, "Quaternion-Based extended Kalman filter for determining orientation by inertial and magnetic sensing," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 7, pp. 1346–1356, 2006.

[24] T. Moore and D. Stouch, "A generalized extended Kalman filter implementation for the robot operating system," in *Intelligent Autonomous Systems 13, Advances in Intelligent Systems and Computing*, E. Menegatti, N. Michael, K. Berns, and H. Yamaguchi, Eds., pp. 335–348, Springer, 2016.

[25] E. A. Wan and R. van der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*, pp. 153–158, Mcmaster Univ, Lake Louise, Canada, 2000.

[26] A. Smyth and M. L. Wu, "Multi-rate Kalman filtering for the data fusion of displacement and acceleration response measurements in dynamic system monitoring," *Mechanical Systems and Signal Processing*, vol. 21, no. 2, pp. 706–723, 2007.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5582646, 12 pages
https://doi.org/10.1155/2021/5582646

Research Article

An Improved Adaptive Clone Genetic Algorithm for Task Allocation Optimization in ITWSNs

Zhihua Zha,[1] Chaoqun Li,[2] Jing Xiao,[2] Yao Zhang,[3] Hu Qin,[2] Yang Liu,[2] Jie Zhou,[2] and Jie Wu[1]

[1]*Mechanical and Electrical Engineering, Shihezi University, Shihezi 832000, China*
[2]*College of Information Science and Technology, Shihezi University, Shihezi 832000, China*
[3]*University of the Cordilleras, Baguio City 2600, Philippines*

Correspondence should be addressed to Jie Wu; wjshz@126.com

Received 18 February 2021; Revised 13 March 2021; Accepted 17 March 2021; Published 7 April 2021

Academic Editor: Bin Gao

Research on intelligent transportation wireless sensor networks (ITWSNs) plays a very important role in an intelligent transportation system. ITWSNs deploy high-yield and low-energy-consumption traffic remote sensing sensor nodes with complex traffic parameter coordination on both sides of the road and use the self-organizing capabilities of each node to automatically establish the entire network. In the large-scale self-organization process, the importance of tasks undertaken by each node is different. It is not difficult to prove that the task allocation of traffic remote sensing sensors is an NP-hard problem, and an efficient task allocation strategy is necessary for the ITWSNs. This paper proposes an improved adaptive clone genetic algorithm (IACGA) to solve the problem of task allocation in ITWSNs. The algorithm uses a clonal expansion operator to speed up the convergence rate and uses an adaptive operator to improve the global search capability. To verify the performance of the IACGA for task allocation optimization in ITWSNs, the algorithm is compared with the elite genetic algorithm (EGA), the simulated annealing (SA), and the shuffled frog leaping algorithm (SFLA). The simulation results show that the execution performance of the IACGA is higher than EGA, SA, and SFLA. Moreover, the convergence speed of the IACGA is faster. In addition, the revenue of ITWSNs using IACGA is higher than those of EGA, SA, and SFLA. Therefore, the proposed algorithm can effectively improve the revenue of the entire ITWSN system.

1. Introduction

Nowadays, with the rapid increase of vehicles, the phenomenon of traffic congestion and pollution is getting worse, which leads to frequent traffic violations and accidents. These have become bottlenecks for the further development of cities [1]. Therefore, it is urgent for the transportation department to apply a more advanced and intelligent data acquisition means to obtain the massive data of the transportation industry. Transportation departments can provide real-time and accurate traffic information services for passengers and provide a reference for staff to deal with emergencies and traffic violations [2, 3], for example, detect vehicles at intersections in all directions to improve simplification, signal control algorithms, and traffic efficiency based on the monitoring results. At the same time, the growth of existing roads and other hardware facilities can no longer satisfy the ever-increasing traffic problems, and intelligent transportation systems are getting more and more attention [4, 5].

An intelligent transportation system mainly includes the collection, transmission, control, and guidance of traffic information. A wireless sensor network can provide an effective means for information collection and transmission of the intelligent transportation system. The effective operation of the intelligent transportation system depends on obtaining comprehensive, accurate, and real-time dynamic traffic information. People process the information collected by sensor nodes to obtain comprehensive traffic condition, which facilitates identification, decision-making, positioning, detection, and tracking of vehicles in the traffic management. Therefore, information collection has become a hot issue in the development of intelligent transportation [6]. The traffic

conditions of different road sections are also different, and the tasks assigned by the sensors are also different. To better detect the traffic condition in the case of limited traffic sensor resources, an efficient task allocation strategy can collect traffic to a greater extent information for monitoring traffic conditions more effectively [7–9]. The focus of the research is to develop new task allocation schemes to maximize the network throughput and revenue of intelligent transportation wireless sensor networks (ITWSNs).

In the whole development process of ITWSNs, efficient QoS task allocation strategies have played an irreplaceable important role and have been widely used in various fields of transportation [10–12]. Scholars have extensively studied the task allocation problem in wireless sensor networks [13–15]. Paper [16] finds the key subtasks based on the estimated completion time of the subtasks and the weight coefficients and preferentially selects node assignments with strong capabilities and high processing efficiency. Paper [17] mixes the adaptability of particle swarm optimization with the flexible ability of dynamic alliance and obtains the fitness value through the weighting method to obtain the global optimal allocation method. Paper [18] allocates tasks to different clusters to achieve the goal of high benefit and then allocates tasks from the clusters to appropriate sensor nodes to balance the energy loss of the network. Paper [19] proposes a dynamic joint task allocation algorithm using linear programming to obtain a more balanced task allocation strategy. However, none of the above methods can better improve the revenue of task allocation in ITWSNs.

Aiming at the problem of maximizing revenue, the task allocation optimization technology is applied to ITWSNs, and it can greatly improve the overall revenue of the network. The main contributions are as follows:

(1) Firstly, this paper proposes an improved adaptive clone genetic algorithm (IACGA) to solve the optimization problem of task allocation, designs the task allocation model of ITWSNs, and designs a new fitness function to evaluate the performance of the algorithm

(2) Secondly, new adaptive operators and clone operators are designed to improve the optimization ability of the algorithm. IACGA combines adaptive operator and clonal operator, which has better performance, enhances global search ability, and avoids falling into local optimum

(3) Finally, the simulation results of IACGA, EGA, SA, and SFLA in ITWSN task allocation are compared to verify the superiority of IACGA in task allocation optimization, and the detailed data and discussion are given

The remaining structure of the paper is shown below. Section 2 introduces related research in the field of task allocation in intelligent transportation. Section 3 shows the task allocation model. Section 4 proposes an improved adaptive clone genetic algorithm to solve the problem of task allocation in ITWSNs. Section 5 illustrates the effectiveness of IACGA in solving the task allocation problem through simulation experiments and discusses it. Section 6 is the conclusion part.

2. Related Work

Task allocation is a classic problem widely studied in the field of ITWSNs, and its application in the field of wireless sensor networks in the intelligent transportation system is also crucial. However, wireless sensor network resources are severely limited, and existing algorithms cannot be directly applied. Paper [20] proposed a nested optimization technology based on a genetic algorithm to perform energy-efficient task allocation in a multihop cluster network. Generalized optimization goals can not only meet the real-time requirements of the application but also achieve energy efficiency. The optimization solution is obtained by combining the processes of task mapping, routing path allocation, and task scheduling based on genetic algorithms. The task graph simulation experiment is randomly generated, and the results show that the nested optimization technology has better performance than the random optimization technology. However, due to the high complexity of the algorithm, the efficiency of the program is not good.

In paper [21], for the task allocation problem of the urban road traffic information collection sensor network for the collaborative collection of complex traffic parameters, the sensor network is mapped to a multiagent system, and the task completion time, node energy consumption, and network load balance are used as the evaluation function. The authors used alliance-based collaborative methods to construct a nonlinear multiobjective optimization model of sensor network task allocation. The authors used genetic simulated annealing to search for the optimal alliance structure to achieve task allocation strategy optimization. Simulation experiments are carried out in the actual scene of road traffic information collection; the results show that the genetic simulated annealing can effectively optimize the alliance structure of task allocation. Compared with other optimization algorithms, the optimized model has low fitness function value, short task completion time, and low network energy consumption. This method can be used for traffic-oriented collaborative detection task allocation problems of the information collection sensor network. But the algorithm has slow convergence speed and poor performance.

Paper [22] proposed a hierarchical optimization task scheduling algorithm based on the characteristics of wireless sensor network task scheduling subject to deadlines and node energy constraints. In the paper, tasks are prioritized according to the threshold established by the deadline. Tasks with more urgent deadlines are assigned first to improve the success rate of scheduling. For tasks with loose deadlines, the goal is to reduce energy consumption and balance the load to increase the network revenue. Simulation experiments show that the algorithm has achieved better results in improving the success rate of task scheduling and balancing network load. However, in the case of limited computing power, the computational complexity of the algorithm increases exponentially with the increase in the number of network nodes.

Regarding the task allocation algorithm of intelligent transportation, there is a lot of progress in research. However, the existing research rarely involves complex perception tasks that require multiperson collaboration in group intelligence perception. Paper [23] studies this type of task. First, a

location-related task allocation problem for collaborative group intelligence perception is presented, and a formal analysis is carried out on it; then, it is proved that the problem is NP-hard to solve, and a greedy strategy and strategy based on this problem are proposed. However, this algorithm has high complexity and poor performance.

In the paper [24], the authors used the SFLA to explore the task allocation problem. Under the premise of task allocation modeling, the improved SFLA is used to solve the model. First, according to the characteristics of the target, the model solution matrix is designed by using a decimal encoding method. From this matrix, the sensor decision matrix can be directly obtained. On the basis of the original SFLA, by introducing a variable step size related to the number of iterations, the execution process of the algorithm is converted from a multipoint mutation mode to a single-point mutation mode, so that the algorithm has stronger robustness. Under the premise of satisfying the task allocation constraints, specific solutions and steps are given according to the principle of the algorithm. The feasibility and effectiveness of the algorithm are verified by example simulation. However, the algorithm is prone to premature convergence.

In the paper [25], the authors established a mathematical model of the road condition target allocation problem through the analysis of the factors affecting the target allocation in the road, and then, based on the idea of a hybrid optimization algorithm, combined the heuristic search mechanism with the simulated annealing, and proposed an improved greedy simulated annealing algorithm. The simulation results show that the algorithm is effective and feasible and can give a better optimal allocation plan. However, it is still easy to fall into a local optimal solution.

In view of the problems existing in the above literature, we have proposed a new solution to solve the problem of task allocation optimization and maximize the overall benefits of the ITWSN system, thereby reducing costs and improving benefits for the collection of information for intelligent transportation systems.

3. System Model

In a complex traffic information collection environment, in order to maximize the network benefits of ITWSNs, a mathematical model of ITWSNs is designed for the constraints of control range and computing power.

This paper proposes a task allocation model for ITWSNs located in the coordinates of the traffic area. This model can realize the continuous coordination of heterogeneous agents between groups under communication constraints, thereby maximizing the network revenue of wireless sensor networks. This model can be simply abstracted into N tasks and S traffic remote sensing sensor nodes in ITWSNs. Suppose S sensor nodes are randomly installed in the road traffic area to perform N tasks of road traffic information collection. The goal of this paper is to obtain the maximum network revenue value by assigning different tasks on different sensor nodes.

It is assumed that the advantages of each task are evaluated before task assignment, and the urgency of task assignment is estimated. In formula (1), the urgency of the n_{th} task assigned to the sensor can be represented by u_n, and the advantage of the n_{th} task assigned to the s_{th} sensor can be represented by $p_{s,n}$. From this, the revenue value of the sensor node performing the task can be represented by r_n. In formula (2), R represents the income of ITWSN task distribution, and it is expected to obtain the maximum revenue max (R).

$$r_{s,n} = u_n p_{s,n}, \tag{1}$$

$$R = \sum_{n=1}^{N} r_{s,n}, \tag{2}$$

$$P_{s,n} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} \end{bmatrix} (s \in [1, 3], n \in [1, 5]). \tag{3}$$

In formula (4), U_n represents the urgency of each task:

$$U_n = [u_1\ u_2\ u_3\ u_4\ u_5]\ (n \in [1, 5]). \tag{4}$$

At this time, suppose there are 1 individual, 5 tasks, and 3 sensor nodes in the model. Individuals naturally generate a coding solution randomly in formula (5), and D represents a task allocation solution. Randomly generate a task allocation solution, and the result is expressed in formula (5), and the corresponding coding method of the allocation plan is given in Section 4:

$$D = [3 \quad 2 \quad 2 \quad 1 \quad 3]. \tag{5}$$

In order to speed up the execution efficiency, the next step here is to convert the real number encoding to the binary encoding method. The binary matrix is shown in

$$D^* = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{6}$$

$$P^*_{s,n} = P_{s,n} D^*, \tag{7}$$

$$P^*_{s,n} = \begin{bmatrix} 0 & 0 & 0 & p_{14} & 0 \\ 0 & p_{22} & p_{23} & 0 & 0 \\ p_{31} & 0 & 0 & 0 & p_{35} \end{bmatrix} (s \in [1, 3], n \in [1, 5]). \tag{8}$$

According to (3) and (6), $P^*_{s,n}$ is calculated in (7), which represents the advantage of assigning the n_{th} task to the s_{th} sensor. The result matrix is shown in formula (8). D^* is a binary task allocation plan matrix. $D^*_{s,n} = 0$ means that the f_{th} task is not allocated to the e_{th} sensor. $D^*_{s,n} = 1$ means that the n_{th} task is assigned to the s_{th} sensor. $P_{s,n}$ is a task allocation advantage matrix, where the value represents the advantage value of the n_{th} task assigned to the s_{th} sensor.

The identity matrix with column s in P_n is generated according to formulas (9) and (10). The number of sensors in the model corresponds to the number of columns in b_s. The calculated P_n is shown in formula (11).

$$b_s = [1 \quad 1 \quad 1], \tag{9}$$

$$P_n = b_s \cdot P^*_{s,n}, \tag{10}$$

$$P_n = [p_{31} \quad p_{22} \quad p_{23} \quad p_{14} \quad p_{35}]. \tag{11}$$

In formula (12), R_n represents the total benefit value of a group of task allocation plans:

$$R_n = P_n \cdot U'_n. \tag{12}$$

In the process of task allocation, due to the computing power and QoS constraints of sensor nodes, the two factors will affect the network revenue in the task allocation model. Bandwidth functions ($B(e)$), delay functions ($D(e)$), delay jitter functions ($D^{-}J(e)$), and loss packet rate functions ($P^{-}L(e)$) are designed. Suppose that there are four factors in a link v_i and v_j in ITWSNs. The QoS constraints are expressed in

$$B(l(v_i, v_j)) = \min\{B(e)\}, \tag{13}$$

$$D(l(v_i, v_j)) = \sum_{e \in l(v_i, v_j)} D(e), \tag{14}$$

$$D\_J(l(v_i, v_j)) = \sum_{e \in l(v_i, v_j)} D\_J(e), \tag{15}$$

$$P\_L(l(v_i, v_j)) = 1 - \prod_{e \in l(v_i, v_j)} (1 - P\_L(e)), \tag{16}$$

where e represents the link transmission energy, $l(v_i, v_j)$ represents a link, $B(l(v_i, v_j))$ represents minimum bandwidth, $D(l(v_i, v_j))$ represents total jitter, $D^{-}J(l(v_i, v_j))$ represents total delay jitter, and $P^{-}L(l(v_i, v_j))$ represents the total packet loss rate.

In ITWSNs, the task allocation of sensor nodes must not only meet actual traffic control requirements but also improve the QoS and revenue of the entire network. The higher the efficiency of task execution, the higher the overall profit of the system. System benefit refers to the sum of the benefits of each sensor node to complete the task. Because the traffic remote sensing sensor nodes in ITWSNs are heterogeneous, the assigned execution efficiency of the nodes is different, so the total task revenue of the sensor nodes may also be different. An excellent task distribution scheme can usually achieve better overall network service quality and higher overall network efficiency. Therefore, the allocation plan needs to consider the urgency and load size of each task and the execution capability of each sensor node to determine a task scheduling strategy.

4. IACGA for Task Allocation Optimization in ITWSNs

Aiming at the task allocation problem in ITWSNs, an optimization algorithm based on IACGA is proposed. This idea comes from biological evolution in nature. In our IACGA strategy, new adaptive strategies and cloning strategies were designed. These strategies enable IACGA to allocate tasks well and quickly find the best solution.

Traditional genetic algorithms tend to converge prematurely and fall into the local optimum. Therefore, in the IACGA we proposed, an adaptive mechanism is designed. In the crossover and mutation stage, the probability is adjusted according to the current algorithm operation, thereby affecting the global search capability of the entire algorithm. We have added a clonal immune mechanism. After the population fitness evaluation is completed, we will find the best individual to clone and then immunize to form the next generation of population. The resulting new population has the advantages of diversity and close to the optimal solution. Due to the complex traffic conditions and the task conditions of various heterogeneous sensors, the use of a clonal immune mechanism can effectively improve the efficiency of the algorithm, better solve the task allocation problem, and increase the system revenue of ITWSNs.

The main execution steps of IACGA are shown in Figure 1.

We can use the following detailed steps to illustrate the algorithm flow shown in Figure 1.

Step 1. Initialize the population. First, construct a parent population that satisfies the conditions of the model. The population can be abstracted as a real matrix. Suppose there are M individuals and N tasks and the size of the matrix is MN.

Step 2. Calculate the fitness of the parent and perform selection operations. Sort according to the fitness of each individual to find the best individual.

Step 3. Clone the best individual. Clone the most adaptable individual, use immunity to decide whether to accept, and recombine a new population.

Step 4. Adjust parameters adaptively. Judge whether the population fitness is clustered, and adjust the parameters of crossover and mutation.

Step 5. Perform crossover and mutation according to the parameters in Step 4.

Step 6. Repeat Steps 2–5, and reach the maximum number of iterations to meet the termination condition.

Step 7. Terminate the algorithm and output the best task allocation plan.

In Algorithm 1, we show the entire IACGA algorithm flow in pseudocode.

This section discusses several parts of IACGA from the aspects of task allocation coding scheme and population

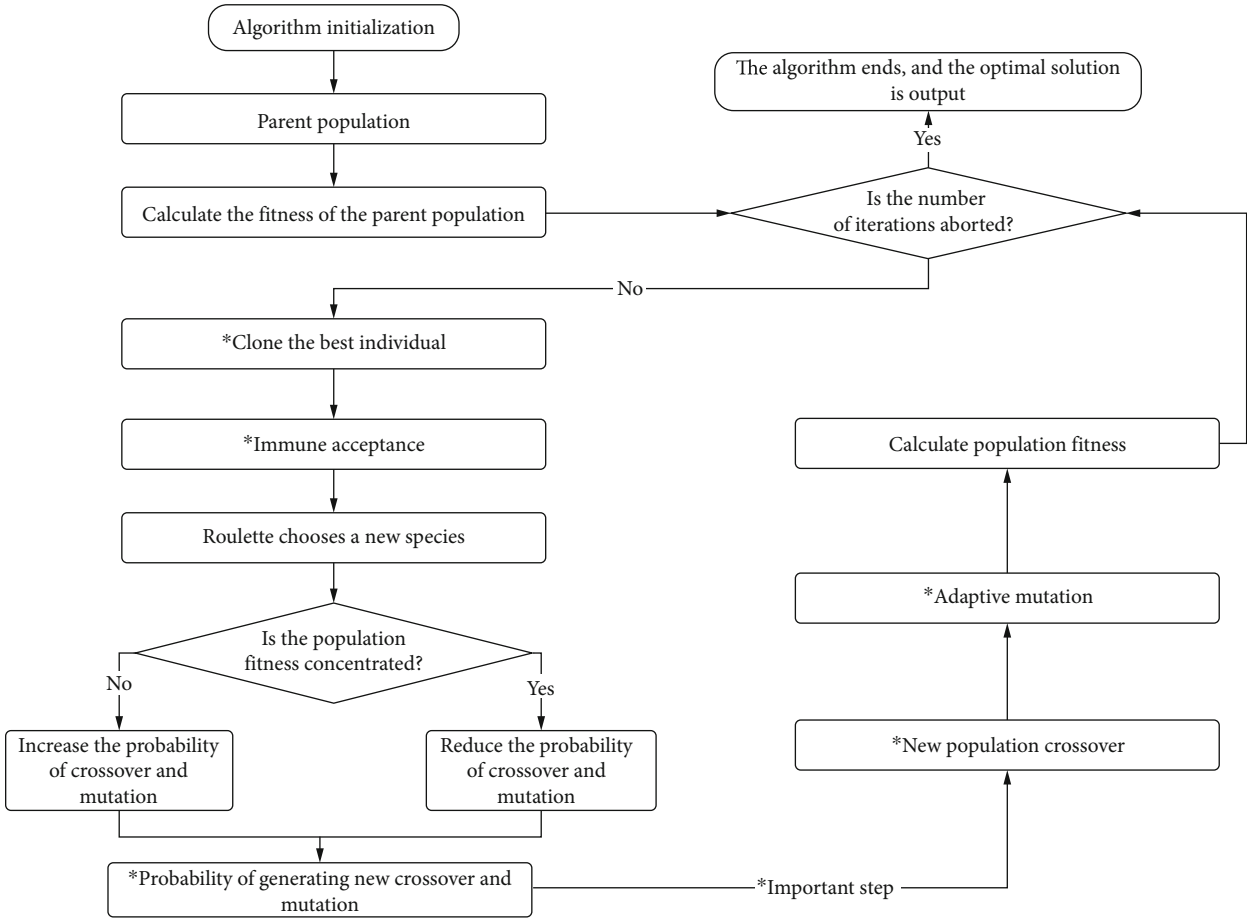


Figure 1: The flow chart of IACGA.

```
IACGA for Task Allocation Optimization in ITWSNs
  BEGIN
    Initialize the population, using random methods;
    Set generations = maximum iteration;
    For generation =1: generations
      Calculate the fitness of the parent;
      Find the individuals with the highest fitness;
      Clone the best individual;
      Immune acceptance;
      if The population fitness converge
        Increase the probability of crossover and mutation;
          else
            Reduce the probability of crossover and mutation;
      end
      Crossover operation;
      Mutation operation;
    end For
    Output the best task allocation scheme;
  END
```

Algorithm 1: IACGA's algorithm pseudocode.

initialization, fitness calculation, selection, crossover, mutation, and optimization operators.

4.1. Coding Scheme. Coding is the first important step to solve the task allocation problem. The task allocation problem of the sensor network is to allocate different tasks to different sensors, so as to get the greatest benefit. The coding idea is to treat a set of allocation schemes as a chromosome with multiple genes. The coding method directly affects the running of the program, the calculation of fitness, and subsequent crossover and mutation operations. Therefore, this article uses integer encoding to facilitate the execution of the algorithm and improve the readability of the program. Suppose there are M individuals in the population, S nodes in the sensor network, and N tasks to be assigned. In formula (17), $d_{m,n}$ represents the situation where the n_{th} task in the m_{th} individual is allocated to the s_{th} sensor:

$$D_{(M,N)} = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,N-1} & d_{1,N} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,N-1} & d_{2,N} \\ \cdots & \cdots & d_{m,n} & \cdots & \cdots \\ d_{M-1,1} & d_{M-1,2} & \cdots & d_{M-1,N-1} & d_{M-1,N} \\ d_{M,1} & d_{M,2} & \cdots & d_{M,N-1} & d_{M,N} \end{bmatrix} \cdot (d_{m,n} \in [1, S], m \in [1, M], n \in [1, N]). \tag{17}$$

The allocation scheme of 4 chromosomes, 8 sensors, and 9 tasks is expressed in

$$D_{(4,9)} = \begin{bmatrix} 1 & 7 & 7 & 3 & 5 & 8 & 2 & 6 & 1 \\ 4 & 2 & 6 & 8 & 8 & 5 & 3 & 3 & 7 \\ 8 & 3 & 1 & 1 & 6 & 3 & 5 & 5 & 4 \\ 2 & 1 & 6 & 5 & 5 & 2 & 7 & 8 & 3 \end{bmatrix}. \tag{18}$$

4.2. Initial Population. The population is coded according to the task allocation model. Its purpose is to establish an abstract connection between task assignment and IACGA. The population initialization can adopt a random scheme to better simulate the natural environment. Therefore, M natural biological individuals are generated in the initialized population. The population can be simply described as $D = \{D_1, D_2, \cdots, D_M\}$. The m_{th} biological individuals can be expressed as $D_m = \{d_{m,1}, d_{m,2}, \cdots, d_{m,N}\}$. The specific gene coding example is shown in formula (18).

4.3. Fitness Evaluation. The fitness function determines the convergence speed of IACGA to a certain extent. Each individual in the population has its own fitness value. IACGA embodies the process of biological evolution in nature, selecting good individuals through fitness and weeding out poor individuals. In this study, individuals are evaluated based on the value of income from task allocation, that is, fitness value. The goal of the research is to maximize network revenue. The profit value of task allocation can be calculated by formulas (1) and (2).

4.4. Selection. The fitness of each individual in the initialized population is very different. In order to speed up the algorithm solving speed, IACGA finds the highest individual through the fitness of each individual, that is, the optimal individual. Our goal is to make highly adaptable individuals inherit as much as possible into the next generation population. The environmental adaptability of the new population will become stronger. The idea of selection operation comes from Darwin's theory of biological evolution. A reasonable selection method can enhance the optimization ability of the algorithm. This paper uses the roulette selection operator to calculate the relative fitness value of all individuals in the population according to formula (2) and then calculates the probability of each individual being selected according to formula (19). We assume that there are M individuals in the population:

$$P(m) = \frac{R(m)}{\sum_{m=1}^{M} R(m)} \ (m \in [1, M]), \tag{19}$$

where $P(m)$ represents the probability of the m_{th} individual being selected, $R(m)$ represents the fitness value of the m_{th} individual, and $\sum_{m=1}^{M} R(m)$ represents the total fitness of the population.

4.5. Crossover. The essence of crossover operation is the exchange of partial structures of two individuals. After the parent population crosses, new offspring populations are produced, which increases genetic diversity. The crossover operation not only guarantees the stable evolution of the population but also makes the algorithm develop in the direction of the optimal solution. The probability of crossover also greatly affects the convergence speed of the algorithm. This paper proposes the strategy of adaptively adjusting the cross probability to ensure the global search ability of the algorithm. At the same time, the single-point crossover method is adopted, and random factors are added to find the crossover point and exchange structure.

In order to better understand the crossover process, we use formulas (20) and (21) to visualize it. We chose the crossover point at the 5_{th} gene sequence and swapped all sequences from then on:

$$D_1 = \begin{bmatrix} 1 & 7 & 7 & 3 & | & 5 & 8 & 2 & 6 & 1 \\ 4 & 2 & 6 & 8 & | & 8 & 5 & 3 & 3 & 7 \end{bmatrix} \text{(before)}, \tag{20}$$

$$D_2 = \begin{bmatrix} 1 & 7 & 7 & 3 & | & 8 & 5 & 3 & 3 & 7 \\ 4 & 2 & 6 & 8 & | & 5 & 8 & 2 & 6 & 1 \end{bmatrix} \text{(after)}, \tag{21}$$

where D_1 represents the gene sequence of the two chromosomes before the crossover representing the task allocation plan and D_2 represents the gene sequence after the two chromosomes cross.

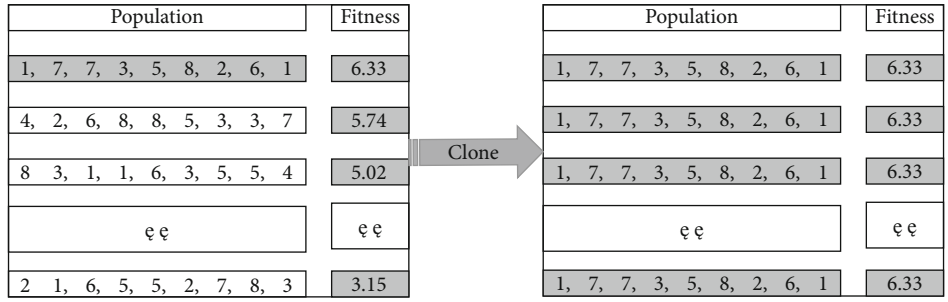


Figure 2: The process of cloning elite individuals.

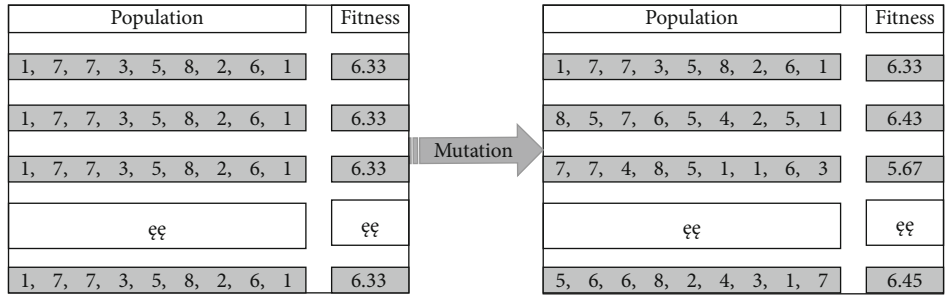


Figure 3: The process of population variation.

4.6. Mutation. Mutation operation is a crucial operation in genetic evolution. Mutation causes random changes in certain parts of some individuals in the population, resulting in new genes. Compared with the crossover operation, mutation is an auxiliary operation, and mutation is used to mutate genes on a single chromosome. Mutation also increases the genetic diversity of the population, increases the search space, and avoids premature convergence of the algorithm. The mutation probability used in the mutation process of traditional genetic algorithms is usually fixed. If the setting is too large, the population structure is unstable, and the optimal solution cannot be found. If the setting is too low, the evolution of the population will be stagnated, falling into a local optimal solution, and the algorithm will converge prematurely. Therefore, this study designed a new adaptive mutation factor in the mutation and adjusted the mutation probability in real time according to the degree of fitness aggregation of the algorithm, so as to achieve the purpose of optimizing task allocation. In order to better understand the mutation process, we use formulas (22) and (23) to visualize it. We selected the mutation points at the 3_{rd} and 7_{th} gene sequences.

$$D_3 = [1 \quad 7 \quad *7 \quad 3 \quad 5 \quad 8 \quad *2 \quad 6 \quad 1] \text{(before)}, \tag{22}$$

$$D_4 = [1 \quad 7 \quad *2 \quad 3 \quad 5 \quad 8 \quad *4 \quad 6 \quad 1] \text{(after)}, \tag{23}$$

where D_3 represents the gene sequence before mutation of a chromosome representing the task allocation plan and D_4 shows the gene sequence after a chromosome mutation.

4.7. Adaptive Operator. The adaptive mechanism designed in this study is mainly used in the crossover and mutation stages. Crossover and mutation will change the original genes of individuals and are the main means to ensure population diversity. In the past classic biological evolution algorithms, scholars usually set some constants, and the constants remain unchanged during the entire algorithm running cycle. This can reduce the difficulty of algorithm design, but the final result of the algorithm is often not good. We set the probability of crossover and mutation in the algorithm as PC and PM, respectively. If the two are set larger, it will affect the stability of the algorithm, and the smaller setting will affect the evolution speed of the algorithm. Therefore, we designed an adaptive strategy to change PC and PM in real time according to the fitness value. In order to prevent PC and PM from approaching 0 and the algorithm from falling into the local optimum, we designed formulas (24) and (25), with the size of the individual fitness value changes nonlinearly to quantify the probability of crossover and mutation:

Table 1: Key experimental parameters of IACGA.

Algorithm	Number of generations	Population size	Adaptive crossover probability	Adaptive mutation probability
IACGA	100	50	0.65~0.88	0.01~0.04

$$PC(m) = PC_{\min} + \frac{PC_{\max} - PC_{\min}}{1 + \exp\left[k \cdot R(m) - R_{\text{avg}}/R_{\max} - R_{\text{avg}}\right]}, \tag{24}$$

$$PM(m) = PM_{\min} + \frac{PM_{\max} - PM_{\min}}{1 + \exp\left[k \cdot R(m) - R_{\text{avg}}/R_{\max} - R_{\text{avg}}\right]}, \tag{25}$$

Table 2: Key experimental parameters of EGA.

Algorithm	Number of generations	Population size	Crossover probability	Mutation probability	Percentage of elites
IACGA	100	50	0.75	0.04	10%

where PC_{max} and PC_{min} are the upper limit of 0.88 and the lower limit of 0.65 for probability adjustment, respectively, k is set to 15, and PM_{max} and PM_{min} are the upper limit of 0.04 and the lower limit of 0.01 for mutation probability adjustment, respectively. R_{max} is the maximum value of fitness of individuals in the population, and R_{avg} is the average fitness value of all individuals in the population, and $R(m)$ represents the fitness value of the m_{th} individual.

4.8. Clone Operator. In order to improve the convergence speed and diversity of the algorithm, we propose a cloning mechanism to be applied to the selection and mutation stages. The advantage of the cloning mechanism is that clonal amplification can increase the convergence speed of optimization calculations, while clonal mutation can maintain the diversity of the population. The main operation of cloning is asexual reproduction, and the performance mainly depends on the quality of the clonal mutation operation. Choose a small mutation probability for operation. At this time, the local search of the algorithm is very fine, but the global search ability of the algorithm is poor, and it is easy to fall into the local optimal solution, resulting in a waste of computing resources. The key of the artificial immune system to solve the problem is to use the principle of immunodominance in immunology to continuously change the antibody itself in response to the stimulation of the antigen. Immune memory cloning basically runs in parallel on two groups, namely, antibody population and memory unit, thus more comprehensively simulating the clonal selection process of the biological immune system. The clonal expansion in this study adopts the elite cloning model and clones the elite individuals with the greatest population fitness. Figures 2 and 3 clearly show the process of clonal expansion and mutation, respectively. The clone mutation operation adopts the adaptive mutation mechanism designed by us. The mutation process is similar to that shown in formulas (24) and (25).

5. Simulation and Discussion

5.1. Experimental Setup. In this section, we will use simulation to test the algorithm performance of IACGA in ITWSN task allocation and compare the simulation results with EGA, SA, and SFLA in ITWSN task allocation. The results of the four algorithms are the average of 100 experiments. Under the same other conditions, different sensor nodes and tasks of different scales are used for simulation comparison in ITWSNs. The hardware computing environment is a PC with Intel® Core™ i5 2.30 GHz CPU. The software computing environment is the Windows 10 operating system of the same version number, MATLAB R2018a.

In this simulation, in order to better compare the performance of IACGA with EGA, SA, and SFLA, we set the population size of the four algorithms of IACGA, EGA, SA, and SFLA to 50, and the maximum number of generations to 100. In IACGA, we set the probability range of adaptive crossover to 0.65~0.88 and set the probability range of adaptive mutation to 0.01~0.04. In EGA, we set the probability of crossover to 0.75, set the probability of mutation to 0.04, and set the percentage of elites to 10%. In SA, we set the initial temperature to 200 and the annealing factor to 0.83. In SFLA, the grouping of frogs is set to 5, the number of frogs in each group is 10, and the maximum step length of frog jumping is set to 1. For IACGA, EGA, SA, and SFLA, the experimental parameters of each algorithm are shown in Tables 1–4, respectively.

Table 3: Key experimental parameters of SA.

Algorithm	Number of generations	Population size	The initial temperature	Annealing factor
SA	100	50	200	0.83

Table 4: Key experimental parameters of SFLA.

Algorithm	Number of generations	Population size	Frog group	Frogs in each group	Maximum step size
SFLA	100	50	5	10	1

5.2. Discussion of Experimental Results. In the part of discussion of experimental results, the simulation data will be analyzed in many aspects. The results of the experiment are shown in the form of a simulation evolution curve, bar chart, and data table and analyzed and discussed.

Figures 4(a)–4(d) intuitively give the simulation results of IACGA, EGA, SA, and SFLA in four different numbers of sensor nodes and tasks. On the whole, in the four different scale situations, the optimization performance of IACGA is better than EGA, SA, and SFLA. EGA performance is better than that of SA and SFLA. It can be seen from Figures 4(a)–4(d) that when the algorithm runs to the 50th generation, IACGA has basically converged and reached a good solution. At this time, IACGA's network revenue is far greater than EGA, SA, and SFLA. SA has converged and fell into a local optimal solution in 30 generations. The convergence speed of SFLA is slow, and the degree of optimization is much lower than that of IACGA, EGA, and SA. Especially in Figure 4(d), when the sensor node is 50 and the task is 120, the network revenue of IACGA increases by 44.0117; EGA, SA, and SFLA are, respectively, 42.7619, 41.7644, and 40.8958. In general, under the conditions of running 150 generations, by proposing adaptive strategies and cloning strategies, IACGA can well solve the problem of revenue from task allocation and has better speed and performance than EGA, SA, and SFLA.

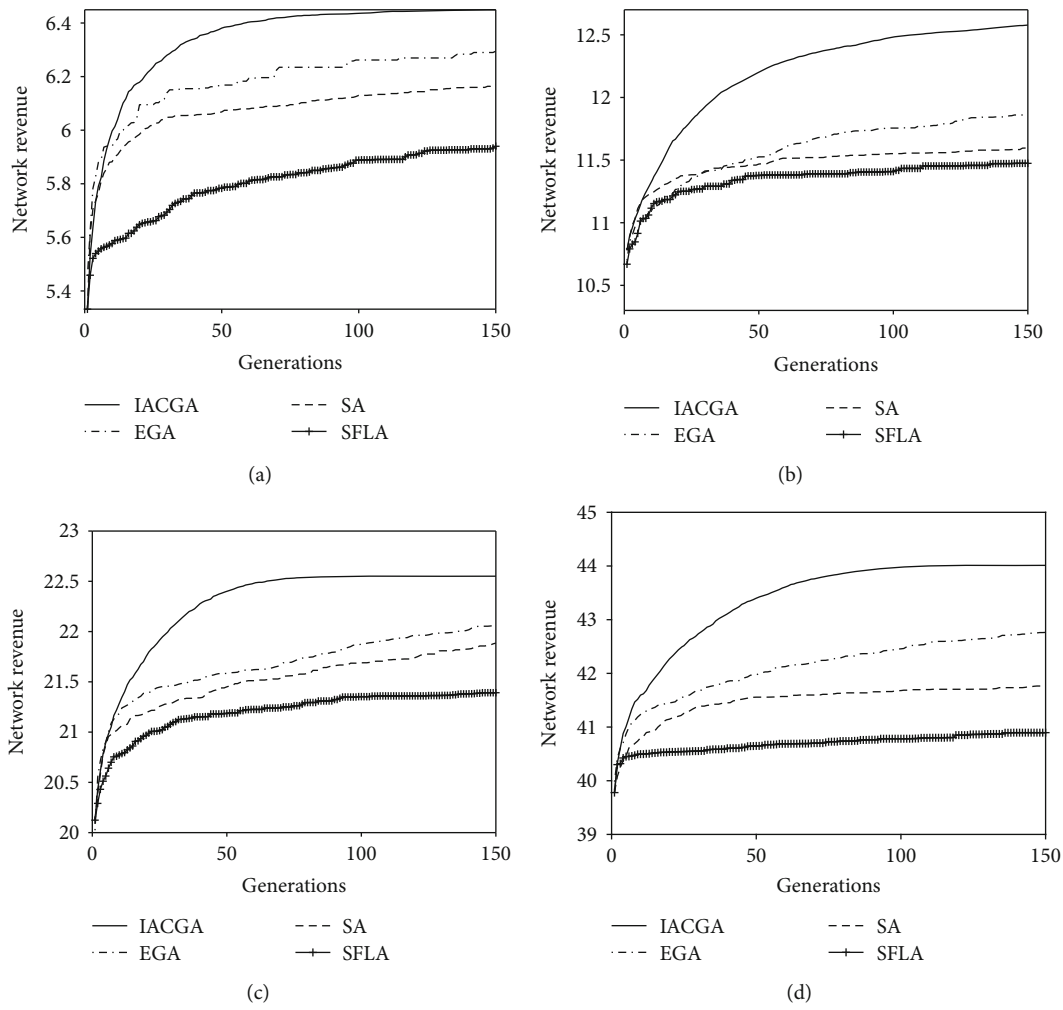


Figure 4: Comparison of network revenue optimization trends of four algorithms: (a) the result of running 150 generations of 15 tasks and 8 nodes; (b) the result of running 150 generations of 30 tasks and 16 nodes; (c) the result of running 150 generations of 60 tasks and 30 nodes; (d) the result of running 150 generations of 120 tasks and 50 nodes.

Figures 5(a)–5(d), respectively, show the network revenue comparison of task allocation when the sensor node is 8, and the task is 15, 25, 40, and 50. Use a histogram to see the gap more intuitively. The results of the four algorithms in Figures 5(a)–5(d) are the results of 150 generations. It can be seen that the total time of IACGA's revenue is greater than EGA, SA, and SFLA in the four cases, and the performance is always the best. When the number of sensor nodes is constant, as the number of tasks increases, the network revenue becomes larger. In Figure 5(d), it can be clearly seen that when the number of tasks is 50, the benefits of IACGA are far greater than EGA, SA, and SFLA. From Figures 5(a)–5(d), the same results can be obtained. IACGA optimizes ITWSN task allocation revenue performance better than EGA, SA, and SFLA.

Figure 6 shows the revenue growth percentages of 15 tasks and 8 nodes, 25 tasks and 8 nodes, 30 tasks and 16 nodes, 40 tasks and 8 nodes, 50 tasks and 8 nodes, 60 tasks and 30 nodes, and 120 tasks and 50 nodes, respectively. The data comes from Table 5. It can be seen from Figure 6 that under seven different tasks, the percentage of optimization of IACGA is greater than that of EGA, SA, and SFLA. In the case of 15 tasks, IACGA has the largest percentage increase in revenue, reaching 21%. When the number of tasks is 30 and 120, the improvement of IACGA is greater than that of EGA, SA, and SFLA. The seven comparison results show that IACGA has a better effect on task allocation network revenue improvement than EGA, SA, and SFLA.

Table 5 shows the network revenue values of IACGA, EGA, SA, and SFLA. It can be seen that as the number of tasks increases, the benefits of the four algorithms will also increase. Under different task allocation parameter settings, IACGA's network benefits are always the largest. The reason is that the adaptive mechanism and cloning mechanism we designed not only increase the global search capability but also speed up the search for the best individual. Compared with IACGA, EGA, SA, and SFLA tend to fall into the local optimum and have poor performance.

Table 6 shows the percentage of increase in IACGA, EGA, SA, and SFLA. When the node is 8, the maximum increase in revenue is 21%, and the task is 15. Followed by 16 nodes and 30 tasks. All the results in Table 5 show that under different conditions, the percentage increase of IACGA's revenue is always greater than that of EGA, SA, and

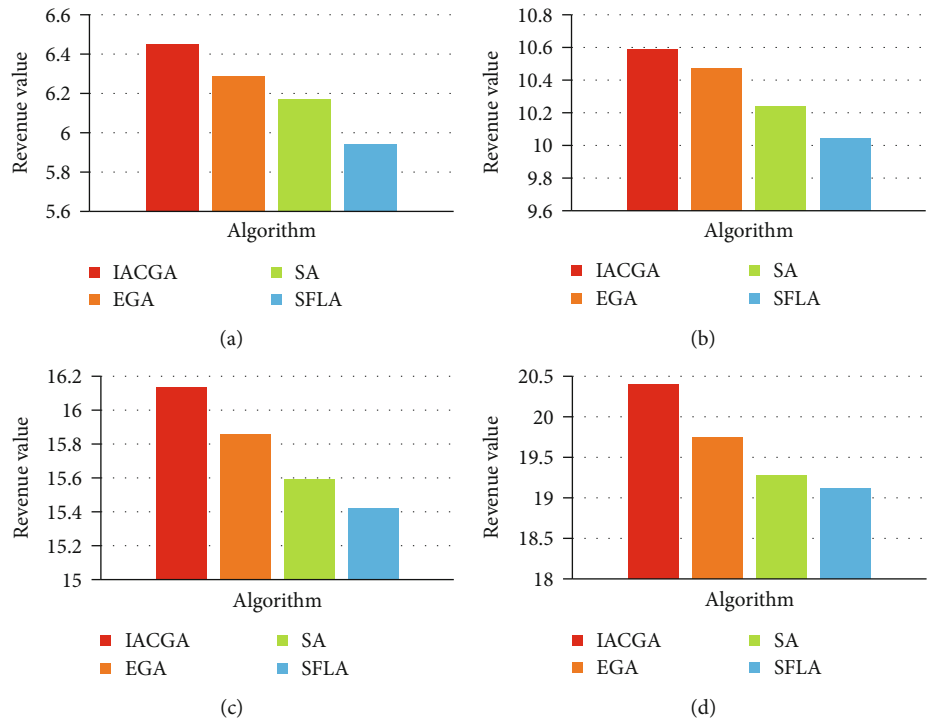


Figure 5: Comparison of the network revenue of different tasks of the four algorithms: (a) revenue of 8 nodes and 15 tasks; (b) revenue of 8 nodes and 25 tasks; (c) revenue of 8 nodes and 40 tasks; (d) revenue of 8 nodes and 50 tasks.

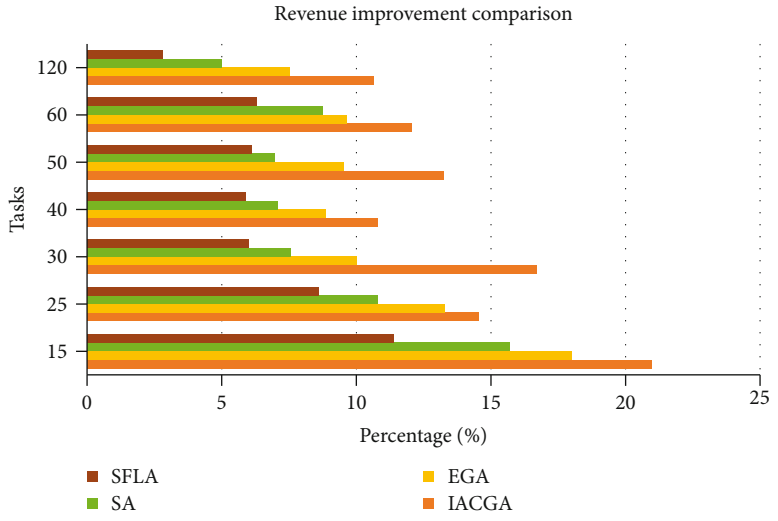


Figure 6: The comparison of revenue improvement of different numbers of tasks.

Table 5: The network revenue of the four algorithms.

Number of nodes and tasks	IACGA	EGA	SA	SFLA
15 tasks and 8 nodes	6.4493	6.2910	6.1700	5.9399
25 tasks and 8 nodes	10.5872	10.4715	10.2425	10.0398
40 tasks and 8 nodes	16.1344	15.8570	15.5931	15.4214
50 tasks and 8 nodes	20.4032	19.7414	19.2759	19.1211
30 tasks and 16 nodes	12.5777	11.8600	11.5961	11.4712
60 tasks and 30 nodes	22.5508	22.0611	21.8853	21.3917
120 tasks and 50 nodes	44.0117	42.7619	41.7644	40.8958

Table 6: The percentage increase of the four algorithms' revenue.

Number of nodes and tasks	IACGA	EGA	SA	SFLA
15 tasks and 8 nodes	21%	18%	16%	11%
25 tasks and 8 nodes	15%	13%	11%	9%
40 tasks and 8 nodes	11%	10%	7%	6%
50 tasks and 8 nodes	13%	9%	7%	6%
30 tasks and 16 nodes	17%	10%	8%	6%
60 tasks and 30 nodes	12%	10%	9%	6%
120 tasks and 50 nodes	11%	7%	5%	3%

Table 7: The running time of the four algorithms' 150 generations.

Number of nodes and tasks	IACGA	EGA	SA	SFLA
8 nodes and 15 tasks	0.10s	0.12 s	0.33 s	0.41 s
16 nodes and 30 tasks	0.15 s	0.15 s	0.47 s	0.58 s
30 nodes and 60 tasks	0.21 s	0.24 s	0.53 s	0.69 s
50 nodes and 120 tasks	0.34 s	0.41 s	0.75 s	0.94 s

Table 8: The computational complexity of the four algorithms is compared.

Algorithm	IACGA	EGA	SA	SFLA
Complexity	$O(n^2)$	$O(n^2)$	$O(n^2)$	$O(n^3)$

SFLA. The results show that IACGA is very effective in solving the task allocation problem.

Table 7 shows the running time of IACGA, EGA, SA, and SFLA under 4 different task allocation experimental parameters. The cycles of the four algorithms are all 150 generations. As the number of nodes and tasks increases, time increases. In the four cases, the running time of IACGA is less than that of EGA, SA, and SFLA. It is proved that the clone operator proposed in this paper accelerates the optimization time of the algorithm and finds the optimal solution quickly.

It can be seen from Table 8 that the algorithm complexity of IACGA, EGA, and SA is $O(n^2)$, and the algorithm complexity of SFLA is $O(n^3)$. IACGA and EGA have the same choice, crossover and mutation process; both are two recycle. In the process of SA temperature drop, each temperature level is also a double cycle. In the iterative process of SFLA, it is necessary to carry out grouping, loop within the group, loop outside the group, and individual coding loop. The complexity is a triple loop. Therefore, SFLA has the highest complexity among the four algorithms. Compared with EGA and SA algorithms, IACGA has the same order of complexity, but it can be seen from Figure 4 that the convergence speed and optimization performance of IACGA are better than those of EGA and SA.

The termination algebra of this algorithm is set to 150 generations. The convergence of the algorithm was discussed in the 150th generation. The convergence criterion of the algorithm adopts the range fluctuation percentage judgment method. The fluctuation range of this experiment is defined as 1%~2%. In the iterative process, we test that within 10 generations, the profit increase percentage ranges from 1% to 2%, which can be considered convergent. The four algorithms use the same convergence criteria. It can be seen from Figures 4(a)–4(d) that when the IACGA algorithm runs to the 60th generation, the upside of network revenue meets the convergence criterion. At this point, the network revenue has reached the highest value of the four algorithms. EGA, SA, and SFLA are also converging at this time, but the return value is very low. In 150 generations, these four algorithms converged to their respective approximate solutions.

In order to better verify the effectiveness of the proposed IACGA, this section compares the algorithm with EGA, SA, and SFLA under different numbers of tasks and experimental conditions of sensor nodes. The simulation data shows that the new IACGA proposed in this paper has great advantages in optimizing the network revenue of ITWSNs, and the algorithm runs fast and has strong optimization capabilities.

6. Conclusion

Aiming at the optimal task allocation scheme of intelligent transportation wireless sensor networks (ITWSNs), this paper proposes a new improved adaptive clone genetic algorithm (IACGA). Before the algorithm design, the task allocation model of the intelligent traffic sensor network is established. We design a new adaptive mechanism to control the probability of crossover and mutation to prevent the algorithm from falling into a local optimum. A new cloning mechanism is designed to select elite individuals to recombine a new population, which increases the diversity of the population and the optimization speed of the algorithm. In addition, we compare IACGA with EGA, SA, and SFLA by simulation and discussed in detail, which prove the superior performance of the proposed new algorithm, effectively solve the task allocation optimization problem in ITWSNs, and successfully maximize the ITWSN revenue.

Data Availability

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Disclosure

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This paper was funded by the Corps innovative talents plan, grant number 2020CB001; project of Youth and middleaged Scientific and Technological In-novation Leading Talents Program of the Corps, grant number 2018CB006; the China Postdoctoral Science Foundation, grant number 220531; the

Funding Project for High Level Talents Research in Shihezi University, grant number RCZK2018C38; and the Project of Shihezi University, grant number ZZZC201915B.

References

[1] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 383–398, 2019.

[2] M. Veres and M. Moussa, "Deep learning for intelligent transportation systems: a survey of emerging trends," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3152–3168, 2020.

[3] D. Li, L. Deng, Z. Cai, B. Franks, and X. Yao, "Notice of retraction: intelligent transportation system in Macao based on deep self-coding learning," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3253–3260, 2018.

[4] Q. Wang, J. Zheng, H. Xu, B. Xu, and R. Chen, "Roadside magnetic sensor system for vehicle detection in urban environments," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1365–1374, 2018.

[5] G. Yan and Q. Qin, "The application of edge computing technology in the collaborative optimization of intelligent transportation system based on information physical fusion," *IEEE Access*, vol. 8, pp. 153264–153272, 2020.

[6] J. Wang, C. Jiang, Z. Han, Y. Ren, and L. Hanzo, "Internet of vehicles: sensing-aided transportation information collection and diffusion," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 3813–3825, 2018.

[7] T. D. T. Nguyen, V. Nguyen, V. -N. Pham, L. N. T. Huynh, M. D. Hossain, and E. -N. Huh, "Modeling data redundancy and cost-aware task allocation in MEC-enabled internet-of-vehicles applications," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1687–1701, 2021.

[8] X. Hou, Z. Ren, J. Wang et al., "Reliable computation offloading for edge-computing-enabled software-defined IoV," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7097–7111, 2020.

[9] L. Sun, J. Wang, and B. Lin, "Task allocation strategy for MEC-enabled IIoTs via Bayesian network based evolutionary computation," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 5, pp. 3441–3449, 2021.

[10] M. Okhovvat and M. R. Kangavari, "A mathematical task dispatching model in wireless sensor actor networks," *Computer Systems Science and Engineering*, vol. 34, no. 1, pp. 5–12, 2019.

[11] S. Khaliq, T. Maqsood, M. Ali, K. Bilal, S. A. Madani, and A. ur Rehman Khan, "A load balanced task scheduling heuristic for large-scale computing systems," *Computer Systems Science and Engineering*, vol. 34, no. 2, pp. 79–90, 2019.

[12] Y. Guo, F. Liu, N. Xiao, and Z. Chen, "Task-based resource allocation bid in edge computing micro datacenter," *Computers, Materials & Continua*, vol. 61, no. 2, pp. 777–792, 2019.

[13] D. Zhu, Y. Wang, C. You et al., "MMLUP: multi-source & multi-task learning for user profiles in social network," *Computers, Materials & Continua*, vol. 61, no. 3, pp. 1105–1115, 2019.

[14] X. Cao, H. Yu, and H. Sun, "Dynamic task assignment for multi-AUV cooperative hunting," *Intelligent Automation & Soft Computing*, vol. 25, no. 1, pp. 1–11, 2019.

[15] T. Ma, S. Pang, W. Zhang, and S. Hao, "Virtual machine based on genetic algorithm used in time and power oriented cloud computing task scheduling," *Intelligent Automation & Soft Computing*, vol. 25, no. 3, pp. 605–613, 2019.

[16] W. Lee, N. Vaughan, and D. Kim, "Task allocation into a foraging task with a series of subtasks in swarm robotic system," *IEEE Access*, vol. 8, pp. 107549–107561, 2020.

[17] C. Wei, Z. Ji, and B. Cai, "Particle swarm optimization for cooperative multi-robot task allocation: a multi-objective approach," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2530–2537, 2020.

[18] X. Tao and W. Song, "Location-dependent task allocation for mobile crowdsensing with clustering effect," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 1029–1045, 2019.

[19] D. Yu, Y. Wang, and Z. Zhou, "Software crowdsourcing task allocation algorithm based on dynamic utility," *IEEE Access*, vol. 7, pp. 33094–33106, 2019.

[20] S. Liu and N. Wang, "Collaborative optimization scheduling of cloud service resources based on improved genetic algorithm," *IEEE Access*, vol. 8, pp. 150878–150890, 2020.

[21] K. Huang, Y. Dong, D. Wang, and S. Wang, "Application of improved simulated annealing genetic algorithm in task assignment of swarm of drones," in *International conference on information science, parallel and distributed systems (ISPDS)*, pp. 266–271, Xi'an, China, 2020.

[22] P. Y. Zhang and M. C. Zhou, "Dynamic cloud task scheduling based on a two-stage strategy," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 2, pp. 772–783, 2018.

[23] J. Zhou, X. Zhao, X. Zhang, D. Zhao, and H. Li, "Task allocation for multi-agent systems based on distributed many-objective evolutionary algorithm and greedy algorithm," *IEEE Access*, vol. 8, pp. 19306–19318, 2020.

[24] D. R. Edla, A. Lipare, R. Cheruku, and V. Kuppili, "An efficient load balancing of gateways using improved shuffled frog leaping algorithm and novel fitness function for WSNs," *IEEE Sensors Journal*, vol. 17, no. 20, pp. 6724–6733, 2017.

[25] U. F. Siddiqi, S. M. Sait, M. S. Demir, and M. Uysal, "Resource allocation for visible light communication systems using simulated annealing based on a problem-specific neighbor function," *IEEE Access*, vol. 7, pp. 64077–64091, 2019.

Hindawi
Journal of Sensors
Volume 2021, Article ID 6647939, 8 pages
https://doi.org/10.1155/2021/6647939

Research Article

Detection of Fatigue Microcrack Using Eddy Current Pulsed Thermography

Xiang Zhang,[1] **Jianping Peng**,[1] **Luquan Du**,[2] **Jie Bai**,[1] **Lingfan Feng**,[1] **Jianqiang Guo**,[1] **and Xiaorong Gao**[1]

[1]*School of Physical Science and Technology, Southwest Jiaotong University, Chengdu 610031, China*
[2]*School of Electronic Information and Automation, ABA Teachers University, ABA, 623002, China*

Correspondence should be addressed to Jianping Peng; adams.peng@swjtu.edu.cn

Received 16 December 2020; Revised 16 January 2021; Accepted 17 March 2021; Published 7 April 2021

Academic Editor: Bin Gao

Microcracks are a common metallic defect, resulting in degradation of material properties. In this paper, specimens with different fatigue microcracks were detected by eddy current pulsed thermography (ECPT). Signal processing algorithms were investigated to improve the detectability and sensitivity; principal component analysis (PCA) and Tucker decomposition were used to compare the performance of microcrack detection. It was found that both algorithms were highly adaptable. A thermal quotient was used to assess the temperature variation trend. Furthermore, the potential correspondence between crack closure and temperature change was investigated.

1. Introduction

In recent years, infrared (IR) technology has been successfully applied to power, rail, and other fields because of its advantages for noncontact, high sensitivity, and visualization [1–4]. As active infrared thermography, eddy current pulsed thermography (ECPT) uses external high-frequency alternating current to excite the detection coil so as to generate heat on the surface or the inside of the measured specimen [5].

In-service equipment may have complex natural defects due to stress and impact loads [6]. Several results based on ECPT in the fields of crack detection have been reported. Specifically, Vrana et al. [7] proposed simplified models for crack detection with induction thermography. Shi et al. [8] proposed a quantitative crack detection method, and cracks of different sizes were analysed. Weekes et al. [9] performed an experimental investigation of fatigue cracks in steel, titanium, and nickel-based superalloy, and results of probability of detection (POD) were established. Tong et al. [10] modified the modelling of ECPT to achieve quantitative evaluation of blade surface fatigue cracks in heavy-duty gas turbines. Netzelmann et al. [11] employed the ECPT method demonstrating the detection of hardening cracks on a large gear tooth. Peng et al. [12] developed solutions for four types of winding defects; the defects under layer insulation can be detected using ECPT. Genest and Li [13] used both experimental and numerical assessments of the induction thermography technique, detecting the microcrack in notched steel coupons.

The abundant transient information in the ECPT has provided grounds for further analysis. In order to enhance the resolution, Maldague and Marinetti [14] proposed pulse phase thermography (PPT), which has an advantage in quantitative inversion. Chen et al. [15] used ICA to identify defect patterns automatically and reduced the influence of the emissivity. Wang et al. [16] used PCA to process the artificial crack in a steel sample and natural fatigue cracks in aircraft brake components. Zhang et al. [17] performed PCA and partial least-squares thermography (PLST) to improve the infrared image performance for defect detection in composite panels. Among the above dimension reduction and matrix decomposition methods, these algorithms provided manifest high levels of flaw contrast relative to that present in the unprocessed data. In addition, tensor decomposition approaches have become an effective tool for feature extraction in infrared thermography crack detection. Gao et al.

[18] developed a spatial transient phase tensor model to extract and separate patterns. Song et al. [19] formulated a tensor decomposition analytical model to identify cracks on samples with different geometry.

However, previous research has rarely discussed the diagnosis of crack closure effect, which is common in the industry. In fatigue tests, it can typically be classified into three categories: plasticity-induced [20, 21], roughness-induced [22], and oxide-induced [23, 24] crack closures. Moreover, Jomdecha et al. [25] proposed a new model to calculate the magnetic flux density over the stress-corrosion crack (SCC) region of different conductivity. Chen et al. [26] found that SCC behaves like a conductive slit in the perspective of eddy current testing. Nevertheless, there remain challenges to distinguish the weak thermal features of closed cracks. In this study, we aim to examine applications of ECPT in fatigue microcrack in a more comprehensive way. Experimental tests on fatigue precrack by three-point bending have been conducted, which is an important step from artificial crack to natural fatigue crack. Compared with PCA, tensor decomposition could preserve more defect information. The performance of detecting fatigue crack is discussed through SNR and thermal quotient. The remainder of this paper is organized as follows: the methodology used in this work is described in Section 2. The experimental setup and specimen are explained in Section 3. The result and discussion are then provided in Section 4. Finally, the conclusion is outlined in Section 5.

2. Methodology

2.1. Induction Heating Theory of ECPT. The main physical process of ECPT involves induced eddy current heating and thermal diffusion. These eddy currents are governed by a subsurface penetration depth, based on an exponentially damped skin effect. According to Joule's law, the thermal power generated by the internal resistance of the material is

$$P_w = \frac{1}{\sigma}|J_e|^2 = \frac{1}{\sigma}|\sigma E|^2, \tag{1}$$

where J_e is the eddy current density and E is the electric field strength; equation (1) determines the resulting temperature field. In general, the environmental temperature or reference temperature is taken as a constant T_0. Thus, the heat conduction equation of a specimen in the company of a defect can be expressed as

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} + \frac{1}{\lambda} q(x, y, z, t) = \frac{\rho C_p}{\lambda} \frac{\partial T}{\partial t}, \tag{2}$$

where $T = T(x, y, z, t)$ denotes the temperature distribution, λ is the thermal conductivity of the material, ρ is the density, and C_p is the specific heat, and $q(x, y, z, t)$ is the internal heat generation function per unit volume and unit time.

2.2. Principal Component Analysis. For raw data, PCA is an unsupervised classification method; its strength is that it reduces the dimensionality of the data while keeping most of the variation in the data set. Each PC could be considered a linear combination of the original thermal sequences and ranked in decreasing order. Singular value decomposition and covariance matrix decomposition methods were used to decompose the 2-D thermal matrix in the following way:

$$T = URV^T, \tag{3}$$

where U is a matrix; it contains a series of empirical orthogonal functions (EOFs). R is a diagonal matrix with the singular values of T; V is an orthogonal matrix.

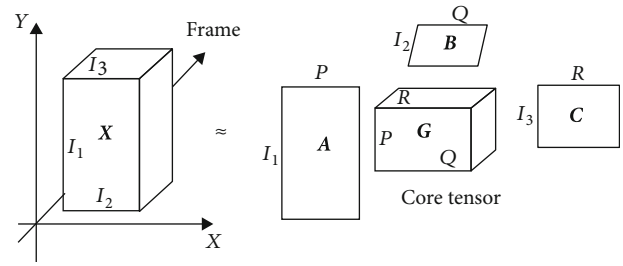


Figure 1: Decomposition of 3rd-order tensor.

2.3. Tensor Decomposition. A tensor is a multidimensional array. The infrared sequences recorded by an IR camera can be represented by a third-order tensor $X \in R^{I_1 \times I_2 \times I_3}$, with two modes representing spatial position and the third mode representing the transient information. We decompose the higher-order tensor into a core tensor multiplied by a matrix along each mode. In the three-way case, the discretized tensor $X \in R^{I_1 \times I_2 \times I_3}$ can be calculated by Tucker decomposition as

$$X \approx G \times_1 A \times_2 B \times_3 C = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} a_p \circ b_q \circ c_r, \tag{4}$$

where $G \in R^{P \times Q \times R}$ is the core tensor; $A \in R^{I_1 \times P}$, $B \in R^{I_2 \times Q}$, and $C \in R^{I_3 \times R}$ are the factor matrices considered as the principal components in each mode; and P, Q, and R are the number of components in the factor matrices. The operator "$\circ$" denotes the vector's outer product. The tensor decomposition method does not perform dimensionality reduction on the thermal imaging high-dimensional image. Therefore, the tensor algorithm can maintain the structural stability of the original data and extract more crack features. This decomposition is illustrated in Figure 1.

2.4. Framework for This Work. Based on the theory introduction in this section, a research approach diagram for ECPT showing defect characterization in two specimens with different shapes was proposed, as shown in Figure 2. It was initiated by using an ECPT platform to acquire the raw data. Then, the thermal sequences in the region of interest (ROI) were preprocessed by PCA and tensor decomposition, respectively, to enhance the thermal contrast. After that, max thermal response and temperature line scan were used

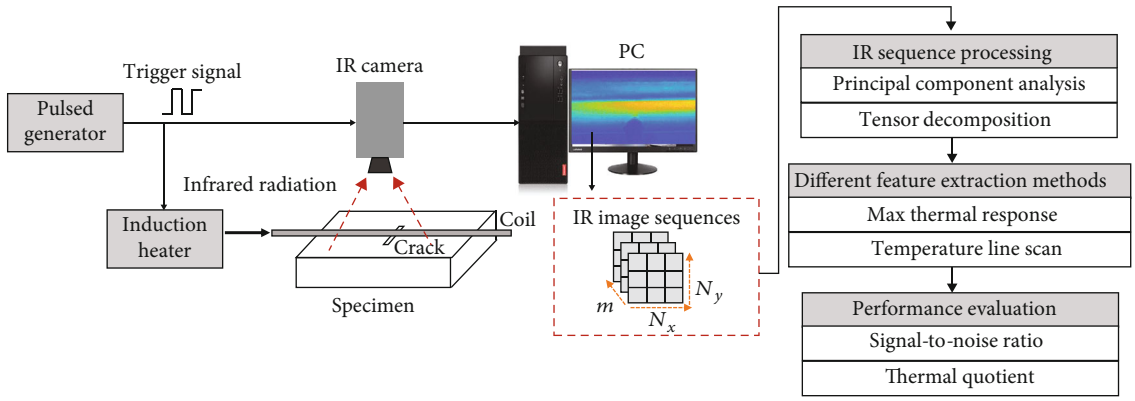


Figure 2: Inductive thermography and research approach diagram.

to extract the thermal features. Finally, the SNR and thermal quotient were used to evaluate the performance.

3. Experiment Setup and Sample

The rail specimen with machined notch (type V) is described in Figure 3. Table 1 gives information about crack length, and Table 2 introduces the material properties. In this study, a straight-through notch of 1 mm in depth was introduced using spark discharge. The straight-through notch was placed connect with the machined notch's tip. Using this method, we were able to generate a well-defined and straight fatigue precrack using three-point bending. The setup of the three-point bending test is shown in Figure 3. Prior to producing the precrack, the specimen surfaces were polished with SiC paper up to #1000. After that, a straight fatigue crack was generated using a fatigue testing machine (MTS809, USA) with a sinusoidal 20 Hz waveform.

The specimen with precrack was then tested using the ECPT method with two conditions each before and after resection along the red dashed line in Figure 4. That is, the specimens were machined into two types (after three-point bending tests) as illustrated in Figure 4, where a is the fatigue precrack length, c is the length of the machined notch (type v), and b and d are the thickness and width, respectively.

The experimental platform for ECPT is shown in Figure 5. It included an infrared camera, an excitation coil, a heating module, and a PC. In the following experiments, the excitation current and frequency were set as 300 A and 286 kHz, respectively. The heating time was set as 200 ms, and the total recording time was 2 s. The FLIR SC650sc IR camera captured thermal images with a spatial resolution of 640 ∗ 120 pixels at a frequency of 200 Hz, then transmitted the information about the thermal sequences to the computer for later analysis and data postprocessing. The measurement was carried out three times, and the mean was taken.

4. Results and Discussion

A typical heating stage experimental result (at 0.2 s) for a specimen with a machined V-shaped notch is shown in

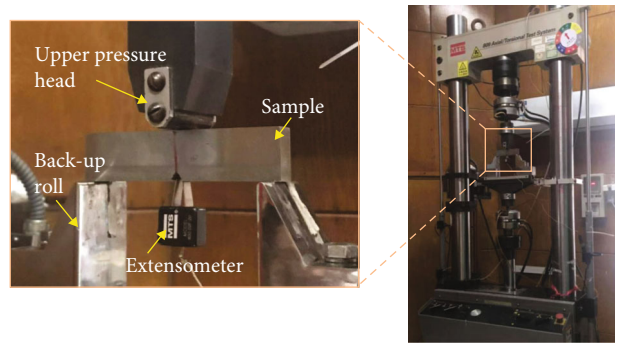


Figure 3: Three-point bending test setup.

Table 1: Crack length information.

Specimen no.	Crack length
#1	1.0 mm
#2	1.5 mm
#3	2.0 mm
#4	2.5 mm

Table 2: Material parameters.

Parameters	Values
Relative permittivity	100
Conductivity (S/m)	5×10^6
Heat capacity (J/(kg K))	490
Thermal conductivity (W/(m K))	50
Thermal diffusivity (m^2/s)	1.172×10^{-5}
Temperature coefficient, $\alpha(1/°C)$	5.0×10^{-3}
Density (kg/m^3)	7.7×10^3
Poisson's ratio	0.3
Elastic modulus (GPa)	200

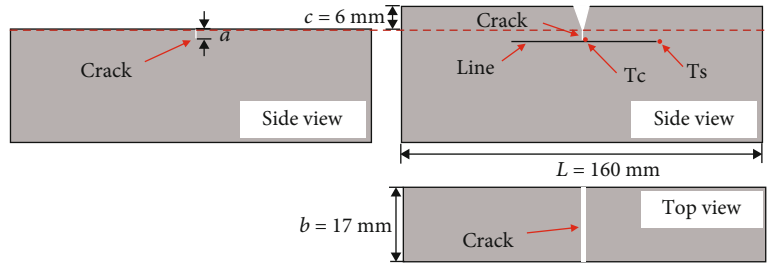


Figure 4: Diagram of specimen.

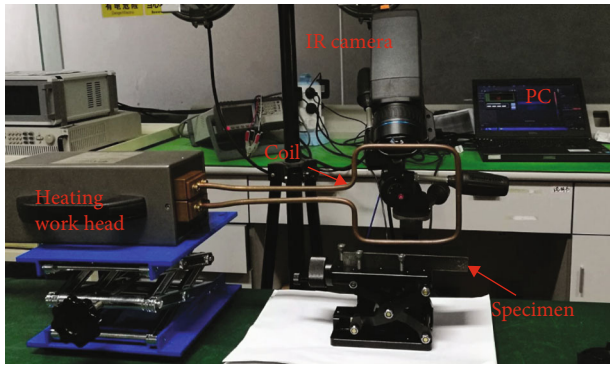


Figure 5: Experimental platform for ECPT.

Figure 6. The region of interest (ROI) including fatigue crack is presented in order to reduce redundant background.

4.1. Temperature Distribution Comparison. The heat distribution on the specimen was uneven due to the defect. In order to investigate the temperature profiles of different types of specimens, a line was created on the surface of specimen #2 (see Figure 4; specimen after resection should extract data with the same position), and the temperature data on it were extracted (Figure 7). This feature was extracted at the pixel level. From the line scans, Figure 7 shows that (1) temperature presents different distribution in the defective and defect-free areas and (2) the width of the heating up stage becomes narrower when the specimen does not have a V-shaped notch. In detail, these phenomena may be explained by the facts that (1) crack closure effect may be influenced by notches and (2) weak conductivity of closed cracks caused a different temperature response in the medial region because the eddy current distribution was inconsistent between specimens.

Figure 8 illustrates the result of maximum thermal response in the defect region along heating periodicities; the effect of the specimen shape has been investigated. The ratio known as the thermal quotient was used to assess the temperature variation trend, which could be expressed as [27]

$$T_q = \frac{\Delta T_{\text{crack}}}{\Delta T_{\text{surface}}}. \quad (5)$$

Based on the result from Figure 8, it can be seen that both distribution laws are basically equal. The R square was then calculated to evaluate the fitted relation. The specimen with a V-shaped notch has a linear relation with a relatively high R square value (90.15%). The fluctuations in the unnotched data may be related to crack closure. In addition, the monotonic increase that denotes the final timepoint of heating has a representative defect contrast throughout the whole image sequence.

4.2. Enhanced Feature. A more detailed account of the defect enhancement method is given in this section. Figures 9 and 10 show reconstructed data from 0 to 2 s using the PCA algorithm for specimen #4. Setting the same principal components for two different specimens, whether they have a V-shaped notch is the difference between two blocks. To be precise, the first two primary components mostly contained information about the coil and heat diffuses, meaning that the contribution of these two parts played an important role throughout the whole experiment. Meanwhile, component three represents the crack, and the last component denotes the background noise. Clearly, the PCA algorithm enhances defect characteristics; fatigue crack could be detected more easily due to there being less redundant information in the third principal component image.

Figure 11 shows the results processed by the tensor decomposition algorithm at the heating stage (specimen #4), where the two specimens both contained the low rank (background) part and the sparse (defect) part. From the results, it can be found that the precrack and background are well distinguished. As we mentioned before, background and thermal diffusion occupy the main part of the contribution rate. Therefore, after subtracting the background in the low rank section, the defect morphology is more clearly seen. Red and white boxes indicate typical positions used for SNR analysis. Red ones reflect the defect area while white ones represent the defect-free region.

To verify the effectiveness of the algorithm, eight rail samples with different crack lengths were tested; they ranged in length from 1 mm to 2.5 mm. Tables 3 and 4 list the results of different signal processing methods for separate trial blocks. In these two tables, component 3 was used as a representative PCA algorithm result. It can be seen in Table 3 that the crack was significantly improved. In Table 4, four specimens after resection followed a similar pattern, with both PCA and tensor making crack identification easier.

4.3. Performance Evaluation and Comparison. Within our study, the signal-to-noise ratio (SNR) [28] was used to

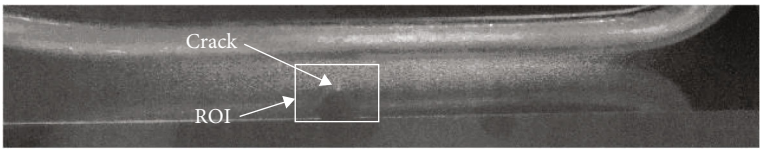


Figure 6: Original thermal data and area of interest.

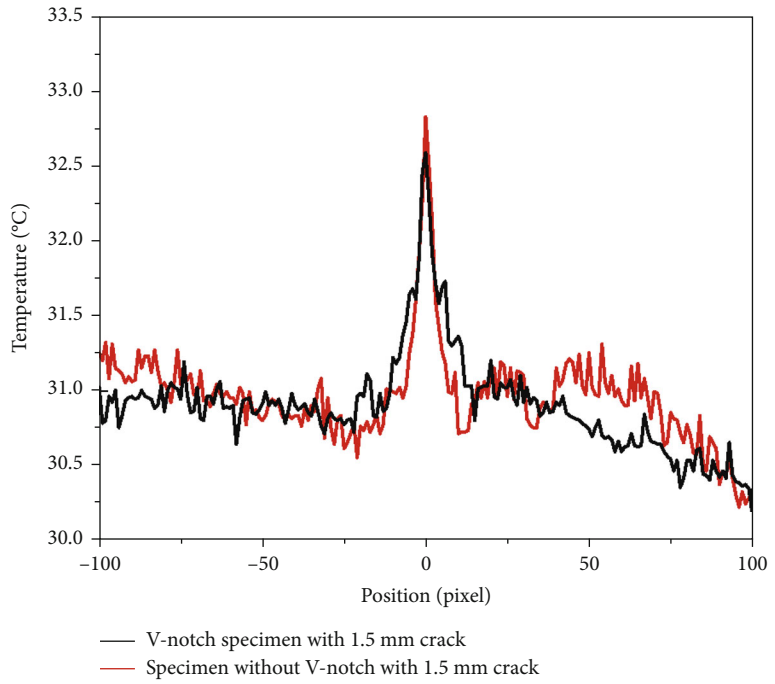


Figure 7: Temperature line scan of different specimens under 200 ms heating pulse for specimen #2.

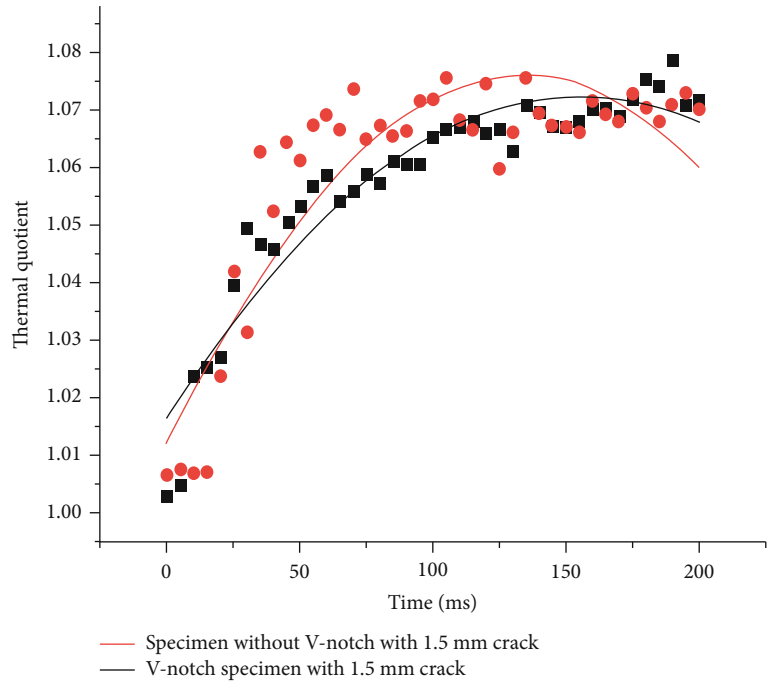


Figure 8: Maximum thermal response versus time with second-order polynomial fit of specimen #2.

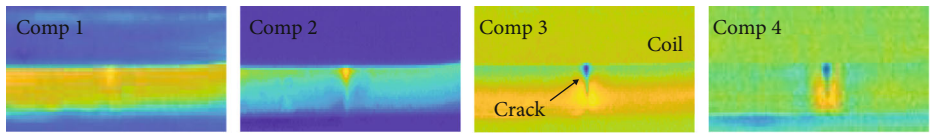


Figure 9: PCA results for specimen #4 without a V-shaped machined notch.

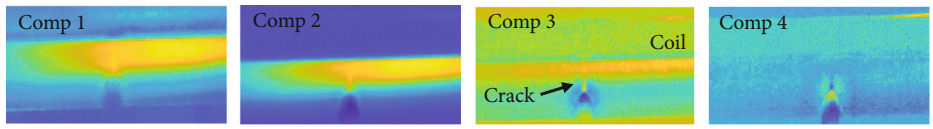
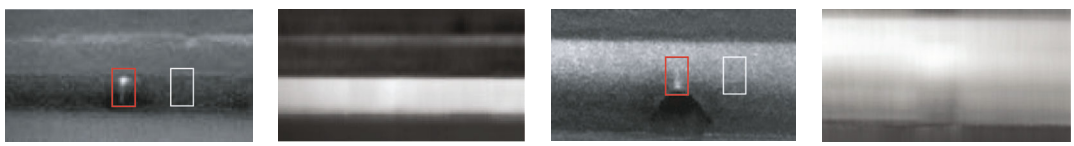


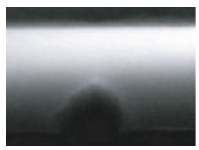
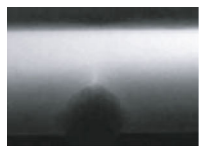
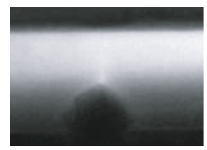
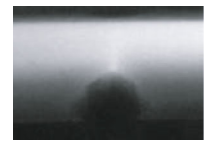
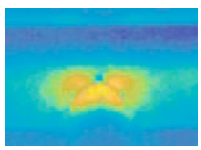
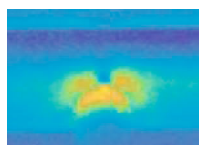
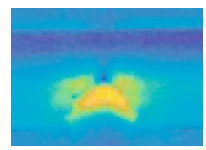
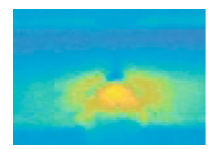
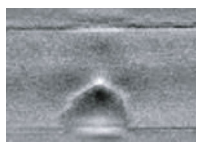
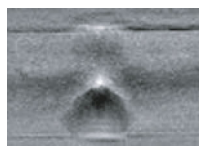
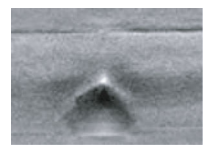
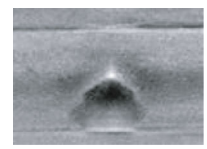
Figure 10: PCA results for notched specimen #4.



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

Figure 11: Tensor decomposition results for specimen #4: (a) sparse rank of sample without notch; (b) low rank of sample without notch; (c) sparse rank of sample with V-shaped notch; (d) low rank of sample with V-shaped notch.

Table 3: The results of different signal processing methods for V-notched specimen.

	Sample #1	Sample #2	Sample #3	Sample #4
Original				
PCA				
Tensor				

evaluate the performance of PCA and tensor decomposition. SNR describes the thermal contrast between the defective and nondefective regions; the prefabricated crack region was selected as "signal" and the defect-free region selected as "noise." The calculation of SNR (dB) can be defined using the equation below:

$$\mathrm{SNR} = 20 \log_{10}\left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{n} T_{d(i,j)}}{\sum_{i=1}^{m}\sum_{j=1}^{n} T_{n(i,j)}}\right), \tag{6}$$

where $\sum_{i=1}^{m}\sum_{j=1}^{n} T_{d(i,j)}$ and $\sum_{i=1}^{m}\sum_{j=1}^{n} T_{n(i,j)}$ are the sum temperature of the crack region and nondefective area, respectively. The original image shows the raw data at 200 ms. The detection performance of various signal processing methods is listed in Table 5; a high SNR value indicates a better crack detection rate. The results show that the Tucker algorithm exhibits a higher crack identification ability than the PCA algorithm in all specimens. Specifically, the result for the original image showed that specimens with a V-shaped notch were more efficient than unnotched samples. This could be

Table 4: The results of different signal processing methods for unnotched specimen.

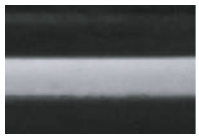
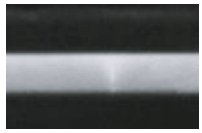
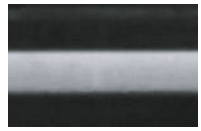
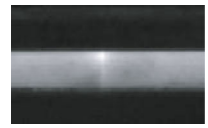
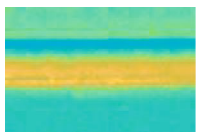
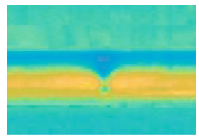
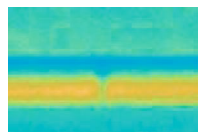
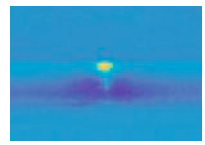
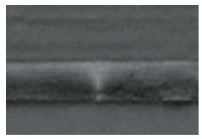
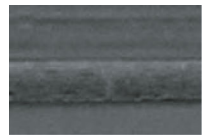
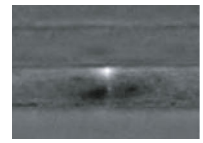
	Sample #1	Sample #2	Sample #3	Sample #4
Original				
PCA				
Tensor				

Table 5: The results of the quantitative analysis with SNR in dB.

	SNR (with machined notch in dB)				SNR (unnotched in dB)			
	1	2	3	4	1	2	3	4
Original	2.14	3.88	3.49	3.10	X	2.26	-0.27	4.02
PCA	2.64	4.54	6.34	4.75	X	2.56	2.26	5.05
Tucker	7.18	8.18	7.65	7.30	X	2.80	1.52	7.67

due to the boundary effect, which generated more heat than in specimens without machined notches. This tendency was also present in the PCA results. In particular, the symbol "X" in the SNR results refers to where the crack could not be detected. Moreover, Table 5 shows that specimens with machined V notches performed better than unnotched blocks. However, combining crack length and SNR is still controversial; even though it does not affect the robust enhancement algorithm, the relationship is not obvious. This could be due to two causes: the inadequate resolution of the IR camera and the complex internal structure of the crack closure.

5. Conclusion

In this paper, eddy current pulsed thermography was investigated to observe fatigue microcracks. Principal component analysis (PCA) and Tucker decomposition were used to extract weak target signals for enhancing detection sensitivity, respectively. Results show that compared with PCA, Tucker decomposition can maintain the structural stability of the original data and extract more crack features that denote a higher SNR. However, it was found that the shape of the sample will influence the detection results. Weak conductivity of closed cracks may change the temperature response in local areas because the crack closure effect may result in localized contact. In addition, the species of closed cracks are not studied in this work, and the linkage between crack length and SNR is not obvious. As such, further detailed analysis of these factors will be undertaken in future studies.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61771409 and the Science and Technology Program of Sichuan under Grant No. 2019YJ0228.

References

[1] J. Wilson, G. Tian, I. Mukriz, and D. Almond, "PEC thermography for imaging multiple cracks from rolling contact fatigue," *NDT & E International*, vol. 44, no. 6, pp. 505–512, 2011.

[2] J. Peng, G. Tian, L. Wang, Y. Zhang, K. Li, and X. Gao, "Investigation into eddy current pulsed thermography for rolling contact fatigue detection and characterization," *NDT & E International*, vol. 74, pp. 72–80, 2015.

[3] Y. He, G. Tian, M. Pan, D. Chen, and H. Zhang, "An investigation into eddy current pulsed thermography for detection of corrosion blister," *Corrosion Science*, vol. 78, pp. 1–6, 2014.

[4] D. Balageas, X. Maldague, D. Burleigh et al., "Thermal (IR) and other NDT techniques for improved material inspection," *Journal of nondestructive evaluation*, vol. 35, no. 1, 2016.

[5] X. P. V. Maldague, "Introduction to NDT by active infrared thermography," *Materials Evaluation*, vol. 60, no. 9, pp. 1060–1073, 2002.

[6] H. Zhang, S. Sfarra, F. Sarasini et al., "Optical and mechanical excitation thermography for impact response in basalt-carbon hybrid fiber-reinforced composite laminates," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 514–522, 2018.

[7] J. Vrana, M. Goldammer, J. Baumann, M. Rothenfusser, and W. Arnold, "Mechanisms and models for crack detection with induction thermography," *AIP Conference Proceedings*, vol. 975, no. 1, pp. 475–482, 2008.

[8] Z. Shi, X. Xu, J. Ma, D. Zhen, and H. Zhang, "Quantitative detection of cracks in steel using eddy current pulsed thermography," *Sensors*, vol. 18, no. 4, 2018.

[9] B. Weekes, D. P. Almond, P. Cawley, and T. Barden, "Eddy-current induced thermography–probability of detection study of small fatigue cracks in steel, titanium and nickel-based superalloy," *NDT & E International*, vol. 49, pp. 47–56, 2012.

[10] Z. Tong, S. Xie, H. Liu et al., "An efficient electromagnetic and thermal modelling of eddy current pulsed thermography for quantitative evaluation of blade fatigue cracks in heavy-duty gas turbines," *Mechanical Systems and Signal Processing*, vol. 142, p. 106781, 2020.

[11] N. Udo, G. Walle, S. Lugin, A. Ehlen, S. Bessert, and B. Valeske, "Induction thermography: principle, applications and first steps towards standardisation," *Quantitative InfraRed Thermography Journal*, vol. 13, pp. 170–181, 2016.

[12] P. Yu, S. Huang, Y. He, and X. Guo, "Eddy current pulsed thermography for noncontact nondestructive inspection of motor winding defects," *IEEE Sensors Journal*, vol. 20, no. 5, pp. 2625–2634, 2020.

[13] M. Genest and G. Li, "Induction thermography of steel coupons with cracks," *Applied Optics*, vol. 57, no. 18, pp. d40–d48, 2018.

[14] X. Maldague and S. Marinetti, "Pulse phase infrared thermography," *Journal of Applied Physics*, vol. 79, no. 5, pp. 2694–2698, 1996.

[15] K. Chen, L. Bai, Y. Chen, Y. Cheng, S. Tian, and P. Zhu, "Defect automatic identification of eddy current pulsed thermography," *Journal of Sensors*, vol. 2014, Article ID 326316, 7 pages, 2014.

[16] Y. Wang, B. Gao, W. L. Woo et al., "Thermal pattern contrast diagnostic of microcracks with induction thermography for aircraft braking components," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 12, pp. 5563–5574, 2018.

[17] H. Zhang, S. Sfarra, A. Osman et al., "Eddy current pulsed thermography for ballistic impact evaluation in basalt-carbon hybrid composite panels," *Applied Optics*, vol. 57, no. 18, pp. 74–81, 2018.

[18] B. Gao, Y. He, W. L. Woo, G. Y. Tian, J. Liu, and Y. Hu, "Multidimensional tensor-based inductive thermography with multiple physical fields for offshore wind turbine gear inspection," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 10, pp. 6305–6315, 2016.

[19] J. Song, B. Gao, W. L. Woo, and G. Y. Tian, "Ensemble tensor decomposition for infrared thermography cracks detection system," *Infrared Physics & Technology*, vol. 105, p. 103203, 2020.

[20] G. Lesiuk, M. Szata, J. A. F. O. Correia, A. M. P. De Jesus, and F. Berto, "Kinetics of fatigue crack growth and crack closure effect in long term operating steel manufactured at the turn of the 19th and 20th centuries," *Engineering Fracture Mechanics*, vol. 185, pp. 160–174, 2017.

[21] J. A. F. O. Correia, S. Blasón, and A. Arcari, "Modified CCS fatigue crack growth model for the AA2019-T851 based on plasticity-induced crack-closure," *Theoretical and Applied Fracture Mechanics*, vol. 85, pp. 26–36, 2016.

[22] G. T. Gray, J. C. Williams, and A. W. Tompson, "Roughness-induced crack closure: an explanation for microstructurally sensitive fatigue crack growth," *Metallurgical Transactions A*, vol. 14, no. 2, pp. 421–433, 1983.

[23] K. Asami and H. Emura, "The influence of moisture in air on fatigue crack propagation characteristics of high-strength steels," *Journal of the Society of Materials Science*, vol. 39, no. 439, pp. 425–431, 1990.

[24] K. Tokaji, Z. Ando, and K. Nagae, "The effect of sheet thickness on near-threshold fatigue crack propagation and oxide and roughness-induced crack closure," *Journal of Engineering Materials and Technology*, vol. 109, no. 1, pp. 86–91, 1987.

[25] C. Jomdecha, W. Cai, S. Xie, Y. Li, and Z. Chen, "Analysis of magnetic flux perturbation due to conductivity variation in equivalent stress-corrosion crack," *International Journal of Applied Electromagnetics and Mechanics*, vol. 59, no. 4, pp. 1385–1392, 2019.

[26] Z. Chen, Z. Chen, N. Yusa, and K. Miya, "A nondestructive strategy for the distinction of natural fatigue and stress corrosion cracks based on signals from eddy current testing," *Journal of Pressure Vessel Technology*, vol. 129, no. 4, pp. 719–728, 2007.

[27] B. Oswald-Tranta, "Thermo-inductive crack detection," *Nondestructive Testing and Evaluation*, vol. 22, no. 2-3, pp. 137–153, 2007.

[28] F. Lopez, C. Ibarra-Castanedo, V. de Paulo Nicolau, and X. Maldague, "Optimization of pulsed thermography inspection by partial least-squares regression," *NDT & E International*, vol. 66, pp. 128–138, 2014.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5559231, 16 pages
https://doi.org/10.1155/2021/5559231

Review Article

Environment Perception Technologies for Power Transmission Line Inspection Robots

Minghao Chen[1,2] **Yunong Tian**,[1,2] **Shiyu Xing**,[1,2] **Zhishuo Li**,[1,2] **En Li**[1,2] **Zize Liang**,[1,2] **and Rui Guo**[3]

[1]*The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing 100190, China*
[2]*The School of Artificial Intelligence, University of Chinese Academy of Sciences, No. 19(A) Yuquan Road, Beijing 100049, China*
[3]*The State Grid Shandong Electric Power Company, 150 Jinger Road, Jinan 250001, China*

Correspondence should be addressed to En Li; en.li@ia.ac.cn

Received 7 January 2021; Revised 27 January 2021; Accepted 11 March 2021; Published 31 March 2021

Academic Editor: Bin Gao

With the fast development of the power system, traditional manual inspection methods of a power transmission line (PTL) cannot supply the demand for high quality and dependability for power grid maintenance. Consequently, the automatic PTL inspection technology becomes one of the key research focuses. For the purpose of summarizing related studies on environment perception and control technologies of PTL inspection, technologies of three-dimensional (3D) reconstruction, object detection, and visual servo of PTL inspection are reviewed, respectively. Firstly, 3D reconstruction of PTL inspection is reviewed and analyzed, especially for the technology of LiDAR-based reconstruction of power lines. Secondly, the technology of typical object detection, including pylons, insulators, and power line accessories, is classified as traditional and deep learning-based methods. After that, their merits and demerits are considered. Thirdly, the progress and issues of visual servo control of inspection robots are also concisely addressed. For improving the automation degree of PTL robots, current problems of key techniques, such as multisensor fusion and the establishment of datasets, are discussed and the prospect of inspection robots is presented.

1. Introduction

Traditional PTL inspection methods include line crawling inspection, ground-based inspection, and manual inspection with telescopes, as shown in Figure 1. Their defects are clearer with the progress of the power system. These methods are slow and dangerous and may not be conducted sometimes. The reasons are as follows: (a) intricate and diverse workspace. The arrangement of PTL corridors includes several scenarios, such as overhead ground wire and multi-bundled conductors. The slope of conductors is different. What is more, a great variety of obstacles are on the PTL, as shown in Figure 2 [1]. (b) Geographical conditions of PTL are various. Part of the PTL is located in some complex areas such as swamps, lakes, and mountains. Although the speedy and maneuverable helicopter inspection method can overcome this difficulty, its detection precision of small-scale objects is affected by the long working distance. Consequently, it is also not the best inspection method. Automatic PTL inspection technology is eager to be developed in these cases, which is also a challenging task. Apart from mechanical design, PTL environment perception and control technologies are the foundation of automatic PTL inspection. And they are research hotspots in PTL inspection.

In the whole process of automatic PTL inspection, the main tasks completed by inspection robots now are as follows: (a) 3D reconstruction of PTL corridors [2, 3]. 3D reconstruction can be employed for the visualization of PTL, which is helpful for inspectors to analyze the excursion trend of PTL corridors, interference of surrounding environment, broken point detection, and snow loading. (b) Fine inspection of pylons [4] and components [5, 6]. Object detection is exploited to recognize and locate them and their defects, such as cracked nut, bolt looseness, and fitting corro-

(a) (b)

(c)

Figure 1: Traditional inspection methods: (a) line crawling inspection; (b) ground-based inspection; (c) manual inspection with a telescope.

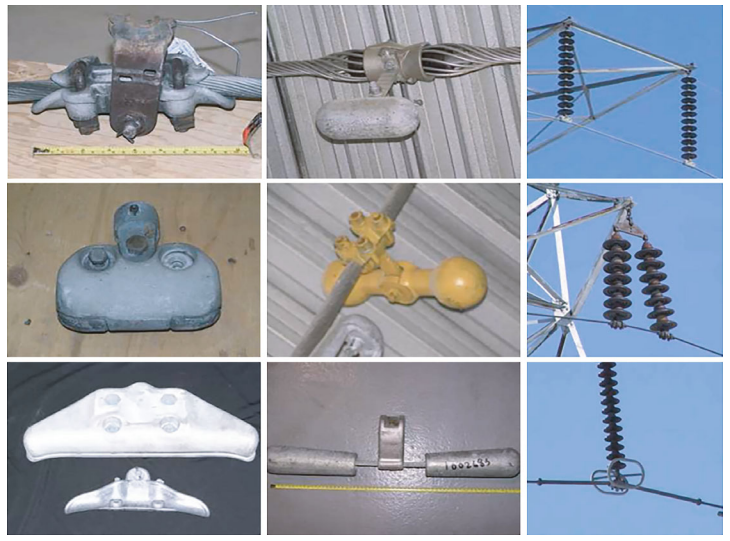


Figure 2: Suspension clamps, vibration dampers, and insulator strings [1].

sion, which are difficult to be found by manual inspection. Besides, object detection can be interfered with by rain, snow, and wind. Therefore, 3D reconstruction is also applied to a local map to acquire a more precise target position. (c) Obstacle crossing and fault maintenance [7, 8]. Obstacle crossing and fault maintenance can be realized by remote teleoperation or autonomous control. Remote teleoperation is easy to realize, which depends on the manual operation of a

ground control unit. However, it is difficult to achieve autonomous operation. Due to much interference existing in the PTL environment, the autonomous operation of inspection robots requires a more robust control system. In contrast to sensor-based control, visual servo control is more robust and flexible. It is suitable for the unpredictable environment of PTL corridors. Aimed at these tasks, 3D reconstruction, object detection, and visual servo of PTL inspection are summarized and analyzed for providing a widely comprehensible review.

The rest of this paper is organized as follows: Section 2 describes the development of the transmission line corridor in 3D reconstruction. Section 3 introduces the progress of traditional and deep learning-based detection algorithms in the field of PTL inspection. Then, their merits and demerits are considered. Section 4 concerns the visual servo of different inspection robots. Section 5 analyzes key technical problems and presents future development directions of inspection robots. Finally, the conclusion is presented in Section 6.

2. 3D Reconstruction of PTL Inspection

3D reconstruction is a means to obtain 3D data of measured objects. It is mainly used to acquire object texture, structure, and scale. There are two approaches to 3D reconstruction: the contact and noncontact methods. In contrast to the former way, noncontact methods are extensively used. Based on whether or not to actively transmit measurement signals, they are divided into active and passive visual means. According to various principles, active visual methods are classified as laser methods [9–13], structured light methods [14], and interferometry [15–17]. On the basis of the number of vision sensors used, passive visual methods can be classified as monocular vision [18, 19], binocular vision [20, 21], and multieye vision [22, 23].

In the field of PTL inspection, active visual methods are chiefly used to construct the 3D model of the PTL corridor. The characteristics of objects can be intuitively expressed by the 3D model. It is helpful for inspectors to discover existing or potential hazards in the transmission line corridor.

2.1. The 3D Reconstruction of Power Lines. The power line plays a key role in the entire PTL corridor as the main carrier of electricity supply. 3D reconstruction of power lines is the foundation for the analysis of wire sag, icing, wind deviation, and distance measurement. Therefore, many studies of it are presented.

2.1.1. Active Visual Methods. Now, laser scanning is normally adopted to complete power line reconstruction. It contains three sections, which are point cloud extraction, point cloud segmentation, and model fitting. Point cloud extraction is divided into simple extraction and fine extraction. The raw data contains point clouds of various objects. The simple extraction is utilized to eliminate point clouds of ground. In addition, fine extraction is employed to further eliminate point clouds that do not belong to lines. Point cloud segmentation falls into determining point clouds of one span and a single conductor. Firstly, the data of different spans are divided according to the highest suspension point. And then, the data of a single power line in one span are confirmed. Finally, the model is selected for fitting.

The extraction of point clouds is generally based on filtering methods. Firstly, DTM [24] is generated by filtering. Then, the points of the ground and objects are differentiated. After that, features such as elevation are employed to obtain the point clouds of power lines. Methods based on TIN such as PTD [25] are used commonly. Yu et al. [26] utilized PTD to eliminate ground points. Furthermore, an angle filter was adopted to acquire point clouds of power lines. In addition to PTD, slope filtering [27] and morphological filtering [28] are also used for eliminating ground points. Nevertheless, they are less applied in the extraction of power lines compared to PTD. The elevation is used in conjunction with DTM to gain the point clouds of power lines. Based on nDSM, Liu et al. [29] used the elevation histogram statistical method to get point clouds of power lines. Besides, there is also direct use of elevation to complete extraction. Shen et al. [30] proposed an elevation threshold segmentation algorithm based on the subspace feature and an elevation density segmentation algorithm. Some other methods for point cloud extraction are also employed. Mclaughlin [31] used a GMM and the EM algorithm for extraction. What is more, Jwa and Sohn [32] adopted Hough transform, eigenvalue analysis, point density analysis, and the one-outlier testing technique for extraction.

In the process of segmentation, point clouds of lines in one span are decided by the position of pylons. The position of pylons can be obtained by the height [26, 33] or the density of point clouds [29]. Moreover, the segmentation of point clouds can be also completed by the minimum linkage hierarchical clustering [34] and the second derivative [35]. There are three types of techniques to determine point clouds of a single conductor. (a) Clustering. To distinguish each power line, point clouds that belong to the same power line in one span are clustered. Lai et al. [36] proposed a cluster analysis method based on spatial distance restriction. What is more, Lin et al. [35] presented the 3D connected component analysis based on fixed-radius near neighbors (FNN) and the k-means algorithm [37] of normalized projection. (b) Hough transform (HT). Yu et al. [26] and Wu et al. [33] employed HT for the segmentation of the single power line. Melzer et al. [34] used an iterative version of the HT, but this method was unstable. (c) Local growth. The means gradually merges or grows to form the entire power line model, such as the local affine model [31] and the piecewise model growing (PMG) [32].

Four types of curve fitting models of power lines are typically employed: (a) the polynomial model [33, 38], (b) the model that combines a line segment and a parabola [36, 37], (c) the model that combines a line segment and the quadratic polynomial with two variables [29, 39], and (d) the model that combines a line segment and a catenary [26, 31, 34, 35]. The first model directly fits the entire power line. However, the other three models split the power line into a line segment and a curved part for fitting. The difference between the latter three models is the expressions of curve fitting. The second model is the approximate expression of the

fourth model. And the performance of the second model is much better because of its simpler formula. Besides, Zhang et al. [40] improved the second model and the fourth model. Overall, the second model is better.

In the above literature, the methods used in each step of the paper are shown in Table 1, which completed the whole process of 3D reconstruction.

Now, 3D reconstruction of power lines is focused on areas with simple structures. The background of PTL, areas of different structures, and various types of power lines are considered much less. It is difficult to identify and reconstruct sleeve nodes and spacer nodes. Moreover, the density of LiDAR point clouds influences firsthand 3D reconstruction.

2.1.2. Passive Visual Methods. Passive visual methods are less employed in the 3D reconstruction of power lines, and their accuracy is lower than that of active visual methods. Zhang et al. [41] developed a multiangle imaging power line detection system to reconstruct 3D power lines, but the method's accuracy was low. Ganovelli et al. [42] presented a means of using a handful of images to reconstruct power lines. This means employed SFM. Because the method was based on specific assumptions, it could not be used in inspection tasks actually. Maurer et al. [43] presented a means that used semantic segmentation based on FCN and multiview geometry. The flowchart is shown in Figure 3. The method was effective but had high requirements on hardware. It is hard to match different images with the lack of effective point features of power lines by passive visual methods. Therefore, active visual methods are mostly employed in the power line reconstruction.

2.2. The 3D Reconstruction of Ground Surface and Pylons. In addition to the reconstruction of power lines, the PTL corridors also need to reconstruct the ground surface and pylons. Nonetheless, their 3D reconstruction is in the preliminary stage. With regard to the ground surface, Maurer et al. [43] completed the 3D reconstruction of power lines and the ground surface simultaneously. In light of the reconstruction of pylons, Chen et al. [44] raised a model-driven approach to reconstruct pylons. The pylon could be divided into head, body, and foot. The three segments were reconstructed by different strategies. And they were assembled to the whole pylon by utilizing the direction and position. The result showed that the method's accuracy could reach the centimeter level. Guo et al. [45] presented a means based on model library and stochastic geometry for pylon reconstruction. This method built an energy model of the correlation between point clouds and pylons. It also used simulated annealing and an MCMC sampler to reconstruct pylons. But this means required plenty of calculations.

Recently, oblique photography has been applied to 3D reconstruction of PTL corridors. Xi et al. [46] designed a UAV power inspection system based on oblique photography. The system used the UAV to obtain multiview images of ground objects and accelerated the 3D reconstruction of the PTL corridors through data processing and model refinement. Pei et al. [47] designed a data acquisition device for 3D reconstruction of PTL corridors, which could obtain the structural information of the pylons' four facades. Oblique photography has some defects in the field of PTL inspection. Oblique photography requires 60%-80% of the air belt overlap. Hence, it is necessary to collect data by multiple flights and the collection efficiency is low. Oblique photography is better for the 3D reconstruction of larger objects but ordinary for small and hollow objects such as power lines and pylons. Consequently, it is imperative to combine oblique photography with other means to gain good reconstruction results. It will be one of the research directions for the 3D reconstruction of PTL corridors in the future.

3. Object Detection of PTL Inspection

Object detection refers to a method of identifying and positioning targets from the background. In nearly two decades, the development of object detection is separated into the traditional detection period and deep learning-based detection period [48]. In the early stage of research on PTL inspection robots, traditional detection algorithms were used to identify and locate power components. With the performance of handcrafted features that tended to be stable, the research on traditional detection algorithms was almost stagnant. The development of inspection robots' object detection fell into a bottleneck. On account of the emergence of R-CNN [49], the progress of object detection algorithms had taken a quantum leap. Neural networks are now used in object detection. They are able to boost detection performance and acquire robust characteristics. As a result, the research on object detection for PTL inspection robots has gradually changed from traditional detection algorithms to deep learning-based detection algorithms.

3.1. Traditional Object Detection Algorithms. Traditional object detection algorithms used on inspection robots were mostly foreground modeling-based methods over the last decades. They are classified as region selection, feature extraction, and classification. Sliding windows are used to acquire many candidate bounding boxes of the entire picture. And then, features of candidate bounding boxes are extracted. Classifiers, such as SVM [50], AdaBoost [51], and multilayer perceptron (MLP), are used to determine objects of candidate bounding boxes as the target or background. At last, NMS [52] is used to complete the detection. In the field of PTL inspection, SIFT [53], SURF [54], LBP [55], and HOG [56] are often used in traditional detection algorithms. Relevant research can be divided into two aspects: (a) pylons and (b) insulators and fittings.

3.1.1. Pylons. While conducting pylon detection, pylons and birds' nests are mainly detected. For pylons, Sampedro et al. [57] advanced a power pylon detection method based on HOG and MLP. Two MLPs were trained in this method. They were used for background-foreground segmentation and the classification of four types of pylons. The system architecture is shown in Figure 4. Cerón et al. [58] advanced a pylon detection method based on a line detection method and a grid of two-dimensional feature descriptors. Pylons could be characterized well by the descriptors. Wang et al.

Table 1: The methods used in each step of 3D reconstruction in different studies.

Reference	Simple extraction	Fine extraction	Determination of point clouds in one span	Determination of point clouds in a single power line	Fitting models
Melzer and Briese [34]	Interpolation method for DTM generation	A grid culling mechanism	A minimum linkage hierarchical clustering approach	Iterative HT	The fourth model
Mclaughlin [31]	Using a Gaussian mixture model and the EM algorithm		A local affine model		The fourth model
Yu et al. [26]	PTD	An angle filer	The position of pylons (by height)	HT	The fourth model
Liu et al. [29]	The elevation histogram statistical method	Using point cloud density	The position of pylons (by density)	A spatial domain segmentation algorithm	The third model
Wu et al. [33]	Echoes of LiDAR data	Grid height elevation and elevation threshold	The position of pylons (by height)	HT	The first model

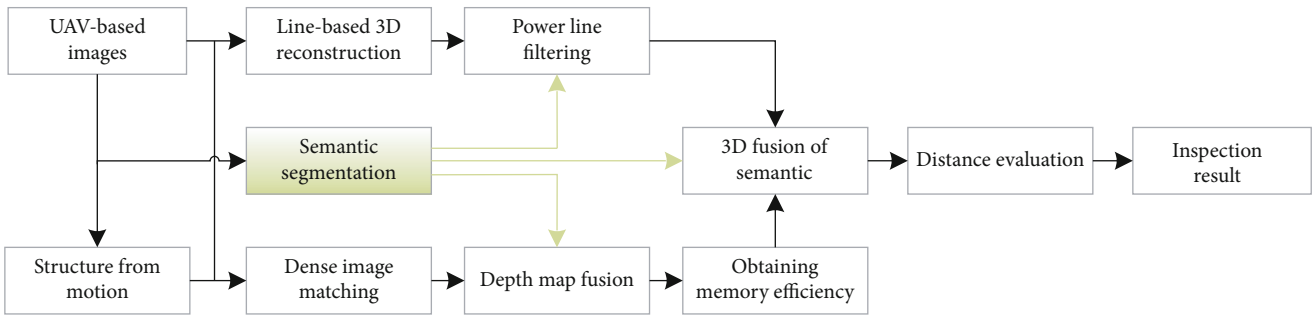


Figure 3: The overview of the method based on FCN and multiview geometry [43].

[59] used HOG and SVM to complete the rapid detection of pylons, but this method's dismissal alarm rate was higher than the false alarm rate. Wang et al. [60] presented a method based on HOG for detecting the pylon components from far to near. This method used HOG features of the pylon in different orientations to train the MLP. It was suitable for pylon detection in an open environment. Aiming at the bird's nest, Zhang et al. [61] raised the means based on the coarse-fine search tactics. It used HOG and AdaBoost to detect pylons' position and roughly determined the candidate bounding boxes of birds' nests. Then, color features were used to subtly detect birds' nests. Xu et al. [62] proposed a method based on HSV and texture features. This method could eliminate much interference caused by areas, which were similar to the color and texture of the bird's nest under the complex background.

3.1.2. Insulators and Fittings. In terms of insulators and fittings, many of the detection algorithms are designed for a single object, while only a few algorithms are for multiobjects. For insulators, Zhao et al. [63] came up with an insulator detection means based on SIFT and RANSAC. This method used RANSAC to remove abnormal SIFT features of insulators and completed the insulator detection through affine transformation. After that, Zhao and Liu [64] presented an approach based on SURF and intuitionistic fuzzy set (IFS). The steps are shown in Figure 5. The insulator's SURF features were divided by IFS. Then, the connected regions of all categories were calculated to obtain the smallest circumscribed rectangle of each region for positioning. Prasad et al. [65] used SVM and local binary pattern histogram Fourier (LBP-HF) to complete health status detection of insulators, with a precision of 93.33%.

For fittings, Zhang et al. [66] proposed a fitting means based on HOG. This method used PCA to reduce the dimension of HOG features. Then, these features were identified by SVM. In contrast to other traditional algorithms, this means is more accurate. Zhang et al. [67] raised a method that could detect vibration dampers under the complex background. This method used Relief-F to weight and merge aggregate channel features (ACF) and complex frequency domain features. Ultimately, AdaBoost and NMS were used to complete the detection of vibration dampers. Feng et al. [68] employed HOG and SVM to accomplish the detection of bolts, but the accuracy of the means was low. Fan et al. [69] presented a method based on an improved HT for bolt detection. The peak selection strategy of HT was ameliorated for the improvement of detection accuracy.

On the grounds of size, the targets of PTL can be classified into the pylon level, the device level, the part level, and the component level [70]. At the moment, object detection research is focused on targets of pylon level and device level. There is less research on detecting targets of the part level and component level. Traditional object detection algorithms

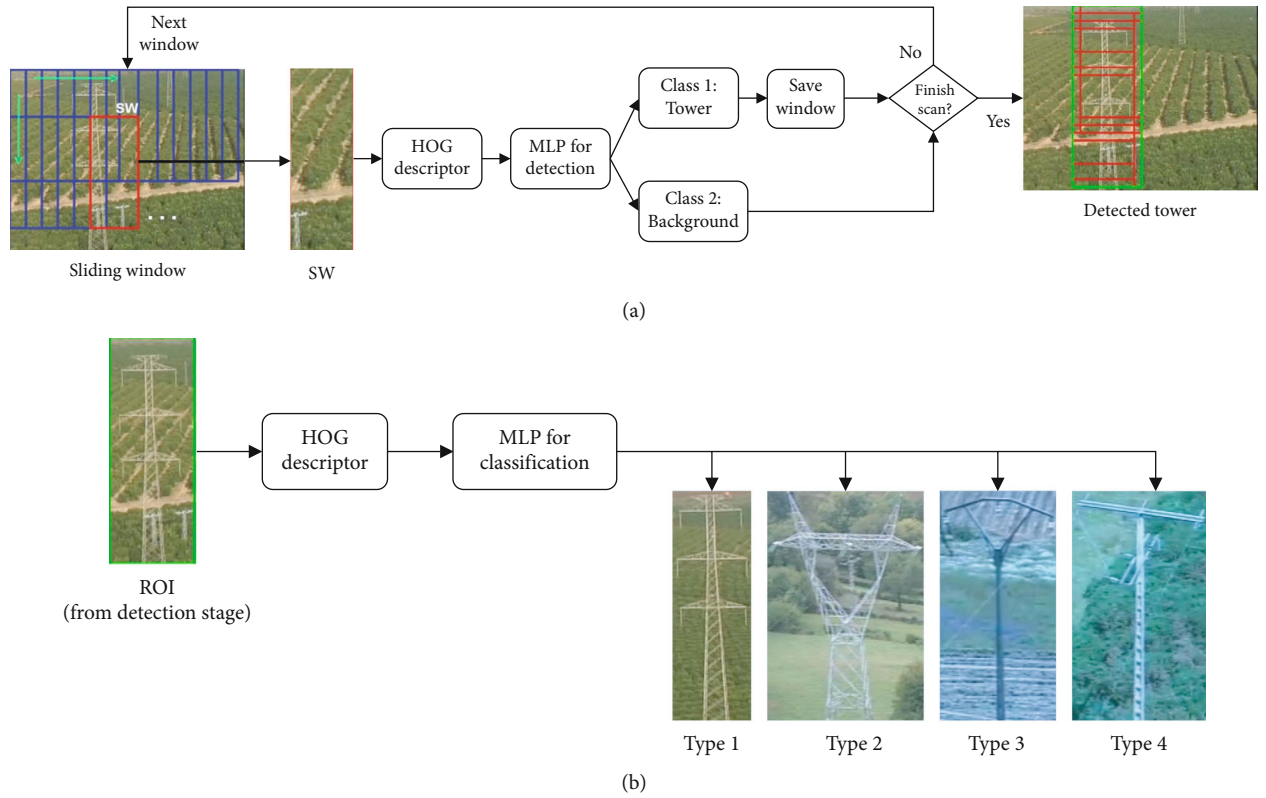


Figure 4: The system architecture: (a) pylon detection stage; (b) pylon classification stage [57].

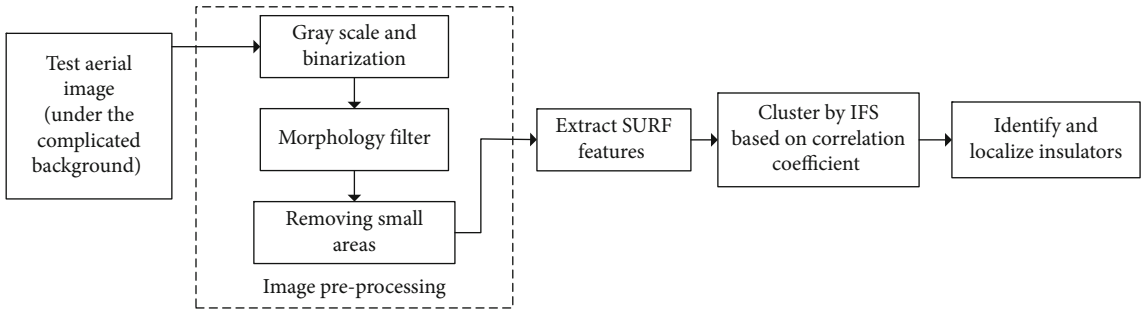


Figure 5: The steps of the identification and localization of insulators [64].

applied to pylon level and device level generally use the framework of "feature expression" + "classifier". Consequently, features and classifiers are vital for the precision of detecting different objects. In traditional algorithms, basic features usually cannot perfectly characterize the detected target. Features need to be improved as the detected object is replaced. This phenomenon is more obvious in the field of PTL inspection. There are nearly two thousand kinds of fittings in different specifications used in the field of PTL inspection. The features acquired from different regions cannot be unified. For the purpose of ensuring precision, it is imperative to design features for specific targets. Although the features designed for specific objects are intuitive and accessible, it has drawbacks such as low efficiency and poor generalization ability. In addition, tasks of PTL inspection require the robustness and real-time performance of algorithms. Facing the changes of light and viewing angle in the environment, handcrafted features are less robust. Besides, detection accuracy will be immensely influenced in the face of harsh operating environments. Moreover, handcrafted features are highly complex. And they are poorly real-time, which cannot satisfy needs of PTL tasks.

3.2. Deep Learning-Based Object Detection Algorithms. Powered by deep learning, R-CNN [49] broke the deadlock of stagnation of object detection. Plenty of similar approaches emerged in the following years, and they are also applied in PTL. These algorithms are classified as the "two-stage detection" and the "one-stage detection". In "two-stage detection" algorithms, since the speed of Faster R-CNN is relatively quick [71, 72], it is often used in PTL inspection. In "one-stage detection" algorithms, the algorithms of the YOLO

series [73–76] are used much more. Object detection algorithms, which are based on deep learning, can also be classified as two aspects like traditional algorithms.

3.2.1. Pylons. In terms of pylons, Guo et al. [77] presented a real-time pylon detection model based on YOLO. The method used *k*-means to improve the anchor parameters in YOLO. Its detection rate was about twenty frames per second, and mAP was 94.09%. Wang et al. [78] verified the performance of different networks for detecting pylons. For birds' nests, Shi et al. [79] proposed a bird's nest detection model by employing RetinaNet [80]. This model was accurate but not real-time. No other neural network models had been tried as the backbone in this model, so there were some limitations in the experiment. Aiming at the same task, Wang et al. [81] proposed an approach, which was used to detect birds' nests in a multiscale of high-voltage pylons. For enhancing the detection ability for birds' nests, multiscale convolution features and region proposals were employed in this model, which was based on Faster R-CNN. The model was evaluated on a test set composed of two thousand images. And the average detection precision reached 84.55%.

3.2.2. Insulators and Fittings. For insulators, Liu et al. [82] employed Faster R-CNN to detect insulators. The precision reached 94%. Zhao et al. [83] proposed an approach based on improved Faster R-CNN. This approach achieved better detection of insulators by improving NMS and anchor generation in RPN. It also solved the problem of missed detection caused by insulator shielding. Han et al. [84] designed a new neural network to detect insulators. There was high average precision but low real-time performance in the model.

In terms of fittings, the detection based on deep learning is no longer limited to a single target. In recent years, research on multiobject detection has emerged. The targets include a variety of fittings and insulators. Lin et al. [85] advanced a means based on improved Faster R-CNN. The means could keep a good detection performance in different resolutions, angles, and positions. Dong [86] came up with a real-time multidevice detection method based on YOLOv3. Data augmentation was used to improve the model's performance. And it had good average accuracy for insulators and vibration dampers. Yang et al. [87] presented an MSFF-KCD algorithm to detect multiscale devices. The precision of this means could maintain 86% on ARM devices, but the speed is not quick. In addition, Yang et al. [88] compared the performance of different backbone networks based on SSD [89] and used feature fusion to improve detection accuracy. For the purpose of detecting fittings, Qi et al. [90] presented an improved SSD model. This model improved the detection performance of small or dense targets in a complex background by improving Intersection over Union (IoU) and using the repulsion loss function. For the same task, Liu et al. [91] proposed a means based on R-FCN. Online hard example mining, sample adjusting, and soft NMS were employed to promote the performance of R-FCN.

The above research of multidevice detection indicates that deep learning-based detection algorithms have a good generalization ability. They do without a complicated process of handcrafted features and can choose suitable and robust features to express targets. For multiscale problems, similar to the traditional algorithms, their performance is poor. So it is necessary to consider the use of multiscale technologies, such as improved anchor generation, SNIP [92], and TridentNet [93], to solve this problem. It is a good direction for the power line inspection study.

Due to the confidentiality of power line inspection datasets, there is no one public dataset now, which hinders the progress of power line inspection. So only some typical deep learning-based detection algorithms for PTL inspection are listed in Table 2 for comparing performance. In China, there are two power grid companies, which are State Grid Corporation of China and China Southern Power Grid. They have massive datasets of the power grid. Consequently, if they cooperatively create a public dataset, which combines data from power grids across the country, researchers could first use public datasets to testify the effectiveness of the means or the model proposed by themselves and show their results based on private datasets.

Although it is impossible to accurately compare the performance between these models, it can be seen that the performance is improved due to new modified models. However, there do not exist neural network models, which achieve simultaneously high precision and speed, in object detection. Just as these two series, the precision and the speed are focused on severally. The improved algorithms based on these two series often sacrifice real time to improve detection accuracy. Now researchers always use YOLO series to detect objects because of its excellent real-time performance. Nevertheless, PTL inspection should be able to fulfill the high requirements in both aspects. Thus, many improved algorithms are only in the research stage. Moreover, limited by computational capabilities of embedded platforms, they cannot be actually used in engineering. Besides, the deep learning model used in engineering needs to be capable of training with new data, but little research is ongoing in this area.

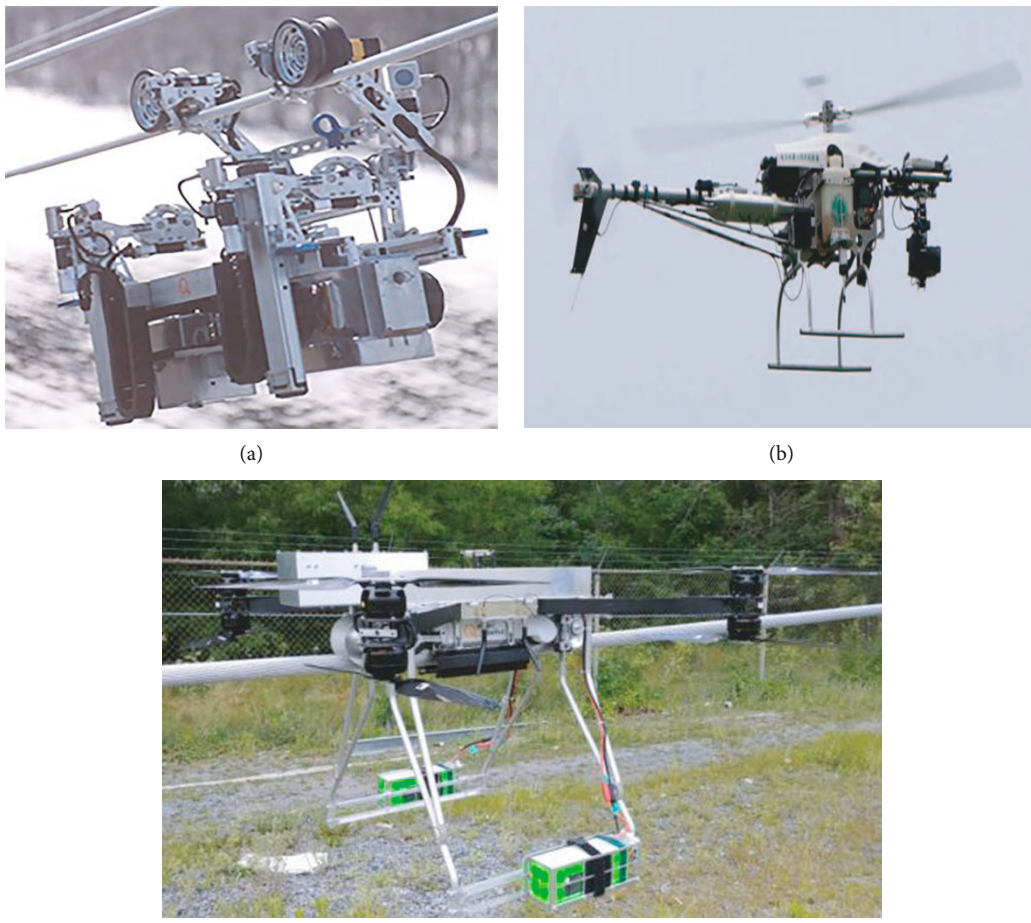
4. Visual Servo of PTL Inspection

The visual servo system takes visual information as feedback. After the relative pose between the robot and the target is computed by visual information, it can realize the robot's pose control. Three classification standards including types of feedback [95–97], camera location [98], and calibration requirements [99] are commonly used in visual servo systems. Among them, PBVS and IBVS are usually employed in the tasks of PTL inspection. The former is based on position, and the latter is based on images. According to operation modes, inspection robots can be divided into climbing robots, UAVs, and hybrid platforms [94], as shown in Figure 6. Research on hybrid platforms is not yet mature, and visual servo control solutions are only used in the first two types of inspection robots.

4.1. Servo Control of Climbing Robots. The visual servo control of climbing robots is most applied to autonomous line-

Table 2: Typical deep learning-based detection algorithms for PTL inspection.

Reference	Object	Network	Dataset	mAP
Guo et al. [77]	Pylon	YOLO+anchor cluster based on *k*-means	Training set: 7636 images, test set: 815 images	88.68%
Wang et al. [81]	Bird's nest	Faster R-CNN+modified ResNet-50	Training set: 9000 images, test set: 3000 images	94.09%
Han et al. [84]	Insulator	A new model based on modified ResNet-50	Training set: 2675 images, test set: 1356 images	98.30%
Yang et al. [88]	Fittings	MSFF-KCD	Training set: 3440 images, test set: 382 images	90.80%



(a) (b) (c)

Figure 6: The PTL inspection robots: (a) LineScout (climbing robots); (b) SmartCopter (UAVs); (c) LineDrone (hybrid platforms) [94].

grasping control in the process of bypassing obstacles. And visual servo is less used in other aspects of control.

In terms of autonomous line grabbing, Wang et al. [100] designed a single-camera-based visual servo control method for line grasping. Although this method could meet the line-grasping requirements, it mainly depended on the mechanical structure of the robot. Zhang et al. [101] designed an IBVS controller for getting across obstacles. The image processing of this scheme was complicated, and it was sensitive to light changes. After that, Zhao et al. [102] proposed a means of combining remote control and IBVS. This method was effectively carried out in the laboratory environment. Guo et al. [103] raised an IBVS method for autonomous line-grasping control. This method used a 2D fuzzy controller. The final deflection error of a grasping line was within two degrees, and the time was less than forty seconds. Wang et al. [104] presented a line-grasping control based on hand-eye-vision. This method used a 2D fuzzy controller as shown in Figure 7. The impacts of changes in lighting, background, and other factors on the line-grasping accuracy were considered comprehensively. And it consumed twenty seconds. Aiming at a new dual-arm inspection robot designed by themselves, He et al. [105] proposed an adaptive visual servo method. This method could provide good characteristics of power lines without hand-eye calibration, and it improved the robustness of autonomously grasping the line.

The above-mentioned studies are aimed at the line-grasping problem of dual-arm PTL inspection robots, and many of them use IBVS. But the inspection robots only complete the autonomous line-grasping function in the laboratory environment, and they are not used in actual application scenarios. The problems of image noise, model

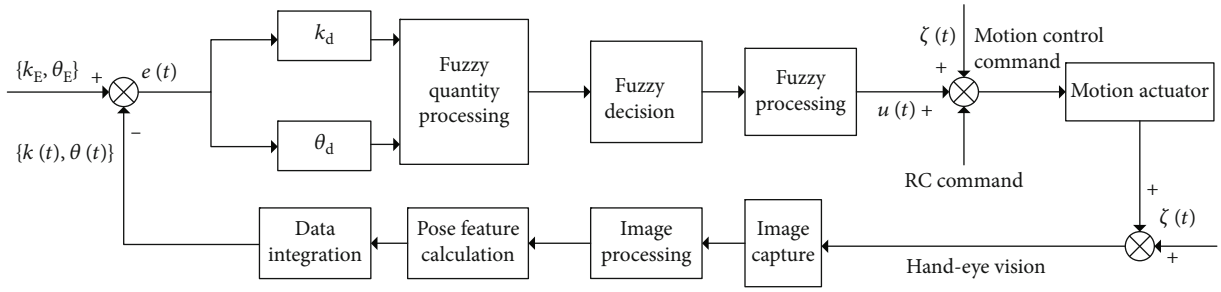


Figure 7: Visual servo fuzzy controller [104].

errors, control delay, and environmental changes are not considered, which may be met in the actual scenarios. Besides, it takes too long for the inspection robots to grasp the line and the efficiency is low. Compared with traditional visual servo, uncalibrated visual servo has higher flexibility and adaptability, which is more suitable for inspection robots. Therefore, uncalibrated visual servo of inspection robots will be the future research direction.

In other ways, aiming at the problem of autodocking charging control for inspection robots, Wu et al. [107] used IBVS for precise positioning control. A variable universe fuzzy control was employed to overcome the difficulty of insufficient control precision. Besides, Jiang et al. [106] presented a double closed-loop autonomous localization control means. It originated from the BP network and visual servo. The BP network and visual servo were, respectively, used to solve the coarse and fine positioning of the manipulator. This method could effectually boost the automation level of inspection robots, but it took too long to complete the bolt alignment. The control architecture diagram of the method is shown in Figure 8.

Intelligent methods have begun to be used in the visual servo of inspection robots. However, they are not fully used in visual servo because of calculation drawbacks of intelligent means. Besides, many of the above-mentioned visual servo studies only concentrate on one task. However, there were many maintenance tasks on the PTL. The visual servo scheme based on single-task control has poor applicability, so it is necessary to focus on the research of multitask visual servo control.

4.2. Servo Control of UAVs. Compared with the visual servo control of climbing robots, the research scope of UAVs' visual servo control is more limited. Currently, only visual servo control of tracking power lines is available.

Aiming at the problem of tracking power lines, Araar and Aouf [108] of Cranfield University proposed two solutions, which, respectively, were an IBVS method combined with the LQ-Servo and a partial PBVS (PPBVS) method. There was a better effect of PPBVS by contrast. However, this method was greatly affected by calibration errors. Xie et al. [109] raised a novel IBVS method. It could be robust to camera calibration errors without depth estimation. However, it was only in the simulation stage. Aiming at the problem of fixed-wing UAV tracking targets in the presence of wind, Mills et al. [110] designed an IBVS method. It considered the wind correction angle as the desired line angle. The result showed that it could boost the tracking response of UAVs. Rafique and Lynch [111] used IBVS combined with output feedback for the motion control of UAVs. Taking into account uncertain factors such as linear acceleration disturbance and quality, it adopted the inner-outer loop structure. The result showed that the stability of the structure is good. However, it was not tested in the outdoor PTL environment.

Many studies of visual servo control of UAVs are in the simulation stage. UAVs are notably influenced by environmental factors. The influence of wind on the motion control of UAVs needs to be considered in tracking power lines. Besides, compared to climbing robots, the inspection speed of UAVs is faster. It is necessary to boost response speed of servo during tracking. On the basis of tracking power lines, UAVs also need to perform path planning. So path planning for visual servo of UAVs is a good research direction.

5. Problems and Prospects

Many research institutions have started to study inspection robots globally. However, the research of inspection robots is only in the preliminary stage. Some key problems remain unsolved in terms of 3D reconstruction, object detection, and visual servo control. At the same time, the research of inspection robots in other areas still needs to develop.

5.1. Pivotal Technical Problems. It is effective to use inspection robots to replace manual inspection. And there are clear application requirements of PTL inspection. Hence, there is a very important engineering value to develop an inspection robot system that can perform efficient operations in a complex PTL environment. At present, studies concentrates on mechanical design, environmental perception, and visual servo control in the field of PTL inspection. Nevertheless, many of the results are mechanical design-related, and few results are given in other areas. Therefore, for boosting the efficiency and operation precision of inspection robots, the following problems need to be solved.

5.1.1. The Problem of Environmental Perception. Both the high-altitude environment and natural interference factors pose challenges to inspection robots. In this scenario, highly accurate environmental information is imperative for task conduction. The use of a single sensor will inevitably have perceptual limitations. For example, illumination variations

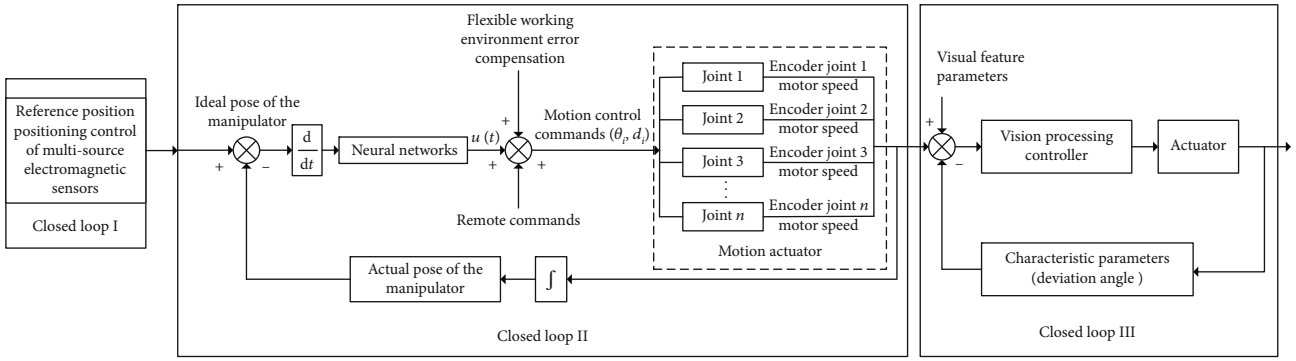


Figure 8: The control architecture diagram of the method based on the BP network and visual servo [106].

affect the camera and fog affects the LiDAR. When power devices are aging or damaged, fault areas on the power devices can be hardly recognized with naked eyes, but heat or discharge will occur in the fault area. These phenomena can be detected by infrared thermal devices and UV cameras. Therefore, multisensor information fusion is indispensable for PTL inspection. As shown in Figure 9, it can be carried out from the following steps:

3D fusion reconstruction of transmission line corridors is used. Firstly, a rough 3D model of the transmission line corridor is constructed by oblique photography. The data registration is completed by the ICP algorithm to form a detailed 3D model. The fault areas are approximately found by the detailed 3D model.

The data fusion of infrared, visible light, and ultraviolet images is employed. After determining approximate fault areas and time synchronization, these sensors are used to shoot for image fusion. At present, image fusion still rests on the pixel stage. Feature fusion should be further considered for it.

The accurate coordinate of discharge, hotspot, and visible defects in fault areas are gained by object detection. The spatial conversion relations are gained by the calibration of the visible-light camera and LiDAR. For getting 3D information, the precise coordinate of the defect area is projected into point clouds according to spatial conversion relations. Then, the 3D information is used for local 3D reconstruction.

Visual servo control is applied. The current pose of the robot and objects can be obtained by local 3D reconstruction. After that, it can be provided to a 3D visual servo controller to complete the control.

5.1.2. The Problem of Datasets Required for Deep Learning. With the development of AI, a growing number of fields commence research by employing deep learning. Deep learning algorithms rely on data, but there is no standard power inspection dataset for researchers to conduct algorithm research. Therefore, the establishment of a standard power inspection dataset for deep learning development promotion is urgently needed. It can be constructed from the following steps:

Standards of inspection images and power devices are made. Aiming at different defects of power devices, a unified defect description and classification rules should be established. The size and sharpness of inspection images should be normalized. Standards provide convenience when instructing personnel in data collection and highly improve the dataset construction.

Datasets of PTL inspection should be divided according to various regions. The geographical environment and weather conditions of each region are different. PTL will be affected by a distinct climate, which makes the type of the same power device be disparate. Different types of power devices may vary greatly in appearance and shape, which results in regional characteristics. Therefore, there are regional characteristics in PTL corridors. In addition, the background of various regions differs greatly. Hence, it is necessary to divide the whole dataset first according to the territories.

According to different kinds of power devices and degrees of defects, datasets of regional PTL inspection should be further divided into different subset datasets. According to specific power devices in PTL, it is helpful to form the corresponding PTL dataset by subset datasets, which can represent the characteristics greatly and be helpful for training models. The steps are shown in Figure 10.

Now, a good deep learning model is always based on high-quality annotated data. However, it is expensive to collect data, especially annotated data. In the field of PTL inspection, this phenomenon is more obvious. Consequently, researchers try to employ weakly supervised data to train models [112, 113]. Furthermore, they employ unsupervised learning to annotate data. These methods can be also used in object detection of PTL inspection [114] for reducing the cost of data annotation. It will be a good research direction of PTL inspection.

5.1.3. The Problem of Applying a Light Network Model Suitable for Embedded Platforms. The training of models usually runs on servers. The computational ability should be considered when deep learning-based object algorithms are used. Therefore, setting up a light-weight network in advance is significant. Then the model, transferred from the server to the embedded platform, can be additionally trained in actual scenarios. At present, different backbones such as SSD and YOLOv4 can be used in MobileNetv3. Furthermore, fine-tuning can also be employed.

5.1.4. The Problem of Control Efficiency and Robustness of the Visual Servo System. Now, the whole control process of

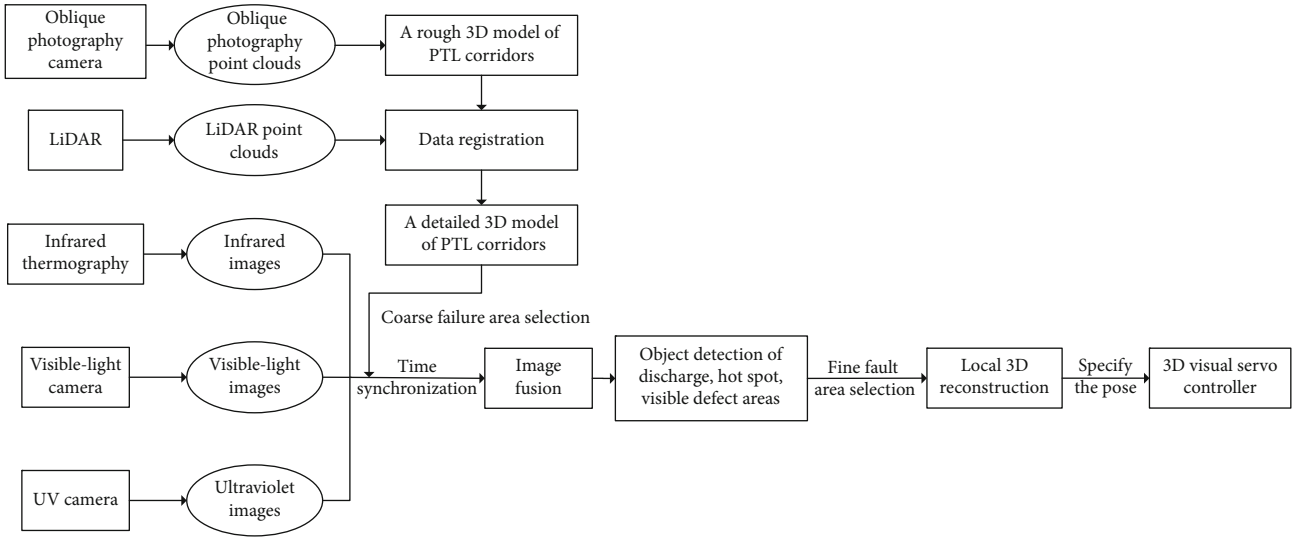


Figure 9: The flowchart of multisensor information fusion.

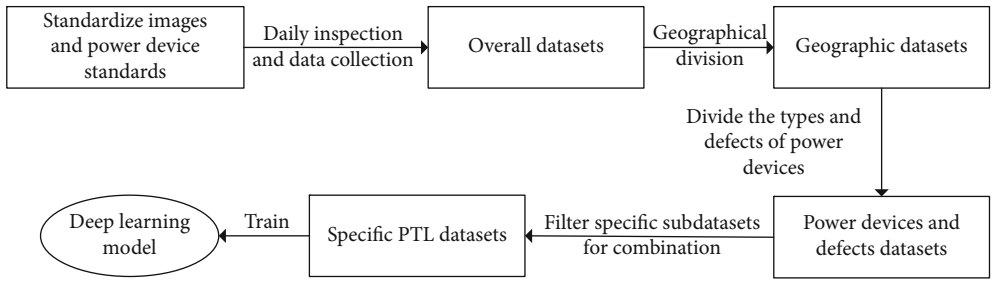


Figure 10: The steps of power inspection dataset establishment.

grasping the line or tracking lines takes much time. And the efficiency is low. In this regard, improving visual sampling frequency and image processing speed is able to ameliorate its real-time capability. In addition, subtle structural changes of inspection robots may occur due to wind and drag of vibration dampers, which leads to errors. The uncalibrated visual servo has a high tolerance for these errors. Therefore, it can be adopted for improving the robustness.

5.2. Prospects. In the application scenarios of PTL inspection, not only the foundation capabilities of inspection robots but also further expansion should be realized according to requirements. It is helpful to realize the vital engineering value of inspection robots. The facets of future directions are as follows:

5.2.1. The Development of Mechanical Design. The core of inspection robots is the mechanical design. And a good one can greatly improve the efficiency. For example, compared to LineScout [115], LineRanger [116] takes much less time to cross obstacles because of the ingenious and passive obstacle-crossing mechanism. However, the mechanical design of PTL inspection robots is not refined now, and robots cannot balance inspection and operation. So far, in contrast to other inspection robots, climbing robots are more practical due to their in-contact measurements. The engineering value of climbing robots is greater when they can succeed in crossing pylons. So the mechanical design of inspection robots still needs to be improved.

5.2.2. The Development of Energy Harvesting. Battery capacities directly influence the duration of PTL inspection. Regarding a certain load capacity of robots, a high-capacity battery, such as lithium battery, can increase maximum endurance of robots, but it limits the weight of other mechanical components. So online charging is important. For increasing the inspection endurance, inspection robots can acquire electric energy from power line autonomously. So an induction power supply system should be developed.

5.2.3. The Development of Electromagnetic Protection. EMI on inspection robots is more intense with increasing voltage, especially the transient discharge process of the equipotential working. However, the present inspection robots are weak to EMI. Therefore, it is an inevitable problem during the inspection. The entire robot should conduct electricity for avoiding electrical discharge within components of robots. Shielding is an effective protection measurement aimed at EMI, and high-quality shields should be designed. However, inspection robots cannot be fully shielded, which means that other optimization technologies, such as routing, filtering, and

grounding techniques, should be considered. Consequently, how to design electromagnetic protection is a crucial challenge.

5.2.4. The Development of Modularity. It is inevitable that inspection robots will be damaged in a bad working environment. And the modularity development of inspection robots is significant for rapid maintenance. The modularity of inspection robot is convenient for the maintainer to find the fault location and replace the fault module of robots. Moreover, it is easy to install and disassemble inspection robots by modularity, which also improves the efficiency of online and offline switchover. Furthermore, it helps to switch robot configurations in different inspection scenarios, such as overhead ground wire and multibundled conductors.

5.2.5. The Development of Reliability. The working environment of inspection robots is harsh, such as the power line area with long span and steep slope. There is a requirement for good robustness to accurately complete the task. Consequently, inspection robots should have the ability to cross the most complex area in PTL. For the purpose of satisfying the needs of multiple and complicated tasks, effective load capacity and operation control technology should be improved. What is more, before online operation, robots should be tested for endurance, such as wear of the mechanical system, battery performance, and EMI robustness.

5.2.6. The Development of Cooperation. This cooperation not only refers to the cooperation among a variety of inspection robots. It also includes the cooperation between inspection robots and the ground control system (GCS). The multirobot system can promote the performance of PTL inspection. In addition, in the case of teleoperation, the efficiency of inspection can be also affected by the cooperation between inspection robots and GCS. Appropriate deployments of GCS points in a long PTL corridor are the key to the cooperation between inspection robots and GCS. The cooperation includes real-time bidirectional data transmission, the energy consumption of disassembly, installation, and teleoperation.

5.2.7. The Development of Friendly Operability and Autonomy. The final goal is to achieve autonomous patrol of inspection robots. However, autonomous control and teleoperation need to coexist in the process of development. Although inspection robots could automatically get across obstacles such as spacers and suspension clamps, there is a lack of autonomous ability in specific operations. The specific operations include dealing with damaged power lines and removing foreign bodies. For fulfilling tasks securely and briskly, teleoperation should be used. Meanwhile, the friendly man-machine interface of inspection robots should be developed for simple teleoperation.

6. Conclusion

For promoting researchers to develop advanced research of PTL inspection, studies of 3D reconstruction, object detection, and visual servo of PTL inspection are summarized. The extant problems, such as perception limitation and lack of datasets, are correspondingly discussed. What is more, several aspects are posed, which are promising research directions in the field of PTL inspection. In a word, the present perception technologies and control technologies cannot meet the requirement of automatic and accurate power line inspection. With the emergence of these research results, the automation degree of inspection robots will be highly boosted.

Conflicts of Interest

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

Acknowledgments

This project was supported by the National Key Research and Development Program of China (Grant 2018YFB1307400) and the National Natural Science Foundation of China (Grant 61873267).

References

[1] N. Pouliot, P. L. Richard, and S. Montambault, "LineScout technology opens the way to robotic inspection and maintenance of high-voltage power lines," *IEEE Power and Energy Technology Systems Journal*, vol. 2, no. 1, pp. 1–11, 2015.

[2] C. Zhang, Z. Liu, S. Yang, and B. Xu, "Key technologies of laser point cloud data processing in power line corridor," in *Conference on LIDAR Imaging Detection and Target Recognition*, Changchun, China, 2017.

[3] X. Mai, C. Chen, and X. Peng, "3D visualization technique of transmission line corridors: system design and implementation," *Electric Power*, vol. 48, no. 2, pp. 98–103, 2015.

[4] B. Chen and X. Miao, "Distribution line pole detection and counting based on YOLO using UAV inspection line video," *Journal of Electrical Engineering and Technology*, vol. 15, no. 1, pp. 441–448, 2020.

[5] F. Gao, J. Wang, Z. Kong et al., "Recognition of insulator explosion based on deep learning," in *14th IEEE International Computer Conference on Wavelet Active Media Technology and Information Processing*, pp. 79–82, Chengdu, China, 2017.

[6] Z. Zhao, A. Jiang, Y. Qi, W. Zhang, and W. Zhao, "Fittings detection in transmission line images with SSD model embedded occlusion relation module," *CAAI Transactions on Intelligent Systems*, vol. 15, no. 4, pp. 1–7, 2020.

[7] K. Toussaint, N. Pouliot, and S. Montambault, "Transmission line maintenance robots capable of crossing obstacles: state-of-the-art review and challenges ahead," *Journal of Field Robotics*, vol. 26, no. 5, pp. 477–499, 2009.

[8] S. Montambault and N. Pouliot, "About the future of power line robotics," in *1st International Conference on Applied Robotics for the Power Industry*, pp. 1–6, Montreal, QC, Canada, 2010.

[9] K. Kraus and N. Pfeifer, "Determination of terrain models in wooded areas with airborne laser scanner data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 53, no. 4, pp. 193–203, 1998.

[10] W. Boehler, M. Vicent, and A. Marbs, "Investigating laser scanner accuracy," *The International Archives of*

Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 34, no. 5, pp. 696–701, 2003.

[11] W. Goebel, B. M. Kampa, and F. Helmchen, "Imaging cellular network dynamics in three dimensions using fast 3d laser scanning," *Nature Methods*, vol. 4, no. 1, pp. 73–79, 2007.

[12] M. C. Amann, T. Bosch, and M. Lescure, "Laser ranging: a critical review of usual techniques for distance measurement," *Optical Engineering*, vol. 40, no. 1, pp. 10–19, 2001.

[13] S. May, D. Droeschel, D. Holz, C. Wiesen, and S. Fuchs, "3d pose estimation and mapping with time-of-flight cameras," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, France, 2008.

[14] S. Zhu and X. Qiang, "Analysis of 3-d coordinate vision measuring methods with feature points on workpiece," *Optics and Precision Engineering*, vol. 8, no. 2, pp. 192–197, 2000.

[15] D. Geng, Y. He, and H. Su, "Study on the measurement of transparent step by white-light interferometer," *Optical Instruments*, vol. 35, no. 6, pp. 74–77, 2013.

[16] L. W. Bingleman and G. S. Schajer, "DIC-based surface motion correction for ESPI measurements," *Experimental Mechanics*, vol. 51, no. 7, pp. 1207–1216, 2011.

[17] L. Sun and Y. Yu, "Transient 3D deformation measurement method by color splitting based on phase shift and ESPI," *Optical Instruments*, vol. 38, no. 1, pp. 20–26, 2016.

[18] M. R. Oswald, E. Töppe, C. Nieuwenhuis, and D. Cremers, "A review of geometry recovery from a single image focusing on curved object reconstruction," in *Innovations for Shape Analysis*, Springer, Heidelberg, Berlin, 2013.

[19] J. Chen, Y. Zhang, P. Song, Y. Wei, and Y. Wang, "Application of deep learning to 3d object reconstruction from a single image," *Acta Automatica Sinica*, vol. 45, no. 4, pp. 657–668, 2019.

[20] M. Sizintsev and R. P. Wildes, "Spacetime stereo and 3d flow via binocular spatiotemporal orientation analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2241–2254, 2014.

[21] F. Qi, D. Zhao, and W. Gao, "Reduced reference stereoscopic image quality assessment based on binocular perceptual information," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2338–2344, 2015.

[22] C. Stewart and C. Dyer, "The trinocular general support algorithm: a three-camera stereo algorithm for overcoming binocular matching errors," in *1988 Second International Conference on Computer Vision*, pp. 134–138, Los Alamitos, USA, 1988.

[23] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz, "Multi-view stereo for community photo collections," in *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, Rio de Janeiro, Brazil, 2007.

[24] D. Yang, "Digital terrain model," *Bulletin of Surveying and Mapping*, vol. 3, pp. 3–5, 1998.

[25] P. Axelsson, "Dem generation from laser scanner data using adaptive tin models," *International Archives of Photogrammetry and Remote Sensing*, vol. 33, pp. 110–117, 2000.

[26] J. Yu, C. Mu, Y. Feng, and Y. Dou, "Powerlines extraction techniques from airborne LiDAR data," *Geomatics and Information Science of Wuhan University*, vol. 36, no. 11, pp. 1275–1279, 2011.

[27] G. Vosselman, "Slope based filtering of laser altimetry data," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 33, pp. 935–942, 2000.

[28] K. Zhang, S. C. Chen, D. Whitman, M. L. SHyu, J. Yan, and C. Zhang, "A progressive morphological filter for removing nonground measurements from airborne LiDAR data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 4, pp. 872–882, 2003.

[29] Z. Liu, J. Liang, and J. Zhang, "Power lines extraction from airborne LiDAR data using spatial domain segmentation," *Journal of Remote Sensing*, vol. 18, no. 1, pp. 61–76, 2014.

[30] X. Shen, C. Qin, Y. Du, and X. Yu, "An automatic power line extraction method from airborne light detection and ranging point cloud in complex terrain," *Journal of Tongji University. Natural Science*, vol. 46, no. 7, pp. 982–987, 2018.

[31] R. A. McLaughlin, "Extracting transmission lines from airborne LiDAR data," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 2, pp. 222–226, 2006.

[32] Y. Jwa and G. Sohn, "A piecewise catenary curve model growing for 3d power line reconstruction," *Photogrammetric Engineering and Remote Sensing*, vol. 78, no. 12, pp. 1227–1240, 2012.

[33] J. Wu, L. Chen, L. Li, J. Yang, and X. Liang, "Power line extraction and reconstruction from airborne LiDAR point cloud," *Laser Technology*, vol. 43, no. 4, pp. 500–505, 2019.

[34] T. Melzer and C. Briese, "Extraction and modeling of power lines from ALS abstract: point clouds," *28th Workshop of Austrian Association for Pattern Recognition*, pp. 47–54, 2004.

[35] X. Lin, M. Duan, J. Zhang, and Y. Zang, "A method of reconstructing 3d powerlines from airborne LiDAR point clouds," *Science of Surveying and Mapping*, vol. 41, no. 1, pp. 109–114,64, 2016.

[36] X. Lai, D. Dai, M. Zheng, and Y. Du, "Powerline three-dimensional reconstruction for LiDAR point cloud data," *Journal of Remote Sensing*, vol. 18, no. 6, pp. 1223–1229, 2014.

[37] X. Lin and J. Zhang, "3d power line reconstruction from airborne LiDAR point cloud of overhead electric power transmission corridors," *Acta Geodetica et Cartographica Sinica*, vol. 45, no. 3, pp. 347–353, 2016.

[38] L. Cheng, L. Tong, Y. Wang, and M. Li, "Extraction of urban power lines from vehicle-borne LiDAR data," *Remote Sensing*, vol. 6, no. 4, pp. 3302–3320, 2014.

[39] J. Liang, J. Zhang, K. Deng, Z. Liu, and Q. Shi, "A new power-line extraction method based on airborne LiDAR point cloud data," in *2011 International Symposium on Image and Data Fusion*, pp. 1–4, Tengchong, China, 2011.

[40] J. Zhang, M. Duan, X. Lin, and Y. Zang, "Comparison and analysis of models for 3d power line reconstruction using LiDAR point cloud," *Geomatics and Information Science of Wuhan University*, vol. 42, no. 11, pp. 1565–1572, 2017.

[41] W. Zhang, G. Yan, N. Wang, Q. Li, and W. Zhao, "Automatic 3d power line reconstruction of multi-angular imaging power line inspection system," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 6752, no. 1, pp. 10–19, 2007.

[42] F. Ganovelli, L. Malomo, and R. Scopigno, "Reconstructing power lines from images," in *2018 International Conference on Image and Vision Computing New Zealand*, pp. 1–6, Auckland, New Zealand, 2018.

[43] M. Maurer, M. Hofer, F. Fraundorfer, and H. Bischof, "Automated inspection of power line corridors to measure

vegetation undercut using UAV-based images," in *International Conference on Unmanned Aerial Vehicles in Geomatics*, pp. 33–40, Bonn, Germany, 2017.

[44] Z. Chen, Z. Lan, H. Long, and Q. Hu, "3d modeling of pylon from airborne lidar data," , Article ID 915807*Remote Sensing of the Environment: 18th National Symposium on Remote Sensing of China*, vol. 9158, 2014International Society for Optics and Photonics, 2014.

[45] B. Guo, X. Huang, Q. Li, F. Zhang, J. Zhu, and C. Wang, "A stochastic geometry method for pylon reconstruction from airborne LiDAR data," *Remote Sensing*, vol. 8, no. 3, p. 243, 2016.

[46] L. Xi, H. Zhao, S. Wang, X. Zhang, and M. Ma, "The design of unmanned aerial vehicle power patrol base on oblique photography technology," *Electronic Science and Technology*, vol. 32, no. 5, pp. 89–92, 2019.

[47] H. Pei, S. Jiang, G. Lin, H. Huang, W. Jiang, and C. Yang, "3d reconstruction of transmission route based on UAV oblique photogrammetry," *Science of Surveying and Mapping*, vol. 41, no. 12, pp. 292–296, 2016.

[48] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: a survey," 2019.

[49] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, 2014.

[50] J. Platt, "Sequential minimal optimization: a fast algorithm for training support vector machines," *Advances in Kernel Methods-Support Vector Learning*, vol. 208, 1998.

[51] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[52] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *18th International Conference on Pattern Recognition*, vol. 3, pp. 850–855, Hong Kong, China, 2006.

[53] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[54] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[55] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[56] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, San Diego, CA, 2005.

[57] C. Sampedro, C. Martinez, A. Chauhan, and P. Campoy, "A supervised approach to electric tower detection and classification for power line inspection," in *2014 International Joint Conference on Neural Networks*, pp. 1970–1977, Beijing, China, 2014.

[58] A. Cerón, I. Mondragón, and F. Prieto, "Real-time transmission tower detection from video based on a feature descriptor," *IET Computer Vision*, vol. 11, no. 1, pp. 33–42, 2017.

[59] X. Wang, D. Li, and L. Zhang, "A supervised approach to electric tower detection for power line inspection," *Northeast Electric Power Technology*, vol. 38, no. 11, pp. 12–14,19, 2017.

[60] Z. Wang, J. Han, X. Sun, and B. Yang, "Method for orientation determination of transmission line tower based on visual navigation," *Laser and Optoelectronics Progress*, vol. 56, no. 8, 2019.

[61] Y. Zhang, Y. Chen, D. Wang, Z. Qian, and C. Ma, "Coarse-to-fine detection for nests on pylon," *Information Technology*, vol. 3, pp. 104–109, 2017.

[62] J. Xu, J. Han, Z. Tong, and Y. Wang, "Method for detecting bird's nest on tower based on UAV image," *Computer Engineering and Application*, vol. 53, no. 6, pp. 231–235, 2017.

[63] Z. Zhao, N. Liu, and Y. Yuan, "The recognition and localization of insulators based on SIFT and RANSAC," in *Proceedings of 3rd International Conference on Multimedia Technology*, pp. 692–699, Guangzhou, China, 2013.

[64] Z. Zhao and N. Liu, "The recognition and localization of insulators adopting SURF and IFS based on correlation coefficient," *Optik*, vol. 125, no. 20, pp. 6049–6052, 2014.

[65] P. S. Prasad and B. P. Rao, "LBP-HF features and machine learning applied for automated monitoring of insulators for overhead power distribution lines," in *2016 International Conference on Wireless Communications, Signal Processing and Networking*, pp. 808–812, Chennai, India, 2016.

[66] F. Zhang, R. Guo, Z. Cheng et al., "Detection for transmission line obstacles based on principal component gradient histogram," *Computer Engineering and Application*, vol. 52, no. 15, pp. 254–259,265, 2016.

[67] D. Zhang, X. Qiu, C. Cao, and J. Zhu, "An vibration damper detection algorithm combined with aggregation channel and complex frequency domain features," *Computer Technology and Development*, vol. 30, no. 3, pp. 147–151, 2020.

[68] M. Feng, W. Luo, L. Yu et al., "A bolt detection method for pictures captured from an unmanned aerial vehicle in power transmission line inspection," *Journal of Electric Power Science and Technology*, vol. 33, no. 4, pp. 135–140, 2018.

[69] S. Fan, D. Yang, D. Zou, and Y. Yan, "Vision-based tracing, recognition and positioning strategy for bolt tightening live working robot on power transmission line," *Journal of Electronic Measurement and Instrumentation*, vol. 31, no. 9, pp. 1514–1523, 2017.

[70] Z. Zhao, H. Qi, and L. Nie, "Research overview on visual detection of transmission lines based on deep learning," *Guangdong Electric Power*, vol. 32, no. 9, pp. 11–23, 2019.

[71] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision*, pp. 1440–1448, Santiago, Chile, 2015.

[72] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[73] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Seattle, WA, USA, 2016.

[74] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6517–6525, Honolulu, HI, USA, 2017.

[75] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," 2018.

[76] A. Bochkovskiy, C. Y. Wang, and H. Y. Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020.

[77] J. Guo, B. Chen, R. Wang, J. Wang, and L. Zhong, "YOLO-based real-time detection of power line poles from unmanned

aerial vehicle inspection vision," *Electric Power*, vol. 52, no. 7, pp. 17–23, 2019.

[78] H. Wang, G. Yang, E. Li, Y. Tian, M. Zhao, and Z. Liang, "High-voltage power transmission tower detection based on Faster R-CNN and YOLO-V3," in *2019 Chinese Control Conference*, pp. 8750–8755, Guangzhou, China, 2019.

[79] L. Shi, H. Yang, Z. Zhou, L. Yang, H. Zhang, and H. Du, "Intelligent detection of bird's nest based on RetinaNet model," *Power Systems and Big Data*, vol. 23, no. 2, pp. 53–58, 2020.

[80] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[81] J. Wang, H. Luo, P. Yu, L. Zheng, and F. Hu, "Bird's nest detection in multi-scale of high-voltage tower based on Faster R-CNN," *Journal of Beijing Jiaotong University*, vol. 43, no. 5, pp. 37–43, 2019.

[82] X. Liu, H. Jiang, J. Chen, J. Chen, S. Zhuang, and X. Miao, "Insulator detection in aerial images based on faster regions with convolutional neural network," in *14th International Conference on Control and Automation*, pp. 1082–1086, Anchorage, AK, USA, 2018.

[83] Z. Zhao, Z. Zhen, L. Zhang, Y. Qi, Y. Kong, and K. Zhang, "Insulator detection method in inspection image based on improved Faster R-CNN," *Energies*, vol. 12, no. 7, 2019.

[84] J. Han, Z. Yang, Q. Zhang et al., "A method of insulator faults detection in aerial images for high-voltage transmission lines inspection," *Applied Sciences*, vol. 9, no. 10, 2019.

[85] G. Lin, B. Wang, H. Peng, X. Wang, S. Chen, and L. Zhang, "Multi-target detection and location of transmission line inspection image based on improved faster-RCNN," *Electric Power Automation Equipment*, vol. 39, no. 5, pp. 213–218, 2019.

[86] S. Dong, "Real-time detection of power transmission line key components based on YOLOv3," *Electronic Measurement Technology*, vol. 42, no. 23, pp. 173–178, 2019.

[87] G. Yang, C. Sun, N. Zhang et al., "Detection of key components of transmission lines based on multi-scale feature fusion," *Electric Measurement and Instrumentation*, vol. 57, no. 3, pp. 54–59, 2020.

[88] G. Yang, C. Sun, D. Wang et al., "Comparative study of transmission line component detection models based on UAV front end and SSD algorithm," *Journal of Taiyuan University of Technology*, vol. 51, no. 2, pp. 212–219, 2020.

[89] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," in *14th European Conference on Computer Vision*, pp. 21–37, Cham, 2016.

[90] Y. Qi, A. Jiang, Z. Zhao, J. Lang, and L. Nie, "Fittings detection method in patrol images of transmission line based on improved SSD," *Electric Measurement and Instrumentation*, vol. 56, no. 22, pp. 7–12,43, 2019.

[91] S. Liu, B. Wang, K. Gao, Y. Wang, C. Gao, and J. Chen, "Object detection method for aerial inspection image based on region-based fully convolutional network," *Automation of Electric Power Systems*, vol. 43, no. 13, pp. 162–168, 2019.

[92] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection - SNIP," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3578–3587, Salt Lake City, UT, USA, 2018.

[93] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," in *2019 IEEE/CVF International Conference on Computer Vision*, pp. 6053–6062, Seoul, South Korea, 2019.

[94] A. B. Alhassan, X. Zhang, H. Shen, and H. Xu, "Power transmission line inspection robots: a review, trends and challenges for future research," *International Journal of Electrical Power and Energy Systems*, vol. 118, 2020.

[95] S. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, 1996.

[96] E. Malis, F. Chaumette, and S. Boudet, "2 1/2 D visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 15, no. 2, pp. 238–250, 1999.

[97] F. Chaumette and E. Malis, "2-1/2-D visual servoing: a possible solution to improve image-based and position-based visual servoings," in *IEEE International Conference on Robotics and Automation*, pp. 630–635, San Francisco, CA, USA, 2000.

[98] Q. Zhao, G. Lian, and Z. Sun, "Survey of robot visual servoing," *Control and Decision*, vol. 16, no. 6, pp. 849–853, 2001.

[99] B. Tao, Z. Gong, and H. Ding, "Survey on uncalibrated robot visual servoing control," *Chinese Journal of Theoretical and Applied Mechanics*, vol. 48, no. 4, pp. 767–783, 2016.

[100] L. Wang, H. Wang, L. Fang, and M. Zhao, "Visual-servo-based line-grasping control for power transmission line inspection robot," *Robot*, vol. 29, no. 5, pp. 451–455, 2007.

[101] Y. Zhang, Z. Liang, M. Tan, W. Ye, and B. Lian, "Visual servo control of obstacle negotiation for overhead power line inspection robot," *Robot*, vol. 29, no. 2, pp. 111–116, 2007.

[102] D. Zhao, G. Yang, E. Li, and Z. Liang, "Design and its visual servoing control of an inspection robot for power transmission lines," in *2013 IEEE International Conference on Robotics and Biomimetics*, pp. 546–551, Shenzhen, China, 2013.

[103] W. Guo, H. Wang, Y. Jiang, and P. Sun, "Visual servo control for automatic line-grasping of a power transmission line inspection robot," *Robot*, vol. 34, no. 5, pp. 620–627, 2012.

[104] W. Wang, G. Wu, Y. Bai et al., "Hand-eye-vision based control for an inspection robot's autonomous line grasping," *Journal of Central South University*, vol. 21, no. 6, pp. 2216–2227, 2014.

[105] T. He, H. Wang, W. Chen, and W. Wang, "Visual servoing of a new designed inspection robot for autonomous transmission line grasping," in *International Conference on Wearable Sensors and Robots*, pp. 553–569, Singapore, 2017.

[106] W. Jiang, Z. Zhou, W. Chen, L. Yu, H. Li, and G. Wu, "Manipulator double close loop autonomous localization control of high-voltage cable mobile operation robot," *Transactions of Beijing Institute of Technology*, vol. 39, no. 6, pp. 589–596, 2019.

[107] G. Wu, Z. Yang, W. Wang, L. Guo, J. Hu, and P. Zhou, "On auto-docking charging control method for the inspection robot," *Journal of Harbin Institute of Technology*, vol. 48, no. 7, pp. 123–129, 2016.

[108] O. Araar and N. Aouf, "Visual servoing of a quadrotor UAV for autonomous power lines inspection," in *22nd Mediterranean Conference on Control and Automation*, pp. 1418–1424, Palermo, Italy, 2014.

[109] H. Xie, A. Lynch, and M. Jagersand, "IBVS of a rotary wing UAV using line features," in *IEEE 27th Canadian Conference on Electrical and Computer Engineering*, pp. 1–6, Toronto, Canada, 2014.

[110] S. Mills, N. Aouf, and L. Mejias, "Image based visual servo control for fixed wing UAVs tracking linear infrastructure in wind," in *2013 IEEE International Conference on Robotics and Automation*, pp. 5769–5774, Karlsruhe, Germany, 2013.

[111] M. A. Rafique and A. F. Lynch, "Output-feedback image-based visual servoing for multirotor unmanned aerial vehicle line following," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 4, pp. 3182–3196, 2020.

[112] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2846–2854, Seattle, WA, 2016.

[113] P. Tang, X. Wang, S. Bai et al., "PCL: proposal cluster learning for weakly supervised object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 176–191, 2020.

[114] C. Ge, J. Wang, J. Y. Wang, Q. Qi, H. F. Sun, and J. X. Liao, "Towards automatic visual inspection: a weakly supervised learning method for industrial applicable object detection," *Computers in Industry*, vol. 121, p. 103232, 2020.

[115] S. Montambault and N. Pouliot, "Design and validation of a mobile robot for power line inspection and maintenance," in *6th International Conference on Field and Service Robotics*, Chamonix, France, 2007.

[116] P. L. Richard, N. Pouliot, F. Morin et al., "LineRanger: analysis and field testing of an innovative robot for efficient assessment of bundled high-voltage powerlines," in *2019 International Conference on Robotics and Automation*, pp. 9130–9136, Montreal, QC, Canada, 2019.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5558643, 12 pages
https://doi.org/10.1155/2021/5558643

Research Article

A Chaotic Elite Niche Evolutionary Algorithm for Low-Power Clustering in Environment Monitoring Wireless Sensor Networks

Bao Liu[1], **Rui Yang**[1], **Mengying Xu**[1] and **Jie Zhou**[1,2]

[1]*College of Information Science and Technology, Shihezi University, Shihezi 832000, China*
[2]*Xinjiang Tianfu Information Technology Co., Ltd., Xinjiang, China*

Correspondence should be addressed to Jie Zhou; jiezhou@shzu.edu.cn

Received 26 February 2021; Revised 6 March 2021; Accepted 17 March 2021; Published 30 March 2021

Academic Editor: Bin Gao

In recent years, as people's demand for environmental quality has increased, it has become inevitable to monitor sensitive parameters such as temperature and oxygen content. Environmental monitoring wireless sensor networks (EMWSNs) have become a research hotspot because of their flexibility and high monitoring accuracy. This paper proposes a chaotic elite niche evolutionary algorithm (CENEA) for low-power clustering in EMWSNs. To verify the performance of CENEA, simulation experiments are carried out in this paper. Through simulation experiments, CENEA was compared with shuffled frog leaping algorithm (SFLA), differential evolution algorithm (DE), and genetic algorithm (GA) in the same conditional parameters. The results show that CENEA balances node energy and improved node energy usage efficiency. CENEA's network energy consumption is reduced by 8.3% compared to SFLA, 3.9% lower than DE, and 4.6% lower than GA. Moreover, CENEA improves the precision and minimizes the computation time.

1. Introduction

The development of human society is inseparable from the support of various resources, but in the process of continuous development of human society, due to the continuous deepening of industrialization, the problem of environmental pollution has become increasingly serious. The human ecological environment has been damaged to varying degrees.

Wireless sensor networks (WSNs) have the disadvantages of short network life and high environmental impact. One of the protocols of WSNs is LEACH, which uses a probability model to determine the cluster head, thereby prolonging the life of the network in a complex environment. In the LEACH protocol, the nodes in the WSNs are divided into multiple clusters. Each cluster is composed of a cluster node and many ordinary nodes. The ordinary nodes collect data and then transmit them to the cluster head nodes of the respective clusters. The node fusion compresses the received data and transmits it to the base station [1].

A widely used energy-saving mechanism in WSNs is the duty cycle scheme to reduce energy waste caused by idle monitoring. However, to coordinate the sleep/wake cycle of sensor nodes, the duty cycle scheme requires more control packages to achieve specific application goals. Under different network mechanisms, the duty cycle of sensor nodes needs to be adjusted as network conditions change during operation to achieve the desired delay and energy efficiency. The sender node and the receiver node need to wake up at the same time during the transmission process to complete the data transmission. If a synchronization mechanism is used for data transmission between nodes, to ensure that the clocks between nodes in the network remain constant, more control packets are required in the process of synchronizing the clocks. If the asynchronous mechanism is used for data transmission of the node, the sending node needs to first send a data packet to inform the receiving node of the length of time that it needs to wake up during the data transmission process and then need to retransmit the data packet to complete the data transmission [2].

EMWSNs provide a fast and convenient optional monitoring program for environmental protection due to their advantages of convenient and flexible deployment. However,

due to the limited energy of nodes, the development and application of EMWSNs are hindered. Wireless sensor nodes are generally deployed in unmanned areas with poor conditions. Node access to the network will increase the difficulty of network maintenance. Most of the node batteries will not be replaced, resulting in a limited lifespan of the node in reality. Clustering in EMWSNs divides the nodes in the network into cluster heads and ordinary monitoring nodes. The cluster head contacts other ordinary nodes covered by it to obtain data and then sends the data to the terminal [3–5]. How to use reasonable clustering to conserve the power usage of the EMWSNs while completing the perception task has attracted more and more attention from researchers. Consequently, it is very urgent to create a new EMWSN clustering algorithm to minimize the power usage in the system, increase the efficiency of data transmission, and extend the living period of the system. The CENEA proposed in this paper can greatly improve the performance of EMWSNs.

Within the traditional clustering algorithm, the choice associated with the cluster head will be arbitrary during the establishment phase regarding the cluster [6–8]. The influence factors such as the point transmission distance are not considered, resulting in an improper selection of the cluster head as well as excessive power usage of the common nodes in EMWSNs. In the stage of cluster establishment, CENEA introduces parameters such as node transmission distance to select cluster head nodes to minimize EMWSNs power usage. With the rapid development of EMWSNs, the research of clustering algorithm technology in the network has also achieved outstanding results. The focus of research on clustering algorithms is to optimize the choice of cluster head nodes and the establishment associated with clusters as well as decrease the power usage of nodes in the system [9–11]. Some scholars have proposed new clustering and routing schemes, which provide excellent ideas for improving the life cycle of networks [12–17]. With these years of research and development, the swarm intelligence majorization model has been well utilized in the clustering technology of EMWSNs. Many scholars have optimized the traditional clustering algorithm to increase efficiency, decrease complexity, and improve the algorithm [18]. SFLA, DE, and GA are currently a research focus of clustering algorithms. Using optimization algorithms can quickly discover the optimum way for data transmission as well as extend the life of the network.

For some EMWSNs, researchers are more concerned about the single-round power usage of the EMWSNs after clustering. It is hoped that the energy consumption of sensors per unit time is minimal to reduce the cost. Assuming that EMWSNs are made up of a huge quantity of sensing nodes, a small number of cluster head nodes, and a single gateway node. The gateway node usually has strong signal transceiving capabilities and a fixed location. How to select a small quantity of cluster head nodes from a huge quantity of EMWSN nodes to minimize the power usage of a data collection on the entire network is an important issue in practical applications.

El Alami et al. [19] proposed an improved routing protocol, which can greatly improve the performance of mobile nodes in WSNs, improve network life and energy efficiency, and reduce packet loss to a large extent. Lee et al. [20] proposed an improved clustering protocol, which has an excellent performance in mobile sensor networks. Even during the movement of the node, the packet loss rate can be kept at a low level. El Alami et al. [21] proposed a new clustering hierarchy algorithm to save network energy by changing the sleep and working time of nodes. It has an excellent performance in homogeneous and heterogeneous networks. Lee et al. [22] proposed an improved clustering protocol. This protocol can effectively increase the lifetime of the network and has a good performance in large-scale WSNs.

Liu et al. [23] proposed a model of cluster head selection and path planning based on DE. Improve the performance of each part of the algorithm by reducing the amount of calculation and unifying system energy consumption. In [24], the neural network algorithm is applied to the WSN data fusion process, and the algorithm significantly improves the data processing efficiency. Fattoum et al. [25] use GA to establish a routing mechanism between cluster heads, which reduces energy consumption between clusters.

However, DE, GA, and neural network algorithms cannot dynamically adjust the crossover and mutation probabilities according to the fitness of individuals and populations and tend to fall into premature convergence, resulting in higher energy consumption for the final clustering scheme.

Islam et al. [26] proposed the idea of the main cluster head and a below cluster head. A below cluster head is selected in a cluster to share the energy consumption of the major cluster head nodes and avoid the main cluster head nodes through perishing because of excessive energy consumption. Huamei et al. [27] proposed a cluster head election mechanism based on SFLA, which allows the cluster heads to be reelected after the previous round of cluster heads meet certain conditions, reducing the energy consumption caused by each round of cluster head selection.

However, the operations of these two algorithms are too complicated, and the change of the cluster head does not take into account the complex environment in the actual situation, so the reliability of the algorithm is poor.

Wang et al. [28] proposed a multihop routing protocol, which allows the distant node to choose a node closest to itself for data forwarding, reducing the distance of direct communication with the terminal. In [29], after the nodes are divided into clusters, they have grouped again in each cluster according to the distance and data similarity between the nodes, and a representative node is selected again in each group to deliver the information to the cluster head nodes.

However, once the power of the inner sensor nodes is tired, the outer sensor nodes with a large amount of energy remaining will not be able to work normally because of the too-long transmission distance because it cannot be relayed. It takes a long time to set up a cluster, has a long delay, and consumes a lot of node energy. Also, multiple detection steps before transmitting data packets will increase the delay of data transmission, so it is not appropriate for EMWSNs that require great live performance.

Majeed et al. [30] combine energy and node location to introduce a cost function and uses GA to perform cluster

head election. So that nodes with higher energy and better locations are elected as cluster heads, making the tasks of each node more reasonable. Ghahramani and Laakdashti [31] proposed a routing protocol based on the DE in WSN to minimize power usage as well as extend the system living cycle.

However, the quality of the cost function will determine the efficiency of the way, and the capabilities of the cost function will be poor in a complex environment. Each time the algorithm needs to send a data packet, the source node will discover and establish a cluster. Therefore, it takes a certain time to establish the corresponding cluster, and the cluster will be removed after a certain time. A lot of node energy will be spent in the process of cluster establishment.

The computational time of the cluster head selection problem increases exponentially with the increment of cluster head nodes. To reduce the network power usage rate, we present a CENEA. An objective function is formulated to maximize the reduced network energy consumption rate under multiple constraints. This paper also gives advanced operators by employing elite operators and chaotic map operators in each iteration of the evolutionary procedure. The CENEA combines the merits of the elite evolution and chaotic map. CENEA is a kind of swarm algorithm, which has a strong global search capacity. To denote the advantages of CENEA, experiments are conducted for the cluster head selection problem and performance comparisons are made with SFLA, DE, and GA. Simulation simulations reveal the superior performance of the presented CENEA in both the reduced network energy consumption rate and fast convergence.

The main contributions of this paper are as follows.

(1) This paper proposes a new clustering algorithm. CENEA can reduce the energy consumption of EMWSNs and improve network performance. CENEA has less time complexity and can complete the selection of EMWSN cluster heads and cluster coverage in a short time

(2) This paper designs a new clustering model of EMWSNs. The model coding adopts the real number coding scheme of sensor position information. In EMWSNs, sensors are randomly placed in the environmental monitoring area to simulate complex situations in reality, so it is closer to the situation of environmental monitoring in reality

(3) This paper verifies the excellent performance of CENEA in EMWSNs through simulation experiments. CENEA can improve the viability of EMWSNs and propose a new clustering scheme for the development of EMWSNs

The structure of this paper is as follows. Section 2 introduces the clustering model of EMWSNs and the selection mechanism of cluster heads. Section 3 uses CENEA to optimize the clustering algorithm. Section 4 gives the results of simulation experiments and discusses the performance of CENEA. Then, Section 5 concludes.

2. EMWSN Clustering Model

The distribution of cluster head nodes in EMWSNs determines the energy consumption of network communication. This section will introduce the network clustering model. The network distribution area studied in this paper is a square, and a certain number of sensor nodes are aimlessly dispersed in the square region. The typical network structure is as follows.

As shown in Figure 1, EMWSNs usually adopt a clustered structure. The sensing nodes are divided into multiple clusters within the monitoring range, and each cluster has a cluster head [32]. In the uplink transmission phase, the sensing nodes randomly distributed within the monitoring range complete the sensing of the monitoring target and gather the sensing outcomes to the head of the cluster. The cluster head node collects relevant information from the sensing nodes in the cluster and uploads the information to the gateway node within a direct or multihop manner [33]. The gateway node summarizes the information through every cluster head and transmits it to the user for further analysis and processing. In the downlink transmission stage, the user releases monitoring tasks in the downlink through the gateway node and uniformly allocates monitoring targets and various resources in the network. The gateway node distributes the monitoring task to the sensing nodes in the cluster through the cluster head node and completes the downlink distribution process of the monitoring task.

The power usage of EMWSNs primarily arrives through delivering data, getting data, and transmission paths. The node has receiving consumption and sending consumption and part of the energy consumption from the amplifier. This part of the energy consumption is closely related to the transmitting range. If a is under a offered parameter a_{set}, the sending amplifier uses the energy of free space. When a is greater than a_{set}, the transmitting amplifier adopts the energy consumption model of multipath attenuation. When the data of n bit needs to be sent and the distance from the sender to the destination is a, the energy consumption of the sender is

$$\begin{aligned} L_{Tx}(n,a) &= L_{(Tx-\text{elec})}(n) + L_{(Tx-\text{amp})}(n,a) \\ &= \begin{cases} n \times L_{\text{elec}} + n \times \beta_{\text{fs}} \times a^2, & a < a_{\text{set}}, \\ n \times L_{\text{elec}} + n \times \beta_{\text{amp}} \times a^4, & a \geq a_{\text{set}}. \end{cases} \end{aligned} \tag{1}$$

In equations (1), a is the data transmission distance on the path, L_{elec} is the power consumed through the transmitting signal, β_{fs} and β_{amp} correspond to different energy transmission models, β_{fs} corresponds to the energy transmission model of free space, and β_{amp} corresponds to the energy transmission model of multipath attenuation, where a_{set} determines which energy transmission model the sender adopts, as shown in

$$a_{\text{set}} = \sqrt{\frac{\beta_{\text{fs}}}{\beta_{\text{amp}}}}. \tag{2}$$

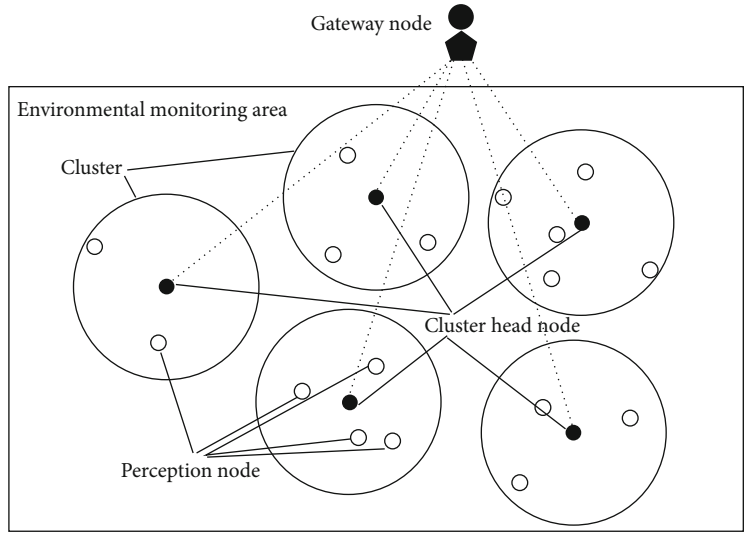


Figure 1: EMWSN cluster structure.

The amount of energy consumed by data transmission on the path is represented by

$$L(n, a) = L_{Tx}(n, a) + L_{Rx}(n). \quad (3)$$

In equations (3), $L_{Tx}(n, d)$ represents the power ingested through the node to deliver information. $L_{Rx}(n)$ represents the power taken through the node to obtain information. Power usage is mainly composed of two parts, the signal energy consumption $L_{T-\text{elec}}(n)$ and the transmission amplifier circuit energy consumption $L_{T-\text{amp}}(n, a)$. The size of $L_{Tx}(n, d)$ is shown in

$$L_{Tx}(n, d) = L_{T-\text{elec}}(n) + L_{T-\text{amp}}(n, a). \quad (4)$$

The receiver is different from the sender, and the energy consumption of the receiver has nothing to do with the distance. Therefore, the energy required to receive the data of *n bit* can be obtained as

$$L_{Rx}(n) = n \times L_{\text{elec}}. \quad (5)$$

3. CENEA for Reducing Network Energy Consumption in EMWSNs

In this work, our CENEA follows the framework of the conventional stochastic methodology. Two novel programs were produced, namely, chaos mapping and elite programs. In the cluster head selection problem of EMWSNs, elite technology can be used to find feasible solutions close to the ideal solution. These procedures are effective strategies, easy to understand and to implement, extensively utilized in the optimizing of cluster head node distributions. The novel procedures of the CENEA are qualified to create a feasible solution for EMWSNs in a computationally acceptable time.

Compared to conventional optimization algorithms like calculus-based techniques as well as exhaustive techniques, CENEA can effectively deal with some complex problems that are not able to be resolved through conventional methods. The standard technique to resolve the optimization issue is to design an objective function so that the objective function can be modeled reasonably while combining various constraints and then transform the optimization problem into finding the maximum value. CENEA simulates the natural biological evolution model, using real number coding to obtain the initial population, cross mutation operation, and group iteration based on the greed criterion to realize the function of search and optimization. Real number coding is more universal in the type of problem solving than binary coding. The greedy selection criterion will retain the best solution individual in the current search space and will not stop until the preset number of iterations is reached. For the problem to be solved, a fitness function is designed, and then the fitness function was taken as the evaluation objective. Retain elite individuals with better fitness values in the iterative process to induce the final solution to approach a better direction. This process makes CENEA have better dynamic tracking. The many advantages of CENEA make it unnecessary to make use of the feature info of the issue to a certain extent, effectively solving the optimization problem in the complex environment.

CENEA is mainly used to solve global optimization problems of continuous variables. It is an intelligent evolutionary algorithm with dynamic tracking and random search. CENEA is aimed at denoting good chromosomes through this evolutionary process. The CENEA utilizes various simple procedures to simulate evolution. The main steps are as follows:

(Step 1) In the initialization stage, the initial population is generated by the chaotic map operator and generally satisfies the condition that it can cover the whole search space

(Step 2) In the mutation stage, a certain number of individuals are selected from the population to produce mutant individuals

(Step 3) In the crossover stage, the target individual and the variant individual are mixed concerning the crossover probability criterion to obtain the test individually

(Step 4) In the individual evaluation stage, the result function value of this offspring is evaluated

(Step 5) In the selection stage, determine the superior new iteration using the greedy method and save elite individuals

(Step 6) In the final stage of the iteration, judge the stop criteria; if termination criterion is attained, then stop the procedure

The flow chart of CENEA is shown in Figure 2.

3.1. Population Initialization Operation of CENEA. The first step of the CENEA is to establish a proper chromosome representation. To utilize a CENEA to match the cluster head node distribution to the EMWSNS, it is necessary to produce an effective encoding scheme and a result function, which will allow the algorithm to choose the fittest solutions. The individual code of CENEA is a string of limited length and limited precision that is usually indicated like $P_k = [p_1, p_2, \cdots, p_M]$ mathematically. K is the quantity of populations, P_k is called a chromosome, M is called the number of genes on the chromosome, and p_M is called a gene on the chromosome. y_m signifies the cluster head node of the m_{th} cluster in the chromosome. In CENEA, each solution is referred to chromosome. Every chromosome represents a feasible solution of cluster head node distribution for the cluster head selection problem. In the d generation, the entire population can be expressed by

$$P(d) = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,M-1} & y_{1,M} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,M-1} & y_{2,M} \\ \vdots & \vdots & y_{z,j} & \vdots & \vdots \\ y_{K-1,1} & y_{K-1,2} & \cdots & y_{K-1,M-1} & y_{K-1,M} \\ y_{K,1} & y_{K,2} & \cdots & y_{K,M-1} & y_{K,M} \end{bmatrix} = \begin{bmatrix} P_1(d) \\ P_2(d) \\ \vdots \\ P_{K-1}(d) \\ P_K(d) \end{bmatrix}. \tag{6}$$

3.2. Chaotic Map Operation of CENEA. The chaotic map operator is used to generate K initial population combinations of length M. The length of the chromosome is the same as the quantity of cluster head nodes within EMWSMs, which means that each value of the final optimized individual maps the selection of each cluster head node.

Chaotic motion is a common nonlinear random phenomenon. It looks complicated and chaotic on the outside. But in reality, it contains exquisite laws, which have good randomness, ergodicity, and regularity. The randomness and ergodicity of chaotic motion can traverse all states within a certain range without repetition. It is these characteristics of chaotic that provide a broad new idea for optimization calculations and various optimization algorithms. Highly sensitive to initial conditions, boundedness, randomness, and ergodicity are the four most distinctive characteristics of chaotic. The chaotic map used in this paper is the sinusoidal map. The chaotic map iteration is

$$s_{x+1} = \theta s_x{}^2 \sin(\pi s_x). \tag{7}$$

In equation (7), x is the number of iterations. When the θ value is 2.3 and the s_1 value is 0.7, it can be expressed as

$$s_{x+1} = \sin(\pi s_x). \tag{8}$$

The value range of the chaotic value jumps out of the special limitation of the sine function range of [-1,1], and its value range becomes (0,1).

3.3. Fitness Calculation of CENEA. In the EMWSM clustering scheme for optimizing the power usage of a solitary circular transmission, the major consideration is how to decrease the power consumption of a single round of communication through clustering. At the same period, the Euclidean length via the common sensor node to the corresponding cluster head node is considered. Assigning a common node to the nearest cluster head node helps sensor nodes consume less energy when communicating with cluster head nodes. Therefore, the common nodes are assigned to their corresponding cluster head nodes; the average length T from the sensor node to the cluster head node is

$$T = \frac{1}{G} \sum_{g=1}^{G} F_g. \tag{9}$$

In equation (9), G represents the overall number of ordinary common nodes distributed in EMWSNs, and F_g represents the length through the sensor node to its corresponding cluster head node.

The smaller the average distance, the better the clustering scheme. The fitness function of CENEA can be expressed as

$$H(P) = \delta \times T. \tag{10}$$

In equation (10), δ is the proportional coefficient. Without loss of generality, set the value of δ to 1.

3.4. Mutation Operation of CENEA. To keep the number of cluster heads constant and introduce randomness, the mutation operation uses a certain mutation probability Q to replace a random bit of the individual with other nonduplicate nodes.

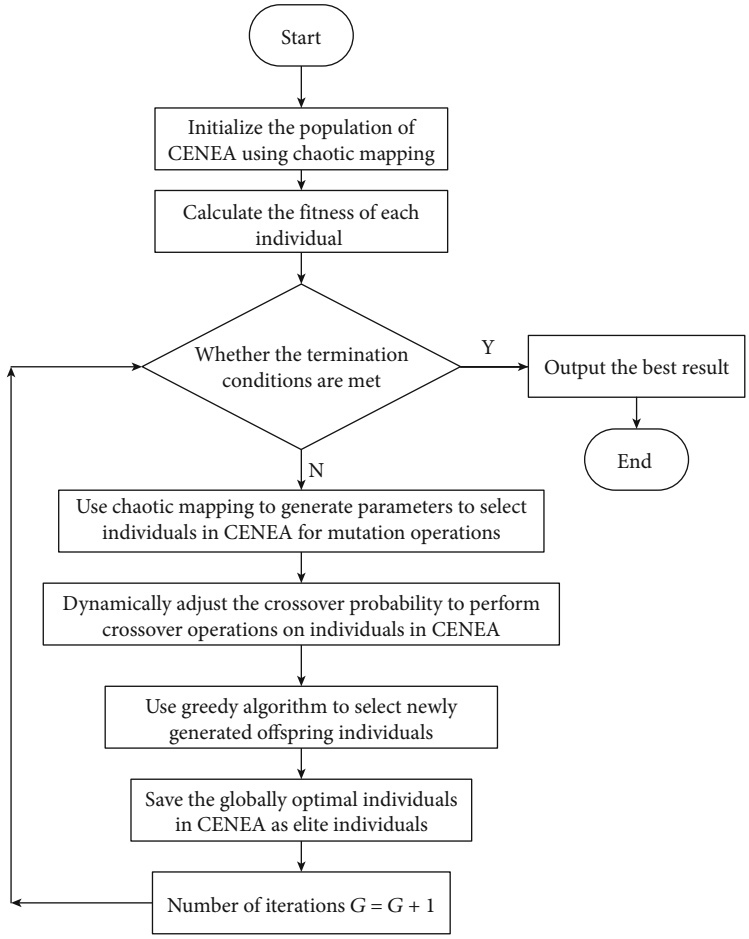


Figure 2: Process of the CENEA.

The newly selected sensor node is regarded as the cluster head node. For a single individual, a random number Q_r is generated according to the chaotic mapping. Then, the operation is performed in

$$y_{z,j} = \begin{cases} y_{z,j}, & Q_r > Q, \\ y_{\text{new}}, & Q_r < Q. \end{cases} \tag{11}$$

In equation (11), y_{new} represents the newly generated cluster head node, and y_{new} does not overlap with the existing cluster head node in the individual.

3.5. Crossover Operation of CENEA. First, perform a logical AND operation on two individuals to get an intermediate individual P'. Secondly, perform the logical XOR operation on the two individuals to obtain another intermediate individual P''. Finally, the cluster head node position in the intermediate individual P'' obtained by the exclusive OR operation is allocated to the new cross individual P''' by the crossover probability C. The fresh individual P_{new} obtained after the crossover could be indicated as

$$P_{\text{new}} = P' + P'''. \tag{12}$$

The crossover probability has an excellent impact on the searchability and convergence efficiency of the formula. In this paper, the relative value depending on the individual fitness benefit of the parent and the average fitness benefit of the group is used for adjustment. The next-generation crossover probability set update strategy is shown in

$$C_k^{d+1} = \begin{cases} {C_k}^d, & H(P_k)^d < H_m^d, \\ {C_k}^d(\mu_2 + (\mu_2 - \mu_1)), & H(P_k)^d > H_m^d. \end{cases} \tag{13}$$

In equation (13), C_k^{d+1} is the crossover probability of the k_{th} individual in the next generation. $H(P_k)^d$ symbolize the fitness value of the present k_{th} individual and H_m^d represent the average fitness value. When the fitness benefit of the current individual is better than the average fitness benefit, the crossover probability of the next-generation individual remains unchanged. Otherwise, the crossover probability will change with a certain coefficient until the fitness benefit of the next generation chromosome is better. μ_1 and μ_2 are the higher and lesser limitations of the crossover probability. When the fitness value of the current individual is higher than the average fitness value, their constituent coefficients affect the

change of the crossover probability of the next generation of individuals.

3.6. Selection Operation of CENEA. The selection operation determines which of the target individual and the newly generated individual will survive to the next generation. Use the greed principle to determine whether the newly generated individual replaces the old individual. The operation rule is shown in

$$P_k = \begin{cases} P_{k,} & H(P_k)^d < H(P_k)^{d-1}, \\ P_{\text{new}}, & H(P_k)^d > H(P_k)^{d-1}. \end{cases} \quad (14)$$

In equation (14), $H(P_k)^d$ represents the fitness of the k_{th} individual in the d generation. $H(P_k)^{d-1}$ represents the fitness of the k_{th} individual in the $d-1$ generation.

3.7. Elite Operation of CENEA. A global variable is set in CENEA to store the optimal individual in the iterative process. The global variable is the elite individual. After the selection operation is performed, CENEA judges whether the fitness of the optimal individual in the population is lower than the fitness of the elite individual. Corresponding operations were performed according to different results, and the operation rules are as in

$$P_{\text{best}} = \begin{cases} P_{\text{best},} & H\left(P_t^d\right) > H(P_{\text{best}}), \\ P_t^d, & H\left(P_t^d\right) < H(P_{\text{best}}). \end{cases} \quad (15)$$

In equation (15), P_{best} represents an elite individual, and P_t^d represents the best individual in the d generation. $H(P_{\text{best}})$ represents the fitness of elite individuals, and $H(P_t^d)$ represents the fitness of the best individuals in the population. If the fitness of the elite individual is smaller than the fitness of the best individual in the population, the elite individual is retained; otherwise, the best individual replaces the elite individual. To make the CENEA iterative process always update the population positively, it is necessary to save the elite individuals in the population. The operation of saving elite individuals is as in

$$P_i^{\text{d}} = P_{\text{best}}. \quad (16)$$

In equation (16), P_i^d represents the i_{th} individual in the d generation population. The value of i is generated by the chaotic map.

3.8. Niche Algorithm of CENEA. The niche algorithm means that in CENEA, the population is divided into several subpopulations, and each subpopulation completes the evolution independently. In every few generations, the best individuals will be exchanged among subpopulations. The niche algorithm can effectively enhance the performance of the formula and prevent the formula from falling into the local optimum.

Table 1: Simulation experiment parameter settings.

Parameter	Value
L_{elec}	50 nJ/bit
β_{fs}	$10\,\text{pJ}/\left(\text{bit}\times\text{m}^2\right)$
β_{amp}	$0.0013\,\text{pJ}/\left(\text{bit}\times\text{m}^4\right)$
a_{set}	87 m
n	3072 bits

4. Results and Discussion

To verify whether the CENEA optimized clustering formula could efficiently save the power usage of EMWSNs in the cluster head selection problem. The simulation software is implemented in this section, and the CENEA optimized routing clustering algorithm is compared with SFLA, GA, and DE, optimization algorithms for comparison and analysis. In the simulation experiment, this paper compared CENEA, SFLA, GA, and DE. In this paper, a simulation experiment is carried out on the MATLAB R2018b software and Intel Core i7 processor platform. After several simulation experiments, the optimal values of the parameters were taken. The environmental parameters of the EMWSNs experiment are shown in Table 1.

The simulation adopted the clustering method based on CENEA, SFLA, GA, and DE. The setting area is 400 m × 400 m. In CENEA, the number of algorithm iterations is 100, and the number of individuals in the population is 100. In SFLA for comparison, the total number of frogs is 100, the population is 20, and the number of frogs in each population is 5. The maximum step size is 40 m. In GA, the population numbers are 100. The roulette method is used for selection, the crossover operation is a single point crossover, and the mutation probability is 0.1. In DE, the population numbers are 100. The crossover factor is 0.3, and the scaling factor is 1.

The cluster head ratio is set to 0.05. The simulation result is shown in Figures 3(a)–3(d), respectively, representing the cases where the number of sensor nodes is 100, 200, 300, and 400.

Figure 3 shows the complete network communication power ingested through all sensor nodes in CENEA, SFLA, GA, and DE when the cluster head ratio is 5% and the quantity of sensor nodes varies with the number of algorithm iterations. From the simulation outcomes, it could be observed which SFLA has a certain decline in the initial stage of algorithm iteration, but after the number of iterations reaches a certain number, it repeatedly falls into evolutionary stagnation, and the finally obtained clustering scheme has a large network communication energy consumption. GA and DE have a relatively stable performance during operation, but fail to dynamically adjust the algorithm parameters during the evolution process, resulting in slower algorithm evolution. The final clustering scheme has a higher network communication energy consumption than CENEA's clustering scheme. CENEA's clustering scheme uses the calculation of the fitness of the individual to be crossed, the fitness of the

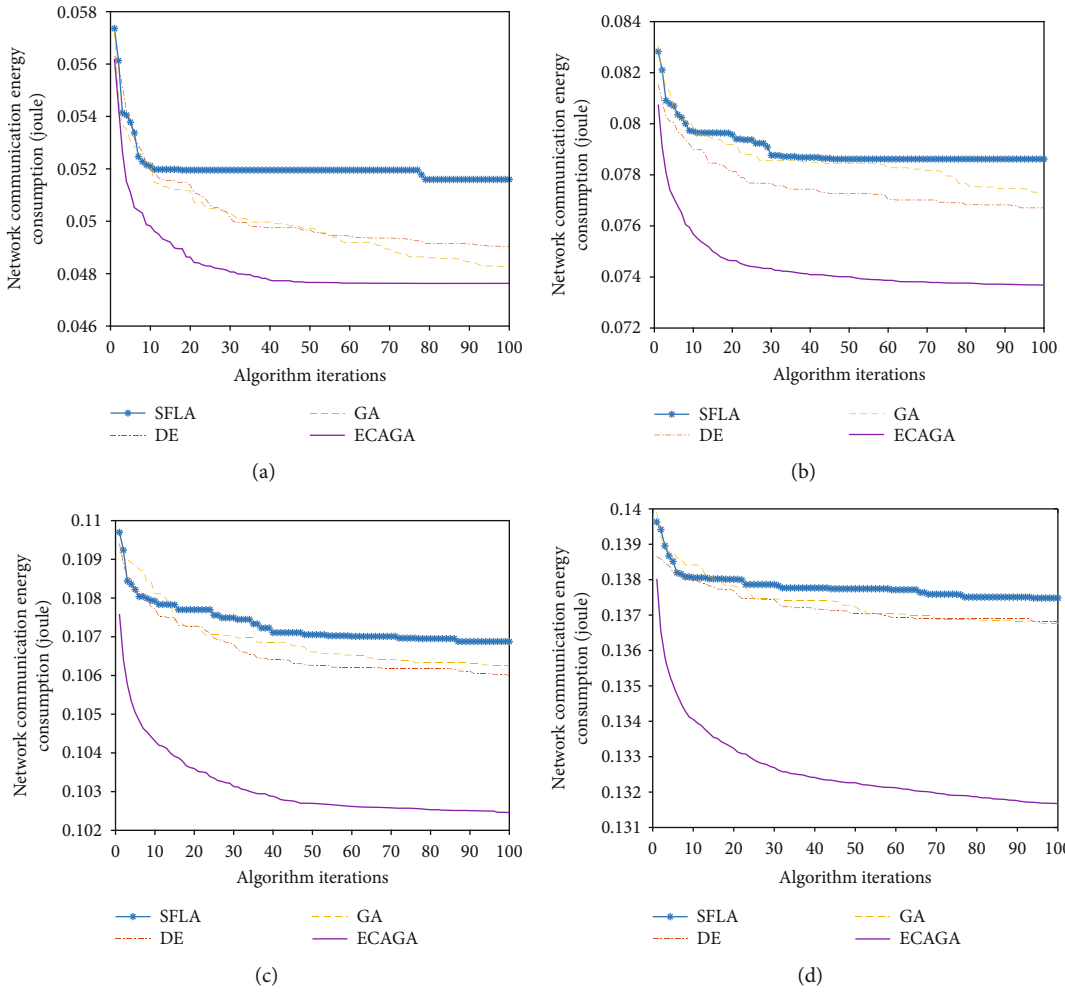


Figure 3: Changes in network communication energy consumption with the number of sensor nodes: (a) 100 sensors; (b) 200 sensors; (c) 300 sensors; (d) 400 sensors.

individual to be mutated, and the average fitness of the population as inputs to avoid the algorithm from falling into the local optimum. The chaotic map is used to obtain random numbers during initialization and selection operations, which ensures the global efficiency of the CENEA. Simultaneously, due to the adoption of the niche method, the crossover and mutation probabilities are dynamically adjusted during operation. CENEA's clustering scheme avoids evolutionary stagnation and premature convergence caused by fixed algorithm parameters. It can be seen from Figure 3 that under the condition of different sensor nodes, the total power usage of EMWSN communication required by CENEA is reduced by 4.1% to 8.3% compared to SFLA and 1.3% to 4.6% compared to GA. DE has dropped by 2.9% to 3.9%, which means that the energy efficiency is higher when the amount of data transmitted is the same.

Figure 4 shows the changes in network communication energy consumption when 400 sensor nodes are within the monitoring area of different area sizes. The cluster head ratio is 5%. Figures 4(a)–4(d), respectively, represent the area sizes which are 100 m × 100 m, 200 m × 200 m, 300 m × 300 m, and 400 m × 400 m.

Figure 4 shows that CENEA has greater performance advantages than SFLA, GA, and DE in both small and large areas. In EMWSNs, the area of environmental monitoring is uncertain. Compared with SFLA, GA, and DE, CENEA has stronger adaptability when EMWSNs change the size of the monitoring area. This is the ability to quickly configure EMWSN cluster heads under different environmental conditions. CENEA optimization is faster. It can be seen that when the number of iterations is 20, the network energy consumption is close to the optimal result. CENEA can use the least time to get the best results. The running time of the CENEA is shown in Table 2.

Table 2 shows that the running time of CENEA's algorithm is much shorter than SFLA and smaller than DE and GA. The final result of algorithm optimization is shown in Figure 5.

Figure 5 shows that the last optimization outcomes of CENEA are optimal compared to SFLA, GA, and DE under different area size conditions. Simultaneously, the CENEA has the shortest running time. It means that CENEA can get the best EMWSN cluster head solution in the shortest time in a complex environment.

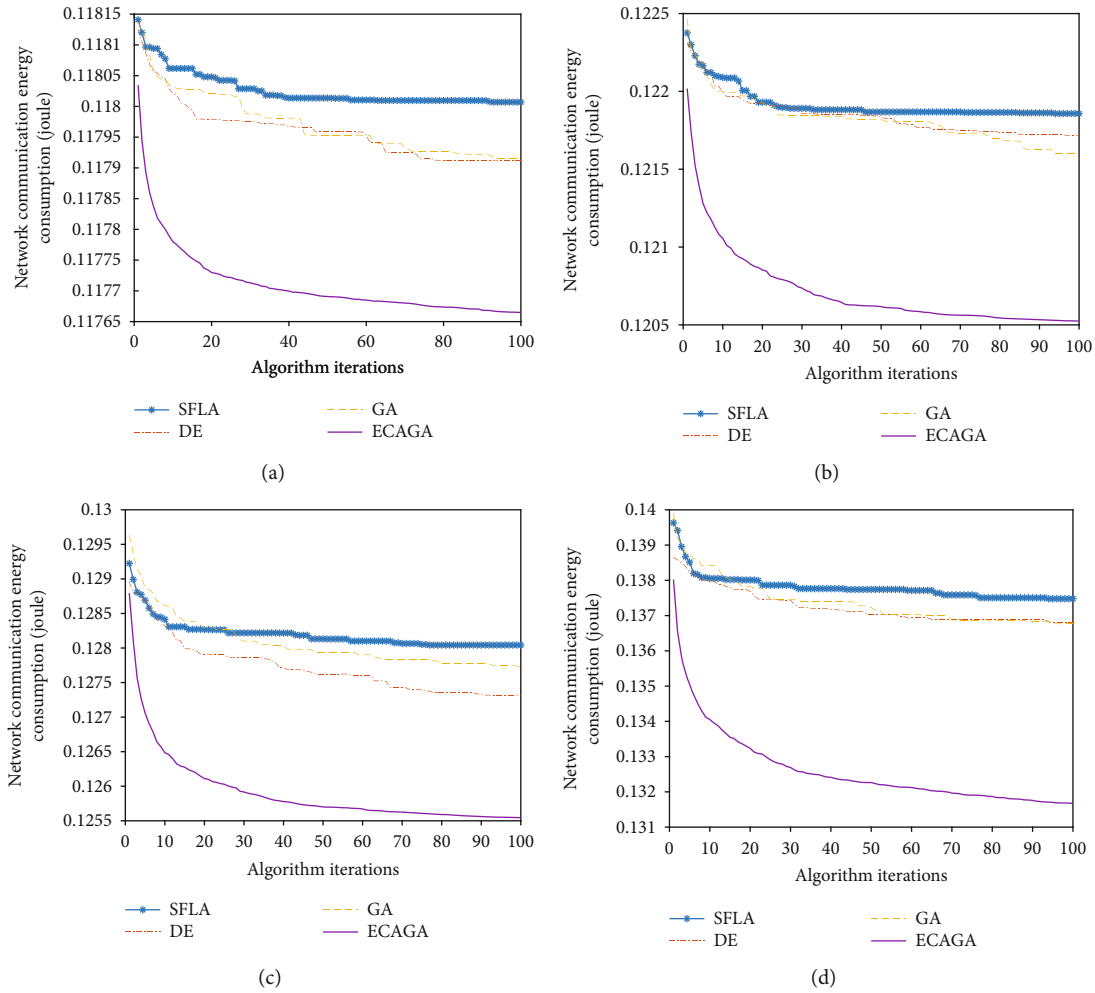


Figure 4: Changes in network communication energy consumption with the following area sizes: (a) 100 m × 100 m; (b) 200 m × 200 m; (c) 300 m × 300 m; (d) 400 m × 400 m.

Table 2: Algorithm running time.

	100 m × 100 m	200 m × 200 m	300 m × 300 m	400 m × 400 m
CENEA	17.3 s	23.2 s	17.8 s	18.8 s
SFLA	207.0 s	275.2 s	205.8 s	214.0 s
GA	17.7 s	24.6 s	19.2 s	19.3 s
DE	30.6 s	46.0 s	33.1 s	34.1 s

Figure 6 shows that when the cluster head ratios in EMWSN are different and the clustering methods of CENEA, SFLA, GA, and DE are used, respectively, the values of SFLA, GA, and DE are more than the overall power usage of the EMWSN communication. The number of sensor nodes is 400, and the area size is 400 m × 400 m. Figures 6(a)–6(d), respectively, represent the proportion of cluster heads is 5%, 10%, 15%, and 20%.

It could be observed from Figure 6 that the CENEA-based clustering technique effectively reduces network communication energy consumption compared to SFLA, GA, and DE. When the proportion of cluster heads is 5%, the advantage of CENEA communication energy reduction is very large compared with other algorithms. When the proportion of cluster heads is 10%, 15%, and 20%, the greater the number of sensor nodes is. Compared with the other three algorithms, the reduction in network communication energy consumption by CENEA is also obvious. The main reason is that SFLA, GA, and DE did not dynamically adjust their parameters in the iterative process, resulting in slow evolution. In the process of evolution, CENEA can dynamically adjust its mutation probability through the fuzzy controller. Use elite operators to ensure the optimization trend, and improve the evolution speed of the algorithm. Like the

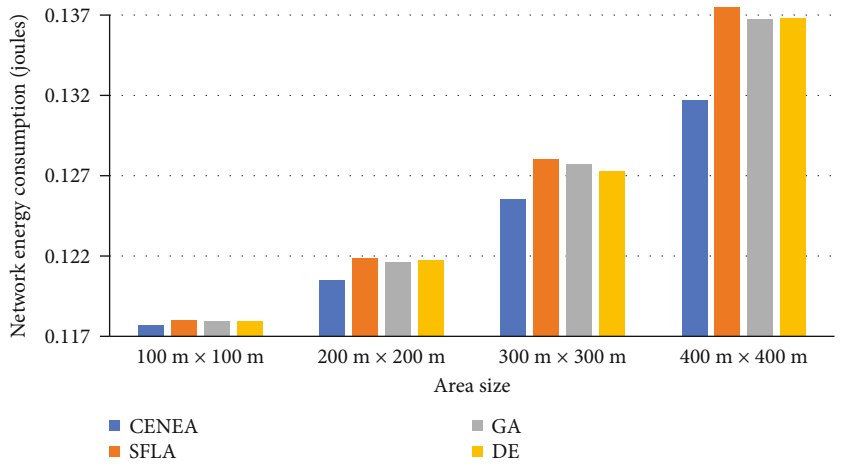


Figure 5: The final optimization results of the algorithm in different area sizes.

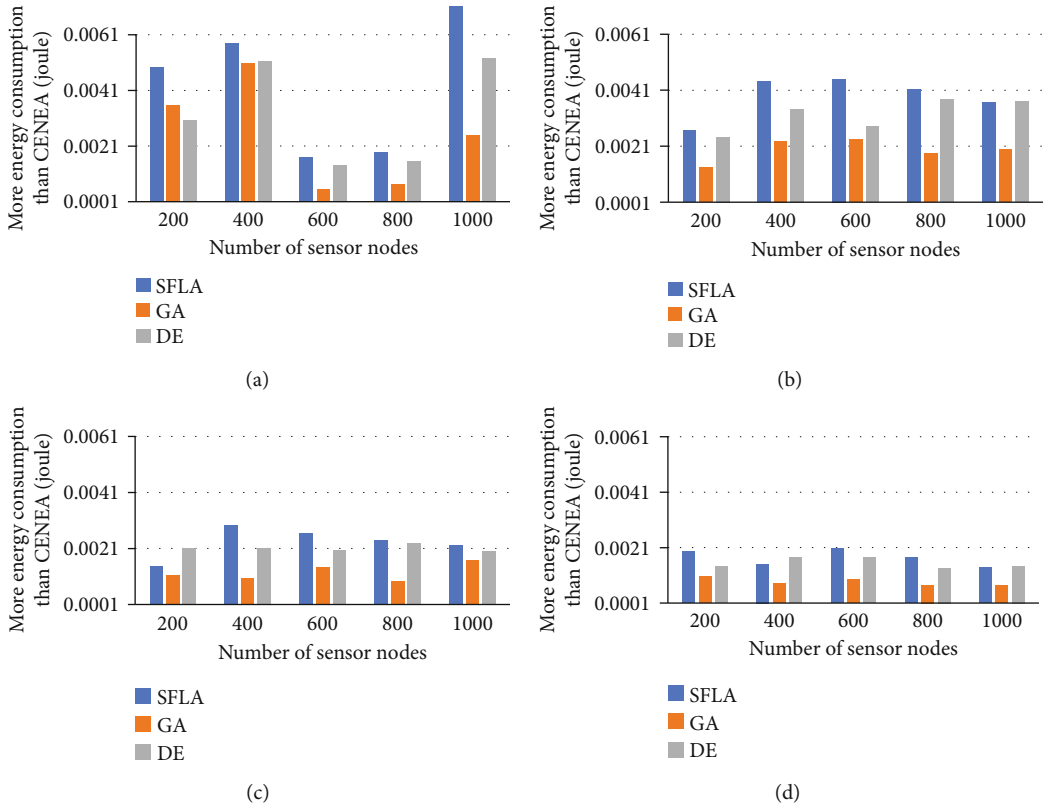


Figure 6: The values of SFLA, GA, and DE are more than the total energy consumption of CENEA network communication: (a) the proportion of cluster heads is 5%; (b) the proportion of cluster heads is 10%; (c) the proportion of cluster heads is 15%; (d) the proportion of cluster heads is 20%.

quantity of sensor nodes raising, the complexity of the problem increases exponentially, and CENEA still has a great performance advantage over the other three algorithms.

5. Conclusion

This paper proposes a CENEA clustering scheme for EMWSNs, which uses a heuristic algorithm to dynamically select the location of the cluster head to decrease the power usage of EMWSNs. The CENEA avoids premature convergence by dynamically changing the algorithm parameters in the iterative process and, at the same time, has a faster convergence speed. The simulation outcomes display that, compared with the other three schemes, the proposed CENEA clustering plan in EMWSNs could efficiently decrease the power usage of a single round of network communication.

Comparing CENEA with SFLA, GA, and DE through simulation, it is verified that CENEA effectively reduces network energy consumption in a small-scale environment (100 m × 100 m) and a large-scale environment (400 m × 400 m). The performance advantage of CENEA means that it can propose more excellent cluster head selection schemes in environmental monitoring. CENEA can meet the actual needs of EMWSNs.

Although the energy-efficient clustering algorithm proposed in this paper has proved its superior performance through simulation, it still has some shortcomings due to the limitation of research ability and environmental conditions. Although the computation overhead of CENEA is greatly reduced compared to SFLA, DE, and GA. However, CENEA still cannot avoid a certain computational overhead, especially when facing large-scale wireless sensor networks or running on low-performance hardware. In this paper, the sensor nodes in EMWSNs are statically and randomly distributed in the monitoring area, but some application scenarios require the sensor nodes to be distributed as mobile monitoring data. Although it was verified and implemented under computer simulation conditions, it did not consider the impact of environmental factors on the communication between EMWSN nodes. The application scenario of CENEA is EMWSNs with homogeneous nodes, and the nodes are randomly distributed in a two-dimensional area, without considering the actual three-dimensional environment. In the future, it will be studied to distribute sensor nodes as mobile monitoring data in certain application scenarios. To improve the practicability and adaptability of the algorithm, the next step of research can place the network in a three-dimensional scene and be heterogeneous. In the future, the impact of noise, temperature, obstacles, and other environmental factors on data transmission between EMWSN nodes will be considered, and it will be close to the actual monitoring site.

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 that they have no conflicts of interest.

Acknowledgments

This paper was funded by the Corps Innovative Talents Plan (grant number 2020CB001), the project of Youth and Middle-Aged Scientific and Technological Innovation Leading Talents Program of the Corps (grant number 2018CB006), the China Postdoctoral Science Foundation (grant number 220531), the Funding Project for High Level Talents Research in Shihezi University (grant number RCZK2018C38) and the Project of Shihezi University (grant number ZZZC201915B), and the Postgraduate Education Innovation Program of the Autonomous Region.

References

[1] T. Nguyen, T. Hoang, V. Pham, T. Nguyen, and N. Nguyen, "Enhancing energy efficiency of WSNs through a novel fuzzy logic based on LEACH protocol," in *2019 19th International Symposium on Communications and Information Technologies (ISCIT)*pp. 108–112, Ho Chi Minh City, Vietnam, 2019.

[2] B. A. Muzakkari, M. A. Mohamed, M. F. A. Kadir, and M. Mamat, "Queue and priority-aware adaptive duty cycle scheme for energy efficient wireless sensor networks," *IEEE Access*, vol. 8, pp. 17231–17242, 2020.

[3] E. Pei, J. Pei, S. Liu, W. Cheng, Y. Li, and Z. Zhang, "A heterogeneous nodes-based low energy adaptive clustering hierarchy in cognitive radio sensor network," *IEEE Access*, vol. 7, pp. 132010–132026, 2019.

[4] T. Zhang, G. Chen, Q. Zeng, G. Song, C. Li, and H. Duan, "Routing clustering protocol for 3D wireless sensor networks based on fragile collection ant colony algorithm," *IEEE Access*, vol. 8, pp. 58874–58888, 2020.

[5] W. Dargie and J. Wen, "A simple clustering strategy for wireless sensor networks," *IEEE Sensors Letters*, vol. 4, no. 6, pp. 1–4, 2020.

[6] F. Lin, W. Dai, W. Li, Z. Xu, and L. Yuan, "A framework of priority-aware packet transmission scheduling in cluster-based industrial wireless sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5596–5606, 2020.

[7] Y. Cao and Z. Wang, "Combinatorial optimization-based clustering algorithm for wireless sensor networks," *Mathematical Problems in Engineering*, vol. 2020, Article ID 6139704, 13 pages, 2020.

[8] A. Pathak, "A proficient bee colony-clustering protocol to prolong lifetime of wireless sensor networks," *Journal of Computer Networks and Communications*, vol. 2020, Article ID 1236187, 9 pages, 2020.

[9] D. Huang, C. -D. Wang, H. Peng, J. Lai, and C. -K. Kwoh, "Enhanced Ensemble Clustering via Fast Propagation of Cluster-Wise Similarities," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 508–520, 2021.

[10] M. Ahmad, B. Shah, A. Ullah et al., "Optimal clustering in wireless sensor networks for the Internet of things based on memetic algorithm: memeWSN," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 8875950, 14 pages, 2021.

[11] K. N. Qureshi, M. U. Bashir, J. Lloret, and A. Leon, "Optimized cluster-based dynamic energy-aware routing protocol for wireless sensor networks in agriculture precision," *Journal of Sensors*, vol. 2020, Article ID 9040395, 19 pages, 2020.

[12] J. Wang, C. Ju, Y. Gao, A. K. Sangaiah, and G. Kim, "A pso based energy efficient coverage control algorithm for wireless sensor networks," *Computers, Materials & Continua*, vol. 56, no. 3, pp. 433–446, 2018.

[13] D. Gao, S. Zhang, F. Zhang, X. Fan, and J. Zhang, "Maximum data generation rate routing protocol based on data flow controlling technology for rechargeable wireless sensor networks," *Computers, Materials & Continua*, vol. 59, no. 2, pp. 649–667, 2019.

[14] A. Janarthanan and D. Kumar, "Localization based evolutionary routing (lober) for efficient aggregation in wireless multimedia sensor networks," *Computers, Materials & Continua*, vol. 60, no. 3, pp. 895–912, 2019.

[15] J. Wang, G. Yu, X. Yin, F. Li, and H.-J. Kim, "An enhanced PEGASIS algorithm with mobile sink support for wireless sensor networks," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 9472075, 9 pages, 2018.

[16] J. Wang, X. Gu, W. Liu, A. K. Sangaiah, and H. J. Kim, "An empower hamilton loop based data collection algorithm with mobile agent for WSNs," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, 2019.

[17] J. Wang, Y. Gao, C. Zhou, R. S. Sherratt, and L. Wang, "Optimal coverage multi-path scheduling scheme with multiple mobile sinks for wsns," *Computers, Materials & Continua*, vol. 62, no. 2, pp. 695–711, 2020.

[18] O. J. Aroba, N. Naicker, and T. Adeliyi, "An innovative hyper-heuristic, Gaussian clustering scheme for energy-efficient optimization in wireless sensor networks," *Journal of Sensors*, vol. 2021, Article ID 6666742, 12 pages, 2021.

[19] H. El Alami and A. Najid, "MS-routing-Gi: routing technique to minimise energy consumption and packet loss in WSNs with mobile sink," *IET Networks*, vol. 7, no. 6, pp. 422–428, 2018.

[20] J.-S. Lee and C.-L. Teng, "An enhanced hierarchical clustering approach for mobile sensor networks using fuzzy inference systems," *IEEE Internet of Things Journal*, vol. 4, no. 4, pp. 1095–1103, 2017.

[21] H. El Alami and A. Najid, "ECH: an enhanced clustering hierarchy approach to maximize lifetime of wireless sensor networks," *IEEE Access*, vol. 7, pp. 107142–107153, 2019.

[22] J.-S. Lee and W.-L. Cheng, "Fuzzy-logic-based clustering approach for wireless sensor networks using energy predication," *IEEE Sensors Journal*, vol. 12, no. 9, pp. 2891–2897, 2012.

[23] X. Liu, K. Mei, and S. Yu, "Clustering algorithm in wireless sensor networks based on differential evolution algorithm," *Chongqing, China, IEEE 4th information technology, networking, electronic and automation control conference (ITNEC)*, vol. 2020, pp. 478–482, 2020.

[24] F. Sanhaji, H. Satori, and K. Satori, *"Cluster Head Selection Based on Neural Networks in Wireless Sensor Networks," 2019 International Conference on Wireless Technologies*, Morocco, Embedded and Intelligent Systems (WITS), Fez, 2019.

[25] M. Fattoum, Z. Jellali, and L. N. Atallah, "A joint clustering and routing algorithm based on GA for multi objective optimization in WSN," in *2020 IEEE Eighth International Conference on Communications and Networking (ComNet)*pp. 1–5, Hammamet, Tunisia, 2020.

[26] S. J. Islam, S. Islam, M. Z. Ferdus, M. N. I. Khan, M. A. Kashem, and M. S. Islam, "Load compactness and recognizing area aware cluster head selection of wireless sensor networks," in *2020 International conference on computing and information technology (ICCIT-1441)*, pp. 1–4, Tabuk, Saudi Arabia,, 2020.

[27] Q. Huamei, L. Chubin, G. Yijiahe, X. Wangping, and J. Ying, "An energy-efficient non-uniform clustering routing protocol based on improved shuffled frog leaping algorithm for wireless sensor networks," *IET Communications*, vol. 15, no. 3, pp. 374–383, 2021.

[28] J. Wang, D. Zhuangzhuang, Z. He, and X. Wang, "A cluster-head rotating election routing protocol for energy consumption optimization in wireless sensor networks," *Complexity*, vol. 2020, Article ID 6660117, 13 pages, 2020.

[29] S. Yu, Z. Liu, and X. He, "Hybrid PSO and evolutionary game theory protocol for clustering and routing in wireless sensor network," *Journal of Sensors*, vol. 2020, Article ID 8817815, 20 pages, 2020.

[30] D. M. Majeed, H. W. Rabee, and Z. Ma, "Improving energy consumption using fuzzy-GA clustering and ACO routing in WSN," in *2020 3rd international conference on artificial intelligence and big data (ICAIBD)*, pp. 293–298, Chengdu, China, 2020.

[31] M. Ghahramani and A. Laakdashti, "Efficient energy consumption in wireless sensor networks using an improved differential evolution algorithm," in *2020 10th international conference on computer and knowledge engineering (ICCKE)*, pp. 18–23, Mashhad, Iran, 2020.

[32] J. Z. Hare, S. Gupta, and T. A. Wettergren, "POSE.3C: prediction-based opportunistic sensing using distributed classification, clustering, and control in heterogeneous sensor networks," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 4, pp. 1438–1450, 2019.

[33] W. He, "Energy-saving algorithm and simulation of wireless sensor networks based on clustering routing protocol," *IEEE Access*, vol. 7, pp. 172505–172514, 2019.

Hindawi
Journal of Sensors
Volume 2021, Article ID 6664776, 16 pages
https://doi.org/10.1155/2021/6664776

Research Article

Sensor Fusion Basketball Shooting Posture Recognition System Based on CNN

Jingjin Fan,[1] Shuoben Bi,[1,2] Guojie Wang,[2] Li Zhang,[1] and Shilei Sun[2]

[1]*Research Institute of History for Science and Technology, Nanjing University of Information Science & Technology, Nanjing 210044, China*
[2]*School of Geographical Sciences, Nanjing University of Information Science & Technology, Nanjing 210044, China*

Correspondence should be addressed to Shuoben Bi; bishuoben@163.com

Received 20 December 2020; Revised 7 March 2021; Accepted 12 March 2021; Published 30 March 2021

Academic Editor: Bin Gao

In recent years, with the development of wearable sensor devices, research on sports monitoring using inertial measurement units has received increasing attention; however, a specific system for identifying various basketball shooting postures does not exist thus far. In this study, we designed a sensor fusion basketball shooting posture recognition system based on convolutional neural networks. The system, using the sensor fusion framework, collected the basketball shooting posture data of the players' main force hand and main force foot for sensor fusion and used a deep learning model based on convolutional neural networks for recognition. We collected 12,177 sensor fusion basketball shooting posture data entries of 13 Chinese adult male subjects aged 18–40 years and with at least 2 years of basketball experience without professional training. We then trained and tested the shooting posture data using the classic visual geometry group network 16 deep learning model. The intratest achieved a 98.6% average recall rate, 98.6% average precision rate, and 98.6% accuracy rate. The intertest achieved an average recall rate of 89.8%, an average precision rate of 91.1%, and an accuracy rate of 89.9%.

1. Introduction

Basketball is one of the most popular sports with a large fan base worldwide. As a competitive sport, basketball requires two teams of players to use various technical actions to compete with each other. A basketball game includes various technical statistics, such as score, rebound, assist, block, and steal, among which score is the central aspect that decides which team is the winner and loser [1–3]. The score in a basketball match is accomplished through players shooting the ball into the basket, and, thus, shooting plays an important tactical role in this game.

In recent years, with the rapid development of wearable sensor technology and the increased demand for basketball worldwide, many researchers have used wearable devices integrated with an internal measurement unit (IMU) to study basketball shooting [4–6]. Bai et al. [7] used Microsoft Band and weSport systems to collect data from two basketball players on both attack and defense, and they used a support vector machine (SVM) to effectively distinguish shooting and defense in basketball games. Aacikmese et al. [8], using an IMU placed on the arm, classified the six technical movements (forward-backward dribbling, left-right dribbling, regular dribbling, two-handed dribbling, shooting, and lay-up) in basketball by SVM. Zhao et al. [9] used four IMUs placed on the left and right upper arms and forearms to collect basketball technical movement data, using SVM to identify dribbling, passing, catching, and shooting. These studies, which use sensors on the arms, address basic shooting postures but ignore composite shooting postures. Composite shooting postures are shooting postures that consist of a series of hand and foot movements [10]. It is not sufficient to study composite shooting postures using only arm sensor data.

Shooting is a technical movement that requires physical coordination. In the shooting process, the movement of the feet is as important as the movement of the arms [1, 10]. Shi et al. [11] used smart insoles integrated with IMUs to distinguish between dribbling, jumping, and turning around during basketball. Peng et al. [12] also used smart insoles to

study the sideslip, back, cross, jab, and jump steps in basketball. These studies, which concern footstep movement in basketball using smart insoles integrated with IMUs, provide a basis for us to carry out research on composite shooting postures.

After more than 100 years of development, basketball has developed many complex and delicate technical movements, such as shooting. Recognizing these technical movements requires powerful recognition tools. SVM is a simple and robust algorithm and is widely used in basketball technical movement recognition [7–9]. However, SVM requires feature extraction and cannot be applied to large-scale training samples. Convolutional neural networks (CNNs) have solved these problems. A convolutional neural network is one of the representative algorithms of deep learning. Research on CNNs started in the 1980s and 1990s. In the twenty-first century, with the introduction of deep learning and the development of computer hardware capable of supporting deep learning, there has been a rapid development of CNNs, and these CNNs have worked well in the fields of computer vision and natural language processing [13–15]. With the development of CNNs, many researchers have studied sports using multiple IMUs combined with a CNN model of deep learning. Lee et al. [16] used the CNN long short-term memory model of deep learning to classify six squat positions (one correct and five incorrect); Kautz et al. [17] used a deep CNN to classify 10 types of beach volleyball technical movements. The effectiveness of these studies illustrates the great potential of deep learning models based on CNNs in the field of sports technical movement recognition.

Although there are many sports monitoring systems based on IMUs [6, 18, 19], there is still no system based on deep learning models to recognize a variety of basketball shooting postures. This study proposes a sensor fusion basketball shooting posture recognition system based on a CNN to recognize multiple types of basketball shooting postures. The main features of this study are as follows:

(1) A sensor fusion framework dedicated to basketball shooting postures is designed to collect shooting posture data and perform sensor data fusion

(2) 10 types of sensor fusion basketball shooting posture datasets are established, which can be used in related studies

(3) The 10 types of sensor fusion basketball shooting posture datasets were trained and tested using the classic visual geometry group network 16 (VGG16) deep learning model based on CNN, and the results verified the effectiveness of the deep learning model in shooting posture recognition

The remainder of this paper is organized as follows. Section 2 briefly summarizes the system framework and methods of data collection, fusion, and classification. Section 3 presents the experiments and results. Section 4 discusses the results of this study. Finally, conclusions are presented in Section 5.

2. Materials and Methods

2.1. System Hardware and Software Design. The sensor fusion basketball shooting posture recognition system consists of two independent wireless sensor modules, a USB dongle, and a laptop computer, as shown in Figure 1. The wireless sensor module is composed of an IMU (mpu-9250, including accelerometer and gyroscope) and a microcontroller unit (MCU, Nordic nrf52832, including Bluetooth functionality). The module is powered by an external 500 mA 3.3 V lithium battery. The IMU is responsible for collecting the players' raw shooting posture data, including accelerometer data and gyroscope data. The sampling rate was 100 Hz, and the collected data were transmitted to the MCU through I2C. The MCU is the core component of the wireless sensor module and is responsible for transmitting the raw shooting posture data from the IMU to the USB dongle via Bluetooth. The USB dongle includes Bluetooth and USB human interface device (HID) functions; it is responsible for transmitting the raw shooting posture data received from the wireless sensor modules to a laptop through a USB HID. The laptop contains a data-processing software developed by MATLAB, which is responsible for receiving, displaying, and fusing the raw shooting posture data to form the sensor fusion basketball shooting posture datasets.

2.2. Sensor Fusion Framework. Consider right-handed players as an example. When a basketball player shoots, his right hand is the main force hand, and at the same time, his left foot is the main force foot. The main force hand and main force foot perform the main tasks in basketball shooting. They make the shooting postures stable and can best reflect the characteristics of the shooting postures [10]. Therefore, the posture data of the right hand and left foot of right-handed players are key data. Correspondingly, the key data of the left-handed players were generated from the left hand and right foot. Modern basketball has several types of shooting posture. For basic shooting postures, the main force hand sensor data reflects the characteristics of the shooting postures. For composite shooting postures, such as stop jump shots and gather step shots, the main force hand sensor data cannot fully reflect the characteristics of the postures. However, the main force hand sensor data fused with the main force foot sensor data can fully reflect the characteristics of composite shooting postures. Therefore, this study proposes a sensor fusion framework for basketball shooting postures. This framework collects accelerometer and gyroscope data using wireless sensor modules placed on the main force hand and the main force foot of the player. Then, the data are fused to form the input of the deep learning model for shooting posture classification. The sensor fusion framework proposed in this study can classify a variety of complex shooting postures without increasing the number of sensors.

As shown in Figure 2, the sensor fusion framework proposed in this study consists of three steps: (1) shooting posture data collection, (2) data alignment and mergence, and (3) data segmentation and exclusion. First, we synchronized the two independent wireless sensor modules and placed them on the player's main force hand and main force foot.

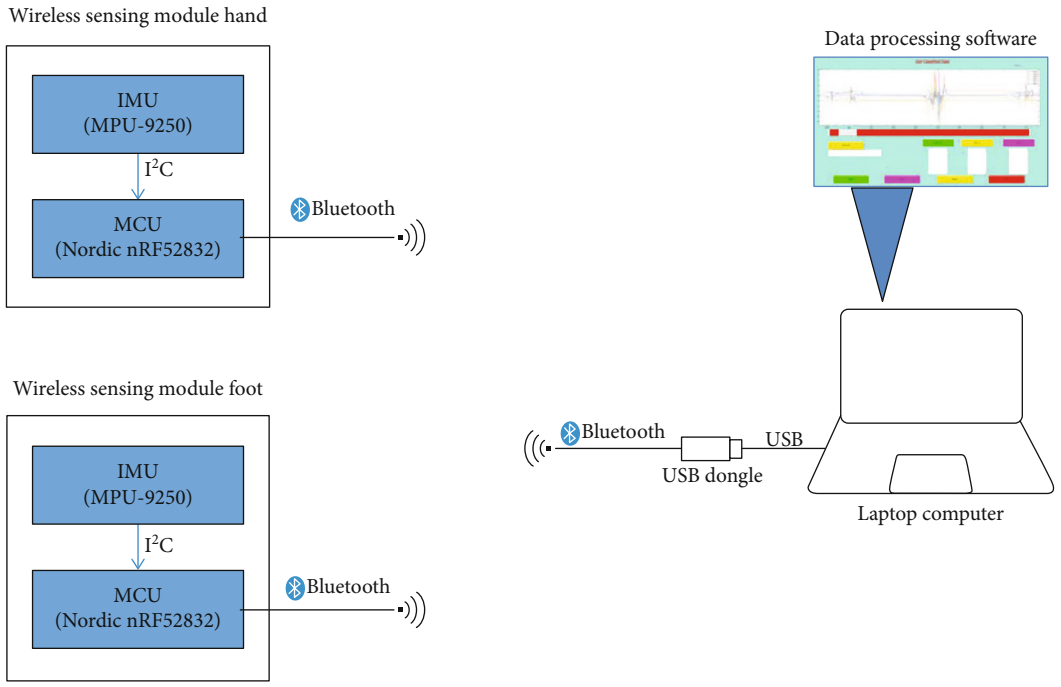


Figure 1: Sensor fusion basketball shooting posture recognition system.

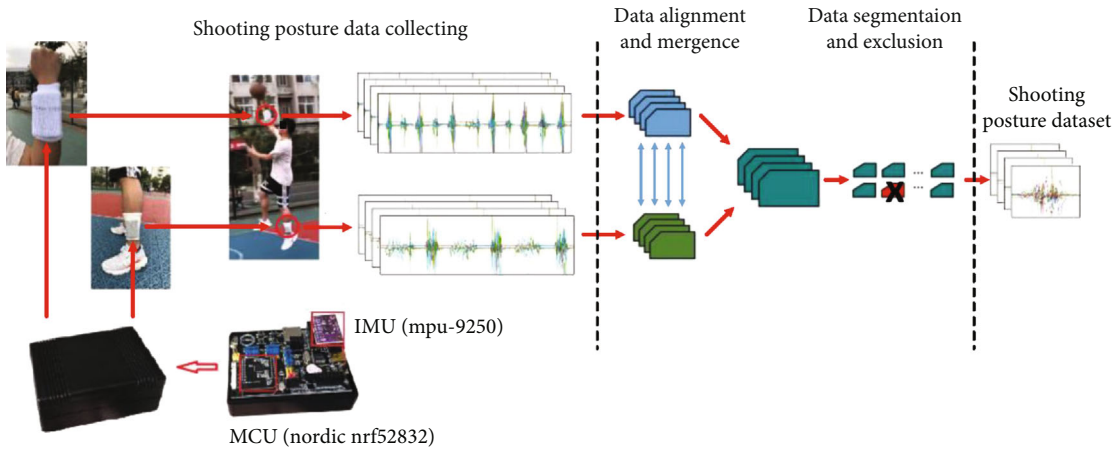


Figure 2: Sensor fusion framework.

Then, we collected the shooting posture data, which contain timestamps and transmit them to the laptop computer, stored in two separate files. Because our sampling rate is 100 Hz, the timestamp is in units of 10 ms. Second, the data in the two data files are aligned according to the timestamps and merged into a sensor fusion data file, as shown in Algorithm 1. File _H, file_F, and file_M represent the main force hand sensor data file, main force foot sensor data file, and sensor fusion data file, respectively. Owing to data loss in the wireless sensor module and other reasons, the shooting posture data of the main force hand and main force foot did not match, and hence, the sensor fusion data frequency suffered a loss of 1.17%. However, the frequency reduction did not affect the recognition of shooting postures. Finally, we divided the sensor fusion data file into independent shooting posture data entries, removed the erroneous posture data, and stored them in the sensor fusion basketball shooting posture dataset, as shown in Algorithm 2, where matrix(i) represents the *i*th shooting posture data matrix. We marked the data generated due to sensor misplacement or incorrect shooting posture in the experimental stage and deleted it in this stage. Thereafter, the sensor fusion basketball shooting posture dataset was finally formed.

2.3. Classification Model. The VGG16 [20–22] deep learning model based on CNN [23–25] is a model for image recognition proposed by the Visual Geometry Group of the University of Oxford in 2014. This model participated in the 2014 ImageNet Image Classification and Positioning Challenge and achieved excellent results. Because the

```
1. Open file_H, file_F, and file_M
2. Send file_H first record to record_H, send file_F first record to record_F
3. while (file_H NOT end) AND (file_F NOT end) do
4. if record_H.timestamp == record_F.timestamp do
5. Merge record_H and record_F to file_M
6. Send file_H next record to record_H, send file_F next record to record_F
7. else if record_H.timestamp > record_F.timestamp do
8. Delete record_F, send file_F next record to record_F
9. else if record_H.timestamp < record_F.timestamp do
10. Delete record_H, send file_H next record to record_H
11. end if
12. end while
13. Close file_H, file_F, and file_M
```

Algorithm 1: Data alignment and mergence.

VGG16 deep learning model showed excellent performance in image classification, and the VGG16 model with one-dimensional convolution kernels had been used to classify the one-dimensional data obtained by using the accelerometer and gyroscope [26], in this study, we used the one-dimensional convolution kernels VGG16 deep learning model to classify sensor fusion basketball shooting postures.

The structure of the VGG16 deep learning model mainly includes convolutional layer, max pooling layer, and fully connected layer, as shown in Figure 3. The function of the convolutional layer, which consists of several convolutional units, is to extract different features of the input data. Adding a greater number of convolutional layers means a greater number of complex features can be extracted. The working mode of the convolutional layer can be expressed by Equations (1), (2), and (3):

$$D = f(S * C), \tag{1}$$

$$S = [S_1, S_2, \cdots, S_m]^T = \begin{bmatrix} s_{11} \cdots s_{1n} \\ \vdots \ddots \vdots \\ s_{m1} \cdots s_{mn} \end{bmatrix}_{m \times n}, \tag{2}$$

$$C = \left[C_1^T, C_2^T, \cdots, C_k^T\right]_{l \times k} (l < n), \tag{3}$$

where S represents the vector of sample data, C represents the vector of the convolution kernel, m is the number of sample data, n is the number of input features, l is the length of the convolution kernel, and k is the number of convolution kernels. The result of the convolution operation of the ith convolution kernel is shown in Equation (4).

$$D_i = \begin{bmatrix} S_1 * C_i^T \\ \vdots \\ S_m * C_i^T \end{bmatrix}_{m \times n}, \tag{4}$$

where $i = 1, 2, 3, \cdots, k$. Because we use padding, the width of the vector after the convolution operation is n.

```
1. Open file_M
2. Send file_M first record to record_M
3. while file_M NOT end do
4. if record_M is shooting interval do
5. Delete record_M
6. else if record_M is shooting posture do
7. Send record_M to the shooting matrix(i)
8. if shooting posture is end do
9. if matrix(i) is not error do
10. Store matrix(i) to posture dataset
11. endif
12. Clear matrix(i)
13. end if
14. end if
15. Send file_M next record to record_M
16. end while
17. Close file_M
```

Algorithm 2: Data segmentation and exclusion.

The max pooling layer is mainly used for reducing feature dimensionality, compressing the number of data and parameters, reducing overfitting, and improving the fault tolerance of the model. The working mode of the max pooling layer can be expressed by Equations (5) and (6), where a is the stride.

$$Q = \max \text{pool}(D) = [Q_1, Q_2, \cdots, Q_k], \tag{5}$$

$$Q_j = \begin{matrix} Q_{11} & \cdots & Q_{1(n-a+1)} \\ \vdots & \ddots & \vdots \\ Q_{m1} & \cdots & Q_{m(n-a+1)} \end{matrix}_{m \times (n-a+1)}. \tag{6}$$

The fully connected layer mainly plays the role of classification, which is used to integrate and map the distributed feature representation extracted by the convolutional layer and the max pooling layer to the sample label space. The output of the fully connected layer is the final classification result.

The weight initialization of the convolutional layer and fully connected layer uses the Kaiming method [27], which can accelerate the convergence speed of the model.

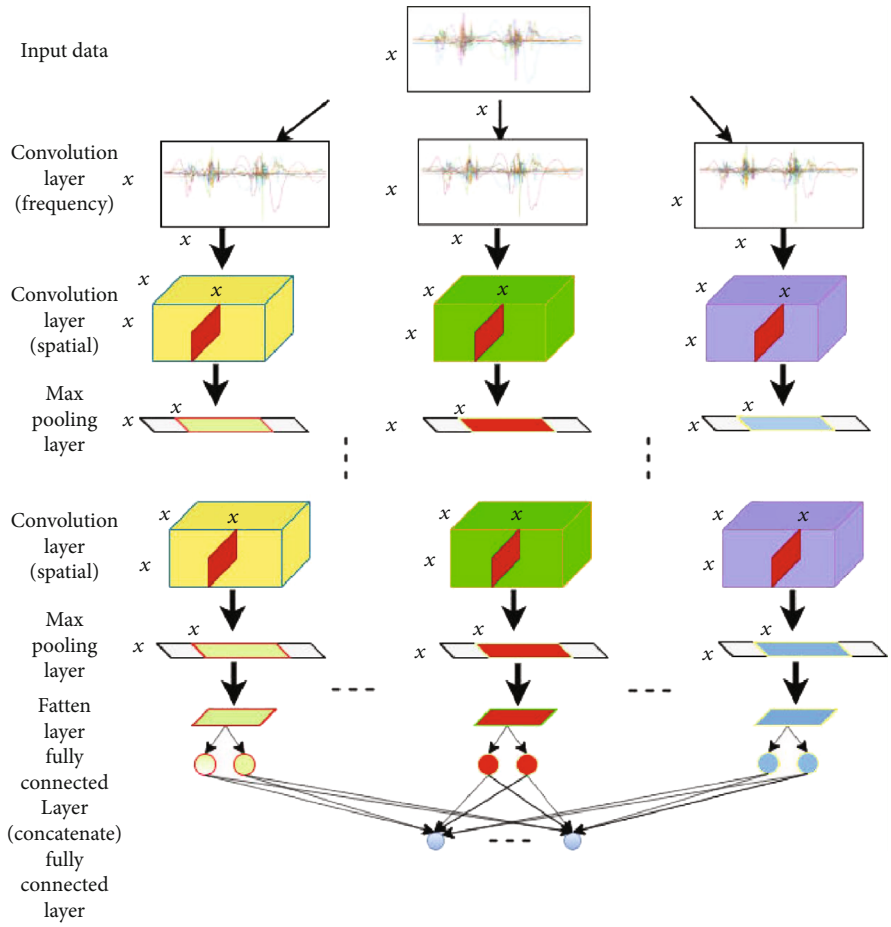


Figure 3: Structure of the visual geometry group network 16 (VGG16) deep learning model.

Table 1: Parameter settings of the visual geometry group network 16 (VGG16) deep learning model.

Parameters	Value
CNN layer weight initialization	Kaiming
Full connection layer weight initialization	Kaiming
Data standardization	*Z*-Score
Optimizer	Adam
Initial learning rate	1E-3
Loss function	Cross entropy
Network layer activation function	Rectified linear unit (ReLU)
Pooling method	Maximum pooling
Training rounds	20
Dropout rate	0.5
Batch size	200
Padding	[1]

The *Z*-Score method [28], which can convert datasets of different measurements into a unified measurement of *Z* -Score for comparison, is adopted for data standardization. The model uses the Adam optimizer [29], which has the advantages of simple implementation, high calculation efficiency, and lower memory requirements, and it is suitable for large-scale data and parameter scenarios. It is often used as the optimization algorithm for stochastic gradient descent (SGD). The minibatch gradient descent algorithm adopted for the model has the high speed of the SGD algorithm as well as the stability of the batch algorithm, which is suitable for deep learning models that need to process large amounts of data [25]; in the proposed model, the batch size is set to 200. The specific parameter settings are listed in Table 1.

Table 2: Characteristics of the subjects.

Parameters	Values
Age (years)	28.5 ± 9.5
Height (cm)	179 ± 14
Weight (kg)	80.5 ± 21.5
Experience (years)	12 ± 9

3. Experimental Method and Results

3.1. Experimental Method. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the IEC for Clinical Research of

Table 3: Basketball shooting postures.

Shooting postures	Explanation
Gather step shot	Outside the painted area, dribble one or two times, and then step back and shoot with both hands.
Hook shot	Inside the painted area, from the side of the basket, do not take off, hook shoot with one hand.
Free throw	After the free throw line, do not take off, shoot with both hands.
Stop jump shot	Outside the painted area, dribble, stop abruptly and jump, and then shoot with both hands.
Pump fake	Inside the painted area, make a fake shot with both hands, and then shoot with both hands.
Inside shot	Inside the painted area, under the basket, facing the basket, do not take off, shoot with both hands.
Jettison throw	From the three-point line, dribble forward, jump, and throw the ball into the basket with one hand.
Lay-up	From the three-point line, dribble forward, and perform a one-hand lay-up.
Jump shot	Outside the painted area, without dribbling, jump and shoot with both hands.
Spin jumper	Inside three-point line, dribble one or two times, and then turn around and shoot with both hands.

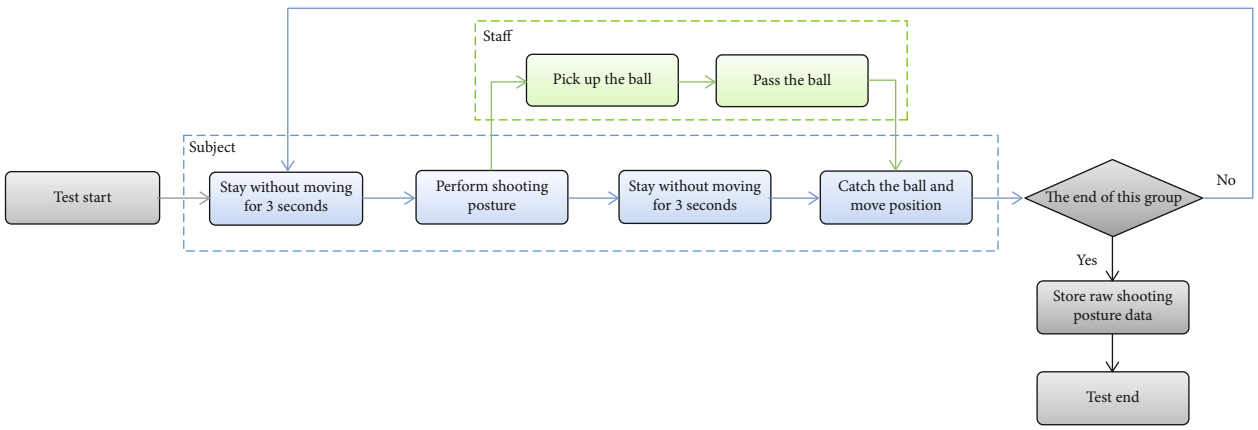


Figure 4: Shooting posture test flowchart.

Zhongda Hospital, affiliated with Southeast University (Project identification code: 2020ZDSYLL151-P01).

A total of 13 Chinese male adults (age 28.5 ± 9.5 years, height 179 ± 14 cm, weight 80.5 ± 21.5 kg) were selected as subjects. Although they had basketball experience (12 ± 9 years), they were not professional players and had no professional training, as reported in Table 2. All subjects were right-handed players. There were two centers (C), two power forwards (PF), three small forwards (SF), three shooting guards (SG), and three point guards (PG). Among them, five subjects had been trained as part of a college team. All subjects gave their informed consent for inclusion before they participated in the study. The subjects were verbally informed of the experiment process and precautions to be taken before the start of the experiment.

We chose 10 types of basketball shooting postures [30–33] for the experiment, as summarized in Table 3. These postures included five basic shooting postures: hook shot, free throw, inside shot, lay-up, and jump shot. In addition, we chose five types of composite shooting postures: gather step shots, stop jump shots, pump fakes, jettison throws, and spin jumpers. The stop jump shot, pump fake, and spin jumper are frequently used in basketball. The gather step shots and jettison throws are new introductions to the sports that have become increasingly popular in recent years.

The experiment was conducted in the basketball court of the Nanjing University of Information Science and Technology. The subjects repeated each of the 10 types of basketball shooting postures, as shown in Tables 3, 50–150 times. Each shooting posture was divided into 1–4 groups according to the physical strength of the subjects, with 25–150 shooting posture cycles in each group. At the beginning of each shooting posture cycle, the subjects held the ball without moving for 3 s and then performed the corresponding shooting posture. When the shooting posture was finished, they did not move until after 3 s had passed. Immediately after the shooting posture was completed, the staff picked up the ball and passed it to the subject after the subject moved. If the testing of the group was not complete, the next shooting posture cycle was started after the subject received the ball. If the testing of the group was complete, the data-processing software stored the raw shooting posture data, which contained 25–150 shooting posture cycles for sensor data fusion. The shooting posture test process is presented in Figure 4.

Finally, 10 types of sensor fusion basketball shooting posture datasets of 13 subjects, a total of 12,210 shooting posture data entries, including 12,177 valid data entries, were formed. The datasets included 1,210 gather step shots, 1,228 hook shots, 1,209 free throws, 1,223 stop jump shots, 1,221 pump fakes, 1,225 inside shots, 1,218 jettison throws, 1,216 lay-

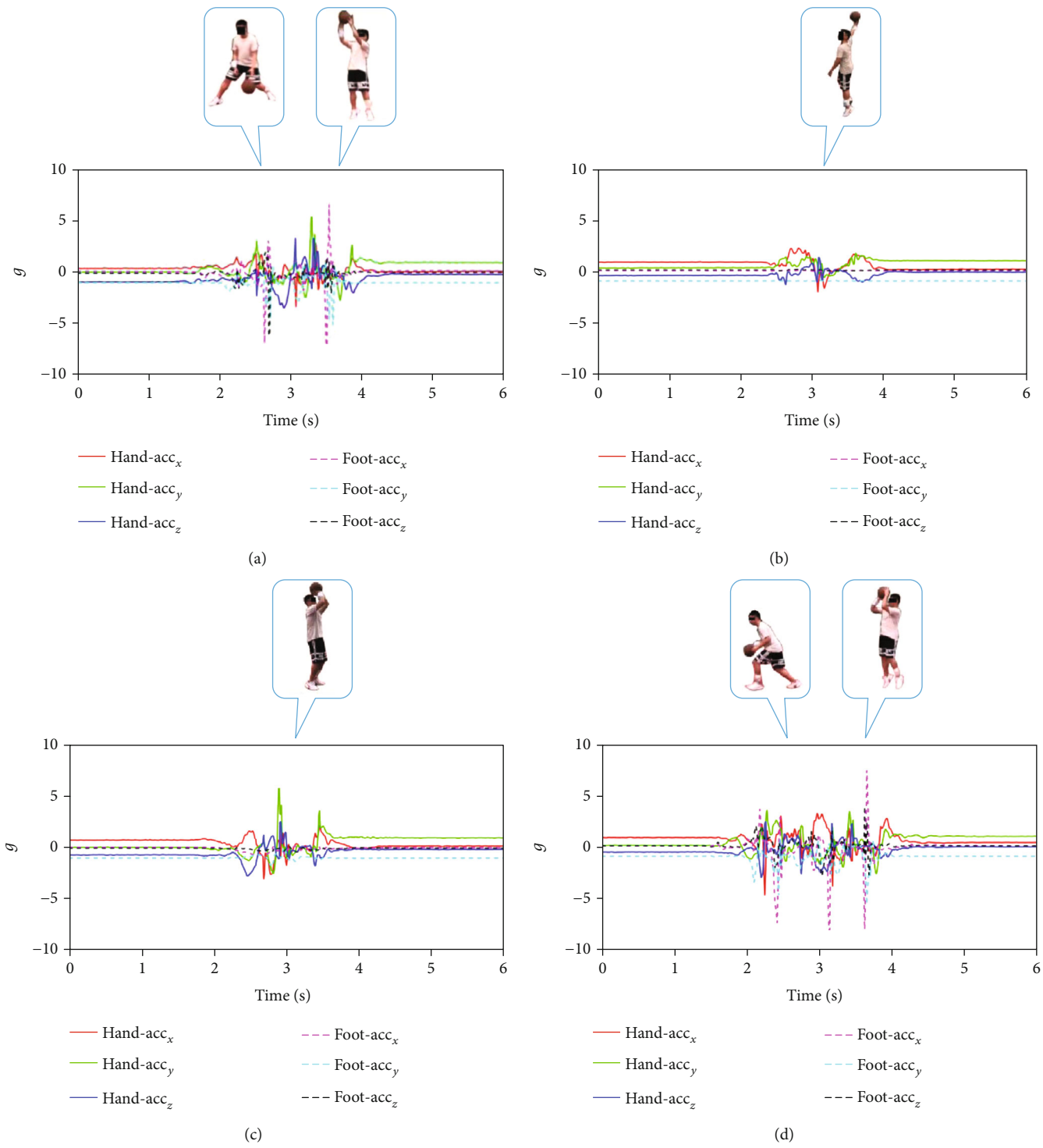


Figure 5: Continued.

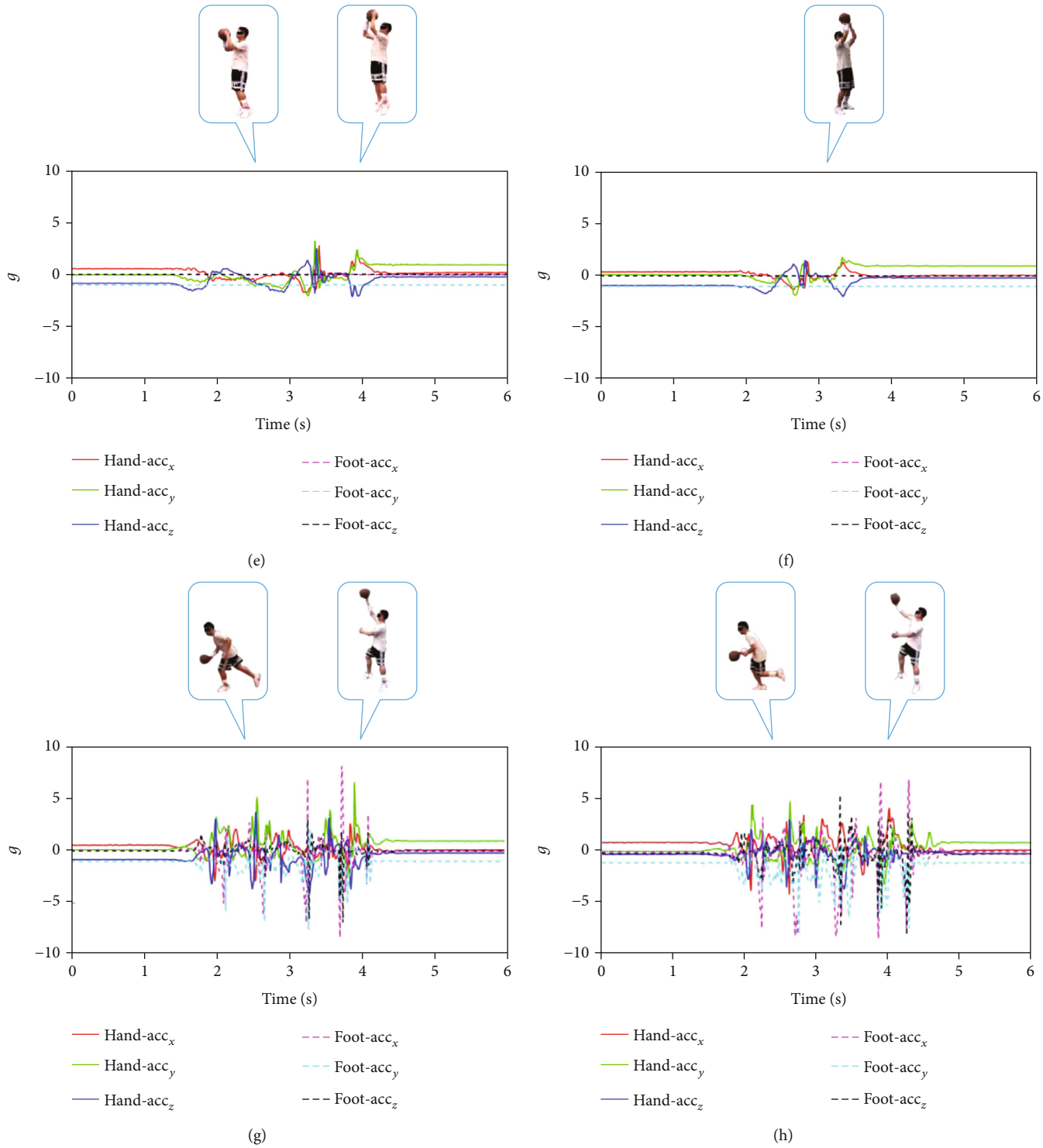


Figure 5: Continued.

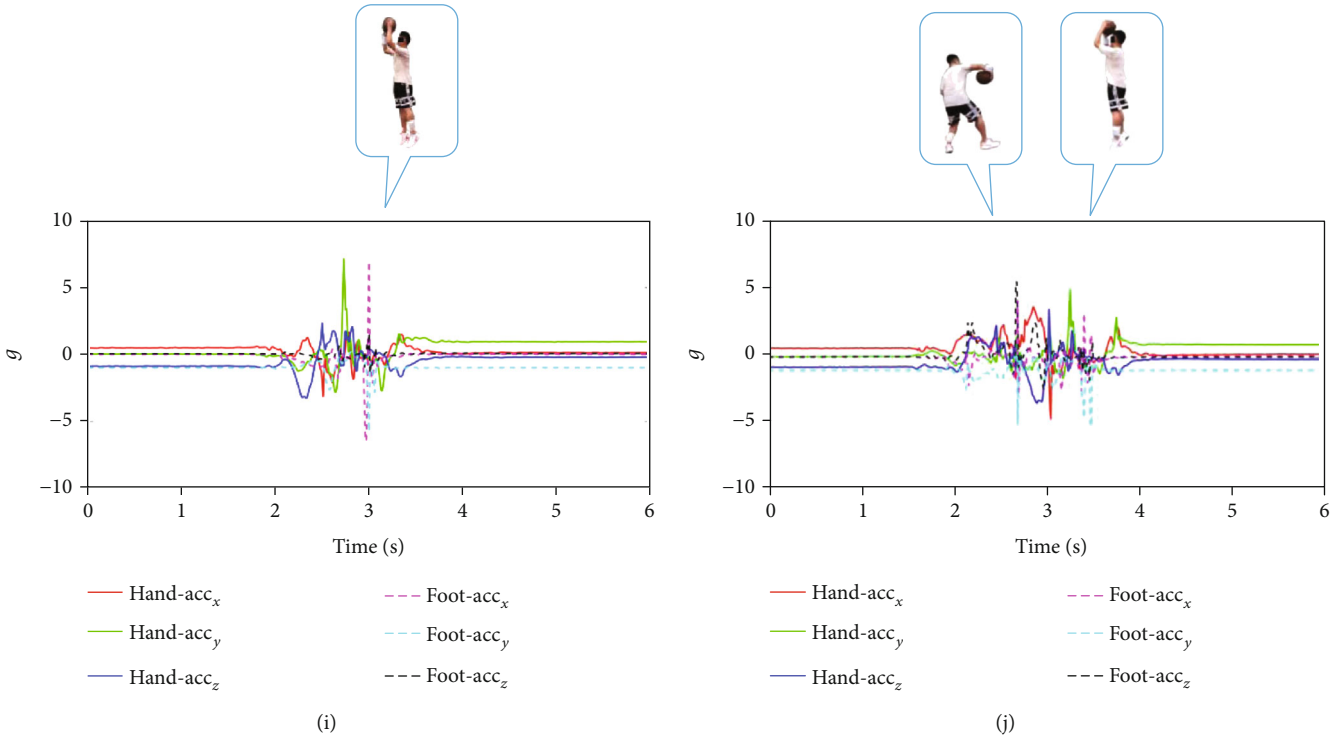


Figure 5: Accelerometer data waveforms of the 10 types of sensor fusion basketball shooting postures: (a) gather step shot, (b) hook shot, (c) free throw, (d) stop jump shot, (e) pump fake, (f) inside shot, (g) jettison throw, (h) lay-up, (i) jump shot, and (j) spin jumper.

ups, 1,207 jump shots, and 1,220 spin jumpers. The accelerometer data waveforms of the 10 types of sensor fusion basketball shooting postures are shown in Figure 5.

3.2. Classification. In this study, intra- and intertraining and testing methods were used for the sensor fusion basketball shooting posture datasets. Both methods were carried out on a computer configured with a Core i5-9400 CPU, 32 GB memory, and a GeForce GT730 graphics card. The operating system was Windows 10 Home, and the model was implemented using the MATLAB 2019b Deep Learning Toolbox.

3.2.1. Intratraining and Testing. All 12,177 sensor fusion basketball shooting posture data entries were randomly arranged, and the training and test datasets were designed with an 8 : 2 ratio, including 9,741 data entries in the training dataset and 2,436 data entries in the test dataset. The training dataset was used to train the model, and the test dataset was used to test the model. Figure 6 presents a comparison between the loss rate and accuracy rate of the intratraining process. As the loss rate decreases, the accuracy rate continuously increases, demonstrating the continuous improvement of the training model. Figure 7 presents the confusion matrix of 10 types of sensor fusion basketball shooting posture test dataset classified by the intratest. The row variables of the matrix represent the recall rate and false negative rate, and the column variables represent the precision rate and false discovery rate.

3.2.2. Intertraining and Testing. The sensor fusion basketball shooting posture data of 13 subjects were randomly arranged; the data of 11 subjects were used to form the training dataset, and the data of two subjects were used to form the test dataset. The total number of training data entries was 10,126, and the total number of test data entries was 2,051. The training dataset was used to train the model, and the test dataset was used to test the model. Figure 8 shows a comparison between the loss rate and the accuracy rate of the intertraining process. Figure 9 presents the confusion matrix of the 10 types of sensor fusion basketball shooting postures test dataset classified by the intertest.

3.3. Results and Analysis. Figure 10 depicts the t-SNE diagram of the intratest dataset. The t-SNE diagram shows the distribution characteristics of the data intuitively by reducing high-dimensional data to two-dimensional data. From the t-SNE diagram of the intratest dataset, it can be observed that the lay-up, jettison throw, and stop jump shot are easily confused, as well as hook shot, inside shot, and free throw.

Table 4 summarizes the recall rate, precision rate, average recall rate, and average precision rate of the intratest classification results. The classification results of the intratest reveal that the average recall rate was 98.6%, and the maximum recall rate was 100% for the jump shot and spin jumper. The minimum recall rate was 96% for the gather step shot. The average precision rate was 98.6%; the maximum precision rate was 100% for the hook shot, and the minimum precision rate was 97.2% for the jump shot. The above data indicate that although the t-SNE diagram shows that there are easily confused shooting postures in the intratest dataset, the sensor fusion basketball shooting posture recognition

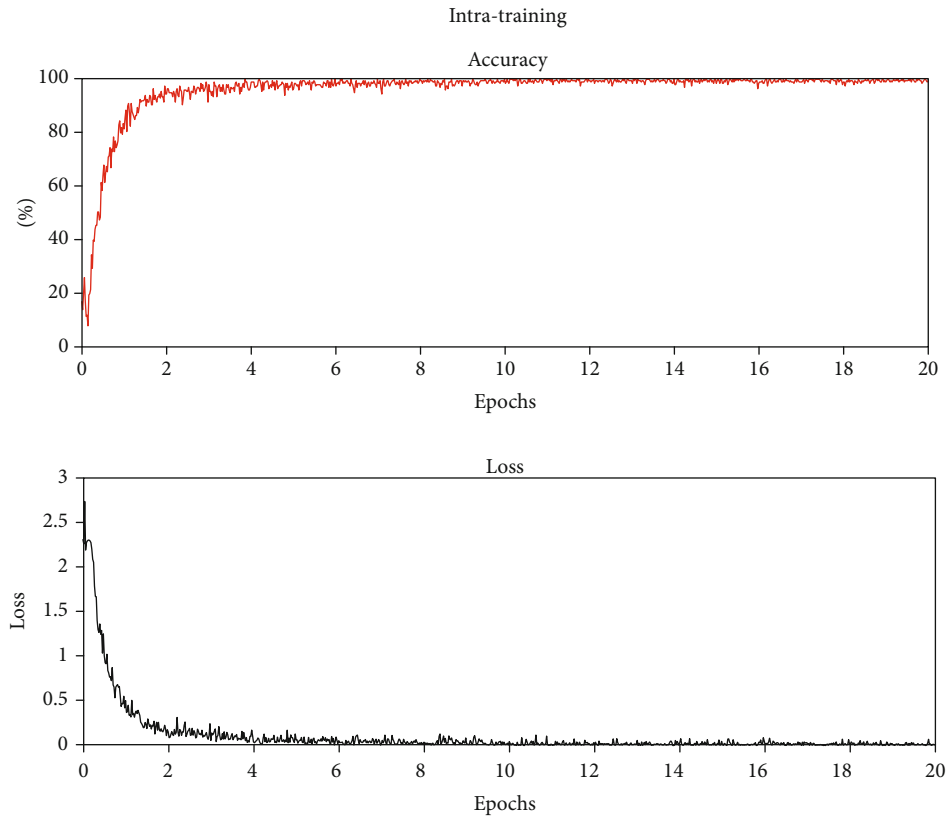


Figure 6: Accuracy rate curve and loss function value curve of the intratraining process.

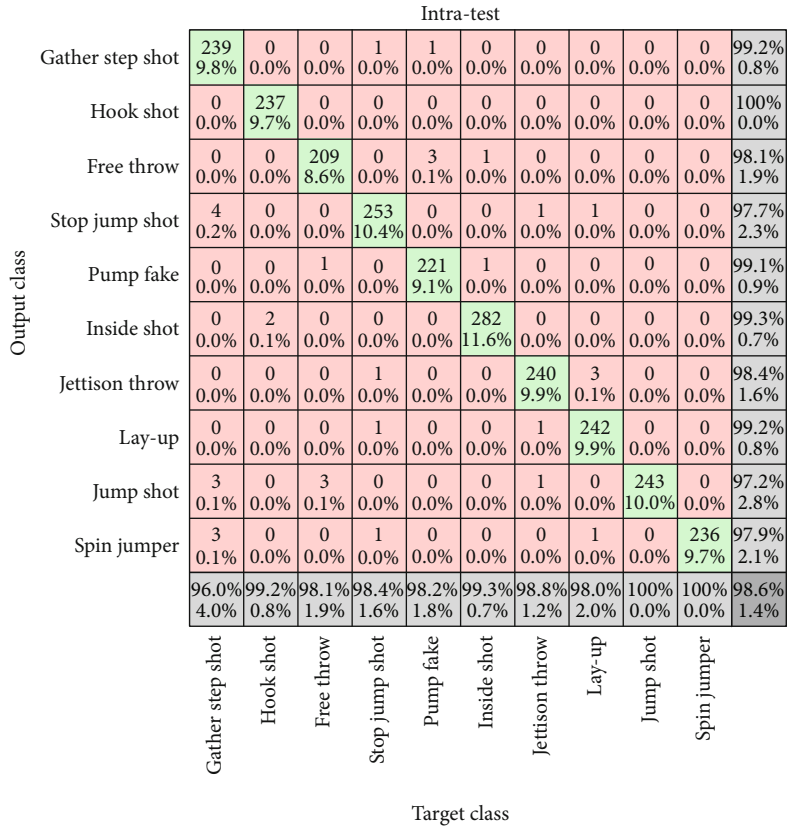


Figure 7: Classification confusion matrix of the intratest.

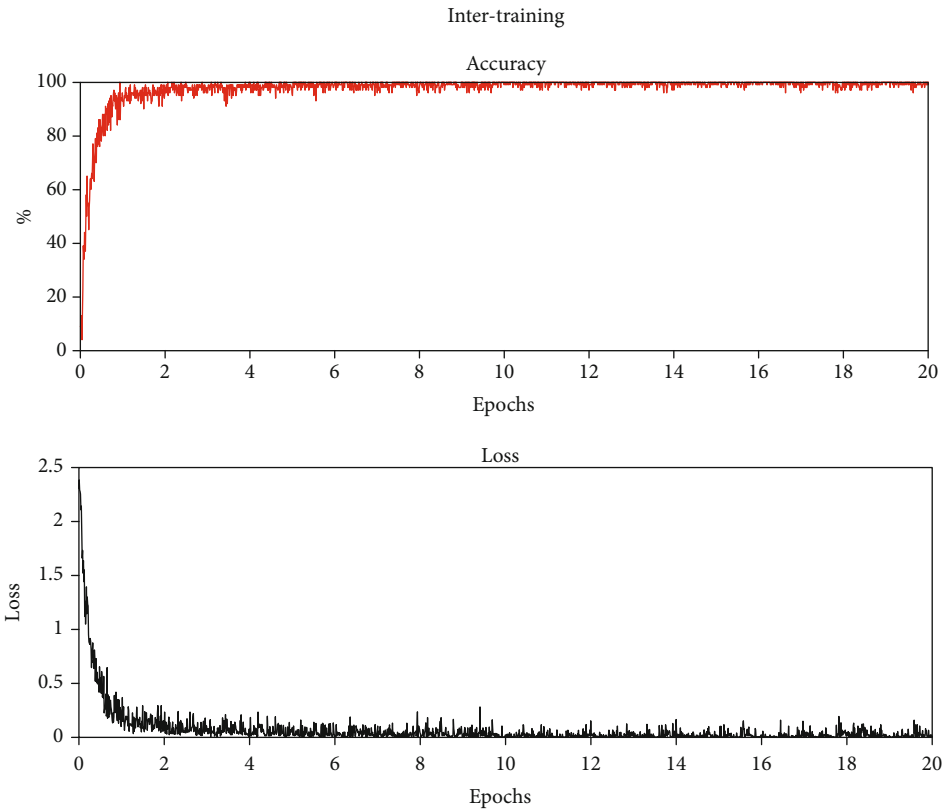


Figure 8: Accuracy rate curve and loss function value curve of the intertraining process.

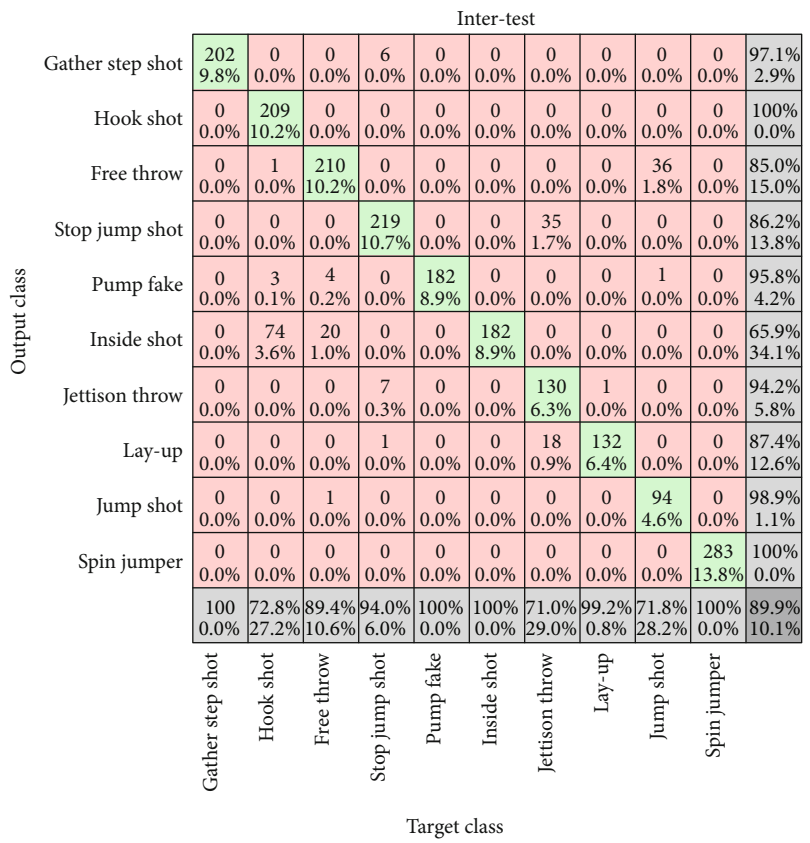


Figure 9: Classification confusion matrix of the intertest.

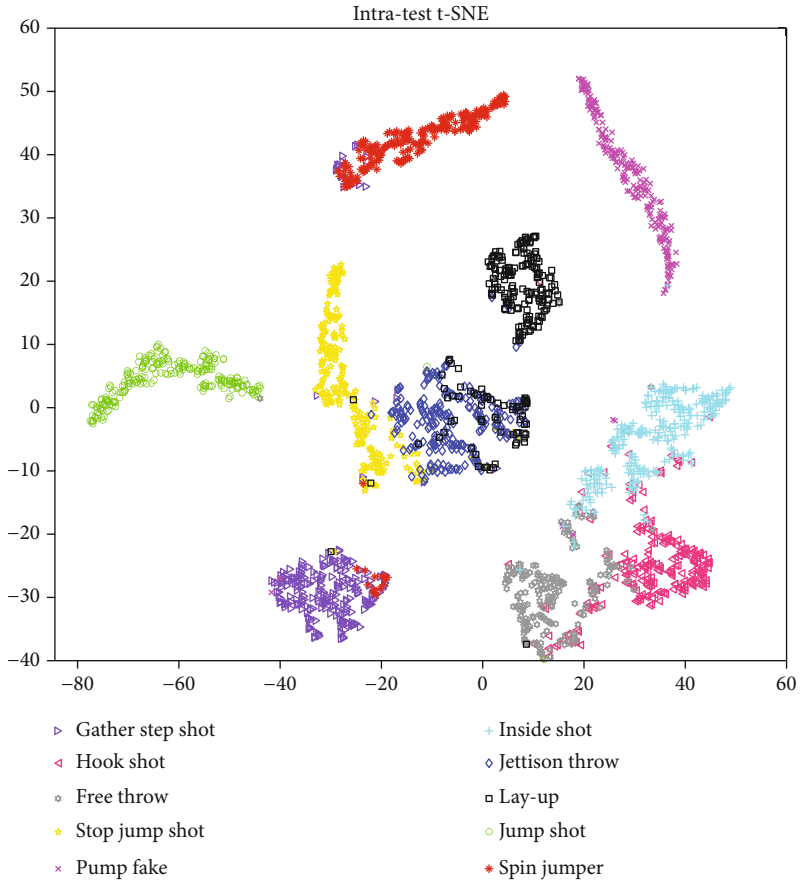


Figure 10: t-SNE intratest diagram.

Table 4: Classification results of intratest.

Number	Shooting postures	Recall	Precision
1	Gather step shot	96%	99.2%
2	Hook shot	99.2%	100%
3	Free throw	98.1%	98.1%
4	Stop jump shot	98.4%	97.7%
5	Pump fake	98.2%	99.1%
6	Inside shot	99.3%	99.3%
7	Jettison throw	98.8%	98.4%
8	Lay-up	98%	99.2%
9	Jump shot	100%	97.2%
10	Spin jumper	100%	97.9%
	Average	98.6%	98.6%

system still performed well in the intratest, as a result of the selection method of the intratraining and intratest datasets.

Figure 11 presents the t-SNE diagram of the intertest dataset. The t-SNE diagram of intertest dataset demonstrated that inside shot, free throw, hook shot, and jump shot are easily confused, as well as stop jump shot and jettison throw.

Table 5 reports the recall rate, precision rate, average recall rate, and average precision rate of the intertest classification results. The classification results of the intertest show that the average recall rate was 89.8%, and the maximum recall rate was 100% for the gather step shot, pump fake, inside shot, and spin jumper. The minimum recall rate was 71% for the jettison throw. The average precision rate was 91.1%; the maximum precision rate was 100% for the hook shot and spin jumper, and the minimum precision rate was 65.9% for the inside shot. As per Figure 10, 20 free throws and 74 hook shots were identified as inside shots. This is because free throws and inside shots are similar in action, differing only slightly in the release angle and speed. In addition, as we did not have wireless sensor modules on both left and right wrists, the ability to discriminate between single-handed and double-handed shooting postures is slight. Thus, some of the one-hand shots, such as hook shots, were identified as double-hand shots, such as inside shots. The 35 jettison throws were identified as stop jump shots for the same reason. There are 36 jump shots identified as free throws owing to the similar shooting distance and shooting angle between free throws and jump shots. Furthermore, there are no barometer data collected in this study; thus, there is no clear distinction between jump shooting posture and non-jump shooting posture. In addition, as the subjects in this experiment included five different play positions (i.e., C, PF, SF, SG, and PG), the heights and weights of the subjects were different, the subjects had very different actions in the same shooting posture, and considering that the subjects were

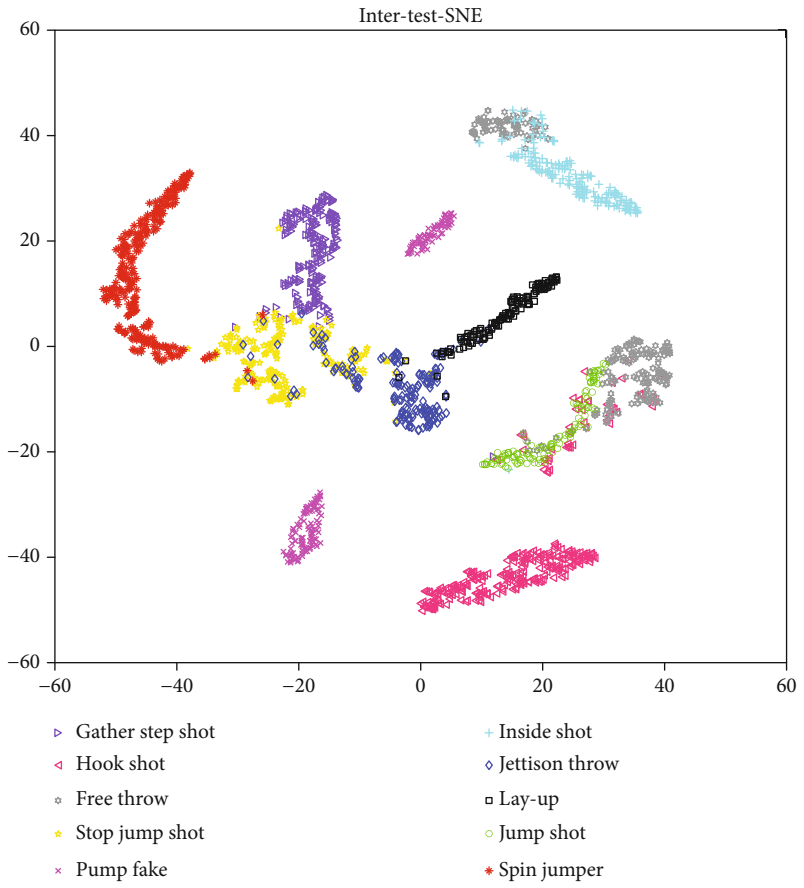


Figure 11: t-SNE intertest diagram.

Table 5: Classification results of intertest.

Number	Shooting postures	Recall	Precision
1	Gather step shot	100%	97.1%
2	Hook shot	72.8%	100%
3	Free throw	89.4%	85%
4	Stop jump shot	94%	86.2%
5	Pump fake	100%	95.8%
6	Inside shot	100%	65.9%
7	Jettison throw	71%	94.2%
8	Lay-up	99.2%	87.4%
9	Jump shot	71.8%	98.9%
10	Spin jumper	100%	100%
	Average	89.8%	91.1%

not professional players, the shooting postures varied considerably. In addition, stability is poor when physical strength is insufficient [34]. These two points also explain the aforementioned low recognition rate of the shooting postures. Finally, the small number of subjects contributes to the low recognition rate.

From the classification results, the VGG16 deep learning model achieved good classification in 10 types of sensor fusion basketball shooting posture recognition experiments.

Table 6: Accuracy comparison of classification models.

Literature	[35]	[36]	[37]	This paper
Model	1D-CNN	CNN	FMS-Net	VGG16
Accuracy	90.8%	91.92%	96.7%	98.6%

In contrast, [35] developed a deep learning model around a one-dimensional convolutional network (1D-CNN) architecture and verified it on the public dataset UTD-MHAD, which contained 27 types of activities. In [36], the CNN model was used to identify six types of pedestrian mode. In [37], a hybrid deep learning model based on the fusion of multiple spatiotemporal networks (FMS-Net) was proposed, which was used to detect the four phases of walking. As all the above research results were achieved only for the intratest, the comparison of the above research results with the intratest results of the VGG16 model used in this study found that the classification results of VGG16 were better than those of the other three classification models. The comparison results are shown in Table 6.

To verify the accuracy and effectiveness of the proposed system, it was compared with references [18, 19, 38]. Reference [18] established a real-time wearable assist system for upper extremity throwing action based on accelerometers, which used the longest common subsequence (LCS) algorithm to recognize the six phases of baseball throwing

Table 7: Systems comparison.

Literature	Sensor number	Sport	Posture number	Algorithm model	Average accuracy
[18]	2	Baseball	6	LCS	93.9%
[19]	3	Indoor rowing	6	ML	92.4%
[38]	2	Basketball	3	SVM	99.5%
This paper	2	Basketball	10	VGG16	94.3%

posture. In [19], an activity assessment chain for evaluating human activity was established using machine learning (ML) to classify six types of indoor rowing stroke postures (one correct and five incorrect). Reference [38] used a wearable and wireless system based on SVM to recognize overhead passes, chest passes, and shooting in basketball. As shown in Table 7, similar to the three systems above, the system proposed in this paper uses a small number of sensors to recognize a number of postures. This system has certain advantages in terms of average accuracy compared with the systems proposed in references [18, 19]. Although the average accuracy is slightly lower than the system proposed in reference [38], the system proposed in this paper recognizes more postures and achieves good recognition, even for easily confused postures. In addition, compared with the ML and SVM models used in references [19, 38], the deep learning model used in this study has greater development potential. Based on the above analysis, the system proposed in this study exhibits certain advantages compared with the other three systems.

4. Discussion

Shooting is an important aspect in basketball matches and training. Correctly distinguishing the shooting posture used by basketball players in a match and during training can help in making a correct evaluation of the technical characteristics of the players, which in turn could prove helpful in carrying out targeted guidance and practice sessions for players. This study proposes a sensor fusion framework for basketball shooting posture. It fuses the sensor data of the main force hand and main force foot to identify and classify the basic shooting postures and composite shooting postures in basketball. The framework proposed here shows a novel development direction for wearable devices in basketball, which is beyond the conventional framework of IMUs placed only on the arms. Although this sensor fusion framework can recognize more composite shooting postures without integrating more sensors, it still has limitations, which are as follows:

(1) The amount of limb data is still limited, which can pose challenges for accurately reflecting the subjects' posture information

(2) While using the proposed method, it is necessary to consider the problem of sensor synchronization and how to align the data of the two sensors if one sensor loses data

(3) The use of two sensors makes our proposed method more costly compared with the method that uses a single sensor

Many shooting postures in basketball have certain similarities, and nonprofessional basketball fans' shooting postures are generally not standard and less stable; hence, their shooting postures are generally confusing. The basketball shooting posture recognition system proposed in this paper selects nonprofessional basketball fans as subjects, uses a deep learning model based on CNN, classifies 10 types of easily confused shooting postures, and obtains a good classification effect, proving the feasibility of the deep learning model for basketball shooting posture recognition and demonstrating the robustness of the proposed system. Moreover, compared with ML models such as SVM, the deep learning model used in this paper has strong development potential in the future, and it is possible to integrate it into a low-cost integrated circuit in the future to reduce the cost of corresponding smart devices. Therefore, the results of this study can be used in the future for the development of low-cost wearable intelligent basketball motion recognition devices for nonprofessional basketball players.

Basketball shooting, especially composite shooting postures, can be divided into a series of decomposition actions. The time-series and attention of each decomposition action are different [10]. Although the VGG16 deep learning model used in this study has achieved good results in classifying 10 types of sensor fusion basketball shooting postures, there remain shooting postures with lower classification accuracy, such as stop jump shots and inside shots. If time-series and attention judgments are added to the deep learning model, the recognition effect could be further improved. Because researchers have used deep learning models that combine time-series and attention judgments for classification [39], we can also add time-series and attention judgments to deep learning models to improve classification accuracy in our future work. Furthermore, lightweight deep learning models such as MobileNet [40] and SqueezeNet [41] will be adopted to ensure the corresponding time and space efficiency after adding time-series and attention judgments.

Reference [42] studied walking and trotting in equestrian sports by calibrating the sensor data accuracy of four coordinate systems. Similarly, the accuracy of sensor data is also important for recognizing basketball shooting postures. Compared with the reference [42], this study still needs to be strengthened in sensor data accuracy calibration. In general, the sensor will suffer from sensor drift after a period of time, which affects the accuracy of the collected data. In addition, sensor misplacement, as the sensor is not firmly fixed on

the limb, can also lead to other accidents. In [43], the authors proposed a method combining zero velocity update (ZUPT) to reduce the sensor drift error. A rotation matrix method was also proposed in [44] that obtained good performance in dealing with sensor misplacement. In future works, to improve the data precision of sensor fusion for long-term data collection, we will attempt to fix sensor misplacement and sensor drift through software calibration, as was performed in references [42–44]. Furthermore, we will enhance the binding of wireless sensor modules and add a software filter to decrease the effect of sensor misplacement and sensor drift and improve the accuracy of the sensor fusion data.

5. Conclusion

In this study, a sensor fusion basketball shooting posture recognition system based on a CNN was designed. The system used a sensor fusion framework to collect the shooting posture data of the players' main force hand and main force foot and performed sensor data fusion. Subsequently, a CNN-based deep learning model was used for classification. A total of 12,177 sensor fusion basketball shooting posture data entries of the right hand and left foot were collected using this system for 13 Chinese adult male subjects aged 18–40 years with at least 2 years of basketball experience but without any professional basketball training. The shooting posture data entries were trained and tested using the classic VGG16 deep learning model based on CNN through intra- and intertraining/testing methods, achieving satisfactory classification results. These classification results are substantially better than those of similar systems, demonstrating the effectiveness and future development potential of the system.

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 conflict of interest.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (Nos. 41971340 and 41271410) and the National Key Research and Development Program of China (Grant no. 2019YFC1510203).

References

[1] M. Sun, *Advanced Course of Modern Basketball*, Peoples Sports Publishing House, Peking, China, 2018.

[2] J. Wang, "Ball games," in *Basketball*, pp. 88–120, Higher Education Press, Peking, China, 2009.

[3] H. Yang, "Essence, characteristics and laws of basketball," *Journal of Chengdu Sport University*, vol. 4, pp. 60–62, 2001.

[4] J. C. Maglott, J. Xu, and P. B. Shull, "Differences in arm motion timing characteristics for basketball free throw and jump shooting via a body-worn sensorized sleeve," in *Proceedings of the 2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 31–34, Eindhoven, Netherlands, 2017.

[5] A. Taniguchi, K. Watanabe, and Y. Kurihara, "Measurement and analyze of jump shoot motion in basketball using a 3-D acceleration and gyroscopic sensor," in *Proceedings of the SICE Annual Conference (SICE)*, pp. 361–365, Akita, Japan, 2012.

[6] M. C. S. Gutiérrez and P. M. V. Castellanos, "Design and validation of a system for improving the effectiveness of basketball players: a biomechanical analysis of the free throw," in *Proceedings of the 2018 IX International Seminar of Biomedical Engineering (SIB)*, pp. 1–8, Bogota, 2018.

[7] L. Bai, C. Efstratiou, and C. S. Ang, "weSport: utilising wristband sensing to detect player activities in basketball games," in *Proceedings of the 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pp. 1–6, Sydney, Australia, 2016.

[8] Y. Acikmese, B. C. Ustundag, and E. Golubovic, "Towards an artificial training expert system for basketball," in *2017 10th International Conference on Electrical and Electronics Engineering (ELECO)*, pp. 1300–1304, Bursa, Turkey, 2017.

[9] L. Zhao and W. Chen, "Detection and recognition of human body posture in motion based on sensor technology," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 15, no. 5, pp. 766–770, 2020.

[10] D. Hopla, *Basketball Shooting*, The People's Posts and Telecommunications Press, Peking, China, 2020, pp. 17-134.

[11] S. Shi, Q. F. Zhou, M. Peng, and X. Cheng, "Utilize smart insole to recognize basketball motions," in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pp. 1430–1434, Chengdu, China, 2018.

[12] M. Peng, Z. Zhang, and Q. Zhou, "Basketball footwork recognition using smart insoles integrated with multiple sensors," in *2020 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 1202–1207, Xiamen, China, 2020.

[13] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pp. 253–256, Paris, France, 2010.

[14] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, 2014.

[15] J. Gu, Z. Wang, J. Kuen et al., "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[16] J. Lee, H. Joo, J. Lee, and Y. Chee, "Automatic classification of squat posture using inertial sensors: deep learning approach," *Sensors*, vol. 20, no. 2, p. 361, 2020.

[17] T. Kautz, B. H. Groh, J. Hannink, U. Jensen, H. Strubberg, and B. M. Eskofier, "Activity recognition in beach volleyball using a deep convolutional neural network," *Data Mining and Knowledge Discovery*, vol. 31, no. 6, pp. 1678–1705, 2017.

[18] K. Y. Lian, W. H. Hsu, D. Balram, and C. Y. Lee, "A real-time wearable assist system for upper extremity throwing action based on accelerometers," *Sensors*, vol. 20, no. 5, p. 1344, 2020.

[19] M. Seiffert, F. Holstein, R. Schlosser, and J. Schiller, "Next generation cooperative wearables: generalized activity assessment computed fully distributed within a wireless body area network," *IEEE Access*, vol. 5, pp. 16793–16807, 2017.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556v5.

[21] H. Qassim, A. Verma, and D. Feinzimer, "Compressed residual-VGG16 CNN model for big data places image recognition," in *Proceedings of the 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 169–175, Las Vegas, NV, USA, 2018.

[22] Y. Osako, H. Yamane, S.-Y. Lin, P.-A. Chen, and R. Tao, "Cultivar discrimination of litchi fruit images using deep learning," *Scientia Horticulturae*, vol. 269, article 109360, 2020.

[23] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5325–5334, Boston, Massachusetts, 2015.

[24] M. B. Priatama, L. Novamizanti, S. Aulia, and E. B. Candrasari, "Hand gesture recognition using discrete wavelet transform and convolutional neural network," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 3, pp. 996–1004, 2020.

[25] P. Kim, *Design Example Based on MATLAB*, Beihang University Press, Peking, China, 2018.

[26] H. Lee and M. Whang, "Heart rate estimated from body movements at six degrees of freedom by convolutional neural networks," *Sensors*, vol. 18, no. 5, p. 1392, 2018.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pp. 1026–1034, Santiago, Chile, 2015.

[28] L. Z. Zhu, D. M. Chen, J. S. Guo, and H. X. Zhang, "Research on human action recognition based on synergistic LSTM neural network," *Computer Technology Development*, vol. 28, no. 12, pp. 79–82, 2018.

[29] X. Jiang, B. Hu, S. Chandra Satapathy, S. H. Wang, and Y. D. Zhang, "Fingerspelling identification for Chinese sign language via AlexNet-based transfer learning and Adam optimizer," *Scientific Programming*, vol. 2020, 13 pages, 2020.

[30] F. Erculj and E. Strumbelj, "Basketball shot types and shot success in different levels of competitive basketball," *Plos One*, vol. 10, no. 6, article e0128885, 2015.

[31] H. Okubo and M. Hubbard, "Dynamics of the basketball shot with application to the free throw," *Journal of Sports Sciences*, vol. 24, no. 12, pp. 1303–1314, 2006.

[32] L. Ning, M. Xiao-man, and Z. Ya-hui, "Research status and comments on technical characteristics of single-handed shoulder shooting," *Journal of Guangzhou Sport University*, vol. 39, no. 3, pp. 94–100, 2019.

[33] J. Krause, D. Meyer, and J. Meyer, *Basketball Skills and Drills*, Posts & Telecom Press, Peking, China, 2017.

[34] G. Marcolin, N. Camazzola, F. A. Panizzolo, D. Grigoletto, and A. Paoli, "Different intensities of basketball drills affect jump shot accuracy of expert and junior players," *PeerJ*, vol. 6, article e4250, 2018.

[35] N. Lemieux and R. Noumeir, "A hierarchical learning approach for human action recognition," *Sensors*, vol. 20, no. 17, p. 4946, 2020.

[36] J. Ye, X. Li, X. Zhang, Q. Zhang, and W. Chen, "Deep learning-based human activity real-time recognition for pedestrian navigation," *Sensors*, vol. 20, no. 9, p. 2574, 2020.

[37] T. Zhen, L. Yan, and J. L. Kong, "An acceleration based fusion of multiple spatiotemporal networks for gait phase detection," *International Journal of Environmental Research and Public Health*, vol. 17, no. 16, p. 5633, 2020.

[38] M. Mangiarotti, F. Ferrise, S. Graziosi, F. Tamburrino, and M. Bordegoni, "A wearable device to detect in real-time bimanual gestures of basketball players during training sessions," *Journal of Computing and Information Science in Engineering*, vol. 19, no. 1, 2019.

[39] X. Li, X. Yi, Z. Liu et al., "Application of novel hybrid deep leaning model for cleaner production in a paper industrial wastewater treatment system," *Journal of Cleaner Production*, vol. 294, article 126343, 2021.

[40] A. Howard, "MobileNets: efficient convolutional neural networks for mobile vision applications," 2017, https://arxiv.org/abs/1704.04861.

[41] F. Iandola, S. Han, M. Moskewicz, K. Ashraf, W. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, https://arxiv.org/abs/1602.07360v4.

[42] Z. Wang, J. Li, J. Wang et al., "Inertial sensor-based analysis of equestrian sports between beginner and professional riders under different horse gaits," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 11, pp. 2692–2704, 2018.

[43] S. Qiu, Z. Wang, H. Zhao, and H. Hu, "Using distributed wearable sensors to measure and evaluate human lower limb motions," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 4, pp. 939–950, 2016.

[44] M. Jiang, H. Shang, Z. Wang, H. Li, and Y. Wang, "A method to deal with installation errors of wearable accelerometers for human activity recognition," *Physiological Measurement*, vol. 32, no. 3, pp. 347–358, 2011.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5554920, 17 pages
https://doi.org/10.1155/2021/5554920

Research Article

A Convolutional Neural Network-Based Classification and Decision-Making Model for Visible Defect Identification of High-Speed Train Images

Zhixue Wang,[1] **Jianping Peng**,[1] **Wenwei Song**,[1] **Xiaorong Gao**,[1] **Yu Zhang**,[1] **Xiang Zhang**,[1] **Longfei Xiao**,[2] **and Li Ma**[2]

[1]*School of Physical Science and Technology, Southwest Jiaotong University, Chengdu 610031, China*
[2]*Chengdu Lead Science & Technology Co. Ltd., Chengdu 610091, China*

Correspondence should be addressed to Jianping Peng; adams.peng@swjtu.edu.cn

Received 11 January 2021; Revised 5 February 2021; Accepted 8 March 2021; Published 28 March 2021

Academic Editor: Bin Gao

In high-speed train safety inspection, two changed images which are derived from corresponding parts of the same train and photographed at different times are needed to identify whether they are defects. The critical challenge of this change classification task is how to make a correct decision by using bitemporal images. In this paper, two convolutional neural networks are presented to perform this task. Distinct from traditional classification tasks which simply group each image into different categories, the two presented networks are capable of inherently detecting differences between two images and further identifying changes by using a pair of images. In doing so, even in the case that abnormal samples of specific components are unavailable in training, our networks remain capable to make inference as to whether they become abnormal using change information. This proposed method can be used for recognition or verification applications where decisions cannot be made with only one image (state). Equipped with deep learning, this method can address many challenging tasks of high-speed train safety inspection, in which conventional methods cannot work well. To further improve performance, a novel multishape training method is introduced. Extensive experiments demonstrate that the proposed methods perform well.

1. Introduction

Traditional classification tasks using supervised learning methods, such as neural networks and support vector machines, generally require that all categories are available and the number of samples is sufficient. However, in the area of high-speed train safety inspection, abnormal targets that indicate there are underlying dangers while the train is running are scarce. Thus, we do not have a sufficient number of samples to implement deep learning to detect the abnormal targets. Instead, we devised a method based on the structural similarity method (SSIM) in the previous work [1]. In this method, the historical train images without malfunction are taken as baselines. When the current images are obtained and compared against the baselines, the changes occurring to the current trains are detected. As trains are exposed to the open air, there are various complex factors that cause the train surface to change. Therefore, most of the changes are not abnormal targets (correct alarm) but safety changes (false alarm) such as stains and marks, as shown in Figure 1(a) (row 1 and row 2). Besides, in order to obtain superior imaging quality, we photograph the train with supplementary lighting that usually leads to luminance difference. The luminance changes are usually mistakenly detected as abnormal targets, as shown in Figure 1(a) (row 3).

Although the previous work [1] saves plenty of manpower, there remains a need for inspectors to spend time classifying which changes are dangerous. To further reduce labor cost, this work is aimed at an automatic identification of the correct alarms with deep learning. There are various challenges in this task. The correct alarms are the components that are either loose (movement or rotation) or lost

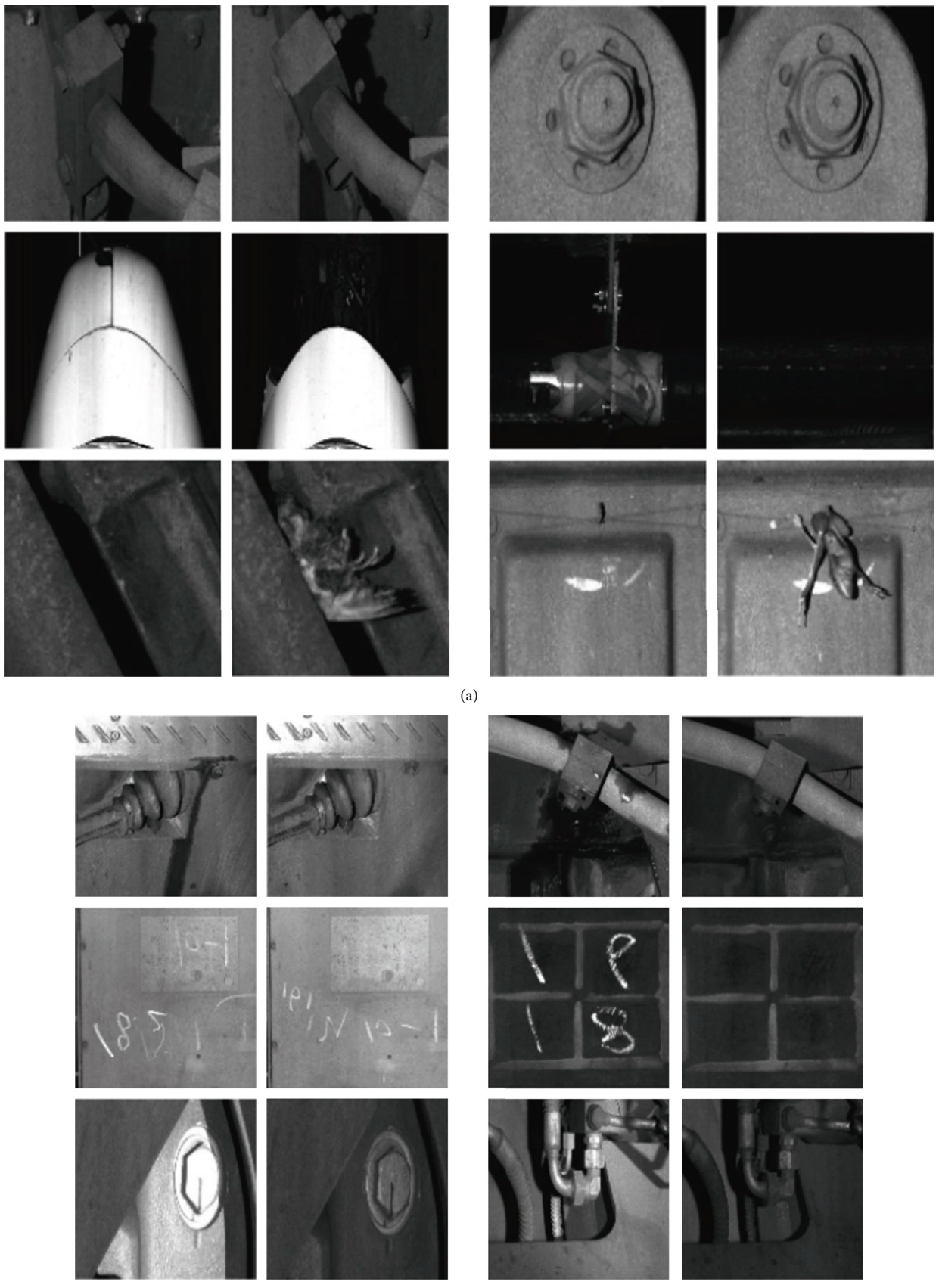


Figure 1: Examples of changes: (a) correct alarms; (b) false alarms.

and foreign bodies that appear on power installations such as the pantograph, as shown in Figure 1(b). These abnormal targets can lead the train to stop or even cause the train to turn over. Therefore, the abnormal target detection is extremely significant. However, correct alarms are incapable to be recognized by traditional classification methods, because their input is only one image (state). According to one state, algorithms cannot analyse whether the components are loose or lost. As for the false alarms, the algorithm should not just judge whether there are stains or luminance differences, because these existing signs do not indicate that there is no looseness, loss, etc. Therefore, the change information between two stages is required to assist the decision-making.

There are many components and equipment fitted on the train, especially at the bottom of the trains. That makes the image information of high-speed trains too complex to extract satisfactory edge information. Furthermore, without stable features, it is difficult to describe stains, luminance, foreign bodies, and component state changes. Actually, we cannot describe all kinds of shapes of the stains, but we can describe an abnormal condition of a component, even if describing all is unwise. Therefore, the manually designed descriptors, such as SIFT [2], may be not considered as an optimal method due to the factors mentioned above. On the other hand, it is effortless to obtain a large dataset that contains corresponding image pairs from the algorithm designed according to the previous work [1]. Therefore, deep learning [3] is adopted.

The paper is organized as follows. Section 2 explores other works that are somewhat like ours. Section 3 describes the two proposed convolutional neural networks (CNNs) for change classification in detail. Then, the experimental results and analyses are presented in Section 4. Finally, Section 5 gives the conclusions drawn from performing this study.

2. Related Works

2.1. Convolutional Neural Networks. CNNs, a family of algorithms especially suited to image analysis, have been applied in different ways, including image classification, object detection, and semantic image segmentation. Due to its strong ability of automatically learning high-level feature representations of images, CNNs can extract enough features for image classification [4–7] and perform better than traditional algorithms such as SIFT, HOG, and SURF. Moreover, it has the unique characteristic of preserving local image relations while performing dimensionality reduction. This makes it easy for CNNs to capture important feature relationships in an image and reduce the number of parameters the algorithm has to compute. CNNs are able to take as inputs and process both 2-dimensional images and 3-dimensional images; Ref. [8] proposed a 3DCNN to classify computed tomography (CT) brain scans which are 3-dimensional volumes. Based on the above, CNNs are the most popular machine learning in image recognition tasks. In object detection, there are also some excellent models such as Faster R-CNN [9], YOLO [10], and SSD [11]. Inspired by the successes of CNNs in above computer vision tasks, many researchers [12–15] make their efforts in different fields by using CNNs and achieve the state-of-the-art.

2.2. Change Classification. CNNs have been applied in different contexts for the comparison of image pairs [11–15]. Despite achieving state-of-the-art results in their tasks with two images, CNNs have yet to be applied to classifying change to the best of our knowledge. In Refs. [16], [17], and [18], CNNs are performed for change detection in the area of earth observation image analysis. By using a pair of coregistered aerial images taken at different times, the networks can infer the change map. It can be used to analyse the evolution of land use, urban coverage, deforestation, etc. In Refs. [19] and [20], the networks are trained to determine if two images correspond to each other by learning their similarity metric. They are widely used in image retrieval and face verification. By leveraging the convolution neural network, a variety of different challenges, for instance, changes in viewpoint, illumination problem, shading, and camera setting difference, are circumvented.

In brief, the first work is to detect where the changes occur, and the second one is to compute how similar they are. Our task is to recognize what kinds of changes happen or judge if these changes are dangerous. It needs to be emphasized that despite our use of change classification to identify abnormal targets for high-speed trains, it can be used for recognition or verification applications where decisions should be made over change information. In addition, although the inference process of the networks is comprised of one stage, they inherently divide the task into two parts, learning change information and classifying the change. We will show it in Sections 4 and 5.

3. Proposed Method

In this paper, two convolutional neural networks with reference to the residual network (ResNet) are presented [21, 22] to perform the change classification. There are two major differences between the change classification task and the traditional ones. First, the input of the networks should be two images instead of one that is required for the traditional classification task. Second, besides extracting image features, the proposed networks should learn to compare the image pairs to detect the change information.

The most straightforward way of improving the performance of the neural networks is to increase the depth [21–26], for which our networks are designed to have 32 layers. By extensive experiments, it is demonstrated that, as for our task, networks going deeper and wider (more units at each layer) cannot bring higher test accuracy but overfitting. The depth and width of our ultimate networks are optimal. The networks are trained end-to-end with 16k image pairs, and the number of samples is enough for a sufficient convergence with 80k iterations. Pretraining with other datasets is not utilized due to the differences between the conventional classification tasks and ours, and the considerable type differences between our dataset and the publicly available datasets such as ImageNet and MS COCO. In addition, according to Ref. [27], if the dataset is large enough (>10k), pretraining only

helps accelerate convergence but does not improve test accuracy or reduce overfitting. Thus, our designs are deemed reasonable. Moreover, a multishape training method is introduced to improve the performance.

3.1. Architectures. As mentioned above, the input to the CNNs is a pair of images. In this case, the main problem is how to integrate the two-image information to feed into the networks. The images we use are 1-channel grayscale images. The first idea is cascading the two images to be a "two-channel image." Although the "two-channel image" does not exist, it is convenient to process in the CNNs using two-channel convolution kernels. This architecture is called a cascaded model as shown in Figure 2(a). In the cascaded model, the two-channel image is processed by convolution layers to obtain the feature maps that contain change information. In order to reduce the overfitting, the global average pooling [16] is used for these feature maps to derive the final feature vector. Finally, the change category is outputted by the fully connected layer (FC).

The second architecture is inspired by Zhan et al. [28], in which two parallel networks are used to learn the pixel domain and wavelet domain information, and are cascaded by a fusion layer to implement image deblocking. Distinct from their work, the two parallel networks are involved to extract feature maps of the historical image (baseline) and current image as shown in Figure 2(b). Identical to the cascaded model, each branch applies a series of convolution layers and global average pooling. Then, the two branch outputs are concatenated and given to the top network that consists of FC. The two branches can be viewed as two feature extractors and the top network as a classifier. Consistent with Siamese and pseudo-Siamese networks [16–20, 26, 27], according to whether the weights of the two branches are shared, this architecture can be categorized into two types. Their performance is shown in Section 4.2.

3.2. Network Details. At present, ResNet [21, 22] and inception network [29–32] are accepted as excellent architectures. Therefore, while designing our networks, we refer to both of them. Owing to efficient convergence performance and concise structure, the residual module is primarily utilized in our networks. We adopt the bottleneck block that consists of two 1×1 and one 3×3 conv kernels [33]. The two 1×1 kernels are involved in dimensionality reduction and increment [21, 22, 29, 31, 32] to reduce the computation workload. The reason why we choose 3×3 conv kernels is that it has been demonstrated that multiple 3×3 conv kernels have the same receptive field as the larger one and have better nonlinear expressiveness due to activation function being used multiple times [24, 31]. Figure 3 presents the details of the block, in which batch normalization (BN) [30] is used as preactivation to improve the regularization of our models. The block can be expressed as

$$x_{i+1} = x_i + F(x_i, W_i), \tag{1}$$

where F indicates a series of BN, ReLU, and convolution operation; x_i and x_{i+1} are the input and output of the block; and W_i is the parameter which the model needs to learn. Recursively, Equation (1) is transformed into

$$x_m = x_n + \sum_{i=n}^{m} F(x_i, W_i). \tag{2}$$

Thus, the feature x_m of any deeper layer can be denoted as the feature x_n of any shallower layer n plus a residual function. Moreover, Equation (2) contributes to nice backward propagation properties. Denoting the loss function as l, we can obtain

$$\frac{\partial l}{\partial x_n} = \frac{\partial l}{\partial x_m}\left(1 + \frac{\partial l}{\partial x_n}\sum_{i=n}^{m} F(x_i, W_i)\right), \tag{3}$$

so that the loss can be directly propagated back to any shallower layer and the gradient of a layer cannot vanish [22]. At the end of the networks, the SoftMax layer is used to generate the pseudoprobability distribution, and by computing the cross-entropy, the loss is obtained to train the networks.

3.3. Multishape Training. The image pairs are directly provided by the previous work [1], and the shape is arbitrary. Namely, the height-width ratio is uncertain, for which our network should adapt to different shapes. To address the issue of different image shapes in training, we utilize three shapes: 180×360, 256×256, and 360×180. Roughly the same total pixels of the three shapes ensure that the computation is approximate in training. Thanks to the global average pooling, before being fed into FC, three shapes can be converted to the same length vector. In training, image pairs are alternately reshaped to the three shapes. While testing, image pairs are reshaped to the closest one.

In doing so, the dataset is augmented to some extent, and the benefit is twofold. In addition to improving the test accuracy due to a larger amount of data, it is conducive to change learning. However, in many previous works, such as R-CNN [34] and SPP-net [34], warping is not recommended due to the veridicality change. In contrast, our task is to recognize not what it is but the changes. Thus, after deformation, the change learning remains unaffected. In particular, stains, luminance changes, and foreign bodies have no stable feature. As a result, if the shape is changed, we could be unable to realize that. The examples are shown in Figure 4.

From Figure 4, it can be seen that after warping, the new stains, luminance changes, and foreign bodies are generated, and they all look natural. Certainly, the backgrounds may be anamorphic. However, they are desirable and make the networks more capable of learning changes instead of background category. For instance, as for the looseness such as what is shown in Figure 1(d), after training, the network may not learn what changes occur but learn that these components are usually in trouble, that is, even though the components in Figure 1(d) do not rotate, it can be recognized as a correct alarm. It is confirmed that this situation does not occur in Section 4.5.

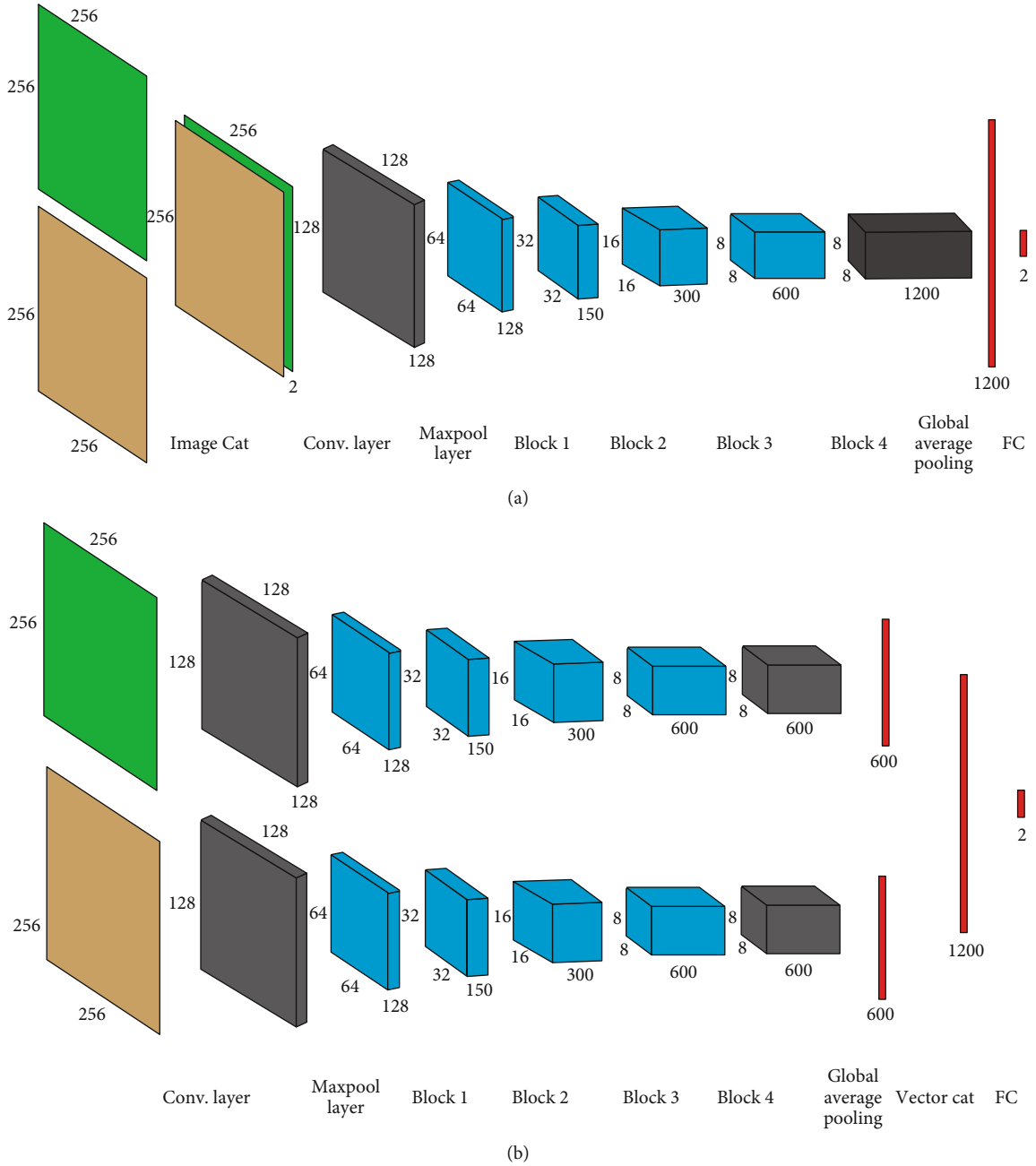


Figure 2: Two architectures: (a) cascaded model; (b) parallel model.

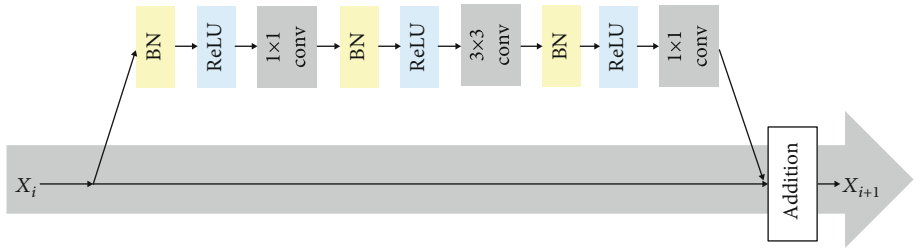


Figure 3: Bottleneck block.

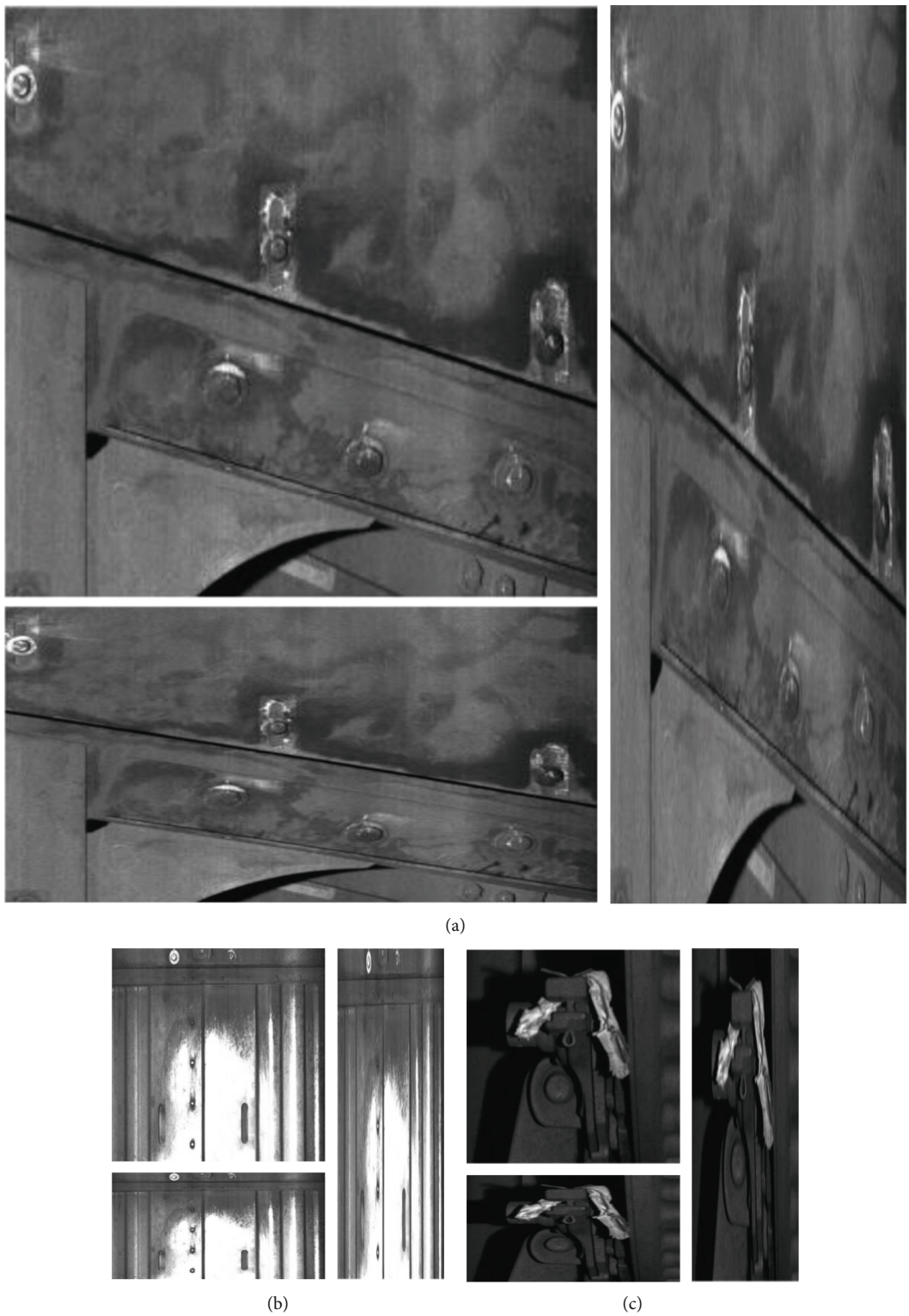


Figure 4: Examples of deformation: (a) stain; (b) luminance change; (c) foreign body.

4. Experimental Results and Analysis

As compared to tradition classification tasks, it has different input and a different goal. Thus, it is meaningless to compare our networks with state-of-the-art networks such as ResNet [21, 22] and inception network [29–32] which are all applied on conventional classification tasks. The point of our experiments is to determine the optimal configuration and to explore the reasonable pretreatment methods for change classification.

All networks are trained with Adam [35]. An exponential decay learning rate is used. The initial value is 0.01, and the

decay rate is 0.99. Except for the first conv layer, before convolution, BN and ReLU are performed in the first place, and the batch size is 96. To prevent overfitting, the L2 regularization is adopted and weights are initialized with the Xavier initializer [36]. All experiments are implemented six times using TensorFlow with an Nvidia GTX1080ti GPU and Intel i7-7700 CPU. The source code is publicly available at https://github.com/vivids/change_classification.

The experimental metrics used in our model are accuracy, precision, recall, and F1 score. The calculation method is shown as follows:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{4}$$

$$\text{presicion} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

$$F = \frac{\left(\alpha^2 + 1\right) \times \text{accuracy} \times \text{precision}}{\alpha^2 \times (\text{accuracy} + \text{precision})}, \tag{7}$$

where TN, TP, FN, and FP are indicated in Table 1. F is a kind of comprehensive evaluation metrics. If α is equal to 1 in (7), it is the F1 score.

4.1. Data Sets and Data Processing. With the assistance of our previous work [1], about 18k image pairs that are the corresponding parts of the same high-speed train were collected at different times. These images are all taken from the high-speed train's body and its key components, such as the locomotive running gear, bogie, wheel, fastening bolt, and pipeline. Due to the different sizes of different key components, the acquired images have different resolution sizes, ranging from tens to thousands. The most defects contained in this dataset are the forebody, which is brought by a tree branch, the body of birds and other animals, plastic bag and other light garbage, and so on. In movement and rotation, one of the defects is a loose fastening component loose and a loose bolt, respectively.

While labeling the image pairs, it was found out that the category of many pairs is ambiguous as a result of cooccurrences of multiple circumstances such as the first image pair in Figure 1(a), where a stain and a luminance change appear simultaneously, so that the exacting correct multiclassification dataset is not available. Considering that we are not concerned about what kinds of changes occur, but in terms of whether they are dangerous, the change classification can be regarded as a binary classification task, a correct alarm or a false alarm. As the correct alarms are all structure changes whereas the false ones are nonstructure changes, the binary classification is feasible. Our experiments are primarily aimed at binary classification. Multiclassification experiments (with unsatisfactory multiclassification dataset) are also executed to demonstrate that the network can recognize different kinds of changes and assess the detail performance of the networks.

Table 1: Confusion matrix.

		Predicted label	
		Positive	Negative
True label	True	TP	FN
	False	FP	TN

As for the multiclassification task, the dataset is split into six categories, stain, luminance, mark, rotation, movement, and foreign body. Concerning binary classification, the first three categories are merged as false alarms while the rest as correct ones. There are more false alarms than correct ones, for which we discard some false ones for equity purpose. In both multiclassification and binary classification, we select about 10% data to test the networks, and the details are shown in Table 2.

In a traditional classification task, before training, the images are usually standardized to the same distribution where the mean is 0 and the variance is 1. However, it can eliminate the brightness difference between an image pair, so it hinders the networks from learning luminance changes. Instead, we merely normalize all image pixel values to [0, 1]. The images are resized to (256×256) for single-shape training and are alternately resized to (180×360), (256×256), and (360×180) for multishape training.

4.2. Two Architectures. The depth and width of neural networks are hyperparameters. To explore the optimal settings, we conduct many experiments. In Table 3, taking the cascaded model as an example, some typical settings are listed. In this section, we design two architectures based on the slim model (see Table 3 for details) to compare their performance and will demonstrate that the slim model can perform well in both speed and accuracy in Section 4.3. The two architectures are shown in Figure 2.

From Table 2, we can clearly see that cascaded model outperforms the parallel ones by a large margin in all metrics, which is attributed to the independent feature extraction of both branches, a result of which the parallel models cannot learn change information well. To further validate that the independent feature extraction hampers the learning for changes, we construct a hybrid model as displayed in Figure 5. In the front part of the network, two images are processed independently. In this way, the network can better extract the fundamental information, such as edges, of the two images. The rest is the same as the cascaded model, so this network has enough layers to detect and process the change information. However, from Table 4, it can be seen that independent feature extraction indeed does a disservice.

The Siamese models used in Refs. 16, 17 and 18 are similar to the parallel models, but they can perform well in change detection which is due to the difference between the two tasks. Moreover, it is also suggested that the change information is primarily extracted in the first few layers, for which these layers are significant to our proposed networks. We will demonstrate that the first convolutional layer is responsible for detecting change information in Section 4.5. Thus, we should integrate the two-image information early.

Table 2: Dataset details.

	2-cls		6-cls					
	False	Correct	Stain	Luminance	Mark	Rotation	Movement	Foreign body
Train	7857	7770	3449	3533	1824	3433	2636	1651
Test	1000	1000	350	350	350	350	350	350

4.3. Architecture Optimization. With networks going deeper, the performance is usually improved [23–26]. However, they are proven by training and testing with some very large datasets such as ImageNet and MS COCO. The reason is not only the deep networks having better nonlinear representation ability but also the shallow networks being underfit for a large dataset. In this section, we demonstrate that as for a small dataset, the deeper and wider network cannot improve the performance but can cause overfitting and bring more computation. We explore six networks of varying depth and width as shown in Table 3. For quantitative analysis of the complexity of the proposed method, we analyse the FLOPs of our networks. In our networks, the 101 layers (deepest) and 32 layers (thin) are the largest and smallest networks with FLOPs 7.5×10^9 and 6.8×10^8, respectively.

Table 5 presents the quantitative evaluation of the above six models. Each of them is tested six times to ensure objectivity. The modified ResNet-50 model [21, 22] is applied in the experiment. The modification is that the channel number of the first conv kernel is modified from 64 to 75 to be consistent with the slim model. The result reveals that the slim model is the most appropriate. Although the fat model achieves an excellent precision rate, it does poorly in the recall metric, as a result of which, the F1 rate is reduced as well. Moreover, the fat model is time-consuming. Owing to overfitting, the models that have more parameters may ignore the learning of category generality but memorize the training images. Thus, while testing, the results are not desirable. On the contrary, if the model is excessively thin or shallow, it is not qualified for the change classification task, that is, the model is underfitting.

4.4. Data Preprocessing. Preprocessing is effective in preventing the model from being affected by the irrelevant factors to some extent. For most recognition tasks, the data preprocessing can augment the dataset to improve the performance of the models. The common pretreatment methods include flipping, grayscale transformation, standardization, cropping [37], etc. However, regarding our task, the learning for different categories can be disrupted by some preprocessing methods. The methods of changing grayscale values are not suitable for luminance changes. Considering that the abnormal targets usually occupy a small part of our images, the cropping is not reasonable. In this section, we first implement training with standardization to validate that it can cause hindrance to the learning for luminance changes. Then, we train our model with image flipping horizontally and vertically to verify that it can improve accuracy. The results are listed in Table 6.

It is obvious that through flipping, the performance of the network is improved, and by means of standardization, the performance is degraded by a large margin. To further explore the influence of standardization, we implement six-category classification experiments. From Table 2, it can be seen that the number of the six categories is uneven, especially for the mark and foreign body. We first select 350 image pairs as the test data and then augment the number of marks, foreign bodies, and movements in the training set by means of grayscale transformation and cropping to equalize the dataset. As aforementioned, these augment methods are not suited to all scenarios, for which it cannot be used in training. However, we can augment the images in the disk and select the ideal ones. Because the test set is selected in advance, the experiments are considered reasonable.

Table 7 reveals that, as predicted, the recall of luminance drops sharply and the recall of stain decreases from 84.00% to 79.71%. Due to the decrease of brightness difference, the model finds it is more difficult to classify stain and luminance. According to Figure 6, we can find that, after standardization, there are more instances of luminance predicted as stains, as well as stains predicted as luminance. For instance, in Figure 6(a), 12.57% of luminance examples are predicted as stains, and 9.43% of stain examples are predicted as luminance. Based on Figure 6(b), after using standardization, the error rate increases to 18% and 12.29%. However, owing to the decrease of brightness difference, the network can learn some categories better, such as the movement and mark. For example, in the confusion matrix (Figure 6), less mark examples are classified as stain and luminance. The rate at which mark examples are classified as stain or luminance has decreased by 1.14% and 1.41%, respectively. Some categories can be targeted to use the standardization method, but it will lead the training set to have nonuniform distributions that are not beneficial for training [30]. Therefore, in our experiments, we do not adopt the standardization method.

4.5. Multishape Training. Multishape training is conducive to learning change information from different shape images. First, it can augment the dataset. Second, owing to being warped, the backgrounds are anamorphic, but the change information is almost unaffected. Certainly, we should reshape the images properly; otherwise, the change information will be harmed as well. Last, while testing, thanks to the ability to process multiple shapes, we can convert the image to the ideal shape for prediction. If high speed is not required, we can make inference with all shapes to vote which category it is. Furthermore, if the precision is pursued, only when the results predicted with all shapes are consistent will the final decision be made. Otherwise, it should be submitted to the inspector to judge.

Table 3: Configurations of cascaded models.

Layer name	Output size	32 layers (fat)	50 layers (ResNet-50)	32 layers (slim)	32 layers (thin)	23 layers (shallow)	101 layers (deepest)
conv1	128×128	7×7, 75, stride 2					
Pooling	64×64	3×3 max pooling, stride 2					
conv2_x	32×32	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 50 \\ 3 \times 3, 50 \\ 1 \times 1, 150 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 32 \\ 3 \times 3, 32 \\ 1 \times 1, 96 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 50 \\ 3 \times 3, 50 \\ 1 \times 1, 150 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 50 \\ 3 \times 3, 50 \\ 1 \times 1, 150 \end{bmatrix} \times 3$
conv3_x	16×16	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 100 \\ 3 \times 3, 100 \\ 1 \times 1, 300 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 192 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 100 \\ 3 \times 3, 100 \\ 1 \times 1, 300 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 100 \\ 3 \times 3, 100 \\ 1 \times 1, 300 \end{bmatrix} \times 4$
conv4_x	8×8	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 200 \\ 3 \times 3, 200 \\ 1 \times 1, 600 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 384 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 200 \\ 3 \times 3, 200 \\ 1 \times 1, 600 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 200 \\ 3 \times 3, 200 \\ 1 \times 1, 600 \end{bmatrix} \times 23$
conv5_x	8×8	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 1$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 400 \\ 3 \times 3, 400 \\ 1 \times 1, 1200 \end{bmatrix} \times 1$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 786 \end{bmatrix} \times 1$	$\begin{bmatrix} 1 \times 1, 400 \\ 3 \times 3, 400 \\ 1 \times 1, 1200 \end{bmatrix} \times 1$	$\begin{bmatrix} 1 \times 1, 400 \\ 3 \times 3, 400 \\ 1 \times 1, 1200 \end{bmatrix} \times 3$
Rest	1×1	Global average pooling, fc, SoftMax, cross-entropy					
FLOPs		2.4×10^9	4.1×10^9	1.5×10^9	6.8×10^8	9.7×10^8	7.5×10^9

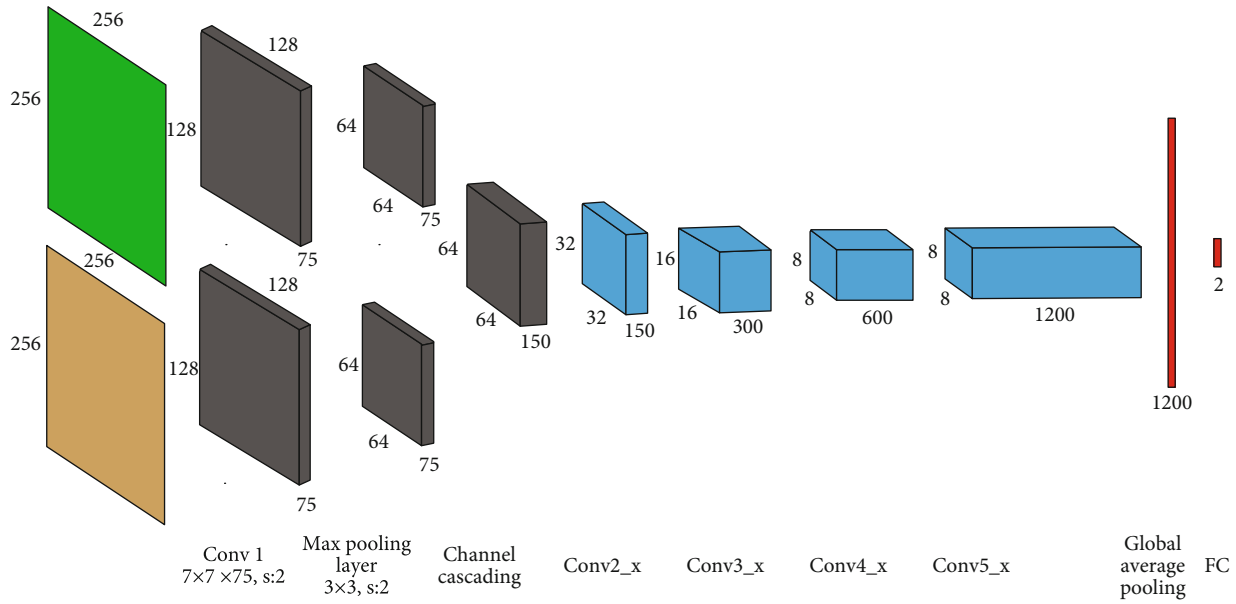


Figure 5: Hybrid model.

Table 4: Results of the three architectures. Both parallel and hybrid models have two versions according to whether the weights are shared (s) or unshared (u).

Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Training time (h)	Inference time (GPU/CPU, s)
Cascaded	92.02	94.08	89.71	91.83	7.05	0.0041/0.0461
Parallel (s)	84.06	88.67	78.15	83.06	17.85	0.0058/0.0792
Parallel (u)	83.94	86.38	80.60	83.38	19.45	0.0057/0.0798
Hybrid (s)	90.75	93.03	88.10	90.50	11.09	0.0039/0.0551
Hybrid (u)	90.90	93.84	87.55	90.58	10.73	0.0039/0.0565

Table 5: Results of cascaded models.

	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
32 layers (fat)	91.67	94.39	88.60	91.40
50 layers (ResNet)	90.82	92.95	88.33	90.58
32 layers (slim)	92.02	94.08	89.71	91.83
32 layers (thin)	90.43	92.47	88.03	90.20
23 layers (shallow)	91.16	93.98	88.00	90.88
101 layers (deepest)	90.57	93.56	87.13	90.22

Table 6: Result of different data preprocessing methods.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Slim	92.02	94.08	89.71	91.83
Slim_std	86.93	88.65	84.90	86.63
Slim_flip	93.07	95.42	90.45	92.89

To have a better understanding of multishape training, we implement the controlled experiments based on the slim model to show how it affects the network performance. Multishape training can improve the performance of our networks. Comparing the experiments 1 and 4 in Table 8 with the slim model in Table 5, it can be discovered that all scores of different metrics are increased. Similar to other data augment methods, multishape training can especially improve the performance for small datasets, the samples of which are not easy to obtain. We implement additional experiments with ideal shape inference on both binary and six-category classification datasets. We halve the data number of the binary classification dataset and carry out experiments without and with multiscale training successively. From Table 9, it can be seen that the performance is improved by a considerable margin. Besides, comparing Figures 7 and 6(a), the rate that the stain and luminance are wrongly predicted as each other goes down further. Comparing the recall of the six-category classification in Table 9 with that in Table 7, the same conclusion can be reached. Moreover, it is also revealed that multishape training is suitable for all categories according to the increase in all category recall rates.

More shapes do not indicate better performance. According to experiments 1, 3, 4, 5, and 6, although (148×442) and (442×148) shapes are included, the performance barely changed. Imagining that if we continue to add shapes such as (128×512), some images with an aspect ratio of 4 : 1 will be reshaped to the ratio 1 : 4. In this case, the change information may be damaged, thus rendering this sample useless

Table 7: Recalls of six-category classification experiments.

Method	Accuracy	Stain	Luminance	Rotation	Movement	Foreign body	Mark
Slim	86.05	84.00	82.29	83.43	86.00	87.14	93.43
Slim_std	85.52	79.71	72.86	84.86	94.00	84.29	97.43

Figure 6: Normalized confusion matrix of six categories: (a) slim method; (b) slim-std method.

Table 8: Effects of various design options on the slim model.

Options	1	2	3	4	5	6	7	8
Include (180,360), (360,180) shapes?	✓	✓	✓	✓	✓	✓	✓	✓
Include (148,442), (442,148) shapes?				✓	✓	✓		
1-shape inference?		✓						
Ideal-shape inference?	✓			✓			✓	
3-shape inference?			✓			✓		✓
5-shape inference?					✓			
Flipping?							✓	✓
Accuracy (%)	93.67	92.90	94.05	93.67	94.00	94.08	94.38	95.08
Precision (%)	94.44	93.16	95.41	95.17	95.18	95.27	94.71	95.73
Recall (%)	92.76	92.60	92.55	92.00	92.71	92.75	94.60	94.21
F1 (%)	93.59	92.88	93.96	93.56	93.92	94.00	94.35	94.96

Table 9: Additional experiments with ideal shape inference.

(a)

	Binary classification with a half dataset			
Method	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Slim	83.13	83.78	81.87	82.81
Slim with 3-shape training	88.86	89.37	88.17	88.76

(b)

	Six-category classification with 3-shape training						
	Accuracy	Stain	Luminance	Rotation	Movement	Foreign body	Mark
Recall (%)	89.78	87.43	83.62	85.98	92.19	91.76	97.71

for training. Therefore, it is mainly the rest shapes that contribute to the better performance of our networks.

Ideal shape inference can help improve the test score. In Table 8, we predict the change category using four strategies: only using shape (256×256) (1-shape inference); using the closest predefined shape adopted in training (ideal shape inference); using shapes (256×256), (180×360), and (360×180) (3-shape inference); and using 5-shape inference that has two additional shapes: (148×442) and (442×148). It is revealed by experiments 1 and 2 that converting the original image to the closest predefined shape is beneficial for the network to recognize changes.

Voting can give the prediction a boost. The idea is similar to multiview testing in SPP-net23. Instead of multiview images cropped from an original image, we feed multishape images reshaped from a test image to the network to predict its category. Finally, the final decision is made according to the majority. From experiments 1, 3, 4, and 5, it is obvious that voting can increase the test scores.

Currently, the best result is outputted by experiment 8 whose F1 score reaches 94.96. According to Tables 7–9, it is demonstrated that as the samples continue to be accumulated, the performance can be further improved.

4.6. Robustness. In order to verify the robustness of our model, we tested twelve pairs of images with luminance or rotation in our cascaded model. The results are shown in Figure 8.

As shown in Figure 8(a), there are 6 pairs of images with different light intensities. For example, in Row 1, the 3 pairs of images are affected by strong luminance, and almost more than half the area is covered by it. The 3 pairs of images with a little luminance in Row 2 are compared. We can see that the confidences of six-pair examples are slightly different with the highest score 99.99% and the lowest score 98.02%. They are all accurately predicted as luminance.

From human knowledge, rotation is easy to classify as movement that is an abnormal change needs to be detected. As well as luminance, in Figure 8(b), we selected six-pair rotation examples with different rotation angle. From the results, it can be discovered that the confidence of per image pairs is very close. It denotes that our cascaded model has strong robustness.

4.7. Analysis. It has been demonstrated that the features extracted by different layers are hierarchical [38]. For example, layer 1 may extract the fundamental features such as edges. Layer 2 responds to corners, and the deeper layers may capture similar textures and more class-specific variation. Usually, the function of the first few layers is uniform, so it is the common practice to freeze them when fine-tuning [9, 39]. To find out what our networks have learned,

Figure 7: Normalized confusion matrix of six-category classification with 3-shape training.

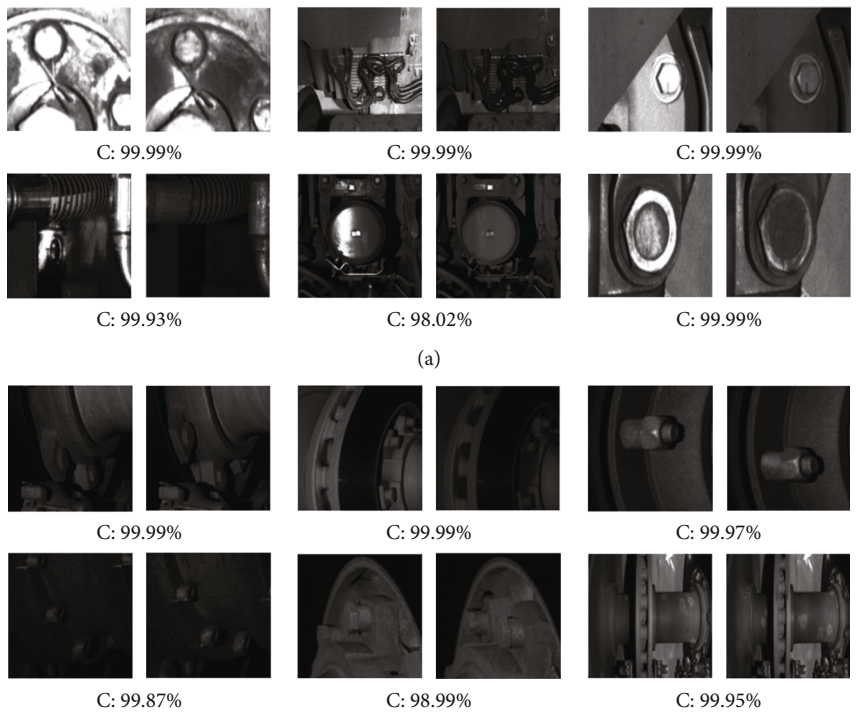


Figure 8: Twelve-pair examples for robustness verification: (a) luminance; (b) rotation. C: confidence.

we visualize all feature maps. Owing to the features extracted by the deep layers being too abstract to understand for us, we show four feature maps extracted by the first layers in Figure 9. We can find that our networks can not only extract the basic edge information but also learn to detect the changes (column 6). For example, in Figure 9(a) (R1, C6 and R3, C6), the component rotation is detected. As for (R2, C6) and (R4, C6), the change parts are segmented. In Figure 9(b) (C6), the position with stains and luminance changes is intensively responsive. Although we intuitively

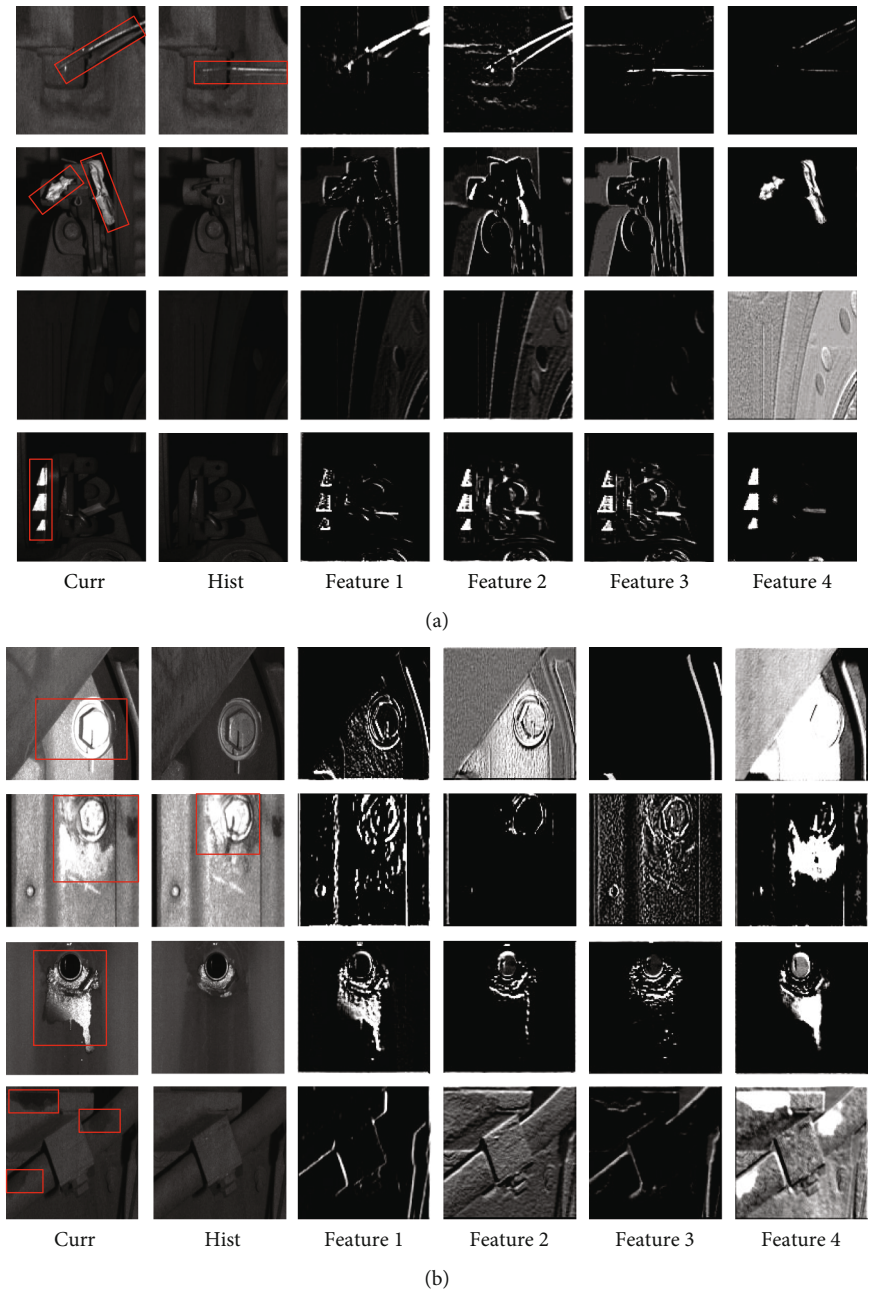


Figure 9: Feature maps extracted by the first convolutional layer for eight particular examples: (a) correct alarms; (b) false alarms. The red rectangles denote the difference of each image, such as, in (a) (R2, C1), there are forebodies; in (b) (R3, C1), the rectangle denotes the area with stain.

think that, compared with the early independent feature extraction methods, cascaded models may suffer from extracting the fundamental information of the two images, with difficulty, from Figure 9, we can see that the cascaded model can perform well in extracting the features of each image and in detecting the change information. Therefore, the cascaded model is superior.

The image content of trains is complex, which makes it unlikely to recite all situations for the networks. However, most of the abnormal targets appear in some fixed place such as the bolt, so it is reasonable to doubt whether our networks indeed learn how to recognize the changes. To verify that our networks do not memorize the components that usually go wrong but can identify the changes, we show three-pair examples that are the same components of the train in Figure 10. We can find that our networks are confident and can make the correct decision. In (R1, C2), even though there exists a luminance difference between the two bolts, the network remains capable to recognize that the bolt is loose. It is demonstrated that the networks have the sense of priority, namely, if the safety change and dangerous change simultaneously occur, the network will judge it as a correct alarm.

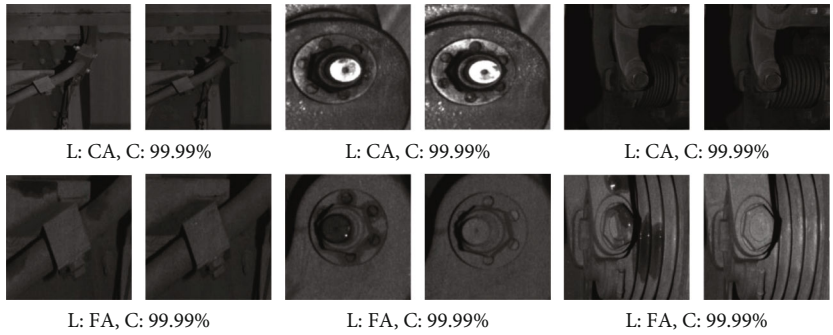
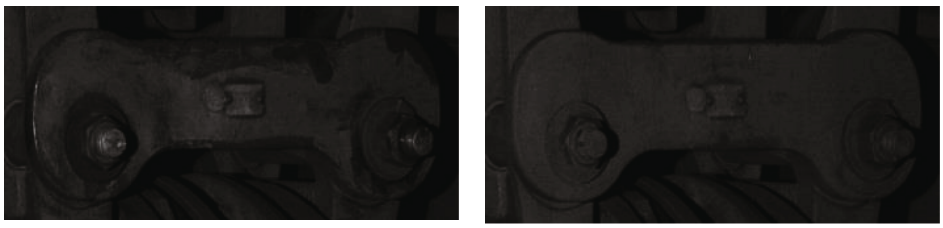


Figure 10: Three pairs of examples that are the same component of the train. L: label; CA: correct alarm; FA: false alarm; C: confidence.



(a) (b)

Figure 11: An example of false negative: (a) current state; (b) historical state.

However, if the dangerous change only occupies a very small part of the whole image while the safety change occupies the majority, such as what is shown in Figure 11, the network may not assess it as dangerous. In Figure 11, after glancing, we may treat it as a false alarm triggered by stains. However, if we look carefully, we will find one bolt loose. We will further study how to address this problem in the subsequent work.

5. Conclusions

In this paper, a cascaded model and a parallel one are presented to achieve change classification. According to the experimental results and network analysis, it is found out that the cascaded model is superior. Based on the cascaded model, extensive experiments are implemented to explore the optimal setting including the depth, width, and pretreatment methods. These experiments also demonstrate the differences among change classification task, traditional classification, and related works such as change detection. In addition, a novel training strategy is tailored to change classification, i.e., the multishape training method. It is experimentally validated that this strategy can improve the performance by a large margin and it is suitable for all categories.

Although we apply change classification to the task of high-speed train safety inspection, it is also suited to other classification scenarios where the decisions cannot be made with a single state. Change classification can also be considered as a solution to the tasks with rare positive samples. Our future direction is to explore how to address the false negative problem caused by structural changes occupying small areas in large images.

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 conflict of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61771409 and the Science and Technology Program of Sichuan under Grant No. 2019YJ0228.

References

[1] W. Song, X. Gao, J. Peng, J. Li, and L. Xie, "Abnormal target detection of high-speed train's roof," in *IEEE Far East Ndt New Technology & Application Forum*, C. Xu, Ed., pp. 143–148, IEEE, Xi'an, China, 2017.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[4] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE*

conference on computer vision and pattern recognition, Providence, RI, USA, 2012.

[5] X. Qiu, M. Li, L. Dong, G. Deng, and L. Zhang, "Dual-band maritime imagery ship classification based on multilayer convolutional feature fusion," *Journal of Sensors*, vol. 2020, 16 pages, 2020.

[6] D. G. Lee, Y. H. Shin, and D.-C. Lee, "Land cover classification using SegNet with slope, aspect, and multidirectional shaded relief images derived from digital surface model," *Journal of Sensors*, vol. 2020, 21 pages, 2020.

[7] B. Basnet, H. Chun, and J. Bang, "An intelligent fault detection model for fault detection in photovoltaic systems," *Journal of Sensors*, vol. 2020, 11 pages, 2020.

[8] J. Ker, S. P. Singh, Y. Bai, J. Rao, T. Lim, and L. Wang, "Image thresholding improves 3-dimensional convolutional neural network diagnosis of different acute brain hemorrhages on computed tomography scans," *Sensors*, vol. 19, no. 9, p. 2167, 2019.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2016.

[11] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," in *European conference on computer vision*, Cham, 2016.

[12] R. Li, H. Liang, Y. Shi, F. Feng, and X. Wang, "Dual-CNN: a convolutional language decoder for paragraph image captioning," *Neurocomputing*, vol. 396, pp. 92–101, 2020.

[13] X. Cui, D. Wang, and Z. Jane Wang, "Multi-scale interpretation model for convolutional neural networks: building trust based on hierarchical interpretation," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2263–2276, 2019.

[14] R. Li, F. Feng, I. Ahmad, and X. Wang, "Retrieving real world clothing images via multi-weight deep convolutional neural networks," *Cluster Computing*, vol. 22, no. S3, pp. 7123–7134, 2019.

[15] Y. Tang and W. Xiangqian, "Salient object detection using cascaded convolutional neural networks and adversarial learning," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2237–2247, 2019.

[16] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation using convolutional neural networks," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2115–2118, Valencia, Spain, 2018.

[17] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *IEEE International Conference on Image Processing*, Athens, Greece, 2018.

[18] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep Siamese convolutional network for optical aerial images," *IEEE Geoscience & Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, 2017.

[19] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4353–4361, Boston, MA, USA, 2015.

[20] S. Chopra, R. Hadsell, and Y. Lecun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE Computer Society Conference on Computer Vision & Pattern Recognition (CVPR)*, vol. 1, pp. 539–546, San Diego, CA, USA, 2005.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision (ECCV)*, Cham, 2016.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[25] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," in *International Conference on Machine Learning, Lille*, 2015.

[26] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," *NIPS*, 2015.

[27] K. He, R. Girshick, and P. Dollár, "Rethinking ImageNet pretraining," 2018, https://arxiv.org/abs/1811.08883/.

[28] W. Zhan, X. He, S. Xiong, C. Ren, and H. Chen, "Image deblocking via joint domain learning," *Journal of Electronic Imaging*, vol. 27, no. 3, 2018.

[29] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE Computer Society*, Boston, MA, USA, 2015.

[30] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, Lille, 2015.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2016.

[32] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, San Francisco, CA, USA, 2017.

[33] M. Lin, Q. Chen, and S. Yan, "Network in network," in *International Conference on Learning Representations (ICLR)*, Banff, 2014.

[34] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pp. 580–587, Columbus, OH, USA, 2014.

[35] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, 2014.

[36] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, Italy, 2010.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[38] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, Cham, 2014.

[39] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, Santiago, 2015.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5511745, 13 pages
https://doi.org/10.1155/2021/5511745

Research Article

Sensor Duty Cycle for Prolonging Network Lifetime Using Quantum Clone Grey Wolf Optimization Algorithm in Industrial Wireless Sensor Networks

Yang Liu, Jing Xiao, Chaoqun Li, Hu Qin, and Jie Zhou

College of Information Science and Technology, Shihezi University, Shihezi 832000, China

Correspondence should be addressed to Jie Zhou; jiezhou@shzu.edu.cn

Received 3 February 2021; Revised 19 February 2021; Accepted 11 March 2021; Published 27 March 2021

Academic Editor: Bin Gao

The application of industrial wireless sensor networks (IWSNs) frequently appears in modern industry, and it is usually to deploy a large quantity of sensor nodes in the monitoring area. This way of deployment improves the robustness of the IWSNs but introduces many redundant nodes, thereby increasing unnecessary overhead. The purpose of this paper is to increase the lifetime of IWSNs without changing the physical facilities and ensuring the coverage of sensors as much as possible. Therefore, we propose a quantum clone grey wolf optimization (QCGWO) algorithm, design a sensor duty cycle model (SDCM) based on real factory conditions, and use the QCGWO to optimize the SDCM. Specifically, QCGWO combines the concept of quantum computing and the clone operation for avoiding the algorithm from falling into a local optimum. Subsequently, we compare the proposed algorithm with the genetic algorithm (GA) and simulated annealing (SA) algorithm. The experimental results suggest that the lifetime of the IWSNs based on QCGWO is longer than that of GA and SA, and the convergence speed of QCGWO is also faster than that of GA and SA. In comparison with the traditional IWSN working mode, our model and algorithm can effectively prolong the lifetime of IWSNs, thus greatly reducing the maintenance cost without replacing sensor nodes in actual industrial production.

1. Introduction

As industrial wireless sensor networks (IWSNs) have more and more applications in the factories, the way to prolong the lifetime of IWSNs without changing the physical facilities has become a hot issue [1, 2]. The content of this paper includes the design of a sensor duty cycle model (SDCM) for IWSN modeling and a novel group intelligence algorithm based on grey wolf optimization for optimizing the SDCM. By using the SDCM, we can conveniently increase the lifetime of the IWSNs in the factory, thus avoiding the very time-consuming, labor-intensive, and sometimes impossible operations in real life [3]. In addition, the use of an artificial intelligence algorithm to optimize the established model can effectively prolong the lifetime of the IWSNs, thereby reducing the maintenance cost of the IWSNs and increasing the benefit of the factory [4, 5]. Furthermore, making full use of existing devices can reduce the generation of discarded equipment and protect the environment on the basis of reducing resources.

In this research, we investigate the IWSNs frequently used in factories, such as chemical sensors that monitor the content of harmful gases, pressure sensors in industrial production, and ultrasonic sensors in the field of industrial automation. We find that these IWSNs are basically placed by using the traditional wide-spreading method and then periodically control some sensors to enter the sleep state for saving energy [6]. This approach has two disadvantages, one is that it cannot meet the requirements of full coverage, another is it cannot minimize the energy consumption of the sensors.

The goals of this article are to increase the lifetime of IWSNs and reduce the cost of factory replacement of sensor devices. The above goals are motivated by the actual needs in the factory [7, 8]. In general, the purpose of using wireless sensors in factories is to monitor the production

environment for ensuring safe production conditions. However, to achieve full coverage of the monitoring in the production workshop, the factory has to place redundant sensors to ensure the performance of the IWSNs, which causes a lot of waste of sensor energy and speeds up the replacement of sensors [9]. In this case, it is necessary to propose a method that can effectively utilize redundant sensors for reducing the number of sensor replacements.

In the issue of improving the lifetime of IWSNs, it is necessary to ensure high coverage of targets first and then perform a sensor node duty cycle [10–12]. When establishing the SDCM, we comprehensively consider the sensor's monitoring range and working time, then give a mathematically measurable lifetime of the IWSNs. Therefore, we can use an artificial intelligence algorithm to optimize the lifetime of IWSNs through the SDCM. The designed model has been verified by a series of simulation experiments. For a given set of IWSN data, it can be input into the SDCM by the sensor's ability of working time and coverage, then use our proposed algorithm to optimize the SDCM for a longer lifetime of the IWSNs.

The innovations of this research are as follows: (i) The sleep mode of industrial wireless sensor nodes is modeled, and the sensor duty cycle model is proposed. Through the SDCM, the lifetime of IWSNs can be prolonged by using an intelligent algorithm, thereby effectively reducing the factory's maintenance costs for IWSNs. (ii) A novel GWO-based intelligent algorithm is proposed, which uses the quantum probability amplitude in quantum computing and the clone concept in biology to avoid falling into local optimum, thereby increasing the usability of the algorithm. In addition, the performance of the proposed algorithm has been compared with GA and SA.

The paper's structure can be expressed as follows. Relative researches on the duty cycle of sensors are given in Section 2. Subsequently, Section 3 shows the evaluation method of IWSN lifetime and the establishment of the SDCM. In Section 4, in order to obtain the optimal IWSN lifetime, we introduce a novel group intelligent algorithm based on grey wolf optimization. Section 5 presents the performance of the proposed model and algorithm through simulation experiments and makes discussion. Finally, in Section 6, the conclusion part is given.

2. Related Work

The current research on the duty cycle of IWSNs can be divided into three types. The first and most used one is to design a routing protocol for reducing the unnecessary communication overhead of sensors; the second type uses artificial intelligence methods to process sensor data for obtaining a suitable mode of duty cycle; the third type applies mathematical approaches to model IWSNs, then optimizes the duty cycle of sensor nodes.

Firstly, a proper routing protocol can reduce communications of sensors in IWSNs. In [13], the authors use a threefold method to improve the lifetime of the IWSNs by adjusting the duty cycle process of sensor nodes, then switch between the active mode and the sleep mode according to the trust value obtained by the nodes. On the other hand, to better improve the service quality of IWSNs, the paper [14] proposes an AODV routing protocol for surplus energy, which realizes the reduction of energy consumption of IWSNs through cross-layer design. Similar to the paper [18], another cross-layer routing method is also proposed. In [15], the authors adjust the wake-up and sleep of nodes in the forwarding stage through the cooperation of routing and MAC layer. Then, the paper [16] proposes a routing protocol for anycast. Each sensor node decides how to transmit data based on its local information and dynamically changes the node's duty cycle status. What is more, for the purpose of solving the problem of excessive energy consumption of the nodes around the sink node in the IWSNs, the paper [17] proposes a method based on the path optimization of the sink node and establishes a corresponding energy consumption model. From other perspectives, the paper [18] uses multihop communication to reduce the long-distance communication overhead of nodes, proposes a routing protocol for clustering nodes, and uses a multihop simulated annealing algorithm to select intermediate nodes. In [19], the authors propose a perceptual routing protocol including network scheduling and task cycle, which helps sensor nodes to continuously monitor. However, the above methods for improving the duty cycle model by using routing protocol do not consider the way of placement for wireless sensors in the real environment. They only reduce the energy consumption of each sensor node but ignore the premise that there is a large quantity of redundant nodes in the IWSNs.

Secondly, artificial intelligence technology can also effectively improve the duty cycle of sensor nodes. In [20], the authors use reinforcement learning to maximize the sensing quality of the sensor nodes, then perform duty cycle based on the available energy, and use the collected energy to make the nodes continuously adapt to the changing environment. With the same purpose of prolonging lifetime of the IWSNs, the paper [21] expresses the position distribution of sensors as an optimization problem and then proposes a cuckoo algorithm to solve the problem, thereby obtaining the optimal position of sensor nodes. In [22], the authors use Q-learning technology and linear regression function to design a MAC protocol. The protocol takes the relationship between load conditions and performance into account and makes up for the disadvantage of Q-learning, and it can do low-latency sensor scheduling. Although these methods can improve the lifetime of the IWSNs, they do not consider the convergence speed of the algorithm, and they also fail to make good use of redundant nodes in the IWSNs.

Finally, there are some researches to optimize sensor duty cycle from other aspects. In [23], the authors consider the asymmetry of the asynchronous duty cycle, obtain the upper and lower limits of the node's duty cycle, and use block design to establish the duty cycle model. The paper [21] uses empirical data to find the noise relationship between the maximum discovery time and the duty cycle by analyzing the error of the proposed model,

and it concludes that the same duty cycle value in an asymmetric scenario can achieve low latency. However, it only considers how to use a node duty cycle to obtain low-latency information transmission and does not pay attention to prolonging the lifetime of the IWSNs.

The previous researches only solve the problem of reducing the communication overhead of each sensor node, thereby achieving the purpose of increasing the lifetime of the IWSNs. They do not make good use of the large quantity of redundant nodes and do not consider solving the problem from the entire network. We can go a step further on the basis of the previous work. On the premise of meeting the requirements of industrial production, we use duty cycle to shut down redundant nodes and reduce the communication overhead of each node, hence the maximum of the lifetime in the IWSNs.

In this paper, we start from the entire wireless sensor network and model the real factory IWSNs. Particularly, our model not only considers the coverage capability of the sensor nodes but also makes the lifetime of the sensor nodes measurable, which is more convenient for subsequent optimization. To reduce sensor communications, we design a novel heuristic optimization algorithm, which can effectively improve the convergence performance and avoid falling into the local optimum, so that the lifetime of the IWSNs can be prolonged.

3. System Model

3.1. Problem Description. In the real factory scene, the target coverage of IWSNs can be divided into two types. One is full coverage, which means every target being covered and monitored by at least one sensor node at every moment. The other is to improve coverage rate, which is often used in environments where full coverage is impossible. In most cases, the placement of factory sensors is redundant, which aims to decrease the occurrence of accidents. Redundant placement often meets the requirement of full coverage. Under the premise of redundant nodes, we can perform duty cycle operation on sensor nodes for saving energy, thereby prolonging the lifetime of the IWSNs.

In this paper, with the aim of facilitating the modeling of the sensor node duty cycle problem from a mathematical perspective, we propose a concept of measurable sensor lifetime, which can be explained as follows: general industrial sensors have their service lifetime, different types of sensors have different values, and even the same type of sensors will have different lifetime values. However, to make the sensors produced by factories more competitive in the market, manufacturers often give the theoretical lifetime of the sensors. Working under more severe conditions than usual, the theoretical lifetime can be obtained by converting the working lifetime with a certain calculation formula. Therefore, with the theoretical lifetime of the sensors, we can divide the lifetime and mathematically model it through the operation of the duty cycle.

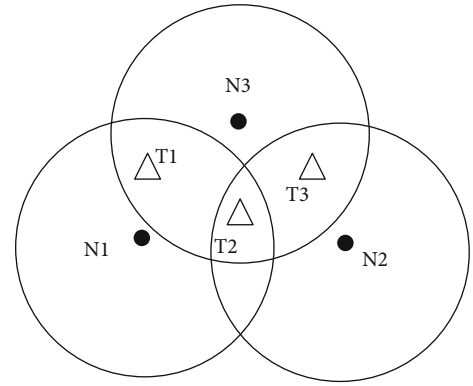


Figure 1: Coverage relationship between sensors and targets.

To better understand the sensor duty cycle model (SDCM), we use an example to illustrate it. In Figure 1, three targets are monitored by three sensor nodes with circular monitoring radii, and the monitor requirement is full coverage. The sensor nodes **N**1, **N**2, and **N**3 are, respectively, represented by points located at the center of the circle, the monitored targets **T**1, **T**2, and **T**3 are symbolized by triangles, and the circle stands for the monitoring radius of the sensor node.

According to Figure 1, we can know that **N**1 covers **T**1 and **T**2, **N**2 covers **T**2 and **T**3, and **N**3 covers **T**1, **T**2, and **T**3. Assume that the lifetime of each sensor node can be divided into two rounds. If the duty cycle operation is not performed, the total working time of the entire IWSNs is two rounds, which are as follows: the first round, we turn on **N**1, **N**2, and **N**3, and the second round, we also make **N**1, **N**2, and **N**3 in active mode. However, under the premise of ensuring full coverage, if we divide the sensors into different coverage sets and only turn on one coverage set in each round, we can prolong the lifetime of the IWSNs through duty cycle operation. Specifically, in the above case, **N**1 and **N**2 can form a coverage set, and **N**3 can be another one. In the first and second rounds, we can only switch on **N**1 and **N**2 and turn off **N**3 for saving energy. At the end of the second round, the energy of **N**1 and **N**2 is exhausted; in the third and fourth rounds, we turn on the **N**3 node. At the end of the fourth round, the energy of the three sensor nodes has been exhausted. Obviously, the lifetime of the IWSNs has increased from the previous two rounds to four rounds through the duty cycle operation while ensuring the full coverage of the targets.

In modern IWSNs, with the increase of redundant nodes and the improvement of sensor coverage, the duty cycle operation of sensor nodes plays a more and more important role. Subsequently, we establish a mathematical model for using an artificial intelligence algorithm to optimize the SDCM.

3.2. Sensor Duty Cycle Model. Suppose there are **X** sensor nodes and **K** monitoring targets in the IWSNs. In the SDCM, in order to ensure that the sensors complete full coverage of the targets, we first create a matrix **S** for expressing the

coverage relationship between sensors and targets, which can be shown as

$$S = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,K-1} & s_{1,K} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,K-1} & s_{2,K} \\ \vdots & & s_{x,k} & & \vdots \\ s_{X-1,1} & s_{X-1,2} & \cdots & s_{X-1,K-1} & s_{X-1,K} \\ s_{X,1} & s_{X,2} & \cdots & s_{X,K-1} & s_{X,K} \end{bmatrix} (s_{x,k} \in \{0, 1\}), \tag{1}$$

where $s_{x,k}$ represents the monitoring relationship between the x_{th} sensor and the k_{th} target. $s_{x,k} = 0$ means that the sensor cannot detect the target node, and $s_{x,k} = 1$ means that the target node is within the monitoring range of the sensor.

However, a coverage relationship matrix is not enough for establishing the SDCM, it is also necessary to create a duty cycle sequence matrix. Combining the previously described concept of measurable lifetime, we divide the theoretical lifetime of each sensor into **N** rounds. It is not difficult to conclude that the maximum life of the entire wireless sensor network is **XN** rounds, where **X** is the number of sensors. The duty cycle sequence matrix can be expressed as

$$T = \begin{bmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,X-1} & t_{1,X} \\ t_{2,1} & t_{2,2} & \cdots & t_{2,X-1} & t_{2,X} \\ \vdots & & t_{i,x} & & \vdots \\ t_{XN-1,1} & t_{XN-1,2} & \cdots & t_{XN-1,X-1} & t_{XN-1,X} \\ t_{XN,1} & t_{XN,2} & \cdots & t_{XN,X-1} & t_{XN,X} \end{bmatrix} (t_{i,x} \in \{0, 1\}), \tag{2}$$

where $t_{i,x} = 1$ indicates that, in the i_{th} round, the x_{th} sensor node is in the active state, and $t_{i,x} = 0$ denotes that the x_{th} sensor node is in the sleep state in the i_{th} round.

To obtain the monitoring matrix between the x_{th} sensor node and the k_{th} monitored target in each round, we need to multiply the matrix **T** and the matrix **C**. The reason for this approach is that **T** is a duty cycle sequence matrix with **XN** rows and **X** columns, and **S** is a sensor coverage matrix with **X** rows and **K** columns. To obtain the monitoring relationship matrix **TS** between the sensor node and the monitored target, it is necessary to multiply the **T** matrix by the **S** matrix to obtain a matrix of **XN** rows and **K** columns. The rows of the **TS** represent the monitoring relationship of the sensor to the target in the round. If the elements in the i_{th} row are all 1, $i \in (1, XN)$, it means that the sensor network has completed full coverage in the i_{th} round and reached the set goal. If there is 0 in the i_{th} row, the task fails in the i_{th} round. The monitoring matrix can be shown as

$$TS = \begin{bmatrix} \sum_{x=1}^{X} t_{1,x}s_{x,1} & \sum_{x=1}^{X} t_{1,x}s_{x,2} & \cdots & \sum_{x=1}^{X} t_{1,x}s_{x,K-1} & \sum_{x=1}^{X} t_{1,x}s_{x,K} \\ \sum_{x=1}^{X} t_{2,x}s_{x,1} & \sum_{x=1}^{X} t_{2,x}s_{x,2} & \cdots & \sum_{x=1}^{X} t_{2,x}s_{x,K-1} & \sum_{x=1}^{X} t_{2,x}s_{x,K} \\ \vdots & & \sum_{x=1}^{X} t_{i,x}s_{x,k} & & \vdots \\ \sum_{x=1}^{X} t_{XN-1,x}s_{x,1} & \sum_{x=1}^{X} t_{XN-1,x}s_{x,2} & \cdots & \sum_{x=1}^{X} t_{XN-1,x}s_{x,K-1} & \sum_{x=1}^{X} t_{XN-1,x}s_{x,K} \\ \sum_{x=1}^{X} t_{XN,x}s_{x,1} & \sum_{x=1}^{X} t_{XN,x}s_{x,2} & \cdots & \sum_{x=1}^{X} t_{XN,x}s_{x,K-1} & \sum_{x=1}^{X} t_{XN,x}s_{x,K} \end{bmatrix}. \tag{3}$$

In equation (3), $\sum_{x=1}^{X} t_{i,x}s_{x,k} = 0$ means that the k_{th} monitored target in the i_{th} round is not monitored by any sensor, and $\sum_{x=1}^{X} t_{i,x}s_{x,k} = p(p > 0)$ represents that the k_{th} monitored target in the i_{th} round is monitored by **p** sensors. The requirement of full coverage shown in the matrix **TS** is that the elements in a row are all positive numbers.

To calculate the lifetime of the IWSNs, we define the fitness function **first¯zero** to represent the number of rows where the first element 0 appears in the matrix **TS** and define the restriction condition. Therefore, the SDCM can be expressed as a fitness function (4) and restriction condition (5).

$$f(T) = \text{first\_zero}(TS) - 1, \tag{4}$$

$$\sum_{x=1}^{XN} t_{i,x} \le N, \quad x = 1 \cdots X, \tag{5}$$

where **T** represents the duty cycle sequence matrix, and **N** denotes the lifetime cycle number of each sensor. The reason of the fitness function (4) is because under the requirement of sensor node full coverage, only the elements in the row of matrix **TS** are all 1, which means that the task is completed. Since the coverage cannot appear gaps, the lifetime value of IWSNs is the number of rows where the first zero element appears minus 1. Equation (5) shows the working lifetime of each sensor does not exceed **N** rounds.

Subsequently, due to the complexity of the sensor duty cycle increases exponentially with the number of sensor nodes and working lifetime, we decide to use a novel heuristic algorithm to solve this problem.

4. QCGWO-Based Duty Cycle in IWSNs to Maximize Network Lifetime

In IWSNs, obtaining the longest network lifetime of sensors in the duty cycle problem is obviously an NP-difficult problem. For the purpose of gaining the optimal solution of the sensor duty cycle, we design a heuristic optimization algorithm based on the gray wolf optimization (GWO) algorithm. The GWO is a group intelligent optimization algorithm proposed in 2014, and it has the characteristics of simple operation and few parameters. However, the traditional GWO is prone to falling into the local optimum, and its global search ability is difficult to control. To overcome these disadvantages, we combine the concept of quantum computing and clone with GWO, then propose a quantum clone gray wolf optimization (QCGWO) algorithm.

The description of the process for QCGWO are problem coding, population initialization, fitness calculation and class division, update wolf population and algorithm parameters, quantum probability amplitude and quantum revolving gate, clone expansion, and termination operation.

4.1. Problem Coding. In the sensor duty cycle problem, the key is to control the sensor to turn on at an appropriate time, so that redundant nodes can be fully utilized and the energy consumption of the sensor network can be reduced. Assuming that different sensors in IWSNs can be divided into the same number of lifetime rounds, since the opening and the closing are a group of Boolean values, we decide to use binary coding in the duty cycle problem. Zero means that the sensor is turned off in this round; otherwise, 1 means that the sensor is turned on in this round. There are two core matrices in the above-mentioned sensor duty cycle problem, one is the sensor coverage matrix **S**, and the other is the sensor duty cycle sequence matrix **T**. The coverage matrix **S** is generated only once in the algorithm, and subsequent duty cycle operations are based on it. Each wolf in QCGWO carries a sequence matrix **T**, and the matrix **T** is optimized through the proposed algorithm until the algorithm ends and an optimal solution is obtained. To enhance the understanding of the sequence matrix **T**, we gave an example for illustrating. Suppose that the working lifetime of each sensor is 2 rounds and there are 2 sensors in IWSNs, then the rows of the matrix **T** are $2 \times 2 = 4$, and

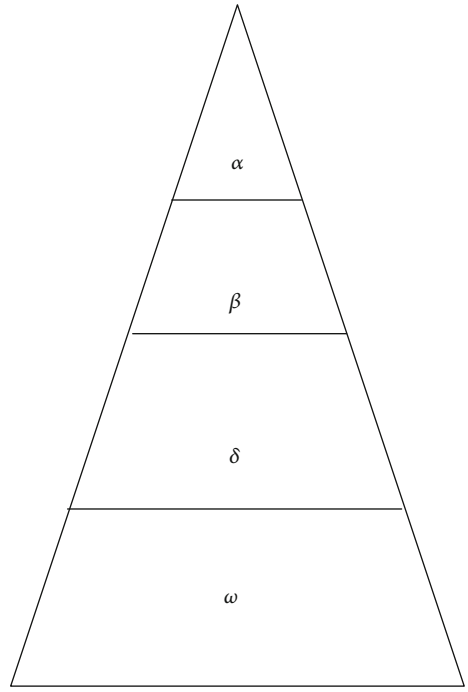


Figure 2: Diagram of wolf population level.

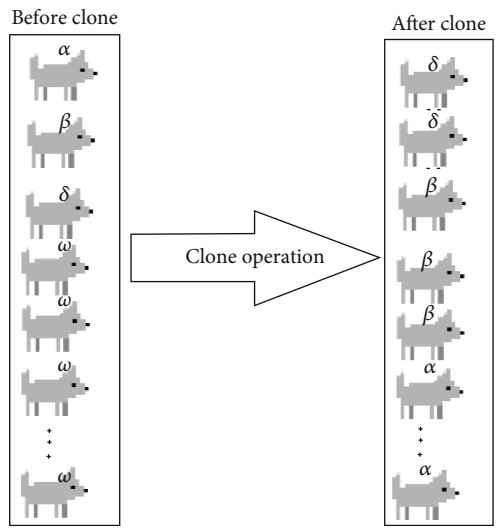


Figure 3: Clone operation.

the columns are 2. The encoding of matrix **T** can be expressed as

$$T = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{6}$$

In (6), the sum of 1 in each column is 2, which means that the lifetime of the sensor is 2 rounds. The first column indicates that the first sensor turns on in the second and third rounds, and the second column represents that the second sensor turns on in the first and second rounds.

4.2. Population Initialization. In the QCGWA, every wolf is a potential optimal solution. Due to each wolf carries a two-dimensional matrix **T**, the population adopts a three-dimensional encoding method. First, it is necessary to calculate the distance between the sensor node and the monitored target; then, the sensor coverage matrix **S** is generated according to the monitoring range of the sensor. In addition, because of the particularity of the sensor duty cycle problem, we need to ensure that each sensor is turned on to its maximum number of lifetime rounds; hence, the initial wolf population is required to meet certain restrictions. Supposing that there are **X** sensor nodes in IWSNs and the maximum lifetime of each sensor node is **N** rounds, then the initialized individual in the population can be represented by

$$\text{individual} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,X-1} & p_{1,X} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,X-1} & p_{2,X} \\ \vdots & & p_{i,x} & & \vdots \\ p_{XN-1,1} & p_{XN-1,2} & \cdots & p_{XN-1,X-1} & p_{XN-1,X} \\ p_{XN,1} & p_{XN,2} & \cdots & p_{XN,X-1} & p_{XN,X} \end{bmatrix} \left(p_{i,x} \in \{0, 1\}\right), \tag{7}$$

$$\sum_{i=1}^{XN} p_{i,m} = N, \quad m \in \{1, 2, \cdots, X\}. \tag{8}$$

In (7) and (8), **p** is a binary bit of the individual, and it indicates the on state of the sensor, 0 means the sensor is off, and 1 means the sensor is on.

4.3. Fitness Calculation and Class Division. The proposal of QCGWO is based on the characteristics of gray wolves in nature. Gray wolves are a group of animals, and they have a strict social hierarchy in the population, as shown in Figure 2. There are differences in quantity and work duty of wolves in different classes. Specifically, the wolf with the highest rank is called the Alpha wolf or the dominant wolf. What is more, the Alpha wolf is not necessarily the strongest wolf in the population, but the optimum at managing the group, and its role is to make decisions about group activities. The wolves in the second class are Beta wolves, which help Alpha wolves make decisions and are potential candidates for Alpha wolves. The wolves in the third class are called Delta wolves, and they usually obey the command of Alpha wolves and Beta wolves. At last, the Omega wolves are in the lowest level; they have the lowest status but help maintain the overall combat ability of the wolves.

In QCGWO, it is necessary to calculate the fitness of each individual in the wolf group for obtaining different levels of wolves. After initializing the population, we can calculate the fitness of each individual according to equation (4), and then, the individual with the highest fitness is divided into Alpha wolves, and the remaining individuals are divided into Beta wolves, Delta wolves, and Omega wolves according to their fitness.

4.4. Update Wolf Population and Algorithm Parameters. In the update mechanism of QCGWO, the process of encircling prey by wolves is imitated; specifically, the QCGWO considers the location of the wolves and the location of the prey. Each update operation of the population is carried out according to the position of the prey, so that the search for the solution space of the problem is realized. In the sensor node duty cycle problem, the prey refers to the individual with the highest fitness. Subsequently, the update process can be represented by

$$L = \left| C * O_{\text{prey}}(\text{gen}) - O(\text{gen}) \right|, \tag{9}$$

$$O(\text{gen} + 1) = O_{\text{prey}}(\text{gen}) - A * L. \tag{10}$$

In (9) and (10), **O** and O_{prey} represent the current positions of the wolf and the prey, respectively, **gen** is the current iteration number, and **L** stands for the distance between the wolf and the prey. **A** and **C** are two vector coefficients, they can be expressed as

$$\begin{aligned} A &= 2a * \text{rand}_1 - a, \\ C &= 2 * \text{rand}_2, \end{aligned} \tag{11}$$

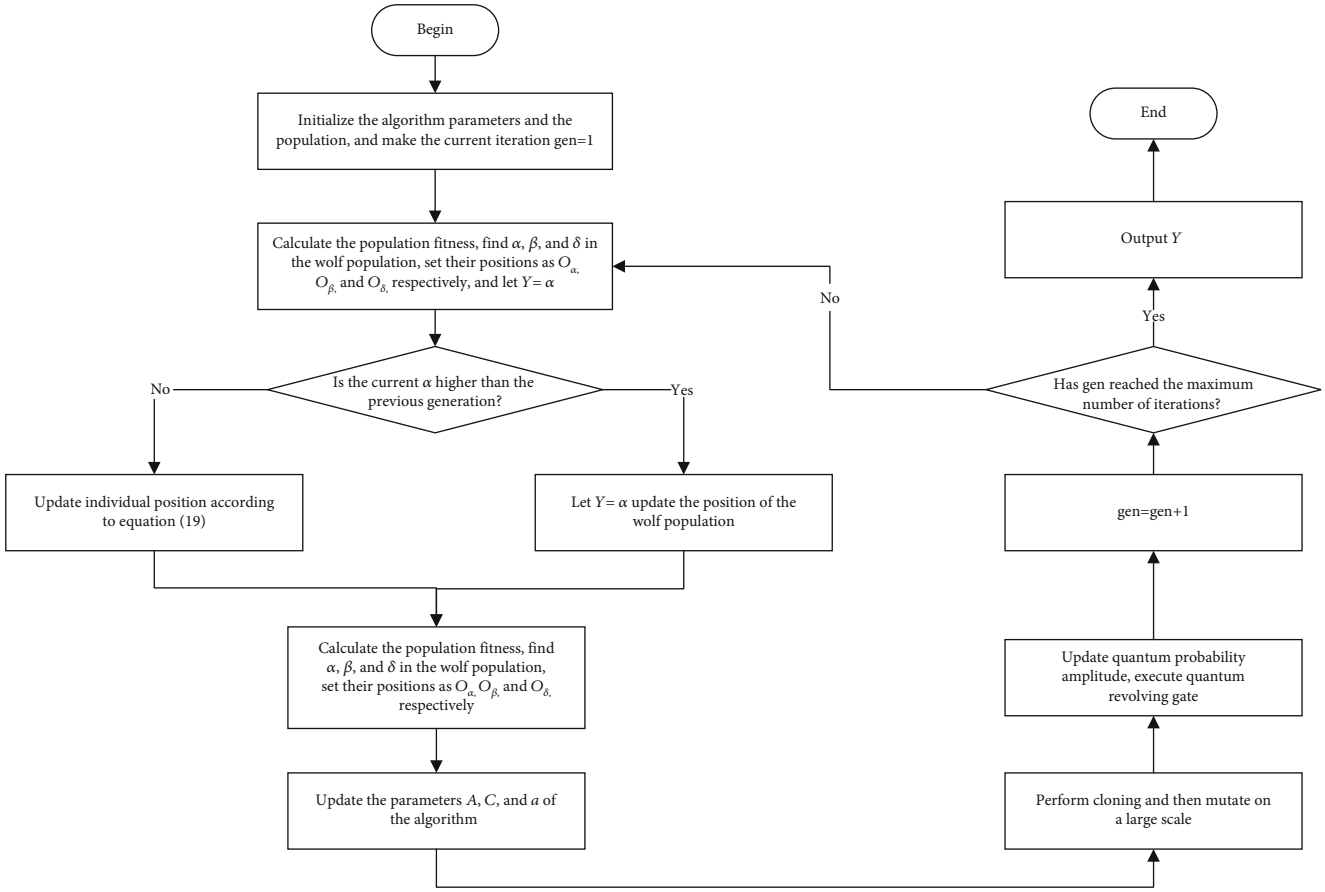


Figure 4: Steps of QCGWO.

Table 1: The experimental conditions in Figure 5.

	Sensors	Targets	Maximum lifetime	Monitoring radius (m)
Figure 5(a)	40	12	10	150
Figure 5(b)	50	15	10	150
Figure 5(c)	60	20	10	150
Figure 5(d)	100	30	10	150

Table 2: The experimental conditions in Figure 6.

	Sensors	Targets	Maximum lifetime	Monitoring radius (m)
Figure 6(a)	30	10	15	180
Figure 6(b)	45	15	15	180
Figure 6(c)	60	20	15	180
Figure 6(d)	75	25	15	180

where **a** is the convergence factor that decreases with the number of iterations from 2 to 0, and rand_1 and rand_2 are random numbers in [0,1].

In QCGWO, the approximate position of the prey is in the middle of Alpha, Beta, and Delta wolves. Subsequently, all wolves in the population surround the estimated position. The movement process of wolves can be expressed as

$$L_{\alpha} = |C_1 * O_{\alpha} - O|, \tag{12}$$

$$L_{\beta} = |C_2 * O_{\beta} - O|, \tag{13}$$

$$L_{\delta} = |C_3 * O_{\delta} - O|, \tag{14}$$

$$O_1 = O_{\alpha} - A_1 * L_{\alpha}, \tag{15}$$

Table 3: The experimental conditions in Figure 7.

	Sensors	Targets	Maximum lifetime	Monitoring radius (m)
Figure 7(a)	30	10	20	130
Figure 7(b)	50	15	20	130
Figure 7(c)	70	25	25	130
Figure 7(d)	100	30	25	130

$$O_2 = O_{\beta} - A_2 * L_{\beta}, \tag{16}$$

$$O_3 = O_{\delta} - A_3 * L_{\delta}. \tag{17}$$

In equations (12)–(17), O_1, O_2, and O_3 represent the positions of Alpha, Beta, and Delta wolves, respectively. L_{α},

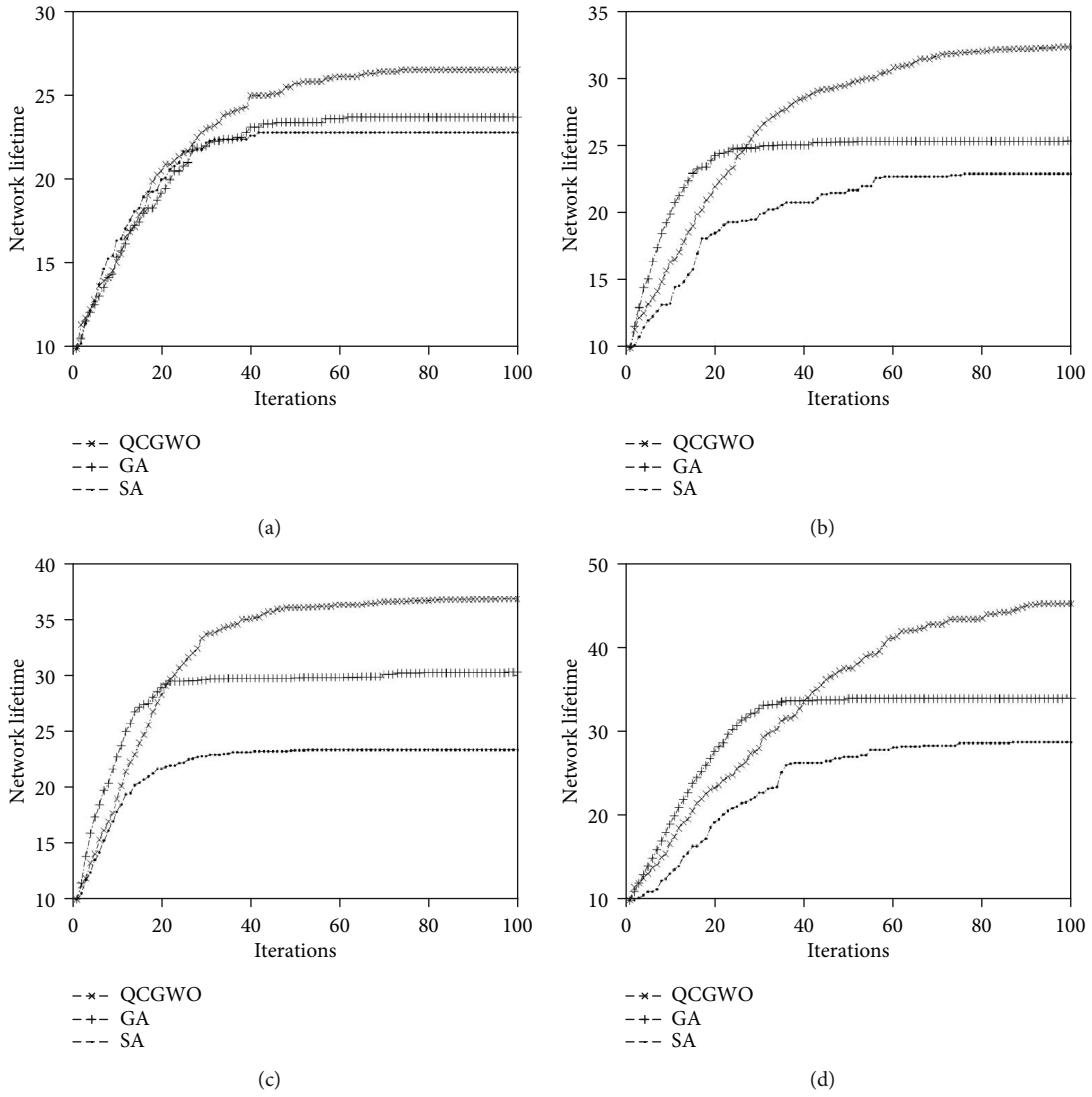


Figure 5: Network lifetime in comparison of the three algorithms: (a) 40 sensors and 12 targets; (b) 50 sensors and 15 targets; (c) 60 sensors and 20 targets; (d) 100 sensors and 30 targets.

L_β, and L_δ denote the distance from the current individual to Alpha, Beta, and Delta wolves, respectively. What is more, the process of wolves rounding up prey is shown in

$$O_{\text{gen}+1} = \frac{[O_1(\text{gen}) + O_2(\text{gen}) + O_3(\text{gen})]}{3}. \quad (18)$$

In (18), **gen** represents the current number of iterations.

4.5. Quantum Probability Amplitude and Quantum Revolving Gate. QCGWO uses the natural parallel mechanism of quantum computing to update the population. Compared with the traditional gray wolf optimization algorithm, the addition of quantum probability amplitude greatly enhances the global parallel search capability of QCGWO, which is a big difference between QCGWO and traditional GWO. By combining qubits and quantum superposition states, the convergence speed can be effectively improved when solving the problem of a large-scale sensor duty cycle. Furthermore, QCGWO represents each wolf in the population with a set of binary quantum probability amplitude bits. In the initial wolf population, all qubit probability amplitudes on each wolf are produced by logistic chaotic mapping, which is shown in

$$q_{z+1} = bq_z(1 - q_z), \quad z = 0, 1, \cdots, X. \quad (19)$$

In (19), $b = 4$ indicates that the mapping is in a chaotic state, **q** represents the generated quantum probability amplitude, and **X** is the number of sensors.

After all the wolves in the population are updated, in order to enhance the population diversity, the quantum revolving gate needs to be updated according to the qubits

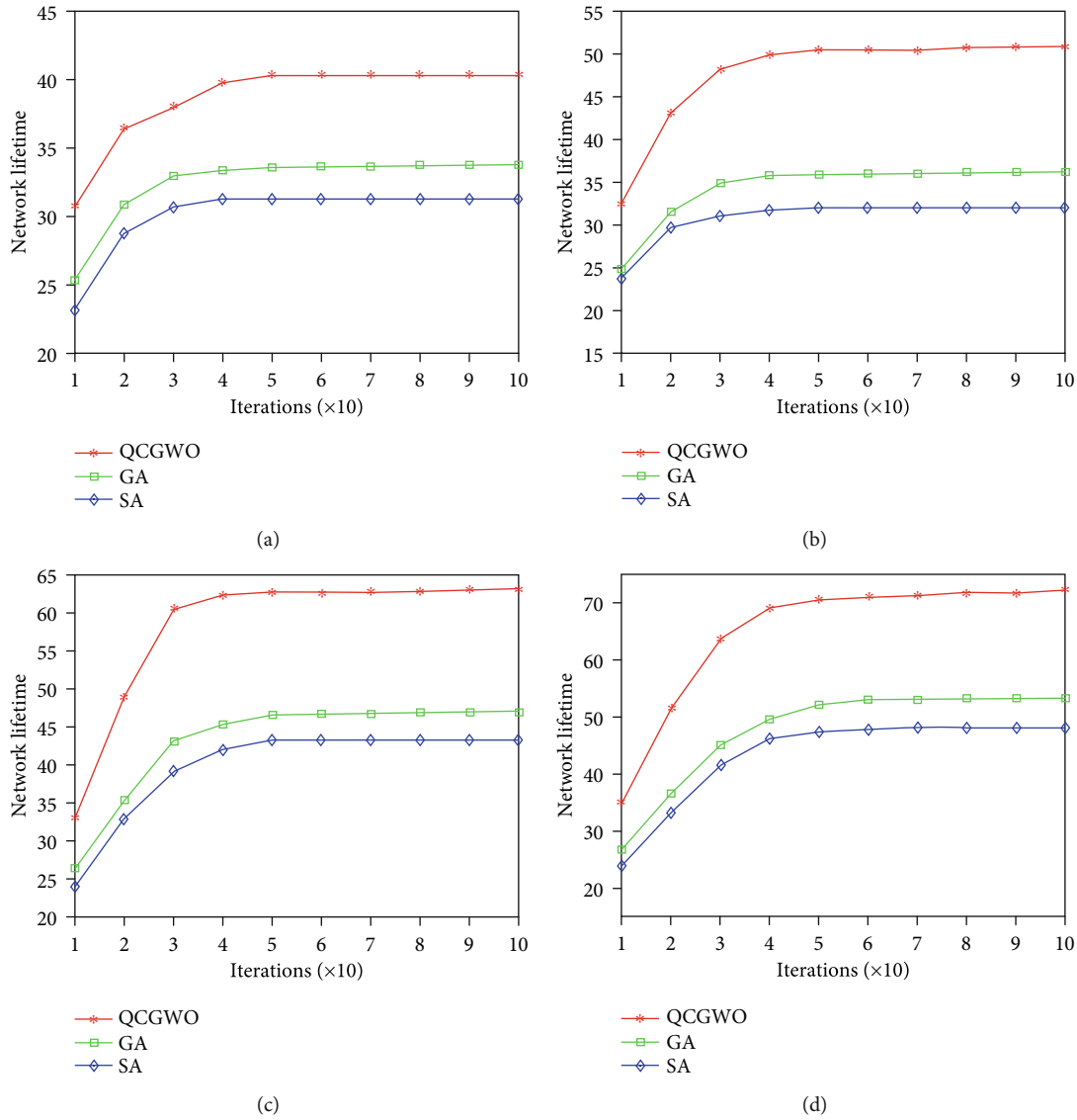


Figure 6: Network lifetime in comparison of the three algorithms: (a) 30 sensors and 10 targets; (b) 45 sensors and 15 targets; (c) 60 sensors and 20 targets; (d) 75 sensors and 25 targets.

on the Alpha wolves in the current population. The update process can be shown as

$$\begin{bmatrix} y_1^{\text{after}} \\ y_2^{\text{after}} \end{bmatrix} = \begin{bmatrix} \cos\varepsilon & -\sin\varepsilon \\ \sin\varepsilon & \cos\varepsilon \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \tag{20}$$

In (20), y_1 and y_2 represent the quantum probability amplitude before revolving, and ε is the angle of quantum revolving.

4.6. Clonal Expansion. The purpose of clonal expansion is to maximize the preservation of individuals with high adaptability, which can obviously increase the convergence performance of the QCGWO. At the same time, under the effect of the quantum probability amplitude, the application of the clone operator will not reduce the algorithm's global search ability, and the combination of clone and quantum effectively improves the performance of QCGWO. In addition, the selection of the clone parent is based on the fitness of the individuals in the current population. The higher the fitness, the more likely the individual is to be selected. For the same purpose of increasing the diversity of the population, the traditional clone operation is updated in QCGWO, and the cloned population is optimized through multilevel cloning. The specific cloning operation can be expressed as Figure 3.

4.7. Termination Operation. In each iteration, QCGWO will repeat the above process. If the QCGWO reaches the specified number of iterations, it will be terminated.

4.8. Steps of the Algorithm. The detailed process of QCGWO is shown below.

Step 1. Initialize the parameters in QCGWO, randomly initialize the positions of the sensors and the targets, then

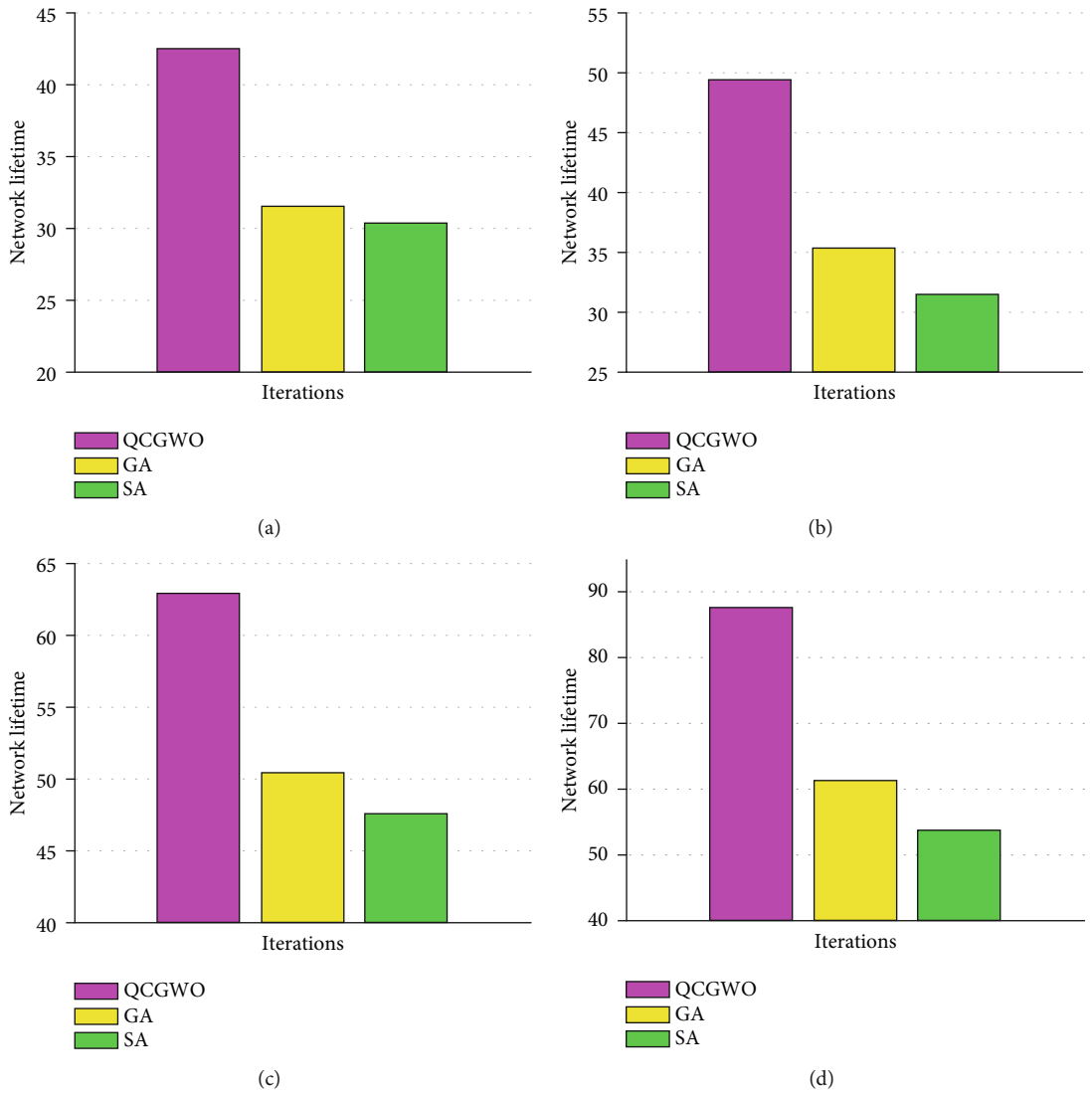


Figure 7: Network lifetime in comparison of the three algorithms: (a) 30 sensors and 10 targets; (b) 50 sensors and 15 targets; (c) 70 sensors and 25 targets; (d) 100 sensors and 30 targets.

generate the sensor coverage matrix **S**, and randomly generate the initial population. The initial quantum probability amplitude is 0.5, and the initial iteration **gen** = 1.

Step 2. Calculate the fitness of each wolf in the population, find **α**, **β**, and **δ** in the current population, and set their positions as O_α, O_β, and O_δ, respectively.

Step 3. Update the positions of all individuals according to equation (18).

Step 4. Calculate the fitness of wolves in the population, find **α**, **β**, and **δ** in the current population, and set their positions as O_α, O_β, and O_δ, respectively.

Step 5. Update the parameters **a**, **A**, and **C** of the algorithm.

Step 6. Sort the fitness, and select the individuals with the highest fitness as the parent to perform the clone operation.

Step 7. Perform large-scale mutation operations on the clonal population. The mutation process uses the quantum probability amplitude.

Step 8. Calculate the fitness of the cloned population, and use the cloned population as the population for the next iteration process.

Step 9. Update the quantum probability amplitude according to the individual with the highest fitness, and execute the quantum revolving gate.

Step 10. **gen** = **gen** + 1, if the maximum number of iterations is reached, terminate the algorithm, otherwise go to step 3.

Table 4: Solution quality comparison.

	The percentage of QCGWO's solution better than that of GA	The percentage of QCGWO's solution better than that of SA
Lifetime in Figure 5(a)	11.81%	16.23%
Lifetime in Figure 5(b)	27.46%	41.19%
Lifetime in Figure 5(c)	21.45%	57.71%
Lifetime in Figure 5(d)	33.04%	57.14%

Table 5: The comparison of convergence speed in Figure 6.

The specified lifetime rounds	The number of iterations of QCGWO	The number of iterations of GA	The number of iterations of SA
30 rounds in Figure 6(a)	10	17	27
30 rounds in Figure 6(b)	8	18	22
40 rounds in Figure 6(c)	14	26	33
40 rounds in Figure 6(d)	14	24	28

The algorithm flow chart of QCGWO is shown in Figure 4.

5. Results and Discussion

The QCGWO method we propose on solving the sensor duty cycle problem will take a series of experiments, and QCGWO has been compared with GA and SA for proving its effectiveness. The comparison for the algorithms is carried out under the conditions of different quantity of sensors, monitored targets, different maximum sensor lifetime, and monitoring area radius. In addition, all test cases are completed on the machine with a R7 4800H 2.9 GHz CPU, and the fitness used in the algorithms is calculated according to formula (4).

With the purpose of enabling the three algorithms to be compared under the same experimental conditions, we uniformly define the parameters commonly used in the sensor duty cycle problem in the IWSNs. The number of iterations is set to 100 generations, and the population size is 40. The monitoring area of IWSNs is set as a square area with a side length of 200, and the coordinates of the sensors and the target nodes are randomly generated in the area. In QCGWO, the initial value of the quantum probability amplitude is set to 0.5, the probability of the quantum revolving gate is set to 0.05, and the mutation rate of the clone operation is 0.3. In GA, the mutation rate of the population is 0.1, and the crossover method is two-point crossover. In SA, the initial temperature is 200, the annealing method is exponentially decreased and the annealing factor is 0.95.

Tables 1–3 show the experimental conditions of Figures 5–7, respectively.

In Figures 5(a)–5(d), the convergence speed of the three algorithms is shown. Specifically, Figure 5(a) indicates that QCGWO converged faster than GA and SA, and QCGWO has maintained a fast convergence rate during the iterative process. In contrast, SA fell into premature convergence in the 40th iteration, and GA also fell into premature convergence in the 60th iteration, so they are unable to find the optimal solution. In Figure 5(b), the maximum network lifetime obtained by QCGWO is 32.56 rounds. However, the optimal solution obtained by GA is 25.08 rounds, and the optimal solution obtained by SA is 23.14 rounds. The solutions of QCGWO are 28% and 39% higher than GA and SA, respectively. In Figure 5(c), GA and SA fell into the local optimum in about 30 iterations, while QCGWO effectively jumped out of the local optimum by its good global search ability. What is more, in Figure 5(d), QCGWO has maintained rapid convergence speed until the 90th generation finds the optimal solution 45.68; however, due to the premature convergence and weak ability for jumping out of the local optimum, the highest solutions obtained by GA and SA are 33.43 and 28.59, respectively, which are lower than the network lifetime of QCGWO. The comparison of the solution quality of the three algorithms in Figure 5 is shown in Table 4.

According to Table 4, it is obvious that the quality of the solution obtained by using QCGWO is better than that of GA and SA. Particularly, when the scale of IWSNs expands, the advantages of QCGWO become more prominent.

Figures 6(a)–6(d) show the trend of the three algorithms more clearly in the form of line charts. According to Figure 6(a), in the 10th generation, QCGWO's network lifetime value is already higher than GA and SA. Since then, QCGWO has maintained a leading position and found the optimal solution 40.10. Subsequently, Figure 6(b) shows that GA and SA fall into premature convergence in the 40th generation, which leads to the local optimal solutions of 36.47 and 27.89, respectively, while QCGWO effectively obtains the optimal solution 51.66. In Figures 6(c) and 6(d), during the early iterations, QCGWO, GA, and SA all converged very quickly, but it is obvious that both GA and SA have fallen into local convergence. Therefore, both GA and SA only got local optimal solutions, while QCGWO obtained better solutions than them. In general, QCGWO has better performance than GA and SA, and the convergence speed of the three algorithms in Figure 6 can be shown as Table 5.

According to Table 5, we can find that in the same experimental conditions, the number of iterations of QCGWO for

reaching the specified number of lifetime rounds is always less than that of GA and SA, which proves the good convergence performance of QCGWO.

With the purpose of making the network lifetimes obtained by the three algorithms more obvious, Figures 7(a)–7(d) use bar charts to display the data. In Figure 7(a), the maximum network lifetime values obtained by QCGWO, GA, and SA are 42.40, 31.50, and 30.30, respectively. In Figures 7(c) and 7(d), the network lifetime obtained by QCGWO is also the highest, with values of 49.35, 62.90, and 87.67, respectively. Moreover, the values obtained by GA are 35.30, 50.40, and 61.33, respectively, and the solutions obtained by SA are 31.50, 47.60, and 53.67, respectively. Therefore, under the specified experimental conditions, QCGWO always performed better than GA and SA.

6. Conclusions

The purpose of this paper is to prolong the lifetime of the IWSNs. Therefore, we modeled the industrial sensor network in the real factory, proposed a quantum clone gray wolf optimization (QCGWO) algorithm, designed the sensor duty cycle model from a different perspective compared with the previous works, and proposed a concept of measurable sensor lifetime. The algorithm we proposed has the advantages of high solution accuracy, strong convergence performance, and strong global search ability. What is more, the QCGWO learns from the traditional gray wolf optimization algorithm (GWO), but we have achieved important innovations of combining the GWO with some current popular technologies, including quantum operator and clone operator, thereby effectively making up for the weakness of GWO that is easy to fall into local optimum.

The effectiveness of the proposed model has been verified by different experimental conditions in Section 5, and the results suggested that the proposed model can achieve a longer network lifetime. In addition, in order to prove the advantages of the proposed algorithm in solving the sensor duty cycle problem, we compare QCGWO with GA and SA. The results show that the QCGWO is more competitive than GA and SA in improving the lifetime of IWSNs. Finally, we would like to highlight that the proposed model and the QCGWO can successfully solve the sensor duty cycle problem, and the approach proposed in this paper provides a new perspective for prolonging the network lifetime in IWSNs.

Data Availability

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Disclosure

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This paper was funded by the Corps Innovative Talents Plan, grant number 2020CB001, the project of Youth and Middle-aged Scientific and Technological Innovation Leading Talents Program of the Corps, grant number 2018CB006, the China Postdoctoral Science Foundation, grant number 220531, the Funding Project for High Level Talents Research in Shihezi University, grant number RCZK2018C38, and the Project of Shihezi University, grant number ZZZC201915B.

References

[1] E. H. Houssein, M. R. Saad, F. A. Hashim, H. Shaban, and M. Hassaballah, "Levy flight distribution: a new metaheuristic algorithm for solving engineering optimization problems," *Engineering Applications of Artificial Intelligence*, vol. 94, p. 18, 2020.

[2] B. Houtan and H. Zarrabi, "Obstacle-aware fuzzy-based localization of wireless chargers in wireless sensor networks," *Canadian Journal of Electrical and Computer Engineering*, vol. 43, no. 1, pp. 17–24, 2020.

[3] E. G. Pardo, J. Blanco-Linares, D. Velazquez, and F. Serradilla, "Optimization of a steam reforming plant modeled with artificial neural networks," *Electronics*, vol. 9, no. 11, 2020.

[4] S. Najjar-Ghabel, L. Farzinvash, and S. N. Razavi, "Mobile sink-based data gathering in wireless sensor networks with obstacles using artificial intelligence algorithms," *Ad Hoc Networks*, vol. 106, p. 12, 2020.

[5] G. S. Gandhi, K. Vikas, V. Ratnam, and K. S. Babu, "Grid clustering and fuzzy reinforcement-learning based energy-efficient data aggregation scheme for distributed WSN," *IET Communications*, vol. 14, no. 16, pp. 2840–2848, 2020.

[6] H. Chen, Y. Qin, K. Lin et al., "PWEND: proactive wakeup based energy-efficient neighbor discovery for mobile sensor networks," *Ad Hoc Networks*, vol. 107, p. 102247, 2020.

[7] A. Janarthanan and D. Kumar, "Localization based evolutionary routing (LOBER) for efficient aggregation in wireless multimedia sensor networks," *Computers, Materials and Continua*, vol. 30, no. 3, pp. 895–912, 2019.

[8] D. M. Gao, S. Zhang, F. Q. Zhang, X. J. Fan, and J. C. Zhang, "Maximum data generation rate routing protocol based on data flow controlling technology for rechargeable wireless sensor networks," *Computers, Materials & Continua*, vol. 59, no. 2, pp. 649–667, 2019.

[9] Q. W. Zhang, D. Z. Li, Y. Fei, J. K. Zhang, Y. Chen, and F. Tong, "RDCPF: a redundancy-based duty-cycling pipelined-forwarding MAC for linear sensor networks," *Sensors*, vol. 20, no. 19, 2020.

[10] M. A. Panhwar, D. Z. Liang, K. A. Memon, S. A. Khuhro, M. A. K. Abbasi, and Z. A. Noor-ul-Ain, "Energy-efficient routing optimization algorithm in WBANs for patient monitoring," *Journal of Ambient Intelligence and Humanized Computing*, 2020.

[11] J. Z. Xie, B. J. Zhang, and C. P. Zhang, "A novel relay node placement and energy efficient routing method for

heterogeneous wireless sensor networks," *IEEE Access*, vol. 8, pp. 202439–202444, 2020.

[12] M. A. Panhwar, Z. Deng, S. A. Khuhro, and D. N. Hakro, "Distance based energy optimization through improved fitness function of genetic algorithm in wireless sensor network," *Studies in Informatics and Control*, vol. 27, no. 4, pp. 461–468, 2018.

[13] A. Latha, S. Prasanna, S. Hemalatha, and B. Sivakumar, "A harmonized trust assisted energy efficient data aggregation scheme for distributed sensor networks," *Cognitive Systems Research*, vol. 56, pp. 14–22, 2019.

[14] U. Patil, A. V. Kulkarni, R. Menon, and M. Venkatesan, "A novel AEB-AODV based AADITHYA cross layer design hibernation algorithm for energy optimization in WSN," *Wireless Personal Communications*, pp. 1–21, 2020.

[15] I. Jemili, D. Ghrab, A. Belghith, and M. Mosbah, "Cross-layer adaptive multipath routing for multimedia wireless sensor networks under duty cycle mode," *Ad Hoc Networks*, vol. 109, p. 102292, 2020.

[16] S. Chen, L. Zhang, Y. Tang et al., "Indoor temperature monitoring using wireless sensor networks: a SMAC application in smart cities," *Sustainable Cities and Society*, vol. 61, article 102333, 2020.

[17] H. Wang, K. Li, and W. Pedrycz, "A routing algorithm based on simulated annealing algorithm for maximising wireless sensor networks lifetime with a sink node," *International Journal of Bio-Inspired Computation*, vol. 15, no. 4, pp. 264–275, 2020.

[18] A. S. Kirsan, U. H. Al Rasyid, I. Syarif, and D. N. Purnamasari, "Energy efficiency optimization for intermediate node selection using MhSA-LEACH: multi-hop simulated annealing in wireless sensor network," *EMITTER International Journal of Engineering Technology*, vol. 8, no. 1, pp. 1–18, 2020.

[19] H. L. Wang, G. B. Zhou, L. Bhatia, Z. C. Zhu, W. Li, and J. A. McCann, "Energy-neutral and QoS-aware protocol in wireless sensor networks for health monitoring of hoisting systems," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5543–5553, 2020.

[20] F. Fraternali, B. Balaji, Y. Agarwal, and R. K. Gupta, "ACES," *ACM Transactions on Sensor Networks (TOSN)*, vol. 16, no. 4, pp. 1–31, 2020.

[21] S. Tahmasebi, M. Safi, S. Zolfi, M. R. Maghsoudi, H. R. Faragardi, and H. Fotouhi, "Cuckoo-PC: an evolutionary synchronization-aware placement of SDN controllers for optimizing the network performance in WSNs," *Sensors*, vol. 20, no. 11, article 3231, 2020.

[22] H. Y. Huang, K. T. Kim, and H. Y. Youn, "Determining node duty cycle using Q-learning and linear regression for WSN," *Frontiers of Computer Science*, vol. 15, no. 1, pp. 1–7, 2020.

[23] D. Passos, C. O. de Sousa, and C. Albuquerque, "An NDT model for block designs operating under asymmetrical duty cycling," *IEEE Wireless Communications Letters*, vol. 9, no. 12, pp. 2116–2120, 2020.

Hindawi
Journal of Sensors
Volume 2021, Article ID 5529527, 12 pages
https://doi.org/10.1155/2021/5529527

Research Article

A Chaotic Parallel Artificial Fish Swarm Algorithm for Water Quality Monitoring Sensor Networks 3D Coverage Optimization

Jie Zhou, Guohong Qi, and Changzheng Liu

College of Information Science and Technology, Shihezi University, Shihezi, China

Correspondence should be addressed to Jie Zhou; jiezhou@shzu.edu.cn and Changzheng Liu; 1823982150@qq.com

Received 18 January 2021; Revised 30 January 2021; Accepted 8 February 2021; Published 26 February 2021

Academic Editor: Bin Gao

In recent years, the increasingly severe water pollution problem encouraged researchers to optimize water quality monitoring sensor networks (WQMSNs) by creating new underwater sensor coverage algorithms. Since the sensor is limited by the monitoring range and the number of targets, optimizing the 3D target coverage of heterogeneous multisensors is essential to maximize the 3D target coverage rate of the monitored waters. To enhance the target coverage rate, the target allocation needs to be searched in all possible combinations. To optimize the 3D coverage of underwater targets, this research proposes a chaotic parallel artificial fish swarm algorithm (CPAFSA). CPAFSA uses chaotic selection to initialize parameters and integrates the global search capabilities of parallel operators. It also applies the elite selection which effectively avoiding local optimization and solving the problem of 3D target coverage. Ultimately, CPAFSA is compared with genetic algorithm (GA) and particle swarm optimization (PSO). The results of the simulation experiment demonstrated the excellent performance of CPAFSA in achieving underwater 3D target coverage.

1. Introduction

In recent years, with economic development and population growth, environmental pollution, especially water pollution, has become increasingly serious. The monitoring of water quality is an important means to prevent and control pollution, so researchers gradually set their sights on real-time water quality monitoring through wireless sensors [1–3]. Since range and target number are the constraints of sensor monitoring, it is necessary to optimize the 3D coverage of heterogeneous multisensors to maximize the target coverage rate of water quality monitoring sensor networks (WQMSNs). However, the optimal coverage of heterogeneous multisensor is an NP-hard problem. The exhaustive method is considered to be a possible way to solve this problem, but its computational complexity is too high to be suitable for actual real-time applications. Most of the research on sensor coverage involves heuristic sensor coverage algorithms [4]. In previous studies, scholars have proposed many optimization algorithms to solve wireless sensor network (WSNs) problems, including genetic algorithm (GA) and particle swarm optimization (PSO) [5–11]. Recently, more and more scholars have proposed various optimization algorithms for sensor optimization [12–16]. The application direction of the optimization algorithm on the sensor has also become broader [17–20].

Water quality monitoring is an important means to prevent and control water pollution. To provide reliable underwater quality monitoring services and data analysis for environmental protection and domestic water use service, many researchers designed and proposed water quality monitoring strategies based on wireless sensor network and Internet of Things.

To improve the coverage of underwater wireless sensor networks (UWSNs) and extend its network lifetime, paper [21] proposed an algorithm combining the virtual force and PSO. The algorithm guides the optimization of the particle swarm, moving the underwater node to a relatively ideal position, thereby accelerating the particle convergence and making the PSO develop toward the target solution. In response to the deployment of wireless sensor nodes, a natural heuristic cuckoo search algorithm was proposed in [22] to find the best deployment position of sensors in a 3D underwater environment. To maximize target coverage with the

least number of sensors, the authors model the deployment of sensors as an optimization problem. Literature [23] studied a method of allocating underwater monitoring coverage resources in sensor networks and proposed positioning and deployment of UWSNs nodes based on GA. The method can maximize the coverage and protection of high-value assets in military applications. In [24], the authors proposed an improved fruit fly optimization algorithm to solve the 3D underwater sensor network coverage optimization problem and designed a 3D space-based network coverage method. The algorithm uses the behavior of fruit flies' preying to optimize global optimal monitoring. This algorithm can quickly obtain the deployment position of sensor nodes, thereby solving the problem of 3D coverage deployment of wireless sensor nodes. In [25], to solve the optimization problem of node redeployment coverage in UWSNs, an underwater sensor network redeployment algorithm based on wolf-pack search technology was proposed. They use sensor nodes to ensure coverage and avoid nodes appearing prematurely.

Research on merging adaptive theory and parallel theory has been carried out since the late 1990s [26]. Developed at the beginning of the 21st century, the artificial fish swarm algorithm (AFSA) is an evolutionary optimization algorithm that tries to find the best solution of optimizing problems by stochastic rules, and it explores the problem region with a probabilistic policy [27]. The theoretical foundations of AFSA were presented by [28]. A chaotic parallel artificial fish swarm algorithm (CPAFSA) is presented for the underwater 3D target coverage problem in this paper. A model of target coverage and monitoring is proposed to maximize the coverage of underwater 3D targets. And, the elite operator is used when updating the individual artificial fish to ensure a better evolution of the fish swarm. The strategy mixes the merits of a parallel selection and a chaotic operator to enhance the global explore capacity. Then, an adaptive adjustment method is used to obtain better experimental results while avoiding local optima.

To illustrate the advantages of CPAFSA in maximizing the underwater 3D target coverage area by WQMSNs, GA and PSO are used for comparisons. On the one hand, compared with genetic algorithms, CPAFSA uses parallel operation to improve optimization capabilities. On the other hand, CPAFSA can overcome the premature problem of traditional genetic algorithms by using chaotic operator. Through these new operations, CPAFSA becomes a suitable global optimization method to find the optimal solution without falling into the local optimum. CPAFSA can create a feasible solution to the underwater 3D target coverage problem of maximizing the underwater 3D target coverage rate of heterogeneous sensors within an acceptable time.

The results, which are simulated based on CPAFSA, GA, and PSO, show that CPAFSA develops a good solution for achieving a higher underwater 3D target coverage rate. CPAFSA has improved the performance of underwater 3D target coverage by combining parallel adjustment and chaotic optimization. It also helps to avoid local optimal.

The structure of this paper is as follows. Section 2 elaborates on the underwater 3D target coverage and surveillance and, then, introduces its target coverage and surveillance model. CPAFSA is used to extend the performance of underwater 3D target coverage monitoring problems in Section 3. Section 4 shows the results of the simulation experiment and discusses the significance of CPAFSA. Then, Section 5 concludes.

2. Underwater 3D Target Coverage and Monitoring Model

The deployment problem of WQMSNs is directly related to the optimal configuration of its communication bandwidth, node power, analysis and calculation capabilities, and other restricted resources. It also affects the quality of communication, monitoring, and perception services to a large extent. How to ensure the target coverage of the monitoring area under the premise of the limited number of sensors and limited monitoring capabilities is a key issue that determines monitoring performance. This section will discuss the mathematical model of underwater 3D target coverage and monitoring by WQMSNs.

Firstly, in real-time monitoring of underwater targets, since the sensing capabilities of sensors are limited, usually sensors can only perceive a limited number of targets in their monitoring area, and each target needs to be monitored by multiple heterogeneous sensors. Suppose there are X target points that need to be monitored in a piece of water. These target points need to be covered and monitored by underwater wireless sensors, and the number of sensors is Y. Equation (1) represents the coverage relationship between each sensor and each target of WQMSNs.

$$FG = \begin{bmatrix} fg_{1,1} & fg_{1,2} & \cdots & fg_{1,Y-1} & fg_{1,Y} \\ fg_{2,1} & fg_{2,2} & \cdots & fg_{2,Y-1} & fg_{2,Y} \\ \vdots & \vdots & fg_{x,y} & \vdots & \vdots \\ fg_{X-1,1} & fg_{X-1,2} & \cdots & fg_{X-1,Y-1} & fg_{X-1,Y} \\ fg_{X,1} & fg_{X,2} & \cdots & fg_{X,Y-1} & fg_{X,Y} \end{bmatrix} \cdot \left(fg_{x,y} \in \{0, 1\}\right). \tag{1}$$

FG represents the coverage relationship matrix between the residual chlorine sensor node and the target node, and $fg_{x,y}$ represents the coverage relationship between the y residual chlorine sensor node and the monitoring target x.

Equation (1) indicates the coverage relationship in WQMSNs, that is, $fg_{x,y} = 1$ means that the monitoring target x is within the y sensor node's monitoring area, and $fg_{x,y} = 0$ indicates that the monitoring target x is outside the monitoring area of the y sensor.

Then, due to the limited monitoring capabilities of sensors, only a limited number of target points within the area of sensor nodes can be monitored. Equation (2) is used to express the monitoring relationship between sensor nodes and covered targets in WQMSNs.

$$JC = \begin{bmatrix} jc_{1,1} & jc_{1,2} & \cdots & jc_{1,V-1} & jc_{1,V} \\ jc_{2,1} & jc_{2,2} & \cdots & jc_{2,V-1} & jc_{2,V} \\ \vdots & \vdots & jc_{u,v} & \vdots & \vdots \\ jc_{U-1,1} & jc_{U-1,2} & \cdots & jc_{U-1,V-1} & jc_{U-1,V} \\ jc_{U,1} & jc_{U,2} & \cdots & jc_{U,V-1} & jc_{U,V} \end{bmatrix} \cdot \left(jc_{u,v} \in \{0, 1\}\right). \quad (2)$$

Equation (2) points out the monitoring relationship in WQMSNs, $jc_{u,v} = 1$ indicates that the monitoring target u is within the monitoring range of the v sensor node and u is monitored. $jc_{u,v} = 0$ suggests that the monitoring target u is outside the monitoring area of the v sensor, or u is within the monitoring range but it is not monitored.

If a sensor node can only monitor N target points in the 3D coverage area, the constraint condition of the monitoring relationship is shown as

$$\sum_{u=1}^{U} jc_{u,v} \le N, v \in V. \quad (3)$$

If each monitored target point must be monitored by at least M sensors and a sensor can merely monitor Ntarget points in the coverage area, the mathematical model of underwater 3D target coverage for monitoring more target points in WQMSNs is

$$jc_{u,v} \le fg_{u,v}. \quad (4)$$

Equation (4) represents the relationship between the monitoring matrix and the coverage matrix. Only when the sensor covers the target point, the corresponding target point in the monitoring matrix may be monitored by the sensor. Considering that the limitation of the number of targets monitored by the sensor, it is also possible that the target points are covered but not monitored.

$$W(u) = \sum_{v=1}^{V} jc_{u,v}, u \in U. \quad (5)$$

$$WM(u) = \begin{cases} 0 & W(u) < M \\ 1 & W(u) \ge M \end{cases}. \quad (6)$$

Equation (5) uses W to store the number of target points monitored by the sensor and, then, determines whether the point is effectively monitored according to the restriction condition (6).

The above clarifies the optimization direction by establishing a mathematical model for underwater sensor coverage and monitoring target points. The next section will design and implement a CPAFSA that expands the underwater 3D coverage and monitoring rate of sensors based on the model.

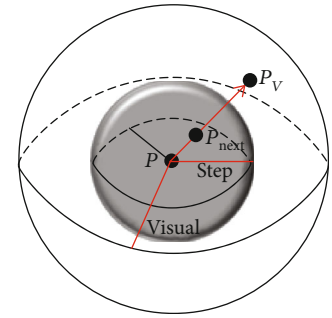


Figure 1: Simulated artificial fish vision.

3. CPAFSA for Maximizing Underwater 3D Coverage and Monitoring Rate in WQMSNs

The basis of the artificial intelligence model based on biological behavior is the bottom-up design method. This model designs the behavioral perception of a single entity and then places the individual or group in the environment so that it can propose solutions to problems in the interaction with the environment. Individuals usually do not have advanced intelligence, but they can show advanced intelligence during group activities. This phenomenon is called swarm intelligence. Individuals with social characteristics can produce group intelligence when they cooperate in certain activities, such as fish swarm.

Artificial fish swarm is an abstraction of a biological fish swarm, which simulates the characteristic behavior of biological fish and their response to environmental stimuli. Individual artificial fish can receive environmental information through vision and respond accordingly, and the action of the individual artificial fish will also affect the other artificial fish individuals. Artificial fish's perception of the environment is realized through vision. But the visual system of biological fish is very complicated, so the concept of the visual field is adopted when simulating artificial fish vision. Figure 1 simulates the relationship between the field of view of an artificial fish and its step length. The current state of the individual fish is set to $P = (p_1, p_2, p_3, \cdots, p_n)$; artificial fish's field of vision is $Visual$and chaotically selects a state in its $Version$at a certain time$P_V = (p_{1V}, p_{2V}, p_{3V}, \cdots, p_{nV})$; if the state P_V is better than P, the fish will take a random step towards P_V to reach state P_{next}. Otherwise, continue to try to randomly select other states in its $Version$. The more the individual fish try, the better they can understand the environmental information in the field of view. This helps to make correct behavior decisions. But the actual search behavior of biological fish cannot increase indefinitely, the number of inspections of artificial fish is also limited. Preserving the uncertain local optimization of artificial fish is conducive to artificial fish searching for the global optimum.

P in Figure 1 represents the current state of the individual fish, including two parameters: visual field of view and step length. Among them, $Visual$ is used for the perception of other fish individuals within the field of view, and $Step$ is the current moveable range of P. In the 3D search space, X = (abscissa, ordinate, vertical). And the fish

swarm wants to find the area with the highest concentration of food, which is also called fitness here, expressed as $Y = f(X)$. The spatial distance between individual fish is expressed as $d_{i,j} = \|X_i - X_j\|$. Besides, fish swarm parameters also include the maximum number of attempts in the fish swarm for preying *TryNumber*, the crowdedness degree of the fish swarm δ, and the number of fish within the field of view of a single fish that is n_f.

In the AFSA, there are mainly four kinds of behaviors: preying behavior, swarming behavior, following behavior, and random behavior to simulate biological fish swarm. Preying behavior is the basic behavior of artificial fish. Artificial fish use the perception parameter *Visual* of the environment to change their position. For different fish state, the food concentration is different. Each fish compares the food concentration at its own location with the randomly selected food concentration in perception *Visual* and selects the moving direction of the larger food concentration so that the entire fish swarm tends to a high concentration position. Swarming behavior means that each fish moves to the center of the neighboring other artificial fish to ensure that the surrounding areas are all partners to reduce the risk. When the swarming behavior is in progress, the artificial fish in the AFSA are all moving to a place where food is concentrated. Each fish can form a group through grouping behavior and move to a high concentration position together. Following is the behavior of artificial fish chasing the fish with the highest food density among other fish nearby. Due to the nature of the fish swarm toward food and away from natural enemies, when some individuals in the group find food and move in a certain direction, other individuals will follow it. The following reduces the time when artificial fish explore the surrounding environment and also reduces the computing time. Random behavior refers to the behavior when the number of times the artificial fish prey on food reaches the specified maximum number of times, and the food concentration is still not increased. At this moment, the fish randomly moves one step to a certain state in its field of view and uses this state as the next state. Random behavior helps the individual fish to escape from the local optimum. Besides, an artificial fish swarm is mainly affected by three items in the optimization process: bulletin board, behavior evaluation, and iteration termination conditions.

The traditional AFSA has strong global convergence, but it is easy to cause the problem which fish swarms difficult to escape from the local optimum in the later stage of execution. This section proposes CPAFSA. It generates fish swarm sequence through chaotic mapping and then generates chaotic adaptive function coefficients to improve the field of view and step length of fish. So that the artificial fish's field of view and step size change consistent with the parameter requirements of different execution stages of the algorithm. Moreover, to enhance the low accuracy of the AFSA optimal solution, parallel technology is used to improve it. We also use elite operators to make the fish swarm search more efficient. Section 5 compares and analyzes the optimization results of the underwater 3D coverage area of CPAFSA, PSO, and GA.

The execution of the CPAFSA algorithm is a process of intelligent optimization. The main steps are as follows:

Step 1: initialize Q individual artificial fish, moving step *Step*, visual field *Visual*, number of attempts *TryNumber*, congestion factor δ, current number of iterations = 0, maximum number of iterations Max, probability factor $\alpha\,(0 < \alpha < 1)$, constant coefficient S

Step 2: calculate the fitness function of Q artificial fish, also called food concentration, obtain the best sensor allocation plan, and assign the fitness function to the bulletin board *BB*

Step 3: if the center position P_c is better than the current artificial fish position P_i and is not too crowded, set $Step = Rand * \|X_c - X_i\|$ and move one step to the center position; otherwise, go to Step5

Step 4: if the optimal fish $P_{\max}$ in the current artificial fish P_i's perception range is better than artificial fish in the current state, and not too crowded, set $Step = Rand * \|X_{\max} - X_i\|$ to move one step to the optimal fish; otherwise, switch to Step 5

Step 5: select a random state P_j in the current field of vision of artificial fish P_i, if $\Delta Y = Y_j - Y_i > 0$, set $Step = Rand * \|X_j - X_i\|$, move to the artificial fish one by one random steps. If after try *TryNumber* times, there is still no result that meets the conditions, a random position is selected as the next position in the field of *Visual*, and a random step is moved to this position. Update the bulletin board

Step 6: if iteration < Max, then set iteration = iteration + 1 and switch to Step 3; if iteration ≥ Max, switch to Step7

Step 7: output the optimal plan information of the bulletin board

Figure 2 shows the process of the CPAFSA algorithm in solving the underwater 3D coverage problem of WQMSNs.

CPAFSA largely depends on the coding method that solves the underwater 3D coverage and monitoring problem of WQMSNs. Because CPAFSA is an optimized algorithm for simulating biological fish swarm, individual artificial fish are the solution to the sensor coverage problem. Each individual fish is an independent solution to the monitoring optimization problem of WQMSNs. The information needed to construct artificial fish collaboration can be expressed as $\partial = [\text{fish swarm size, prey, follow, swarm}]$. The individual artificial fish is coded as a one-dimensional array $P = (p_1, p_2, p_3, \cdots, p_n)$, and it has a 3D coordinate. Determine the swimming direction of the individual fish by the target value corresponding to each element in the array, so as to maximize the underwater 3D coverage. The coding scheme of the one-dimensional array actually limits the fish; then, it can reflect the 3D coverage of the water quality monitoring sensor network. In CPAFSA, a swarm of fish is composed of a certain number of individual fish, and this fish swarm represents all the solutions to the underwater 3D coverage problem of the sensor.

The other parameters of CPAFSA are set as follows.

Under the premise of system resources and running time permitting, using a large population can improve the accuracy of the optimal solution and enhance the ability of the fish swarm to jump out of the local optimal.

The artificial fish's field of view mainly affects its foraging situation. To understand the information of the optimization

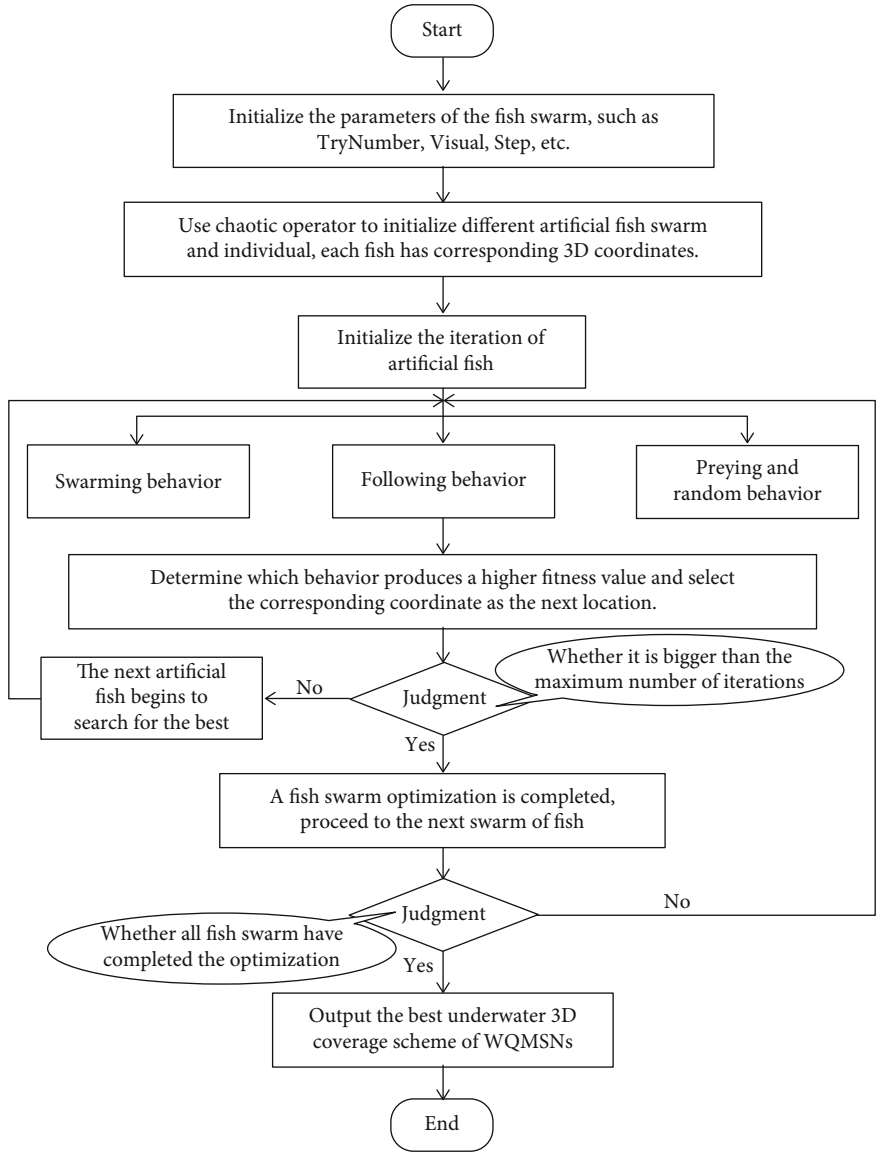


Figure 2: Process of the CPAFSA.

space as much as possible, a larger field of view should be used, but a larger field of view will cause the problem of reduced optimization accuracy.

When setting the step length of the artificial fish, it is necessary to consider the size of the fish field of view. As the field of view increases, the step length should also increase accordingly; otherwise, the convergence speed will be slower, but the optimal solution accuracy will decrease. The smaller the step size set by the artificial fish, the lower the convergence speed of the algorithm, which may fall into the local optimum.

The congestion factor parameter is integrated according to the literature [27].

Too many *TryNumber* attempts will cause the artificial fish to be trapped by local extremes, resulting in premature algorithm maturity and increased optimization time. Fewer preying times will reduce the probability of individual artificial fish preying successfully, causing artificial fish to perform more random behaviors, which is not conducive for algorithm convergence. Generally, the preying attempts of artificial fish do not exceed 100 times, and usually 5 – 50 times. When the local optimum is not significant and the algorithm complexity is not high, increasing the number of preying is an effective means to improve the convergence efficiency of the algorithm.

The bulletin board records the best solution to the WOMSNs 3D monitoring problem in the current artificial fish swarm. When there is a record of the best solution on the bulletin board, the individual artificial fish in this iterative process needs to compare the solution at its location with the best solution on the bulletin board. If the solution represented by the fish is better than the best solution recorded on the bulletin board at this time, the value of the best solution of the artificial fish on the board will be rewritten after the end of this iteration process, so that the bulletin board stores the optimal solution in the optimization process. The

The swarming behavior of CPAFSA

```
CPAFSA swarm () {
    X_c = 0; n_f = 0;
    for (j = 0; j < fishnum; j+ +) {
        if (d_{i,j} < Visual)
            n_f++; X_c=X_c+X_j;
        X_c = X_c / n_f ;
    }
    if (Y_c/n_f > δ * Y_i)
        X_next = X_i + (X_c−X_i)/||X_c−X_i|| * Rand() * Step;
    else
        CPAFSA prey () :
}
```

Figure 3: Pseudocode of swarming behavior.

historical optimal solution was produced. $BB = 0$ when initializing the bulletin board.

Assume that the state of the individual artificial fish in CPAFSA at this moment is X_i and its fitness function is Y_i, each individual fish perceives the number of other individuals n_f according to its visual field $Visual$ and calculates the average state X_c as the center point

$$X_c = \frac{X_1 + X_2 + \cdots + X_{n_f}}{n_f}. \tag{7}$$

Define the fitness value of the center position as Y_c/n_f. If $Y_c/n_f > \delta * Y_i$, it means that the center has more food and is not too crowded, so it will randomly move one step toward the center. Otherwise, the individual fish will perform prey behavior. The mathematical expression of clustering behavior is shown in equations (8) and (9).

$$X_{\text{next}} = X_i + \frac{X_c - X_i}{\|X_c - X_i\|} * \text{Rand}() * \text{Step}, \frac{Y_c}{n_f} > \delta * Y_i. \tag{8}$$

$$X_{\text{next}} = \text{prey}(X_i). \tag{9}$$

The pseudocode of CPAFSA's swarming behavior is shown in Figure 3.

The following behavior simulates the process of a swarm of biological fish moving toward the food source. When a certain fish finds food, the surrounding fish will follow it to swim toward the food. In artificial fish swarm of CPAFSA, this behavior is imitated as the current artificial fish X_{i} searching for the position $X_{\max}$ of the fish with the largest fitness value in its $Version$. When there is $X_{\max}$, and the corresponding fitness value $Y_{\max}/n_f > \delta * Y_i$ means that the location is not currently crowded, the artificial fish will move to the optimal direction by a random step. Otherwise, the individual artificial fish will prey. Equations (10) and (9) express the mathematical following behavior.

$$X_{\text{next}} = X_i + \frac{X_{\max} - X_i}{\|X_{\max} - X_i\|} * \text{Rand}() * \text{Step}, Y_{\max}/n_f > \delta * Y_i. \tag{10}$$

The following behavior of CPAFSA

```
CPAFSA follow () {
    Y_max = 0; n_f = 0;
    for (j = 0; j < fishnum; j++) {
        if (d_{i,j} < Visual && Y_max < Y_j){
            Y_max = Y_j; X_max = X_j;
        }
        if (d_{i,j} < Visual) n_f++;
    }
    if (Y_max/n_f > δ * Y_i)
        X_next = X_i + (X_max−X_i)/||X_max−X_i|| * Rand() * Step;
    else
        CPAFSA prey () :
}
```

Figure 4: Pseudocode of following behavior.

The preying and random behaviors of CPAFSA

```
CPAFSA prey () {
    for (j = 0; j < fishnum; j++) {
        X_j = X_i + Rand() * Visual;
        if (Y_i < Y_j)
            X_next = X_i + (X_j−X_i)/||X_j−X_i|| * Rand() * Step;
        else
            X_next = X_i + Rand() * Step;
    }
```

Figure 5: Pseudocode of preying and random behaviors.

Figure 4 shows the pseudocode of CPAFSA's following behavior.

Preying is the instinct of animals and the basis of biological evolution. When biological fish find an area with a higher food concentration, they will instinctively swim towards there. This behavior is manifested in the artificial fish swarm of CPAFSA. The fish individual randomly selects the state position X_j within its perception range Visual, and obtains the fitness value Y_j at this moment. If $Y_j > Y_i$, move the individual fish to the X_j position by a random step. If $Y_j < Y_i$, the individual fish randomly selects a state position in the Visual again for judgment. When the iteration reaches the maximum number of attempts $TryNumber$ and still does not find a qualified Y_j, the artificial fish will execute into a random state, that is, randomly select a state in the field of $Version$ and move to that state, thereby avoiding local optimal. Preying and random behaviors are expressed in mathematical language as equations (11)–(13).

$$X_j = X_i + \text{Rand}() * Visual. \tag{11}$$

$$X_{\text{next}} = X_i + \frac{X_j - X_i}{\|X_j - X_i\|} * \text{Rand}() * \text{Step}, Y_j > Y_i. \tag{12}$$

$$X_{\text{next}} = X_i + \text{Rand}() * \text{Step}, Y_j < Y_i. \tag{13}$$

The preying and random behaviors of CPAFSA uses pseudocode as Figure 5.

Table 1: Parameters settings of CPAFSA.

	Generations	Population	Visual	Step	Congestion factor	Preying attempts
CPAFSA	100	20	12	20	0.618	10

Table 2: Parameters settings of PSO.

	Generations	Population	Maximum speed	Individual and social learning factors
PSO	100	20	1	2

Table 3: Parameters settings of GA.

	Generations	Population	Mutation probability
GA	100	20	0.01

As a universal phenomenon in nonlinear systems, chaos has diversity and multiscale. These characteristics make chaos theory have great application potential in the field of algorithm optimization. For the nonlinear engineering optimization problem of underwater 3D sensor coverage problem in WQMSNs, CPAFSA uses chaotic mapping to determine the initial coordinates of underwater sensors and targets, which saves workload and improves randomness. Then, CPAFSA applied chaos search to the individual swimming process of individual artificial fish to enhance the search ability of individual fish and the ability of local optimization.

CPAFSA solves the WQMSNs problem from the angle of the fish swarm, and adding parallel operators can greatly improve the solution quality and accuracy of the algorithm. Parallel models include the fine-grained model, master-slave model, and coarse-grained model. The fine-grained model is mainly used in large-scale computer systems. It can maximize the parallelization capability of the algorithm but has very high hardware requirements. The master-slave model has a master processor and multiple slave processors. In this model, the global processing operations of the fish swarm are executed in the main processor, and the following behavior, swarming behavior, random behavior, and preying behavior are executed in the secondary processor. The coarse-grained model is an overall parallel model. The model divides the entire large fish swarm into multiple scattered fish swarms. Each scattered fish swarm evolves independently. After a certain algebraic evolution, compare different fish swarms and copy excellent fish individuals to other fish swarms, thereby improving the search ability of the algorithm in the local environment. CPAFSA uses a coarse-grained parallel model to improve the overall optimization capabilities of WQMSNs in this paper.

4. Results and Discussion

In this section, we compare the CPAFSA, PSO, and GA algorithms and analyze the performance of the algorithms in solving WQMSNs underwater 3D coverage and monitoring optimization. Then, the advantages of CPAFSA in solving this problem were demonstrated through simulation experiments with different parameter settings. The three algorithms judge the pros and cons of the optimization effect through a common fitness function.

In the general parameter setting, the monitoring range of the sensor is $400 \times 400 \times 400\ m^3$. The number of sensors is 35 and the number of monitored targets is 100. The positions of monitored targets and sensor nodes are randomly distributed due to the chaos operator. At the same time, each target point needs to be monitored by 3 sensors, and one sensor can only monitor 4 target points in the coverage area. Besides, CPAFSA, PSO, and GA use the same algebra and scale for comparison. Tables 1–3 show the parameter values of CPAFSA, PSO, and GA.

In CPAFSA, the number of individual fish in the artificial fish school is set to 40. The setting of step length and field of view parameters has an important impact on the performance of this algorithm. According to the CPAFSA optimized for underwater 3D coverage problem in WQMSNs in Section 4, the field of view and step length parameters are set. And based on the comprehensive evaluation of relevant literature, the crowding factor and the maximum number of swarming are set [29–31]. Table 1 shows these parameters.

PSO is a swarm optimization algorithm inspired by the phenomenon of bird swarm preying. It realizes the calculation of spatial solutions through collaboration and imitation between individuals. The particle has a velocity vector feature, which determines the state of the bird's flight. The particles follow the best particles in the current state to search in space and optimize the initialized random particles to find the best result through iteration. In each generation, the process of updating particles is carried out by tracking two extreme values. Extreme value 1 is the historical optimal solution of the particle itself, and extreme value 2 is the optimal solution of the overall particle. Each particle gathers at a certain speed to the best position in its own history and the best position in the neighborhood history to realize the evolution of candidate solutions. Table 2 shows the parameters of PSO.

GA is an optimization algorithm that combines genetic theory with computer technology. Many terms in this algorithm are derived from natural evolution theory. The chromosomes carry the genetic material of the organism, which controls the traits of the organism, and the gene is the functional and combined representation of the genetic material, and the value of a gene is called an allele. A certain number of genes make up a chromosome. The position on the chromosome is called a locus. The combination of genes and locus determines the characteristics of the chromosome, and the traits of an organism are external manifestations. Genetic operators such as selection probability and crossover

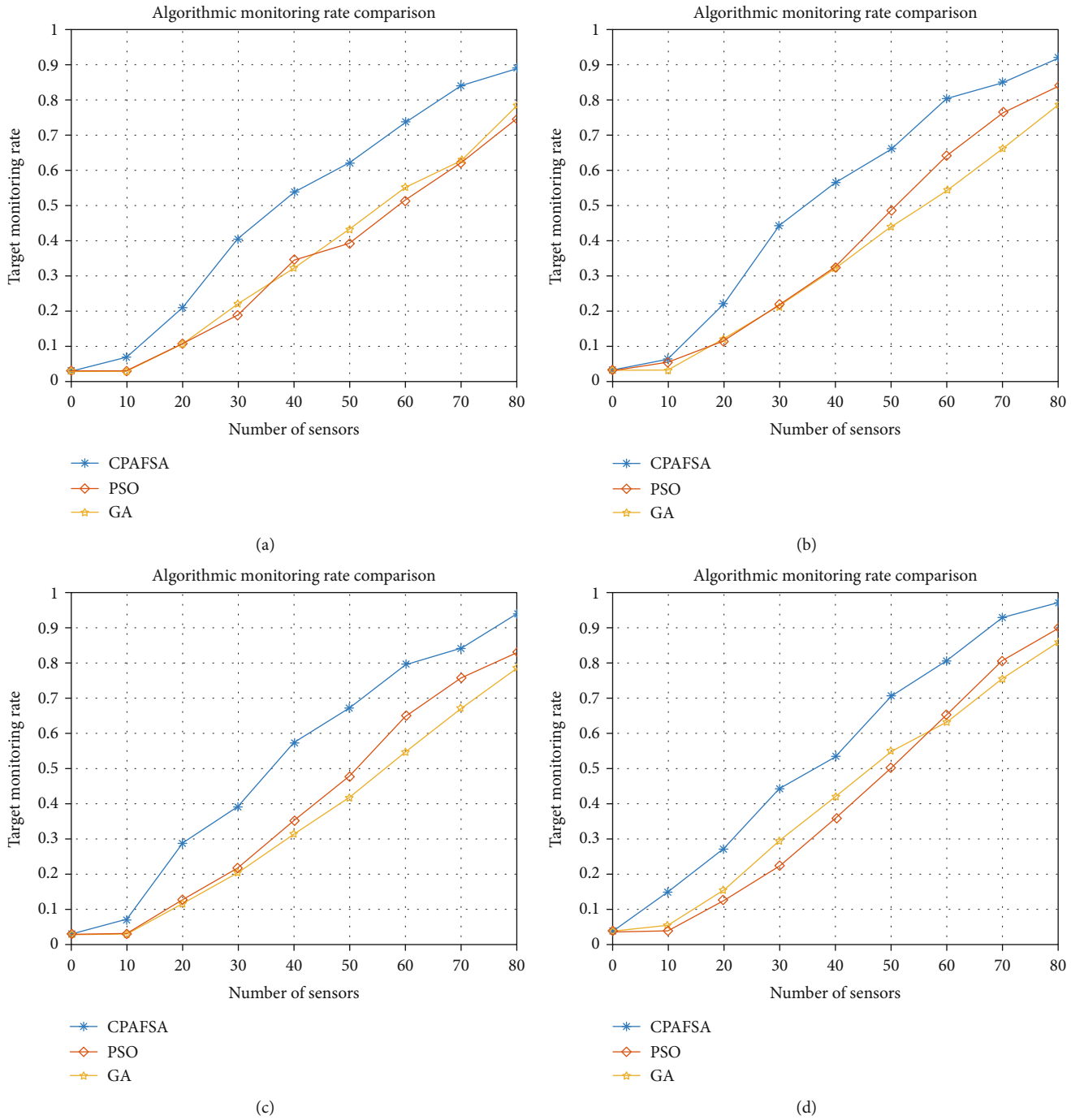


Figure 6: The change curve of the target ratio of successful monitoring as the number of sensor nodes increases: (a) radius of 120 meters; (b) radius of 140 meters; (c) radius of 160 meters; (d) radius of 180 meters.

probability need to be set in GA. Moreover, in the process of running the algorithm, GA also needs to set parameters such as the length of the individual code string and the group size. Table 3 shows the parameter settings of GA.

Figure 6 shows the change in the number of targets successfully monitored underwater caused by the increase in the number of sensor nodes. (a), (b), (c), and (d), respectively, represent the cases where the sensor node monitoring radius is 120 meters, 140 meters, 160 meters, and 180 meters. It can be seen from the figure that as the number of sensors increases, the proportion of targets successfully monitored for GA and PSO has increased, but they are always lower than CPAFSA. CPAFSA optimizes through parallel comparison, which increases the search range compared to the previous two algorithms. Meanwhile, CPAFSA will go through the update process of the optimal solution in each iteration process and can monitor more targets more efficiently. According to Figure 3, Tables 4 and 5, respectively, show the percentage increase of target nodes successfully monitored by CPAFSA compared to GA and PSO. It can be seen

Table 4: The percentage increase of CPAFSA's successful monitoring targets compared to PSO.

Number of sensors	Monitoring radius of 120 meters	Monitoring radius of 140 meters	Monitoring radius of 160 meters	Monitoring radius of 180 meters
10	4.20%	1.00%	4.20%	11.00%
20	10.20%	10.60%	16.40%	14.60%
30	21.80%	22.60%	17.40%	21.80%
40	19.40%	23.80%	22.00%	17.40%
50	23.00%	17.40%	19.60%	20.40%
60	22.20%	16.20%	14.40%	15.20%
70	22.00%	8.60%	8.40%	12.40%
80	14.40%	8.00%	11.20%	7.20%

Table 5: The percentage increase of CPAFSA's successful monitoring targets compared to GA.

Number of sensors	Monitoring radius of 120 meters	Monitoring radius of 140 meters	Monitoring radius of 160 meters	Monitoring radius of 180 meters
10	4.00%	3.20%	4.20%	9.40%
20	10.20%	9.80%	17.40%	11.60%
30	18.40%	23.00%	18.80%	14.60%
40	21.60%	24.20%	25.80%	11.20%
50	18.80%	22.00%	25.40%	15.80%
60	18.60%	26.00%	24.60%	17.20%
70	21.20%	18.80%	17.00%	17.60%
80	10.60%	13.40%	15.40%	11.00%

from the table that the percentage of CPAFSA successfully monitored targets are always 1.00% to 23.80% higher than that of PSO, and 3.20% to 25.80% higher than that of GA. CPAFSA can significantly improve the monitoring effect.

Figure 6 also shows that when the number of sensors increases, the targets successfully monitored by CPAFSA generally increase more than the other two algorithms. The reason is that as the number of sensors and targets continues to grow, the complexity of the 3D coverage of underwater sensors continues to increase. The traditional GA and PSO are prone to fall into premature convergence, while CPAFSA has a parallel operator added to the code, which can perform a broad global optimization. In addition, the clustering and foraging search can transform the global search into the local search, which improves the balance between coarse search and fine search, also improves the speed of algorithm convergence, and avoids falling into the local extremum.

Figure 7 shows the number of targets successfully monitored when the number of target points is 100, the number of sensors is 80 and 100, and the corresponding radius is 130 meters and 150 meters, respectively. The underwater 3D coverage monitoring method is based on CPAFSA, GA, and PSO in WQMSNs. The definition of target success monitoring rate is the percentage of the number of targets successfully monitored to the total number of targets within the sensor coverage. To test the actual effect of the algorithm under different types of sensors, we set the number of sensors in (a) to 80, the radius to 130 meters. Set the number of sensors in (b) to 80, and the radius to 150 meters. Then, the number of sensors in (c) is set to 100, the radius is set to 130 meters. And the number of sensors in (d) is set to 100; the radius is set to 150 meters. The results of the simulation experiment explain that as generations of the algorithm increases, the PSO-based target coverage monitoring rate quickly stabilizes, which indicates that the PSO will quickly reach a local extreme and it is difficult to jump out of the optimization stage. The coverage rate of the GA-based target coverage method is steadily increasing. However, because GA uses a fixed zero-one code, it is difficult to introduce new genes when the mutation rate is low, and it is easy to fall into premature maturity. When the probability of mutation is high, individual adaptability will fluctuate greatly, and the target success monitoring rate will also be high. It is difficult to improve. The single fish in CPAFSA adopts zero-one coding based on chaos to avoid evolutionary stagnation caused by local optimization. It can be seen from the figure that CPAFSA can always monitor more targets than PSO and GA regardless of the same number of sensors but different radii or the same number of sensors and different radii.

Figure 8 shows the percentage of successful surveillance targets based on CPAFSA, GA, and PSO in WQMSNs. To determine the utilization performance of the algorithm under different conditions, the number of sensors in (a) is set to 90, and the radius is set to 190 meters. Set the number of sensors in (b) to 80 and the radius to 200 meters. The number of sensors in (c) is set to 70, and the radius is set to 210 meters. The number of sensors in (d) is set to 60, and the radius is set to 220 meters. It can be seen from the figure that the percentages of successfully monitoring targets based on CPAFSA in (a), (b), (c), and (d) are all above

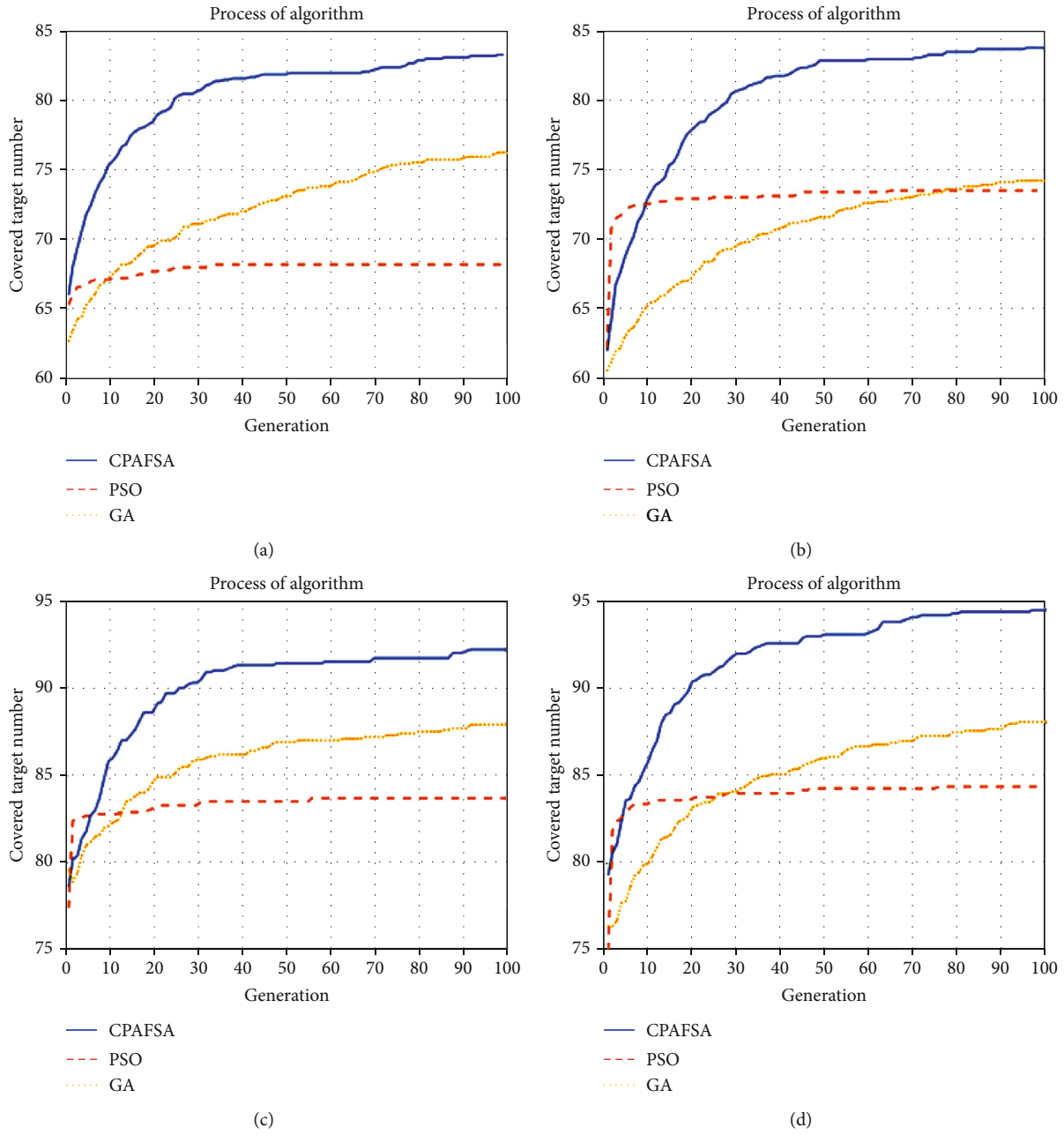


Figure 7: The number of successfully monitored targets varies with algorithm iterations: (a) 80 sensors, 130 meters radius; (b) 80 sensors, 150 meters radius; (c) 100 sensors, 130 meters radius; (d) 100 sensors, 150 meters radius.

90%, and the percentages of PSO and GA successfully monitoring targets decrease as the number of sensors decreases. Under certain conditions, compared to PSO and GA, CPAFSA can use sensors for more effective coverage and monitoring in WQMSNs.

To verify the performance of CPAFSA, this section compares this algorithm with PSO and GA in terms of the number and proportion of successfully monitored targets. The simulation results show that the newly proposed CPAFSA has stronger global optimization capabilities than PSO and GA in solving the underwater 3D coverage optimization problem in WQMSN, and it has not converged prematurely.

5. Conclusion

Aiming at the optimal coverage of the water quality monitoring sensor networks, this paper proposed a new chaotic parallel artificial fish swarm algorithm. Before the algorithm design, an underwater 3D coverage model was established to solve the sensor coverage and monitoring problem. This model optimizes the 3D coverage of underwater wireless sensors to target points when the sensors are limited. To prevent AFSA from converging too fast in the optimization process and falling into a local optimum, CPAFSA uses a chaos operator to enhance the randomness of CPAFSA's initial setting.

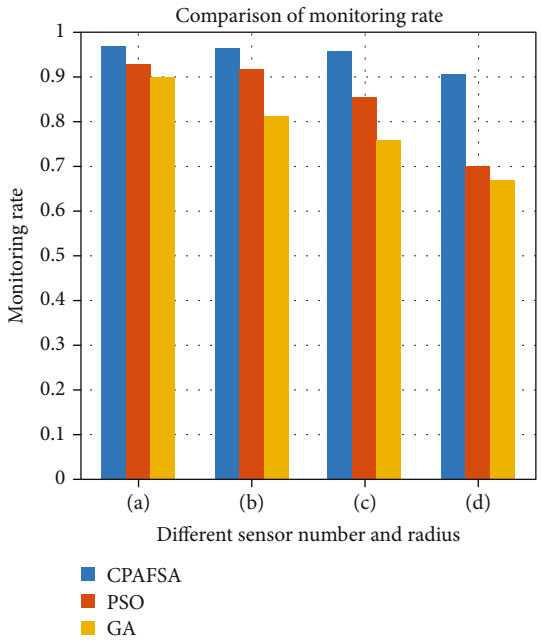


Figure 8: Percentage of successfully monitored targets: (a) 90 sensors with a radius of 190 meters; (b) 80 sensors with a radius of 200 meters; (c) 70 sensors with a radius of 210 meters; (d) 60 sensors with a radius of 220 meters.

And this algorithm enhances the ability of global fish swarm optimization through parallel strategies; thus, AFSA is optimized and the shortcomings of insufficient precision are solved. Besides, this algorithm also combines the elite operator and the adaptive operator in the optimization process, thereby improving the optimization efficiency of a single artificial fish and preventing the ineffective search.

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 that they have no conflicts of interest.

Acknowledgments

This paper was funded by the Corps innovative talents plan, grant number 2020CB001; the project of Youth and Middleaged Scientific and Techno-logical Innovation Leading Talents Program of the Corps, grant number 2018CB006; the China Postdoctoral Science Foundation, grant number 220531; the Funding Project for High Level Talents Research in Shihezi University, grant number RCZK2018C38; and the Project of Shihezi University, grant number ZZZC201915B.

References

[1] H. Jafari, T. Rajaee, and S. Nazif, "An investigation of the possible scenarios for the optimal locating of quality sensors in the water distribution networks with uncertain contamination," *Journal of Water and Health*, vol. 18, no. 5, pp. 704–721, 2020.

[2] K. S. Adu-Manu, F. A. Katsriku, J.-D. Abdulai, and F. Engmann, "Smart river monitoring using wireless sensor networks," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8897126, 19 pages, 2020.

[3] S. Cheng and C. Li, "Research and design of water quality monitoring system based on automatic monitoring of mobile aeration equipment," in *AIP conference proceedings*, Chongqing, China, 2019no. 1, Article ID 020049.

[4] A. B. S. Yıldız, N. Pholdee, S. Bureerat, A. R. Yıldız, and S. M. Sait, "Sine-cosine optimization algorithm for the conceptual design of automobile components," *Materials Testing*, vol. 62, no. 7, pp. 744–748, 2020.

[5] S. E. Bouzid, Y. Seresstou, K. Raoof, M. N. Omri, M. Mbarki, and C. Dridi, "MOONGA: multi-objective optimization of wireless network approach based on genetic algorithm," *IEEE Access*, vol. 8, pp. 105793–105814, 2020.

[6] M. Elhoseny, A. Tharwat, A. Farouk, and A. E. Hassanien, "K-coverage model based on genetic algorithm to extend WSN lifetime," *IEEE Sensors Letters*, vol. 1, no. 4, article 7500404, pp. 1–4, 2017.

[7] J. Li, Z. Luo, and J. Xiao, "A hybrid genetic algorithm with bidirectional mutation for maximizing lifetime of heterogeneous wireless sensor networks," *IEEE Access*, vol. 8, pp. 72261–72274, 2020.

[8] J. Wang, Y. Cao, B. Li, H. J. Kim, and S. Lee, "Particle swarm optimization based clustering algorithm with mobile sink for WSNs," *Future Generation Computer Systems*, vol. 76, pp. 452–457, 2016.

[9] S. C. Manju and B. Kumar, "Genetic algorithm-based metaheuristic for target coverage problem," *IET Wireless Sensor Systems*, vol. 8, no. 4, pp. 170–175, 2018.

[10] Z. Jiao, L. Zhang, M. Xu, C. Cai, and J. Xiong, "Coverage control algorithm-based adaptive particle swarm optimization and node sleeping in wireless multimedia sensor networks," *IEEE Access*, vol. 7, pp. 170096–170105, 2019.

[11] T. Qasim, M. Zia, Q. A. Minhas et al., "An ant colony optimization based approach for minimum cost coverage on 3-D grid in wireless sensor networks," *IEEE Communications Letters*, vol. 22, no. 6, pp. 1140–1143, 2018.

[12] N. Panagant, N. Pholdee, S. Bureerat, A. R. Yıldız, and S. M. Sait, "Seagull optimization algorithm for solving real-world design optimization problems," *Materials Testing*, vol. 62, no. 6, pp. 640–644, 2020.

[13] B. S. Yıldız, A. R. Yıldız, N. Pholdee, S. Bureerat, S. M. Sait, and V. Patel, "The Henry gas solubility optimization algorithm for optimum structural design of automobile brake components," *Materials Testing*, vol. 62, no. 3, pp. 261–264, 2020.

[14] B. S. Yıldız, "Optimal design of automobile structures using moth-flame optimization algorithm and response surface methodology," *Materials Testing*, vol. 62, no. 4, pp. 371–377, 2020.

[15] B. S. Yıldız, "The mine blast algorithm for the structural optimization of electrical vehicle components," *Materials Testing*, vol. 62, no. 5, pp. 497–502, 2020.

[16] B. S. Yıldız, A. R. Yıldız, E. İ. Albak, H. Abderazek, S. M. Sait, and S. Bureerat, "Butterfly optimization algorithm for optimum shape design of automobile suspension components," *Materials Testing*, vol. 62, no. 4, pp. 365–370, 2020.

[17] H. Abderazek, A. R. Yildiz, and S. Mirjalili, "Comparison of recent optimization algorithms for design optimization

of a cam-follower mechanism," *Knowledge-Based Systems*, vol. 191, p. 105237, 2020.

[18] E. Kurtuluş, A. R. Yıldız, S. M. Sait, and S. Bureerat, "A novel hybrid Harris hawks-simulated annealing algorithm and RBF-based metamodel for design optimization of highway guardrails," *Materials testing*, vol. 62, no. 3, pp. 251–260, 2020.

[19] A. R. Yıldız, B. S. Yıldız, S. M. Sait, S. Bureerat, and N. Pholdee, "A new hybrid Harris hawks-Nelder-Mead optimization algorithm for solving design and manufacturing problems," *Materials Testing*, vol. 61, no. 8, pp. 735–743, 2019.

[20] F. Hamza, H. Abderazek, S. Lakhdar, D. Ferhat, and A. R. Yıldız, "Optimum design of cam-roller follower mechanism using a new evolutionary algorithm," *The International Journal of Advanced Manufacturing Technology*, vol. 99, no. 5-8, pp. 1267–1282, 2018.

[21] Y. Sun, Y. Hu, L. Chen, H. Liu, J. Chen, and B. Lv, "The coverage optimization method for underwater sensor network based on VF-PSO algorithm," in *2020 Chinese Control And Decision Conference (CCDC)2020 Chinese Control And Decision Conference (CCDC)*, pp. 2008–2013, Hefei, China, 2020.

[22] D. Arivudainambi, S. Balaji, and T. S. Poorani, "Sensor deployment for target coverage in underwater wireless sensor network," in *2017 International Conference on Performance Evaluation and Modeling in Wired and Wireless Networks (PEMWN)*, pp. 1–6, Paris, 2017.

[23] S. Iyer and D. V. Rao, "Genetic algorithm based optimization technique for underwater sensor network positioning and deployment," *IEEE.2015 IEEE underwater technology (UT)*, pp. 1–6, Chennai, India, 2015.

[24] Y. Zhang, M. Wang, J. Liang, H. Zhang, W. Chen, and S. Jiang, "Coverage enhancing of 3D underwater sensor networks based on improved fruit fly optimization algorithm," *Soft Computing*, vol. 21, no. 20, pp. 6019–6029, 2017.

[25] P. Jiang, Y. Feng, and F. Wu, "Underwater sensor network redeployment algorithm based on wolf search," *Sensors*, vol. 16, no. 10, article 1754, 2016.

[26] C. W. Ou and S. Ranka, "Parallel remapping of adaptive problems," *Journal of Parallel and Distributed Computing*, vol. 42, no. 2, pp. 109–121, 1997.

[27] X. L. Li and J. X. Qian, "Studies on artificial fish swarm optimization algorithm based on decomposition and coordination techniques," *Journal of Circuits and Systems*, vol. 1, pp. 1–6, 2003.

[28] C.-R. Wang, C.-L. Zhou, and J.-W. Ma, "An improved artificial fish-swarm algorithm and its application in feed-forward neural networks," in *2005 International conference on machine learning and cybernetics*, vol. 5, pp. 2890–2894, Guangzhou, China, 2005.

[29] X. Zhou, Z. Wang, D. Li, H. Zhou, Y. Qin, and J. Wang, "Guidance systematic error separation for mobile launch vehicles using artificial fish swarm algorithm," *IEEE Access*, vol. 7, pp. 31422–31434, 2019.

[30] L. I. Zhanwu, C. Yizhe, K. O. Yingxin, Y. A. Haiyan, X. U. An, and L. I. You, "Approach to WTA in air combat using IAFSA-IHS algorithm," *Journal of Systems Engineering and Electronics*, vol. 29, no. 3, pp. 519–529, 2018.

[31] G. Zhou, Y. Li, Y. C. He, X. Wang, and M. Yu, "Artificial fish swarm based power allocation algorithm for MIMO-OFDM relay underwater acoustic communication," *Iet Communications*, vol. 12, no. 9, pp. 1079–1085, 2017.

Hindawi
Journal of Sensors
Volume 2021, Article ID 6613574, 16 pages
https://doi.org/10.1155/2021/6613574

Research Article

System-Level Temperature Compensation Method for the RLG-IMU Based on HHO-RVR

Hao Liang[1,2] **Yumin Tao**[2] **Meijiao Wang**[2] **Yu Guo**[1] **and Xingfa Zhao**[1,2]

[1]*School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China*
[2]*Beijing Aerospace Times Laser Inertial Technology Company, Ltd., Beijing 100094, China*

Correspondence should be addressed to Yu Guo; guoyu@njust.edu.cn

Received 22 December 2020; Revised 10 January 2021; Accepted 21 January 2021; Published 13 February 2021

Academic Editor: Bin Gao

The ring laser gyro inertial measurement unit has many systematic error terms and influences each other. These error terms show a complex nonlinear drift that cannot be ignored when the temperature changes, which seriously affects the stability time and output accuracy of the system. In this paper, a system-level temperature modeling and compensation method is proposed based on the relevance vector regression method. First, all temperature-related parameters are modeled; meanwhile, the Harris hawks optimization algorithm is used to optimize each model parameter. Then, the system compensation is modeled to stabilize the system output to the desired temperature. Compared with the least square method, the fitting performance comparison and the system dynamic compensation experiment prove this method's superiority. The root mean square error, the mean absolute error, the *R*-squared, and the variance of residual increased by an average of 35.27%, 39.29%, 2.29%, and 30.34%, respectively.

1. Introduction

In the inertial sensors, the ring laser gyroscope (RLG) [1–3] has many advantages such as small random drift, wide dynamic range, fast startup speed, and high reliability. It has a high value in the field of inertial navigation. As a representative of strap-down inertial navigation, the ring laser gyroscope inertial measurement unit (RLG-IMU) [4–6], which uses RLG as its core measurement component, is widely used in the field of inertial navigation for its outstanding advantages of small size, high precision, strong environmental adaptability, and low cost. How to improve the performance of the RLG and the RLG-IMU has always been the focus of researchers. After years of research, researchers have done much work on various physical fields that affect the system performance, such as temperature, magnetic field, vibration environment, and how to improve the performance of the RLG-IMU under the influence of various physical fields.

Temperature is an important factor affecting the output precision and startup stability time of the RLG-IMU. On the one hand, the operating temperature range of the RLG-IMU is extensive, which requires strong adaptability of the system. On the other hand, many RLG-IMU devices have a temperature control system composed of the heating part, temperature feedback part, and temperature control part. At the same time, the RLG-IMU has many circuit components, which also generate heat. Other components inside the IMU, such as magnetic coils, also generate heat. After the RLG-IMU is started, a large amount of heat will be generated inside the system. The internal temperature changes, and IMU's output is unstable due to the internal heating part. Such changes in the system's internal and external temperature affect the RLGs and quartz flexible accelerometers' performance and the RLG-IMU internal structure's characteristics. The change of structural characteristics leads to the thermal deformation of inertial devices, which causes a lot of fluctuation and error in output data of the RLG-IMU and limits the startup time and performance of the RLG-IMU. Therefore, it is necessary to study the temperature effect of the IMU system to improve its temperature adaptability and reduce the system stability time.

At present, there are two methods to solve the influence of temperature on the IMU system: temperature control

and temperature compensation. The method of temperature control makes the system quickly heat up to reach the temperature control point, to make the system stable as soon as possible. After the RLG-IMU is started, the internal temperature control system starts to work. The system temperature generally reaches a relatively stable state after a relatively long time (generally dozens of minutes). At this time, the output of the RLG-IMU is stable, and the output accuracy can meet the requirements. Before that, the output of the IMU was unstable due to the drastic temperature change. However, this method's problem is that temperature control requires a "long setting time" or called a "long startup time." The accuracy of the temperature control system will also affect the IMU's performance, and additional hardware needs to be added, and the system's power consumption increased. The method of temperature compensation only needs to obtain the error model under the temperature, and the compensation can be started at the startup of the IMU. Theoretically, the system's output can be stable as soon as it is powered on, but the accuracy of the model establishment will affect the accuracy of the compensation. How to reduce the startup stable time by using the system temperature compensation method is the purpose of this paper. By modeling the system measurement equation's parameters under different temperatures, the output of the RLG-IMU at different temperatures can be converted to the output at a stable temperature. In this way, the operation of the RLG-IMU will not be affected by the change of temperature so that IMU can meet the requirements without "long setting time."

Several methods have been used for the RLG's or quartz flexible accelerometer's zero position temperature compensation, such as stepwise regression [7, 8], artificial neural network [9–12], support vector machine [13–15], and K-mean [16]. The IMU system's various error terms are very complicated, including the zero position and the influence of the scale factor. The system structure changes will also affect the zero position and scale factor of the sensors. These factors not only are affected by temperature but also have a coupling relationship with each other. Compensating for the RLG or quartz flexible accelerometer zero position alone, without considering the effects of other error terms, cannot cover the system's error, which will affect the system's compensation accuracy. Therefore, all drifts must be compensated at the system level. At present, the least square method (LSM) is usually used for system-level compensation in engineering [17–20]. The IMU has more than a dozen parameter items related to temperature. When these parameters change with temperature in a sophisticated nonlinear manner, this traditional modeling method's fitting accuracy is limited.

Machine learning has powerful capabilities in the regression and prediction of complex functions. Based on the Bayesian framework, Tipping proposed the relevance vector machine [21–23]. Compared with the support vector machine, the relevance vector machine uses fewer vectors and has stronger sparsity. Although the training time is longer than that of the support vector machine, the prediction time is much less than that of the support vector machine. The kernel function does not need to meet Mercer's condition: in a finite input space, the function K is a map. If the kernel matrix is positive semidefinite, then the function K can be a kernel function. In SVM theory, the kernel function must satisfy Mercer's condition. In relevance vector machine theory, because of the difference between the relevance vector machine and the SVM architecture, the kernel function does not have to satisfy Mercer's condition, so more kernel functions can be selected. In the case of fewer training data samples, it can ensure excellent generalization ability. The method of using the relevance vector machine for regression is called relevance vector regression (RVR) [24, 25].

In order to solve the problem of high-precision compensation of the RLG-IMU system-level temperature, this paper proposes a system-level temperature error model and compensation method for the RLG-IMU based on the RVR. According to the IMU system's input-output model, all the parameters that affect the output are modeled at the system level and compensated so that the system-level temperature error compensation is more comprehensively achieved. Since the setting of the relevance vector machine kernel function's width parameter has a severe impact on the regression accuracy, it is necessary to optimize this parameter. The Harris hawks optimization (HHO) [26] is a novel metaheuristic [27, 28] optimization algorithm. Compared with the genetic algorithm [29] and particle swarm optimization algorithm [30], it has fast optimization speed and high precision. In this paper, the HHO algorithm is used to optimize the kernel width parameter in the relevance vector machine to improve the regression accuracy of the model, so the method is called HHO-RVR.

For the RLG-IMU system, the influence of temperature on the RLG-IMU system is multifaceted. Although the influence of the temperature on the RLG-IMU is multifaceted, all the effects will be reflected in the measured output pulse at different temperatures. According to the pulse-angular velocity equation and the pulse-apparent acceleration equation, if the objective angular velocity and apparent acceleration are considered real and do not change with temperature, the pulse change caused by temperature is caused by other equation parameters with the change of temperature. This paper is aimed at obtaining the fitting model of the parameters in the equation varying with temperature. When the system works at all temperatures, the system's output pulse can be converted into a pulse output at a selected temperature. Thus, the output of the system is more stable when the temperature changes. This compensation method does not need to consider every factor affected by temperature change in the system. Therefore, we call this compensation method to be system-level temperature compensation.

This paper's structure is as follows: Section 1 introduces the background, current problems, and this paper's work. Section 2 analyzes the influence of temperature on the IMU system. Section 3 introduces the relevance vector machine regression theory and the HHO optimization process and establishes the system parameters' temperature and compensation models. Section 4 introduces the experimental methods, experimental results, and analyses to verify the method's effectiveness and superiority. Section 5 summarizes the whole paper.

2. Analysis of Temperature Effects

As shown in Figure 1, a typical RLG-IMU system consists of three orthogonal RLGs and three orthogonal quartz flexible accelerometers mounted on the base. The IMU measures angular velocity in three directions with three RLGs (G_x, G_y, and G_z) and apparent acceleration in three directions with three quartz flexible accelerometers (A_x, A_y, and A_z).

The influence of temperature on IMU mainly includes three aspects:

(1) Impact on the RLG

The laser gyroscope is based on the Sagnac effect principle and uses the optical path difference to measure the rotational angular velocity. The ring laser gyro is essentially an active ring laser. It is a laser source filled with a helium-neon mixture. The effect of temperature on the gyro is comprehensive [31], such as the change of material characteristics of the laser gyroscope, the influence on the length of the resonant cavity and its coplanarity, the change of the gas flow rate, the change of the characteristics of the mirror, the deformation of the capillary, changes of lock-in characteristics, Langmuir flow, and discharge symmetry. These changes are reflected in changes in the scale factor and zero bias [32].

(2) Impact on the quartz flexible accelerometer

When the quartz flexible accelerometer is started, the fluctuation of the internal temperature field will cause the differential capacitance sensor to generate a larger current, further increasing the moment coil's temperature. The increase in temperature affects the magnetic materials and torque coils in the accelerometer, causing changes in the magnetic flux and causing the accelerometer scale factor's temperature drift. Simultaneously, due to the thermal imbalance in the quartz flexible accelerometer, the arm will produce slight distortion, which will affect the stability of the accelerometer zero bias. Besides, changes in the arm length can affect the scale factor of the accelerometer.

(3) Impact on IMU system structure

For the IMU system, the three RLGs and three quartz flexible accelerometers are fixed on the base and cannot be strictly orthogonal. They have relative installation errors. Due to the generally strong rigidity of the base, the installation errors are relatively stable. However, the geometry of the base is also a small change that will occur with the temperature change, and this change will also be reflected in the parameter drift of the single sensor.

From the above analysis, it can be concluded that the influence of temperature on the RLG-IMU system is multifaceted and complex. Temperature affects not only the individual sensor but also the system structure. Structural changes are also reflected in the output of the sensor. It is not straightforward to analyze and compensate for these factors separately. Therefore, this paper uses the system-level compensation method to integrate all the influencing factors to model and compensate directly.

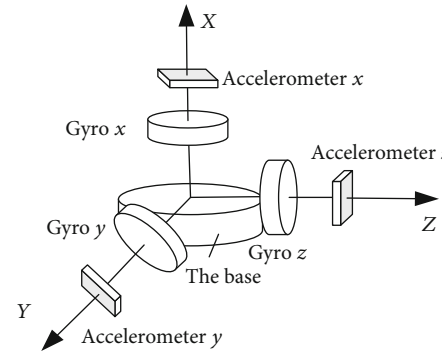


Figure 1: RLG-IMU diagram.

3. Relevance Vector Regression and Temperature Compensation Modeling

3.1. Function Regression Theory of the Relevance Vector Machine. The relevance vector machine is based on the Bayesian theory. In linear regression problems, the input and output can be described as $y = \mathbf{w}^T\mathbf{x} + c$, where $\mathbf{w}$ is the weight vector, $\mathbf{x}$ is the input vector, and c is the offset. When the relationship is nonlinear, it can be expressed as $y = \mathbf{w}^T\phi(\mathbf{x})$ or $y(\mathbf{x}; \mathbf{w})$. Given a set of input target data sets: $\mathbf{S} = \{\mathbf{x}_n, t_n\}_{n=1}^N$, $\mathbf{x}_n \in R^n$, $t_n \in R$, N is the number of data samples. Consider that the target set is data samples superimposed with noise, expressed as $t_n = y(\mathbf{x}_n; \mathbf{w}) + \epsilon_n$, where ϵ_n is the noise of the Gaussian distribution satisfying the mean of 0 and the variance of σ^2, that is, $\varepsilon_n \sim N(0, \sigma^2)$; the conditional probability of the target is [33–35]

$$p(t_n \mid \mathbf{x}) = \mathcal{N}\left(t_n \mid y(\mathbf{x}_n), \sigma^2\right). \tag{1}$$

t_n is independent of each other; then, the likelihood functions of all corresponding data sets are expressed as

$$\begin{aligned} p\left(\mathbf{t} \mid \mathbf{w}, \sigma^2\right) &= \prod_{i=1}^{N} N\left(t_i \mid y(\mathbf{x}_i; \mathbf{w}), \sigma^2\right) \\ &= \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2\right\}, \end{aligned} \tag{2}$$

where $\mathbf{t} = (t_1 \cdots t_N)^T$, $\mathbf{w} = (w_0 \cdots w_N)^T$, and $\mathbf{\Phi}$ is the $N \times (N+1)$ matrix with $\mathbf{\Phi} = [\phi(x_1), \phi(x_2), \cdots, \phi(x_N)]^T$, wherein $\phi(\mathbf{x}_n) = [1, K(\mathbf{x}_n, \mathbf{x}_1), K(\mathbf{x}_n, \mathbf{x}_2), \cdots, K(\mathbf{x}_n, \mathbf{x}_N)]^T$. $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function.

In order to establish the relationship model between input and output, the weight $\mathbf{w}$ in the model is given prior probability.

$$p(\mathbf{w} \mid \boldsymbol{\alpha}) = \prod_{i=0} N\left(w_i \mid 0, \boldsymbol{\alpha}_i^{-1}\right). \tag{3}$$

Hyperparameter $\boldsymbol{\alpha}_i$ is used to describe the inverse variance of each $\mathbf{w}_i$.

The posterior probability of the weight is

$$p\left(\mathbf{w} \mid \mathbf{t}, \boldsymbol{\alpha}, \sigma^2\right) = \frac{p\left(t \mid w, \sigma^2\right) p(w \mid \alpha)}{p\left(t \mid \alpha, \sigma^2\right)} = (2\pi)^{-(N+1)/2} |\boldsymbol{\Sigma}|^{-1/2} \exp \cdot \left\{ -\frac{1}{2} (\mathbf{w} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{w} - \boldsymbol{\mu}) \right\}, \tag{4}$$

where the covariance and mean are $\boldsymbol{\Sigma} = \left(\sigma^{-2} \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} + \mathbf{A}\right)^{-1}$, $\boldsymbol{\mu} = \sigma^{-2} \boldsymbol{\Sigma} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$, and $\mathbf{A} = \operatorname{diag}(\alpha_0, \alpha_1, \cdots, \alpha_N)$.

Use the maximum likelihood method to obtain the optimal solution $\boldsymbol{\alpha}_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}}$ of $p(\mathbf{t} \mid \boldsymbol{\alpha}, \sigma^2)$.

For new inputs $\mathbf{x}_*$, the predicted distribution can be calculated.

$$p\left(t_* \mid \mathbf{t}, \boldsymbol{\alpha}_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}}\right) = \int p\left(t_* \mid \mathbf{w}, \sigma^2_{\mathrm{MP}}\right) p\left(\mathbf{w} \mid \mathbf{t}, \boldsymbol{\alpha}_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}}\right) d\mathbf{w}. \tag{5}$$

Since both of the integrands are Gaussian, this can be calculated.

$$p\left(t_* \mid \mathbf{t}, \boldsymbol{\alpha}_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}}\right) = \mathcal{N}\left(t_* \mid y_*, \sigma^2_*\right) \cdot \begin{cases} y_* = \boldsymbol{\mu}^{\mathrm{T}} \phi(\mathbf{x}_*), \\ \sigma^2_* = \sigma^2_{\mathrm{MP}} + \phi(\mathbf{x}_*)^{\mathrm{T}} \boldsymbol{\Sigma} \phi(\mathbf{x}_*). \end{cases} \tag{6}$$

In equation (2), the kernel function is used to map the feature vector to the high-dimensional space, which can reduce the calculation complexity. Commonly used are the linear kernel function, polynomial kernel function, sigmoid kernel function, and radial basis kernel function. The linear kernel function is mainly used in the case of linear separability. The application effect of the polynomial kernel function on nonlinear data is not ideal, and there are many parameters to be adjusted, which increases the complexity of the model. The sigmoid kernel function has good performance in dealing with nonlinear data, and two parameters need to be adjusted. The radial basis function is a kind of kernel function with robust localization, which can realize nonlinear mapping. The radial basis function is better than the linear kernel function in nonlinear data processing. The shape of the function is a bell-shaped curve, and only one parameter controls the width of the curve. Compared with the polynomial kernel function and the sigmoid kernel function, the radial basis function has fewer parameters, which greatly reduces the model's complexity. The radial basis function is the most widely used kernel function. It has good performance in dealing with large and small samples. This paper uses the radial basis kernel function, whose expression is $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\gamma^2\right)$, where γ denotes the width parameter of the kernel function. Different γ affects the performance of the relevance vector machine, so this parameter must be optimized, and the optimization algorithm based on metaheuristic can significantly improve the optimization efficiency.

3.2. Parameter Optimization Based on HHO. HHO is a novel metaheuristic swarm intelligence algorithm. Compared with the genetic algorithm and the particle swarm optimization algorithm, it has better exploration capabilities and convergence speed [26]. When the hawks catch their prey, they adopt different strategies in different stages. The HHO algorithm can find the optimal value by modeling hawks' behavior in each stage of the predation process.

The HHO algorithm includes the exploration phase and the exploitation phase. In the process of predation, with the increase in the moving distance of prey, the physical strength of prey will gradually decrease, which is called escape energy. The algorithm can transfer from the exploration phase to the exploitation phase through the value range of prey's escape energy. The escape energy is modeled as [26]

$$E = 2E_0 \left(1 - \frac{t}{T}\right), \tag{7}$$

where E denotes the escaping energy of the prey, T denotes the maximum number of iterations, t denotes the current number of iterations, and E_0 denotes the initial energy. E_0 randomly changes inside the interval $(-1, 1)$ at each iteration.

When $|E| \geq 1$, enter the exploration phase. The hawks inhabit different positions randomly and use two strategies to detect prey. An equal probability q is used to distinguish the two strategies. They perch based on the positions of other hawks and the prey when $q < 0.5$ and perch on random tall trees when $q \geq 0.5$. The exploration phase can be expressed as

$$X(t+1) = \begin{cases} X_{\mathrm{rand}}(t) - r_1 |X_{\mathrm{rand}}(t) - 2r_2 X(t)|, & q \geqslant 0.5, \\ \left(X_{\mathrm{prey}}(t) - X_m(t)\right) - r_3(\mathrm{LB} + r_4(\mathrm{UB} - \mathrm{LB})), & q < 0.5, \end{cases} \tag{8}$$

where $X(t+1)$ denotes the position vector of the next-generation hawks, $X(t)$ denotes the position vector of the current-generation hawks, $X_{\mathrm{prey}}(t)$ denotes the position of the prey, r_1, r_2, r_3, r_4, and q are random values between $(0, 1)$ and are updated every generation, LB and UB denote the boundary of the variable, $X_{\mathrm{rand}}(t)$ denotes a random position selected from the current-generation hawks, and $X_m(t)$ denotes the average position of the hawks.

When $|E| < 1$, enter the exploitation phase. At this time, use four siege strategies. r is a random number between $(0, 1)$ and indicates the chance for the prey to escape. $r < 0.5$ indicates that the prey can escape, and $r \geq 0.5$ indicates that the prey cannot escape successfully [26]. The use of soft and hard besiege strategies depends on the prey's energy E. When $|E| \geq 0.5$, it means that the prey still has high energy. At this time, the Hawks adopt the soft besiege strategy, constantly approaching and chasing the prey to consume the energy of the prey. When $|E| < 0.5$, it indicates that the prey's energy is weak. At this time, the Hawks can easily catch the prey by using the hard besiege strategy.

(1) Soft Besiege. When $r \geq 0.5$ and $|E| \geq 0.5$, it means that the prey has enough escape energy and jumps randomly. At this time, the hawks encircle it softly to exhaust the prey, expressed as

$$X(t+1) = \left(X_{\text{prey}}(t) - X(t)\right) - E\left|JX_{\text{prey}}(t) - X(t)\right|, \quad (9)$$

where J is a random number between $(0, 2)$, denoting the random jumping energy of prey, which is used to simulate the prey's activity.

(2) Hard Besiege. When $r \geq 0.5$ and $|E| < 0.5$, it means that the prey is weak and has low escape energy. At this time, the following equation is used to update the position.

$$X(t+1) = X_{\text{prey}}(t) - E\left|X_{\text{prey}}(t) - X(t)\right|. \quad (10)$$

(3) Soft Besiege with Progressive Rapid Dives. When $r < 0.5$ and $|E| \geq 0.5$, the prey has enough energy to escape. At this time, equations (11) and (12) are used to update the position, and the Lévy flight is used to simulate the sudden movement of the hawks and the prey. The Lévy flight is expressed as equation (13), which can further enhance the ability of the algorithm to jump out of the local optimum.

$$Y = X_{\text{prey}}(t) - E\left|JX_{\text{prey}}(t) - X(t)\right|, \quad (11)$$

$$Z = Y + S \times \text{LF}(D), \quad (12)$$

$$\text{LF}(x) = 0.01 \times \frac{u \times \sigma}{|v|^{1/\beta}}, \quad \sigma = \left(\frac{\Gamma(1+\beta) \times \sin(\pi\beta/2)}{\Gamma((1+\beta)/2) \times \beta \times 2^{((\beta-1)/2)}}\right)^{1/\beta}, \quad (13)$$

where u and v are random values inside $(0, 1)$, β is a default constant set to 1.5, and $\Gamma(z)$ is the Gramma function: $\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$.

Compare the two update methods, and select the most suitable position update method according to the fitness function value.

$$X(t+1) = \begin{cases} Y, & \text{if } F(Y) < F(X(t)), \\ Z, & \text{if } F(Z) < F(X(t)). \end{cases} \quad (14)$$

(4) Hard Besiege with Progressive Rapid Dives. When $r < 0.5$ and $|E| < 0.5$, it means that the prey does not have enough energy. At this time, equations (15) and (16) are used to update the position and equation (14) is used to select the appropriate update method.

$$Y = X_{\text{prey}}(t) - E\left|JX_{\text{prey}}(t) - X_m(t)\right|, \quad (15)$$

$$Z = Y + S \times \text{LF}(D). \quad (16)$$

The relevance vector machine is used for data fitting and regression. HHO is used to optimize the kernel width parameter of the relevance vector machine.

The flow of each HHO-RVR is as follows:

Step 1. Initialize the number of HHO populations and the number of iterations. The hawk's position in the HHO algorithm represents the kernel width parameter of the relevance vector machine. The position of each hawk in the first-generation group is randomly distributed.

Step 2. Determine the fitness function. The sample data is randomly divided into five parts, 4 of which are used for the relevance vector machine training, and the remaining 1 part is used for verification, and the test accuracy of the relevance vector machine is used as the fitness function of HHO.

Step 3. Calculate the fitness value of each hawk, and use the hawk corresponding to the optimal fitness as the prey.

Step 4. Update the positions of the hawks, according to equations (7)–(16).

Step 5. Perform iterative operations. Repeat steps 3 and 4 until the number of iterations is satisfied and the calculation is stopped. At this time, the prey position is the optimal kernel width parameter.

3.3. RLG-IMU Temperature Modeling. The physical input of the RLG is the angular velocity, and the measured output is the number of pulses. The computer counts the number of pulses in the sampling time. According to the measurement equation of the system, the angular velocity of the RLG is obtained. The pulse-angular velocity measurement model of the RLGs can be expressed as

$$\frac{\mathbf{N_g}}{\tau} = \mathbf{K_g}\left(\mathbf{E_g}\boldsymbol{\omega}_{\mathbf{gm}} + \mathbf{D_g}\right). \quad (17)$$

$\mathbf{N_g} = [N_{gx}\, N_{gy}\, N_{gz}]^T$ is the vector of the number of pulses output by the RLGs on three orthogonal axes within the sampling time τ. $\mathbf{K_g} = \text{diag}\,(K_{gx}, K_{gy}, K_{gz})$ is the scale factor matrix.$\mathbf{D_g} = [D_{gx}\, D_{gy}\, D_{gz}]^T$ is the bias matrix. $\boldsymbol{\omega}_{\mathbf{gm}} = [\omega_{gxm}\, \omega_{gym}\, \omega_{gzm}]^T$ is the angular velocity component vector of the three axes in the IMU measurement coordinate system.

$$\mathbf{E_g} = \begin{bmatrix} E_{gxx} & E_{gyx} & E_{gzx} \\ E_{gxy} & E_{gyy} & E_{gzy} \\ E_{gxz} & E_{gyz} & E_{gzz} \end{bmatrix} \quad (18)$$

is the installation error matrix of the RLG relative inertial coordinate system for each axis.

In equation (17), the output vector $\mathbf{N_g}$ of the RLGs changes with temperature and $\mathbf{E_g}$ is very little affected by temperature. Because the influence of temperature on the RLG is multifaceted, the change of the sensor's output to the measured angular velocity caused by the temperature change is directly reflected in the scale factor matrix $\mathbf{K_g}$ and

bias matrix $\mathbf{D_g}$, and the structural changes will also be reflected on $\mathbf{K_g}$ and $\mathbf{D_g}$, so equation (17) can be expressed as

$$\frac{\mathbf{N_g}(T)}{\tau} = \mathbf{K_g}(T)\left(\mathbf{E_g}\boldsymbol{\omega}_{\mathbf{gm}} + \mathbf{D_g}(T)\right). \tag{19}$$

The physical input of the quartz flexible accelerometer is the apparent acceleration, and the measured output is the number of pulses. The computer counts the number of pulses in the sampling time. According to the measurement equation of the system, the apparent acceleration of the quartz flexible accelerometer is obtained. The pulse-apparent acceleration measurement model of the quartz flexible accelerometers is expressed as

$$\frac{\mathbf{N_a}}{\tau} = \mathbf{K_a}(\mathbf{E_a}\mathbf{A_{am}} + \mathbf{D_a}). \tag{20}$$

$\mathbf{N_a} = [N_{ax}\, N_{ay}\, N_{az}]^T$ is the vector of the number of pulses output by the quartz flexible accelerometers on three orthogonal axes within the sampling time τ. $\mathbf{K_a} = \text{diag}\,(K_{ax}, K_{ay}, K_{az})$ is the scale factor matrix of the quartz flexible accelerometers. Since the quartz flexible accelerometer is very sensitive to positive and negative scale factors, it is divided into positive and negative: the positive scale factor matrix is expressed as $\mathbf{K_{ap}} = \text{diag}\,(K_{apx}, K_{apy}, K_{apz})$ and the negative scale factor matrix is expressed as $\mathbf{K}_{\text{an}} = \text{diag}\,(K_{anx}, K_{any}, K_{anz})$. $\mathbf{D_a} = [D_{ax}\, D_{ay}\, D_{az}]^T$ is the bias matrix of the quartz flexible accelerometers. $\mathbf{A_{am}} = [A_{axm}\, A_{aym}\, A_{azm}]^T$ is the apparent acceleration component vector of the three axes in the IMU measurement coordinate system.

$$\mathbf{E_a} = \begin{bmatrix} E_{axx} & E_{ayx} & E_{azx} \\ E_{axy} & E_{ayy} & E_{azy} \\ E_{axz} & E_{ayz} & E_{azz} \end{bmatrix} \tag{21}$$

is the installation error matrix of the quartz flexible accelerometer relative inertial coordinate system for each axis. Similarly, equation (20) is expressed as

$$\frac{\mathbf{N_a}(T)}{\tau} = \mathbf{K_a}(T)(\mathbf{E_a}\mathbf{A_{am}} + \mathbf{D_a}(T)). \tag{22}$$

In equations (19) and (22), the input parameter affected by temperature is called the temperature-related parameter, and there are 15 temperature-related parameters in total.

Define a set $\mathbf{P} = \{P_i \mid i = 1 \cdots 15\}$ to denote 15 temperature-related parameters. Make $\{P_i \mid i = 1 \cdots 6\}$ denote each element in $\{\mathbf{D_g}, \mathbf{K_g}\}$ and $\{P_i \mid i = 7 \cdots 15\}$ denote each element in $\{\mathbf{D_a}, \mathbf{K_{ap}}, \mathbf{K}_{\text{an}}\}$.

Since the temperature measurement position of each RLG and quartz flexible accelerometer is different, the temperature measurement curve of each temperature-related parameter will be different. Define a temperature set $\mathbf{T} = \{T_i \mid i = 1 \cdots 15\}$. T_i denotes the temperature change vector of each temperature-related parameter and corresponds one-to-one with P_i; then, the corresponding relationship between set P and temperature can be expressed as $\mathbf{P}(\mathbf{T}) = \{P_i(T_i) \mid i = 1 \cdots 15\}$. From this, 15 models can be built as follows:

$$P_i = M_i(w_i, T_i), \quad i = 1, 2, \cdots, 15, \tag{23}$$

where $M_i(w_i, T_i)$ is the model corresponding to each temperature-related parameter of the IMU and w_i is the optimal width parameter of the kernel function of each model.

3.4. System Compensation Model. Ideally, in equations (19) and (22), angular velocity $\boldsymbol{\omega}_{\mathbf{gm}}$ and apparent acceleration $\mathbf{A_{am}}$ are independent of temperature. First, the output models of angular velocity and apparent acceleration can be obtained according to IMU measurement equations (19) and (22) and the models of temperature-related parameters. The values of the temperature-related parameters at the desired temperature T_E are then selected as constants in the angular/apparent acceleration measurement equation to obtain the compensation pulse. The desired temperature T_E is a hypothetical stable working state. The purpose is that when the system works at any temperature, the measured value of physical quantity can be converted to the value at the desired temperature so that the system will not be affected by the temperature change. The desired temperature is not unique. It can be set according to the actual needs and requirements. No matter how much the desired temperature is set, the measured value of different temperatures can be converted to the measured value at the desired temperature by compensating for the measurement equation's parameters to realize the temperature compensation.

For the RLGs, according to equation (19), the output model of $\boldsymbol{\omega}_{\mathbf{gm}}$ is

$$\boldsymbol{\omega}_{\mathbf{gm}} = \mathbf{E_g^{-1}}\left[\left(\frac{\mathbf{N_g}(T_g)}{\tau}\right)\mathbf{K_g^{-1}}(T_g) - \mathbf{D_g}(T_g)\right]. \tag{24}$$

T_g corresponding to each RLG is T_{gx}, T_{gy}, and T_{gz}.

The compensation equation of RLG output pulse at the desired temperature T_E is

$$\frac{\mathbf{N_g}(C)}{\tau} = \mathbf{K_g}(T_E)\left(\mathbf{E_g}\boldsymbol{\omega}_{\mathbf{gm}} + \mathbf{D_g}(T_E)\right). \tag{25}$$

$\mathbf{N_g}(C) = [N_{gx}(C)\, N_{gy}(C) N_{gz}(C)]^T$ is the output pulse of the three RLGs after compensation. $\mathbf{K_g}(T_E)$ and $\boldsymbol{D_g}(T_E)$ are, respectively, the scale factor matrix and bias matrix of the three RLGs at the desired temperature T_E.

By bringing equation (24) into (25), there is

$$\frac{\mathbf{N_g}(C)}{\tau} = \mathbf{K_g}(T_E)\left\{\mathbf{E_g}\left\{\mathbf{E_g^{-1}}\left[\left(\frac{\mathbf{N_g}(T_g)}{\tau}\right)\mathbf{K_g^{-1}}(T_g) - \mathbf{D_g}(T_g)\right]\right\} + \mathbf{D_g}(T_E)\right\}. \tag{26}$$

Corresponding to different RLGs, T_g includes T_{gx}, T_{gy}, and T_{gz}. The data set $\{T_g, \mathbf{K_g}\}$ can be obtained by

temperature experiment. The data set was used to train the relevance vector machine. At the same time, the optimization process introduced in Section 3.2 is used to optimize the kernel width parameter of the relevance vector machine, so as to obtain the optimized $\mathbf{K_g}(T_g)$ model. Similarly, the model of $\mathbf{D_g}(T_g)$ can be obtained. Since the temperature models of $\mathbf{K_g}(T_g)$ and $\mathbf{D_g}(T_g)$ have been obtained, it is possible to compensate for the output $\mathbf{N_g}(T_g)$ at all temperatures, thereby compensating for $\mathbf{N_g}$ to $\mathbf{N_g}(T_E)$ at different temperatures.

For the quartz flexible accelerometers, according to equation (22), the output model of $\mathbf{A_{am}}$ is

$$\mathbf{A_{am}} = \mathbf{E_a^{-1}}\left[\left(\frac{\mathbf{N_a}(T_a)}{\tau}\right)\mathbf{K_a^{-1}}(T_a) - \mathbf{D_a}(T_a)\right]. \tag{27}$$

T_a corresponding to each quartz flexible accelerometer is T_{ax}, T_{ay}, and T_{az}.

The compensation equation of quartz flexible accelerometer output pulse at the desired temperature T_E is

$$\frac{\mathbf{N_a}(C)}{\tau} = \mathbf{K_a}(T_E)(\mathbf{E_a}\mathbf{A_{am}} + \mathbf{D_a}(T_E)). \tag{28}$$

$\mathbf{N_a}(C) = [N_{ax}(C)\, N_{ay}(C)\, N_{az}(C)]^T$ is the output pulse of the three quartz flexible accelerometers after compensation. $\mathbf{K_a}(T_E)$ and $\mathbf{D_a}(T_E)$ are, respectively, the scale factor matrix and bias matrix of the three quartz flexible accelerometers at the desired temperature T_E.

By bringing equation (27) into (28), there is

$$\frac{\mathbf{N_a}(C)}{\tau} = \mathbf{K_a}(T_E)\left\{\mathbf{E_a}\left\{\mathbf{E_a^{-1}}\left[\left(\frac{\mathbf{N_a}(T_a)}{\tau}\right)\mathbf{K_a^{-1}}(T_a) - \mathbf{D_a}(T_a)\right]\right\} + \mathbf{D_a}(T_E)\right\}. \tag{29}$$

Corresponding to different quartz flexible accelerometers, T_a includes T_{ax}, T_{ay}, and T_{az}. Similarly, the models of $\mathbf{K_a}(T_a)$ and $\mathbf{D_a}(T_a)$ can be obtained. Since the temperature models of $\mathbf{K_a}(T_a)$ and $\mathbf{D_a}(T_a)$ have been obtained, it is possible to compensate for the output $\mathbf{N_a}(T_a)$ at all temperatures, thereby compensating for $\mathbf{N_a}(T_a)$ to $\mathbf{N_a}(T_E)$ at different temperatures.

Equations (26) and (29) are the compensation model of the IMU.

4. Experiment and Analysis

4.1. Data Acquisition and Modeling. The IMU used in the experiment uses three RLGs and three quartz flexible accelerometers as measurement sensors. The scale factor of the selected RLG on the IMU system is approximately 2.14 pulses/arcsec. The scale factor of the selected accelerometer on the IMU system is approximately $2.4e + 4$ pulses/(s.g0). The RLG-IMU is fixed on a two-axis rotating platform in the temperature chamber. Set the temperature chamber to -5°C, and keep it for 4 hours to ensure the temperature inside the IMU and the chamber temperature are consistent. The temperature chamber is set to 7000 minutes from -5°C to 65°C. In the process of temperature rise in the temperature chamber, the IMU continuously carries out rotation calibration work in the temperature chamber. The IMU uses the systematic calibration method, and the rotation speed is 20 deg/s. Various errors will be excited by the continuous rotation of multiple positions. Using the Kalman filter, various error parameters of the RLGs and accelerometers can be calculated, that is, $\mathbf{K_g}$, $\mathbf{D_g}$, $\mathbf{E_g}$, $\mathbf{K_a}$, $\mathbf{D_a}$, and $\mathbf{E_a}$ in Equations (17) and (20). Each calibration takes approximately 40 minutes. A total of 175 calibrations have been completed within the set temperature range and time. The IMU temperature change is caused not only by the temperature change in the temperature chamber but also by the heating part inside the IMU. The temperature sensor measurement shows that the temperature change between every two calibrations is about 0.4°C. The temperature change is tiny. Therefore, the temperature is considered to be stable in one calibration process. Through the experimental process described above, 175 sets of calibration data of the IMU can be obtained over the entire experimental temperature range. After offline calibration data processing, 175 sets of temperature-related parameters at 175 temperature points can be obtained.

Through experiments in the temperature chamber, the data set P_i of each temperature-related parameter and the corresponding temperature set T_i are obtained. There are 15 such sets in total. It can be expressed as $\mathbf{S} = \{\mathbf{S_i}\}_{i=1}^{15}$, $\mathbf{S_i} = \{T_{ij}, P_{ij}\}_{j=1}^{175}$. Each $\mathbf{S_i}$ is used as the relevance vector machine training data. Use the HHO-RVR process in Section 3.2 to train each relevance vector machine. The kernel width parameter of the relevance vector machine is used as the input variable of the HHO. The total population size is set to 30, the max number of iteration is set to 500, and the value range is set to $[0.01, 200]$. The kernel width parameter with the highest prediction accuracy in the range of values is found using the 5-fold cross-validation method. Each temperature-related parameter is modeled according to this process. The set of models $\mathbf{M} = \{M_i\}_{i=1}^{15}$ in equation (23) can be obtained.

For comparison, the least square method is used to model the 15 temperature-related parameters. The model equation is expressed as

$$\mathbf{P} = \mathbf{A}\mathbf{T}, \tag{30}$$

where $\mathbf{P} = [P_1(t)\, P_2(t) \cdots P_{15}(t)]^T$ is the output vector of the temperature-related parameters,

$$\mathbf{A} = \begin{bmatrix} a_{1,0} & a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,0} & a_{2,1} & a_{2,2} & a_{2,3} \\ \cdots & \cdots & \cdots & \cdots \\ a_{15,0} & a_{15,1} & a_{15,2} & a_{15,3} \end{bmatrix} \tag{31}$$

is the coefficient matrix of the equation, and $\mathbf{T} = [1\ t\ t^2\ t^3]^T$ is the temperature input vector.

Table 1: Model parameters.

The temperature-related parameter	Parameters of the RVR models			Coefficients of the least square method			
	Kernel function width parameter (γ)	Number of relevance vectors	The proportion of relevance vectors (%)	a_0	a_1	a_2	a_3
P_1	32	3	1.71	$-2.11E-08$	$-2.17E-06$	$3.69E-04$	$2.15E-02$
P_2	4.41	7	4	$-1.10E-07$	$5.82E-06$	$-7.12E-05$	$5.32E-02$
P_3	10.69	5	2.86	$-1.06E-07$	$8.23E-06$	$-8.18E-05$	$-5.50E-02$
P_4	6.67	5	2.86	$3.64E-11$	$-2.88E-09$	$9.01E-08$	$2.14E+00$
P_5	11.23	5	2.86	$2.21E-10$	$-2.26E-08$	$8.34E-07$	$2.14E+00$
P_6	25.65	2	1.14	$-1.48E-11$	$2.02E-09$	$4.57E-08$	$2.14E+00$
P_7	32	5	2.86	$4.68E-10$	$-6.57E-07$	$1.67E-04$	$-1.13E-02$
P_8	7.71	8	4.57	$3.72E-09$	$-2.46E-07$	$4.04E-06$	$6.17E-03$
P_9	11.08	9	5.14	$3.10E-09$	$-6.43E-07$	$6.02E-05$	$7.01E-03$
P_{10}	28.92	5	2.86	$-2.50E-05$	$1.11E-02$	$-1.16E+00$	$2.40E+04$
P_{11}	19.42	7	4	$-2.51E-05$	$1.24E-02$	$-4.87E-01$	$2.42E+04$
P_{12}	27.74	5	2.86	$-1.32E-05$	$9.90E-03$	$-9.13E-01$	$2.38E+04$
P_{13}	31.09	6	3.43	$-2.55E-05$	$1.12E-02$	$-1.16E+00$	$2.40E+04$
P_{14}	19.4	7	4	$-2.55E-05$	$1.25E-02$	$-4.93E-01$	$2.42E+04$
P_{15}	15.42	8	4.57	$-1.59E-05$	$1.02E-02$	$-9.19E-01$	$2.38E+04$

Fifteen temperature-related parameters were modeled by the RVR method and the least square method, respectively. In the RVR model, because the radial basis function is used as the kernel function, there is only one kernel width parameter. According to equation (30), the least square method of a cubic fitting polynomial is used. Each fitting polynomial needs to determine four coefficients: the row vector of matrix **A** in equation (30). The kernel width parameters, the number of relevance vectors, the proportion of relevance vectors of the 15 RVR models, and the coefficients of 15 least square fitting polynomials are listed in Table 1.

4.2. Analysis of the Regression Performance of the Models. To compare the fitting and regression performance of HHO-RVR and LSM, HHO-RVR and LSM were used to model the temperature value of 15 temperature-related parameters, with a total of 30 models. Three indicators are used to measure the fitting and regression performance of the two modeling methods:

(1) Root Mean Square Error [36]. It is used to measure the error and dispersion between the regression value and the real value. The smaller the value, the smaller the estimation error and the smaller the error dispersion. The equation is as follows:

$$f(\mathbf{X}, \Phi) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\Phi(x_i) - y_i)^2}. \tag{32}$$

(2) Mean Absolute Error [37]. It is used to measure the average absolute errors. Reflect the actual situation of the predicted value error. The smaller the value, the smaller the estimated error. The equation is as follows:

$$f(\mathbf{X}, \Phi) = \frac{1}{m}\sum_{i=1}^{m}|\Phi(x_i) - y_i|. \tag{33}$$

(3) R-Squared. It is used to measure the fitting accuracy of the model; the closer to 1, the higher the fitting accuracy. The equation is as follows:

$$f(\mathbf{X}, \Phi) = 1 - \frac{\sum_{i=1}^{m}(\Phi(x_i) - y_i)^2}{\sum_{i=1}^{m}(y_{\text{mean}} - y_i)^2}. \tag{34}$$

In equations (32), (33), and (34), $\mathbf{X} = \{x_i \mid i = 1 \cdots m\}$ is the sample input, Φ is the regression model, $\Phi(x_i)$ is the regression output, y_i is the sample output value, y_{mean} is the mean of the sample output, and m is the number of samples.

The results of the three indicators are shown in Table 2. Compared with the least square method, all the regression test performances of the 15 models were improved by using the HHO-RVR method. The root mean square error value increased by a maximum of 80.12% and an average of 35.27%. The mean absolute error value increased by a maximum of 83.51% and an average of 39.29%. The R-squared value increased by a maximum of 9.15% and an average of 2.29%. The results show that the RVR has better performance than the least square method in terms of error, error dispersion, and fitting accuracy.

As shown in Figure 2, the original data is plotted against the data calculated by regression using HHO-RVR and LSM

Table 2: Performance comparison.

The temperature-related parameter	Root mean square error			Mean absolute error			R-squared		
	The least square method	HHO-RVR	Improvement (%)	The least square method	HHO-RVR	Improvement (%)	The least square method	HHO-RVR	Improvement (%)
P_1	$6.1037E-04$	$5.7205E-04$	6.28	$4.7865E-04$	$4.5685E-04$	4.56	$9.6287E-01$	$9.6739E-01$	0.47
P_2	$8.5874E-04$	$6.7484E-04$	21.42	$6.8025E-04$	$5.4583E-04$	19.76	$9.0215E-01$	$9.3957E-01$	4.15
P_3	$6.6615E-04$	$5.4146E-04$	18.72	$5.3673E-04$	$4.1850E-04$	22.03	$7.8803E-01$	$8.5995E-01$	9.13
P_4	$9.9417E-07$	$9.7066E-07$	2.37	$7.9998E-07$	$7.6307E-07$	4.61	$4.3971E-01$	$4.6591E-01$	5.96
P_5	$1.4264E-06$	$1.0149E-06$	28.85	$1.0347E-06$	$7.8771E-07$	23.87	$9.0700E-01$	$9.5292E-01$	5.06
P_6	$8.3064E-07$	$8.2362E-07$	0.84	$6.4871E-07$	$6.4206E-07$	1.03	$8.9077E-01$	$8.9261E-01$	0.21
P_7	$1.4118E-05$	$8.0060E-06$	43.29	$1.0219E-05$	$6.3037E-06$	38.31	$9.9997E-01$	$9.9999E-01$	0.002
P_8	$1.9559E-05$	$4.3197E-06$	77.91	$1.5416E-05$	$3.3696E-06$	78.14	$9.1229E-01$	$9.9572E-01$	9.15
P_9	$2.3627E-05$	$5.1660E-06$	78.14	$1.9734E-05$	$4.0025E-06$	79.72	$9.9843E-01$	$9.9992E-01$	0.15
P_{10}	$8.5537E-02$	$7.4064E-02$	13.41	$7.0884E-02$	$5.5104E-02$	22.26	$9.9995E-01$	$9.9996E-01$	0.001
P_{11}	$8.6871E-02$	$1.7273E-02$	80.12	$7.5096E-02$	$1.2384E-02$	83.51	$9.9975E-01$	$9.9999E-01$	0.02
P_{12}	$6.4535E-02$	$5.9523E-02$	7.77	$5.2699E-02$	$3.5745E-02$	32.17	$9.9991E-01$	$9.9992E-01$	0.001
P_{13}	$6.7920E-02$	$4.1029E-02$	39.59	$5.5552E-02$	$2.6900E-02$	51.58	$9.9997E-01$	$9.9999E-01$	0.002
P_{14}	$7.5613E-02$	$1.9322E-02$	74.45	$6.5216E-02$	$1.3913E-02$	78.67	$9.9980E-01$	$9.9999E-01$	0.02
P_{15}	$5.8027E-02$	$3.7193E-02$	35.90	$4.8269E-02$	$2.4574E-02$	49.09	$9.9993E-01$	$9.9997E-01$	0.004

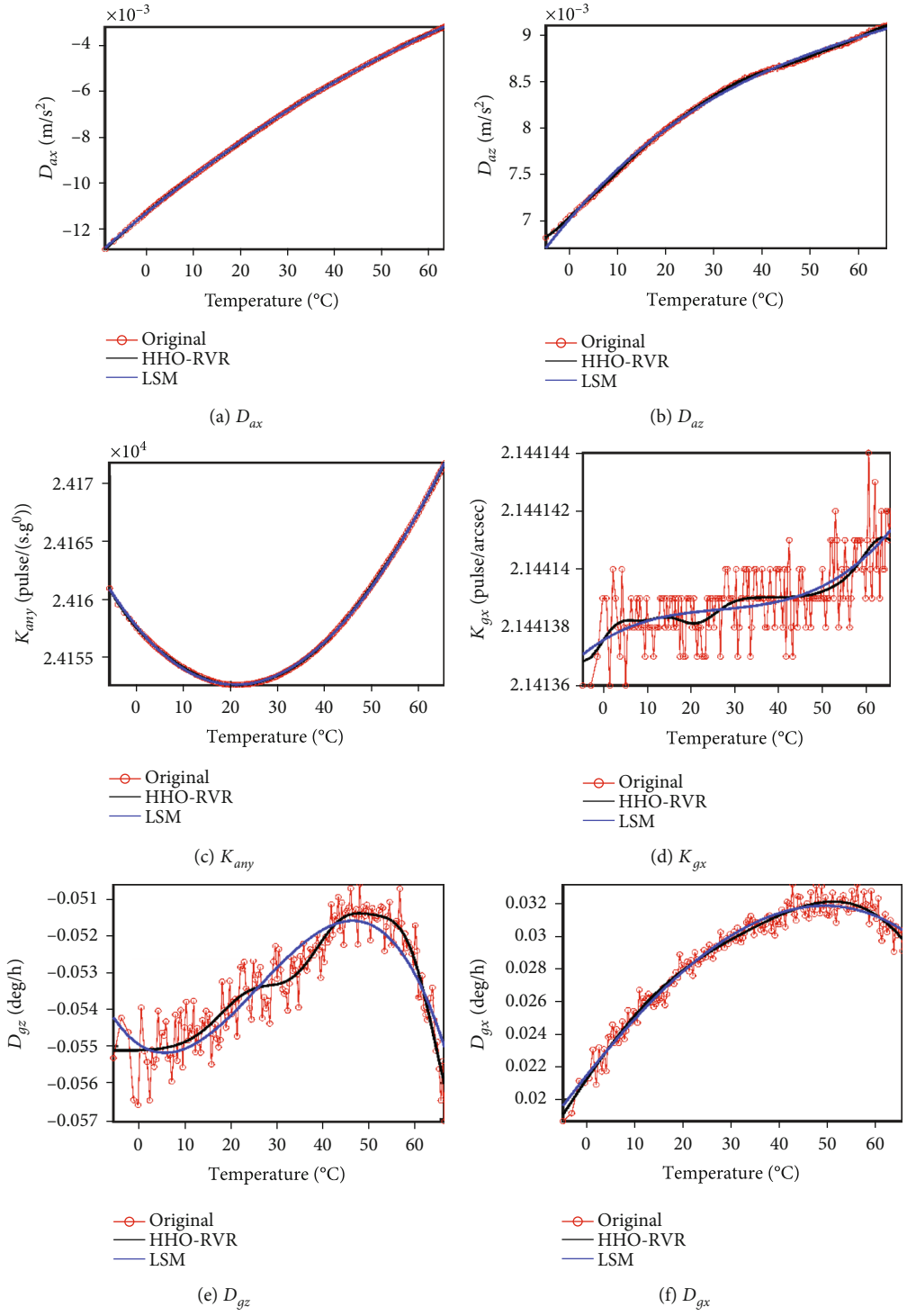


Figure 2: Continued.

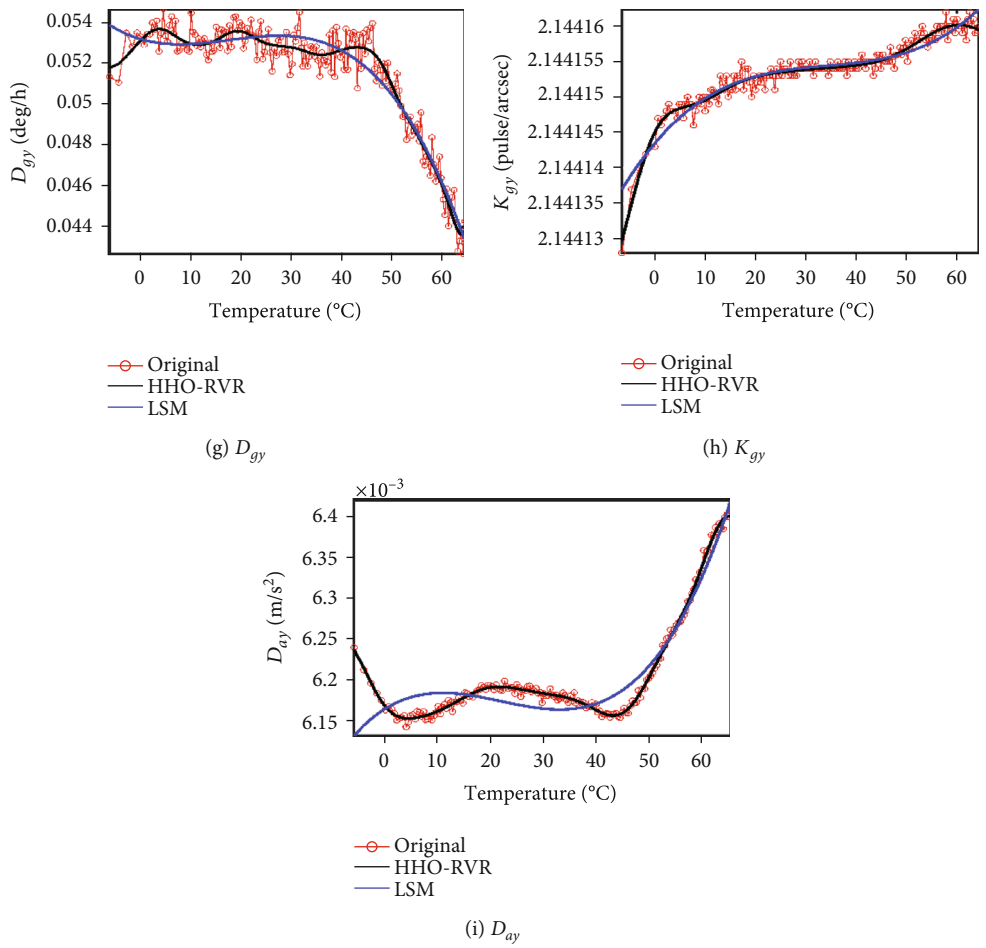


Figure 2: Comparison curve of two regression models.

models, respectively. Observe the goodness of fit between the regression curve using both methods and the original data curve. The horizontal axis is temperature, and the vertical axis is the value of each temperature-related parameter. Since some of the curves do not show any significant differences from the graph, nine representative curves are selected for comparison. According to the order of the figures, they are D_{ax} (the bias of the accelerometer x), D_{az} (the bias of the accelerometer z), K_{any} (the negative scale factor of the accelerometer y), K_{gx} (the scale factor of the gyro x), D_{gz} (the bias of the gyro z), D_{gx} (the bias of the gyro x), D_{gy} (the bias of the gyro y), K_{gy} (the scale factor of the gyro y), and D_{ay} (the bias of the accelerometer y). Figures D_{ax}, D_{az} and K_{any} show that when the original data trend is relatively simple and the fluctuation and dispersion are small, the performance difference between the fitting accuracy of the two methods is not very obvious. Figures K_{gx}, D_{gz}, D_{gx}, and D_{gy} show that the data are characterized by large dispersion, obvious fluctuations, and complex nonlinearity. In this case, the curve fitted by the RVR can accurately fit the complex fluctuations of the original data. However, the least square method can only fit the trend of the whole data and cannot accurately represent the complex fluctuation of the original data, thus affecting the performance of regression and prediction. Figures K_{gy} and D_{ay} show that when the original data dispersion is small, but the data change trend is obvious, the RVR fits the data more accurately than the least square method.

4.3. System Dynamic Compensation. In order to observe the system's dynamic performance after temperature compensation, the IMU dynamic data is used for verification, which can better characterize the compensation performance when the IMU is in motion.

Take 50°C as the desired compensation temperature. First, the temperature compensation model of the temperature-related parameters is obtained in Section 4.1, and the gyro pulse data at the corresponding temperature are brought into equations (26) and (29) to get the compensated pulse data at all temperature points. Secondly, the compensated data are used for calibration operations to obtain the temperature-related parameters of compensated dynamic data. Finally, each temperature-related parameter calculated after compensation is subtracted from the temperature-related parameter at the desired temperature. Two methods are used to model and regress the temperature-related parameters, and the temperature-related parameters are compensated to the desired temperature. The compensated value is subtracted from the measured value to obtain the compensation residual. The smaller the fluctuation of the

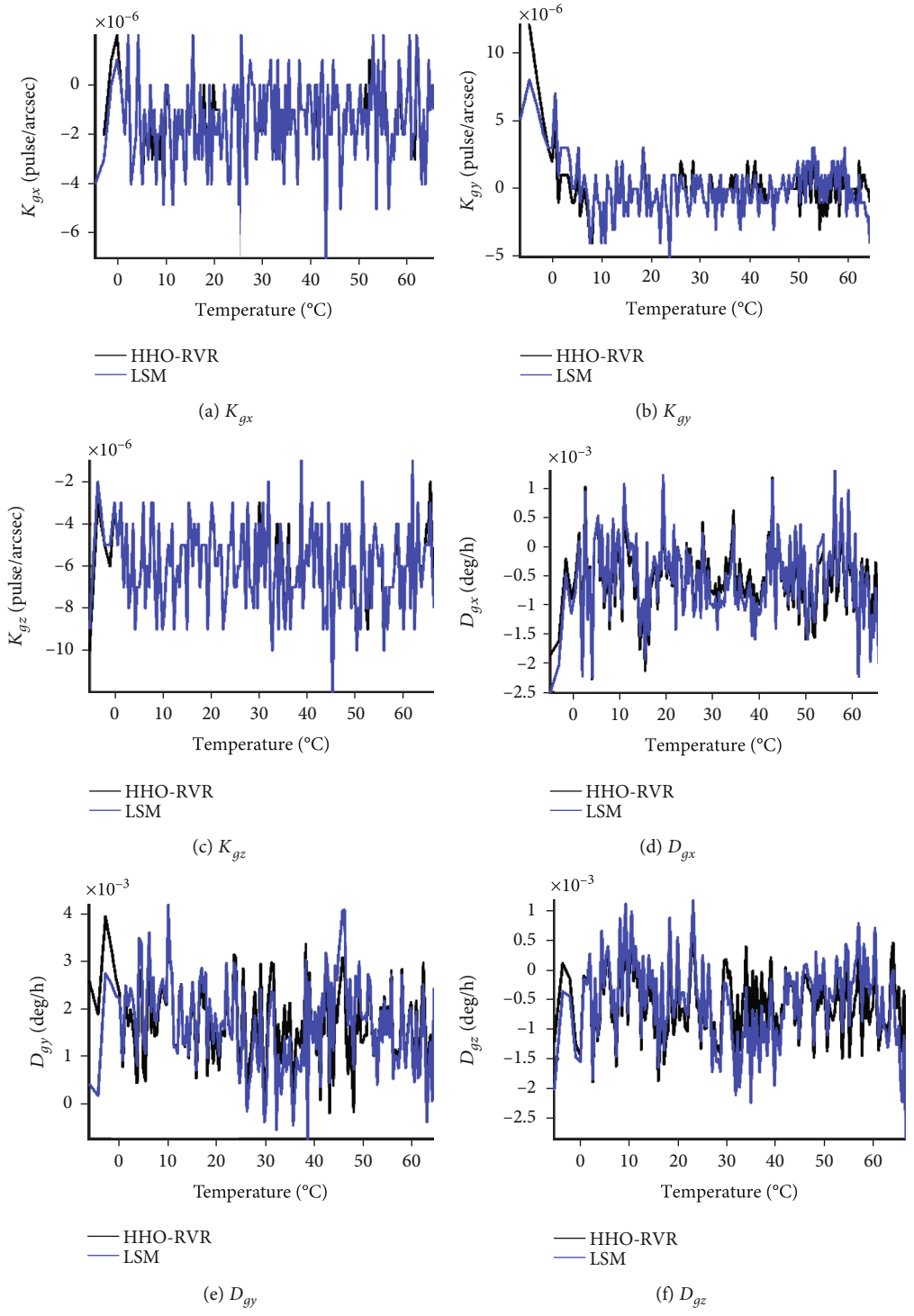


Figure 3: Continued.

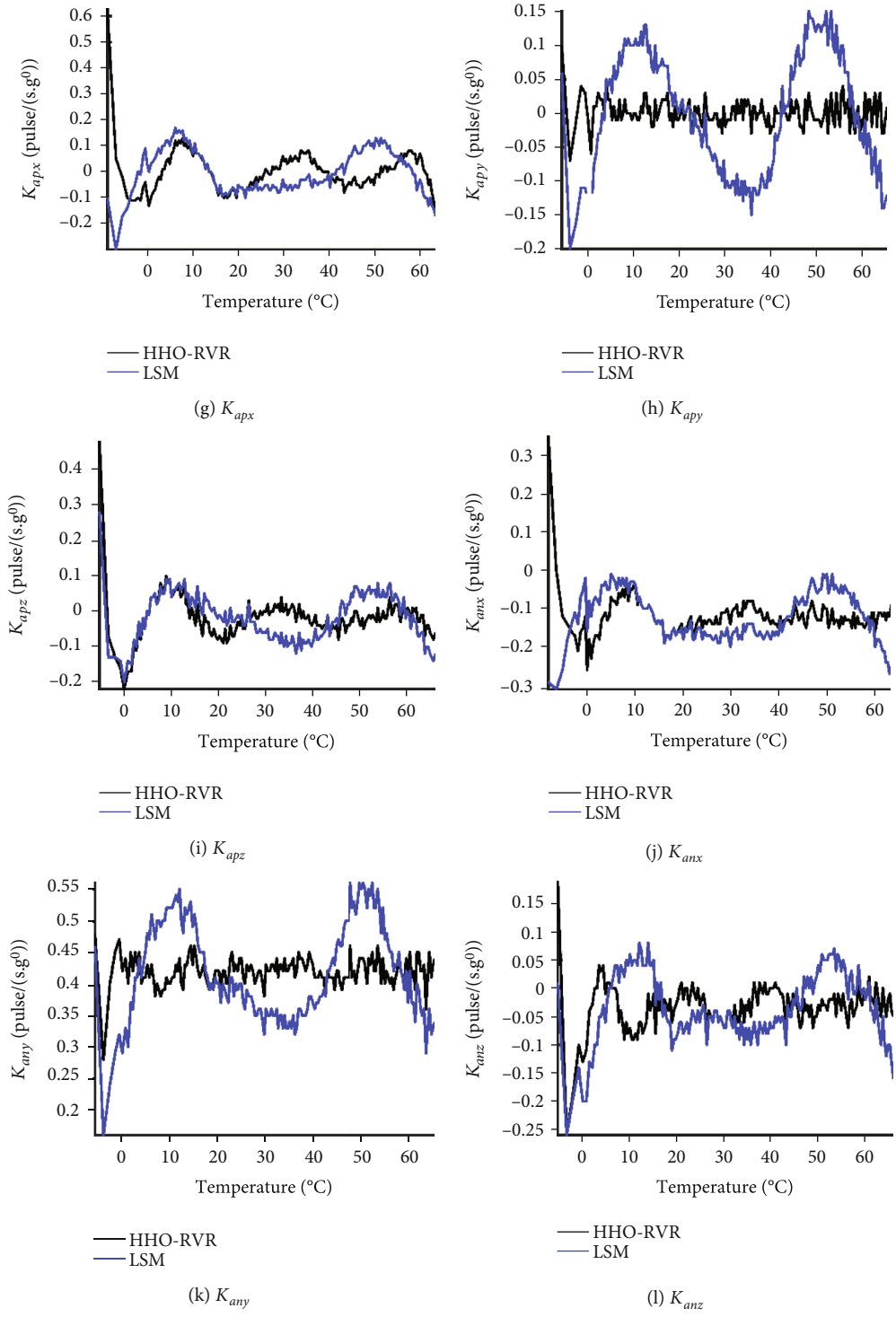


Figure 3: Continued.

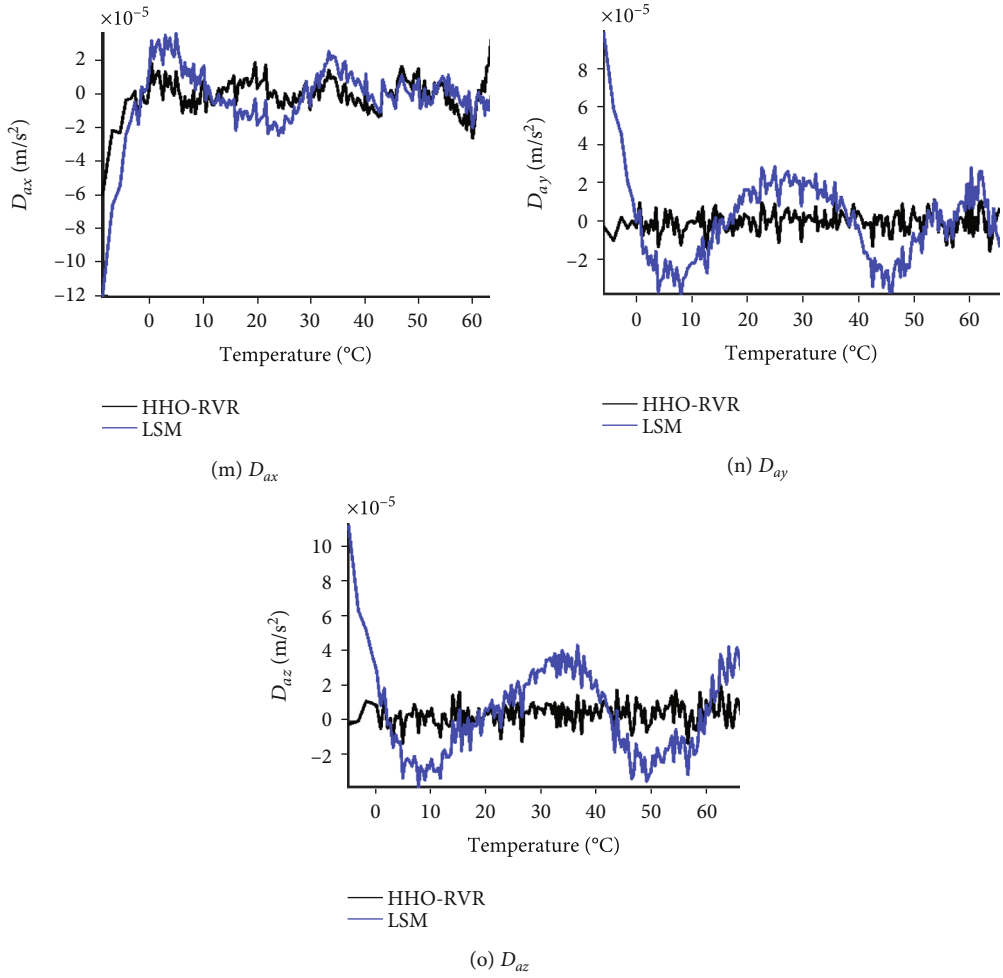


Figure 3: Comparison of residuals after the temperature-related parameter compensation.

Table 3: Variance comparison data for residuals.

The temperature-related parameter	Original data	HHO-RVR	The least square method	Improvement (%)
P_1	$3.16773E-03$	$6.42336E-04$	$6.94344E-04$	7.49
P_2	$2.74528E-03$	$8.32613E-04$	$9.54550E-04$	12.77
P_3	$1.44687E-03$	$5.85158E-04$	$7.18217E-04$	18.53
P_4	$1.32818E-06$	$1.72603E-06$	$1.73323E-06$	0.42
P_5	$4.67740E-06$	$1.91785E-06$	$2.11570E-06$	9.35
P_6	$2.51331E-06$	$2.00860E-06$	$2.03553E-06$	1.32
P_7	$2.56482E-03$	$9.96659E-06$	$1.71954E-05$	42.04
P_8	$6.60438E-05$	$5.72778E-06$	$2.01836E-05$	71.62
P_9	$5.95731E-04$	$6.61036E-06$	$2.44531E-05$	72.97
P_{10}	$1.19157E+01$	$7.68137E-02$	$8.15592E-02$	5.82
P_{11}	$5.44946E+00$	$2.00643E-02$	$8.83737E-02$	77.30
P_{12}	$6.83441E+00$	$6.02220E-02$	$6.64657E-02$	9.39
P_{13}	$1.19019E+01$	$4.94948E-02$	$6.28477E-02$	21.25
P_{14}	$5.39584E+00$	$2.22470E-02$	$7.57132E-02$	70.62
P_{15}	$6.72286E+00$	$3.89916E-02$	$5.99048E-02$	34.91

residual curve, the smaller the residual variance is and the higher the compensation accuracy will be.

Figure 3 shows the residuals of 15 temperature-related parameters compensated by RVR and the least square method. Because the magnitudes of K_{gx}, K_{gy}, K_{gz}, D_{gx}, D_{gy}, and D_{gz} are small and the trend is not apparent, the two curves overlap more. From the curve of 9 parameters of K_{apx}, K_{apy}, K_{apz}, K_{anx}, K_{any}, K_{anz}, D_{ax}, D_{ay}, and D_{az}, it can be seen that the temperature-related parameters compensated by the RVR method have smaller fluctuations and are more stable than those compensated by the least square method. Figure D_{ay} is a typical curve. It can be seen that the residual curve after the LSM compensation has a large fluctuation. The residual curve of the HHO-RVR compensation is very stable. It shows that the original data has a strong nonlinearity. The fitting performance of the LSM is not as good as that of the HHO-RVR, which leads to a large fluctuation of residual after compensation. Other similar figures also illustrate this conclusion.

As shown in Table 3, the variances of the residuals are compared in a table. It can be seen that the residual variance after compensation using the HHO-RVR method is lower than that using the least square method, indicating that the performance of using RVR to compensate for the temperature-related parameters is more superior.

5. Conclusion

In order to compensate for the influence of temperature on the RLG-IMU system and reduce the system startup stability time, an RLG-IMU system-level temperature compensation method is proposed. The RVR was used to model the 15 temperature-related parameters of the RLG-IMU, and the HHO algorithm was used to optimize the width parameters of the RVR kernel function. A system compensation model was established. The test results show that the fitting accuracy and compensation accuracy are more excellent than those of the least square method. In terms of fitting and regression performance, the root mean square error value increased by an average of 35.27%, the mean absolute error value increased by an average of 39.29%, and the R-squared value increased by an average of 2.29%. In terms of system compensation performance, the variance of the residual after compensation increased by 30.34% on average. The system-level compensation method can effectively reduce the startup stability time of the RLG-IMU, which has high application value and practical engineering value.

Data Availability

The data used to support the findings of this study have not been made available because other research works that need the same data have not been completed.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61973167).

References

[1] S. Wang and Z. Zhang, "Research on principle, application and development trend of laser gyro," *Journal of Physics: Conference Series*, vol. 1549, p. 022118, 2020.

[2] A. Sapegin and M. Norenko, "Strapdown inertial attitude system based on ring laser gyro algorithm," *Kyiv Politechnic Institute*, vol. 2, no. 2, pp. 108–113, 2017.

[3] R. B. Hurst, M. Mayerbacher, A. Gebauer, K. U. Schreiber, and J.-P. R. Wells, "High-accuracy absolute rotation rate measurements with a large ring laser gyro: establishing the scale factor," *Applied Optics*, vol. 56, no. 4, pp. 1124–1130, 2017.

[4] W. Q. Sun, L. W. Zhang, W. Xiong, G. Chen, and N. Qu, "Systematic calibration method for RLG inertial measurement unit," *Journal of Chinese Inertial Technology*, vol. 24, pp. 9–13, 2016.

[5] N. Ahmad, R. A. R. Ghazilla, N. M. Khairi, and V. Kasi, "Reviews on various inertial measurement unit (IMU) sensor applications," *International Journal of Signal Processing Systems*, vol. 1, pp. 256–262, 2013.

[6] S. C. Wu, S. Q. Qin, X. S. Wang, and C. S. Hu, "Research on calibration method for dither RLGs inertial measurement unit with shock absorbers," *Journal of National University of Defense Technology*, vol. 33, pp. 33–37, 2011.

[7] P. F. Zhang, Y. Wang, J. X. Tang, and X. D. Yu, "Research on methods for compensating temperature of mechanically dithered RLG," *Binggong Xuebao/Acta Armamentari*, vol. 31, pp. 562–566, 2010.

[8] X. Yu, Y. Xu, G. Wei, and X. Long, "Temperature compensation method for bias of ring laser gyroscope based on artificial fish swarm algorithm," *Hongwai Yu Jiguang Gongcheng/Infrared Laser Engineering*, vol. 43, pp. 81–87, 2014.

[9] X. Li, D. Li, J. Gao, and M. Pang, "Temperature drift compensation algorithm based on BP and GA in quartzes flexible accelerometer," *Applied Mechanics and Materials*, vol. 249–250, pp. 95–99, 2012.

[10] Z. Shi, S. Chen, J. Zhang, L. Zhao, and Q. Sun, "Temperature compensation of laser gyro based on improved RBF neural network," *Guangxue Jingmi Gongcheng/Optics Precision Engineering*, vol. 22, no. 11, pp. 2975–2982, 2014.

[11] Y. Pan, L. Li, C. Ren, and H. Luo, "Study on the compensation for a quartz accelerometer based on a wavelet neural network," *Measurement Science and Technology*, vol. 21, 2010.

[12] J. Ding, J. Zhang, W. Huang, and S. Chen, "Laser gyro temperature compensation using modified RBFNN," *Sensors*, vol. 14, no. 10, pp. 18711–18727, 2014.

[13] G. Li, P. Zhang, G. Wei, Y. Xie, X. Yu, and X. Long, "Multiple-point temperature gradient algorithm for ring laser gyroscope bias compensation," *Sensors*, vol. 15, no. 12, pp. 29910–29922, 2015.

[14] J. Cheng, J. Fang, W. Wu, and J. Li, "Temperature drift modeling and compensation of RLG based on PSO tuning SVM," *Measurement*, vol. 55, pp. 246–254, 2014.

[15] G. Wei, G. Li, Y. Wu, and X. Long, "Application of least squares-support vector machine in system-level temperature

compensation of ring laser gyroscope," *Measurement*, vol. 44, no. 10, pp. 1898–1903, 2011.

[16] S. Y. Li, G. L. Yang, E. K. Yuan, and L. Zhang, "Temperature compensation for mechanically dithered RLG bias based on K-means clustering," *Advanced Materials Research*, vol. 765, pp. 147–150, 2013.

[17] G. Chen, Z. Ren, and H. Sun, "Curve fitting in least-square method and its realization with Matlab," *Ordnance industry automation*, vol. 3, p. 63, 2005.

[18] L. Chen and Y. Zheng, "Study on curve fitting based on least square method," *Journal of Wuxi Institute of Technology*, vol. 11, pp. 52–55, 2012.

[19] S. Denis Ashok and G. L. Samuel, "Least square curve fitting technique for processing time sampled high speed spindle data," *International Journal of Manufacturing Research*, vol. 6, no. 3, pp. 256–276, 2011.

[20] W. Xu, W. Chen, and Y. Liang, "Feasibility study on the least square method for fitting non-Gaussian noise data," *Physica A: Statistical Mechanics and its Applications*, vol. 492, pp. 1917–1930, 2017.

[21] A. C. Faul and M. E. Tipping, "Analysis of sparse Bayesian earning," in *Advances in Neural Information Processing Systems*, 2002.

[22] M. E. Tipping, "The relevance vector machine," *Advances in Neural Information Processing Systems*, pp. 653–658, 2000.

[23] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2000.

[24] L. Liu, J. Wang, and R. Liu, "Infrared small target detection based on relevance vector regression," *Procedia computer science*, vol. 111, pp. 315–322, 2017.

[25] F. Chen, Y. Yang, B. Tang, B. Chen, W. Xiao, and X. Zhong, "Performance degradation prediction of mechanical equipment based on optimized multi-kernel relevant vector machine and fuzzy information granulation," *Measurement*, vol. 151, p. 107116, 2020.

[26] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: algorithm and applications," *Future Generation Computer Systems*, vol. 97, pp. 849–872, 2019.

[27] O. Ramos-Figueroa, M. Quiroz-Castellanos, E. Mezura-Montes, and O. Schütze, "Metaheuristics to solve grouping problems: a review and a case study," *Swarm and Evolutionary Computation*, vol. 53, p. 100643, 2020.

[28] B. Morales-Castañeda, D. Zaldívar, E. Cuevas, F. Fausto, and A. Rodríguez, "A better balance in metaheuristic algorithms: does it exist?," *Swarm and Evolutionary Computation*, vol. 54, p. 100671, 2020.

[29] Z. Sheikh Khozani, H. Bonakdari, and A. H. Zaji, "Application of a genetic algorithm in predicting the percentage of shear force carried by walls in smooth rectangular channels," *Measurement*, vol. 87, pp. 87–98, 2016.

[30] R. Eberhart and J. Kennedy, "New optimizer using particle swarm theory," in *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, pp. 39–43, Nagoya, Japan, Japan, 1995.

[31] W. Tan, W. Wu, D. Dai, X. Wang, Y. Zhao, and S. Qin, "Time series modeling of the ring laser gyroscope's bias considering the temperature delay effect," *Applied Optics*, vol. 57, no. 16, pp. 4551–4557, 2018.

[32] P. Zhang, Y. Wang, X. Yu, G. Wei, and J. Tang, "Effect of temperature characteristic of light path on RLG's bias," *Hongwai Yu Jiguang Gongcheng/Infrared Laser Engineering*, vol. 40, pp. 2393–2397, 2011.

[33] J. Yuan, K. Wang, T. Yu, and M. Fang, "Integrating relevance vector machines and genetic algorithms for optimization of seed-separating process," *Engineering Applications of Artificial Intelligence*, vol. 20, no. 7, pp. 970–979, 2007.

[34] E. Zio and F. Di Maio, "Fatigue crack growth estimation by relevance vector machine," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10681–10692, 2012.

[35] X. Wang, M. Ye, and C. J. Duanmu, "Classification of data from electronic nose using relevance vector machines," *Sensors and Actuators B: Chemical*, vol. 140, no. 1, pp. 143–148, 2009.

[36] W. Wang and Y. Lu, "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model," in *2017 the 5th International Conference on Mechanical Engineering, Materials Science and Civil Engineering*, Kuala Lumpur, Malaysia, December 2017.

[37] P. Maragos, "Morphological correlation and mean absolute error criteria," in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 1568–1571, Glasgow, UK, UK, 1989.

Hindawi
Journal of Sensors
Volume 2021, Article ID 6657709, 14 pages
https://doi.org/10.1155/2021/6657709

Research Article

Indoor Detection and Tracking of People Using mmWave Sensor

Xu Huang[1,2] **Hasnain Cheena**,[1] **Abin Thomas**,[1] **and Joseph K. P. Tsoi**[1]

[1]*Department of Electrical and Computer Engineering, The University of Auckland, Auckland 1010, New Zealand*
[2]*Faulty of Information Engineering, University of Shandong Ying Cai, Shandong, Jinan 250104, China*

Correspondence should be addressed to Joseph K. P. Tsoi; ktso005@aucklanduni.ac.nz

Received 9 December 2020; Revised 3 January 2021; Accepted 12 January 2021; Published 3 February 2021

Academic Editor: Bin Gao

This paper proposes a new indoor people detection and tracking system using a millimeter-wave (mmWave) radar sensor. Firstly, a systematic approach for people detection and tracking is presented—a static clutter removal algorithm used for removing mmWave radar data's static points. Two efficient clustering algorithms are used to cluster and identify people in a scene. The recursive Kalman filter tracking algorithm with data association is used to track multiple people simultaneously. Secondly, a fast indoor people detection and tracking system is designed based on our proposed algorithms. The method is lightweight enough for scalability and portability, and we can execute it in real time on a Raspberry Pi 4. Finally, the proposed method is validated by comparing it with the Texas Instruments (TI) system. The proposed system's experimental accuracy ranged from 98% (calculated by misclassification errors) for one person to 65% for five people. The average position errors at positions 1, 2, and 3 are 0.2992 meters, 0.3271 meters, and 0.3171 meters, respectively. In comparison, the Texas Instruments system had an experimental accuracy ranging from 96% for one person to 45% for five people. The average position errors at positions 1, 2, and 3 are 0.3283 meters, 0.3116 meters, and 0.3343 meters, respectively. The proposed method's advantage is demonstrated in terms of tracking accuracy, computation time, and scalability.

1. Introduction

Indoor detection and tracking of people are useful solutions to energy assignment, health, and safety [1]. Studies show that an indoor detection and tracking system can reduce energy usage for lighting and Heating, Ventilation, and Air Conditioning (HVAC) systems by more than 30% [2]. Additionally, these systems can also improve security applications by giving emergency systems the ability to make more well-informed decisions. So that can enhance the response of emergency systems by providing them with real-time location information of people, where they are going, and the densities of people at different sites to decide whether they are safe or not. Moreover, indoor detection and tracking systems could also help health care businesses monitoring the elderly when they fall. For example, based on location information, nursing staff could make a decision ensuring their safety.

Researchers have studied various types of sensing technologies for indoor object detection measurements, such as passive infrared (PIR) [3], optical cameras [4, 5], LIDAR [6], Wi-Fi [7], and 10 GHz-to-24 GHz microwave [8]. However, all these technologies have challenges with inaccuracy, privacy, environmental robustness, and system complexity [9]. For instance, the HD camera system and other technologies such as Wi-Fi, Bluetooth, and UWB are used for positioning [10–12]. In these studies, Machine Learning (ML) methods are employed to detect people. These methods include decision trees, hidden Markov models, and convolutional neural networks. Machine learning is a computationally intensive process and cannot readily be implemented onto an embedded system.

Moreover, typically camera-based tracking systems require a clear view and the right lighting conditions displayed in [13], where the system uses background subtraction and the Lucas and Kanade tracking algorithm to determine indoor human counting. The system had an experimental accuracy of 97% under lab conditions, but the accuracy during field-testing dropped significantly. Additionally, another critical problem with camera systems is their intrusive nature, leading to privacy concerns. However,

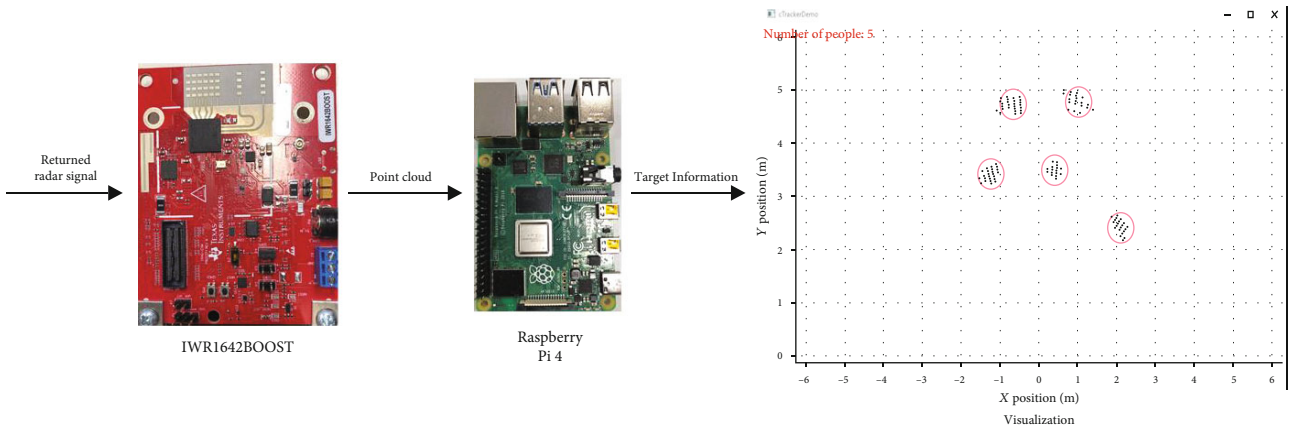


Figure 1: Flow of hardware information.

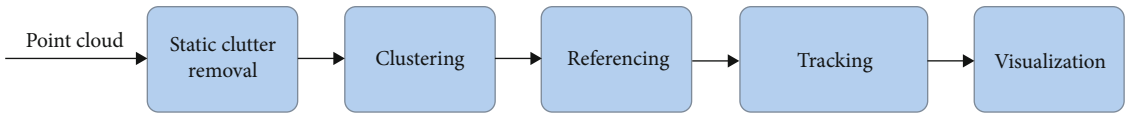


Figure 2: Framework of the data process.

research is being done to discover whether it is possible to use lower resolution cameras to circumvent privacy issues [14].

Motivated by this, this research chose the onboard millimeter-wave (mmWave) radar sensor (IWR1642-BOOST) [15] as the sensing technology [16]. mmWave is a remote wireless sensing technology that has raised lots of attention from academia and industry due to its exceptional advantages. Compared to the existing wireless sensing technologies, this particular radar technology can overcome environmental occlusion problems. We aim to explore fast and robust people detection and tracking models, algorithms, and application guidance using mmWave sensors for indoor applications.

Ongoing research in object detection and tracking data process technologies are mainly focused on vision-based methods [4]. There are currently only limited studies using mmWave radar data for indoor detection and tracking of people. In [17], Wei and Zhang set up a new high-precision passive tracking method (mTrack) and used highly directional 60 GHz millimeter-wave radios to run a discrete beam scanning mechanism to pinpoint the object's initial location and track its trajectory. However, it is based on a signal-phase model. Hence, it is not suitable for applying detection and tracking indoors. In [18], Palacios et al. developed two indoor localization algorithms tailored to mm-wave propagation characteristics based on commercial 60 GHz mm-wave hardware. However, its experiment results only considered location error, and the system computation load is not mentioned. The most related work is people counting and tracking using a mmWave radar sensor by Texas Instruments (TI) [19]. The system employs density-based clustering (DBSCAN) with an extended Kalman filter (EKF), and it reported an accuracy of 51% to 99% between 1 and 5 people. However, its accuracy is questionable since only DBSCAN [20] is used to clustering the varying density data. Moreover, its portability and scalability are limited due to the use of EKF to convert the polar measurement to Cartesian coordinates. The conversion is taken for ease of use, yet it brings additional computation load and process noise.

This paper includes two main contributions. Firstly, we present a systematic approach to detecting and tracking people indoors using a mmWave radar sensor. Two efficient clustering algorithms are proposed here to provide high accuracy and shallow processing time; the recursive Kalman filter (RKF) tracking algorithm performs much better than the EKF in algorithmic complexity and computation time. Moreover, a fast indoor people detection and tracking system was designed based on our proposed algorithms. Furthermore, the system can operate on an embedded platform, the Raspberry Pi 4, creating computing constraints to introduce a portability and scalability aspect. Comparing the results to the commercially available system from TI shows that the method is faster, more accurate, and less heavy than the TI system.

2. Methodology

2.1. Hardware Framework. The hardware system consists of the millimeter-wave radar sensor (IWR1642BOOST), a Raspberry Pi 4 (1.5GHz, 4GB RAM), and a monitor. The flow of hardware information is shown in Figure 1. The sensor emits a radar signal, taking a snapshot of the indoor location at a given point in time. The returned radar signal undergoes preliminary processing on the sensor, the output of which is a point cloud. This point cloud is a collection of points that represents detected people. The point cloud is then processed on the Raspberry Pi 4. The output of the processing is information on identified targets, which is then displayed on a monitor.

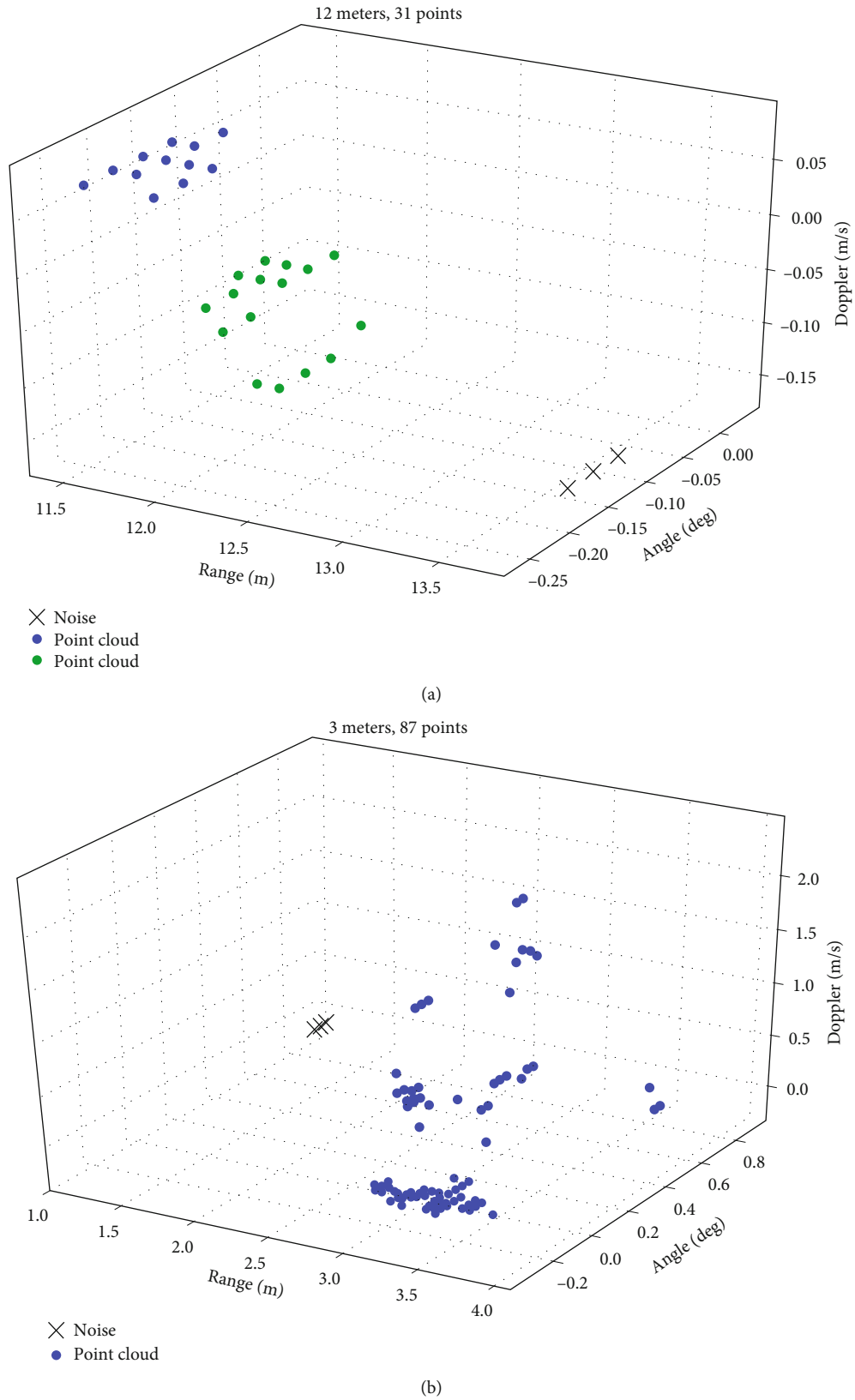


Figure 3: Comparison of the numbers of clustered points at different distances: (a) data points at 12 meters from mmWave sensor; (b) data points at 3 meters from mmWave sensor.

Parameters:
maxDistance- the largest searching distance
minClusterSize- the minimum number of points to classify as a cluster
Steps:
(1) Set the maxDistance and minClusterSize parameters for the clustering algorithm.
(2) Randomly select a point c that has not been marked a cluster or been designated as an outlier (noise).
(3) Compute its neighborhood to determine if it is a core point. If yes, start a cluster around this point.
If no, mark the point as an outlier.
(4) If p is a core point, a cluster is formed, expand the cluster by adding all directly reachable points to the cluster.
(5) If an outlier is added, change that point's status from outlier to border point.
(6) Repeat steps 2-5 until all points are either assigned to a cluster or designated as an outlier.
(7) Calculate the mean of each cluster.

Algorithm 1: Clustering: DBmeans.

Parameters:
maxDistance- the largest searching distance
minClusterSize- the minimum number of points to classify as a cluster
Steps:
(1) Set the maxDistance and minClusterSize parameters for the clustering algorithm.
(2) Randomly select a point p that has not been marked a cluster or been designated as an outlier (noise).
(3) Compute its neighborhood to determine if it is a core point. If yes, start a cluster around this point.
If no, mark the point as an outlier.
(4) If p is a core point, a cluster is formed, expand the cluster by adding all directly reachable points to the cluster.
(5) If an outlier is added, change that point's status from outlier to border point.
(6) Repeat steps 2-5 until all points are either assigned to a cluster or designated as an outlier.
(7) Randomly select a point c in each cluster as a medoid.
(8) Assign each of the remaining points (nonmediod) in every cluster represented by the nearest medoid.
(9) Randomly select a nonmedoid point O_{random} in every cluster.
(10) Consider each of the current medoids O_j in every cluster: compute the total cost S of swapping O_j with O_{random}, includes the cost contributions of reassigning nonmedoid points caused by the swap.
If $S < 0$, then swap O_j with O_{random} to form the new medoid.
(11) Repeat steps 8-10, until no change.

Algorithm 2: Clustering: DBmedoids.

2.2. System Data Process Framework. The flow chart in Figure 2 depicts the systematic approach used in this paper to process and analyze mmWave sensor raw data (point cloud), then the sensor transmits the point cloud data to the Raspberry Pi. This paper mainly focuses on the clustering and referencing algorithms, tracking algorithm, and timing analysis of the merged approach for real time and portability, then compares it to the TI method. Firstly, the point cloud information from the mmWave sensor is parsed and then processed for static clutter removal. The points are then grouped in the clustering+referencing module, and then finally, the people's points are tracked in the tracking module, from which the people number is derived.

2.3. Static Clutter Removal. The static clutter removal model is aimed fat excluding as many as static points as possible from the background. It requires range information since it filters out nonrange changing (static) objects from the scene. The steps of the static clutter removal algorithm are listed as follows.

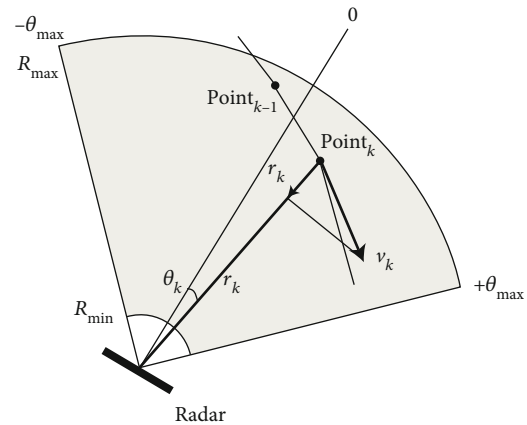


Figure 4: Radar geometry in 2D.

Step 1. Range processing performs Fast Fourier Transform (FFT) on Analog to Digital Converter (ADC) samples per antenna per chirp. FFT output is a set of range bins.

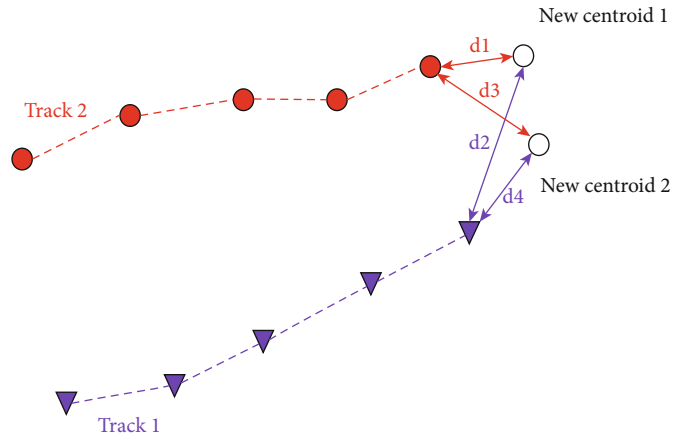


Figure 5: Simplified GNN diagram.

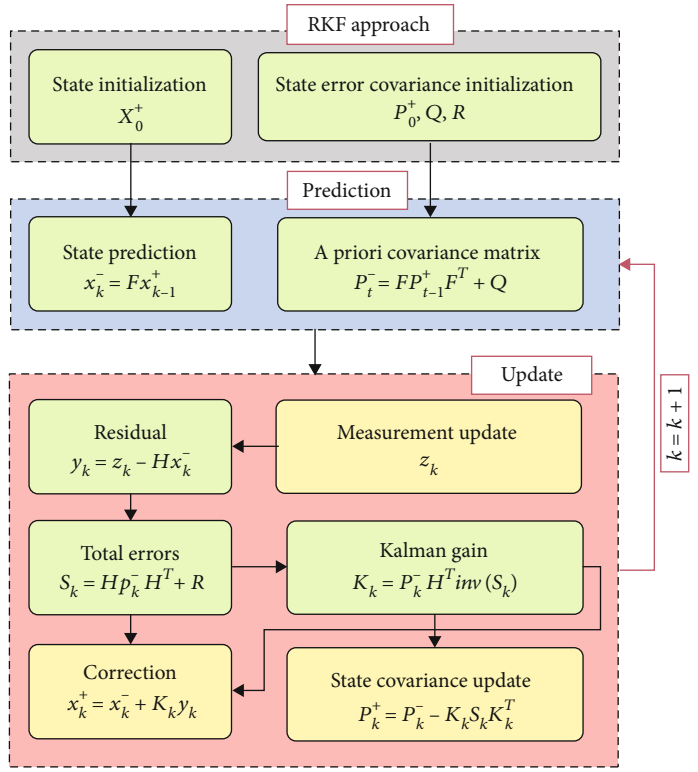


Figure 6: Implementation flowchart of the proposed RKF.

Step 2. Perform static clutter removal by subtracting the estimated Direct Current (DC) component from each range bin.

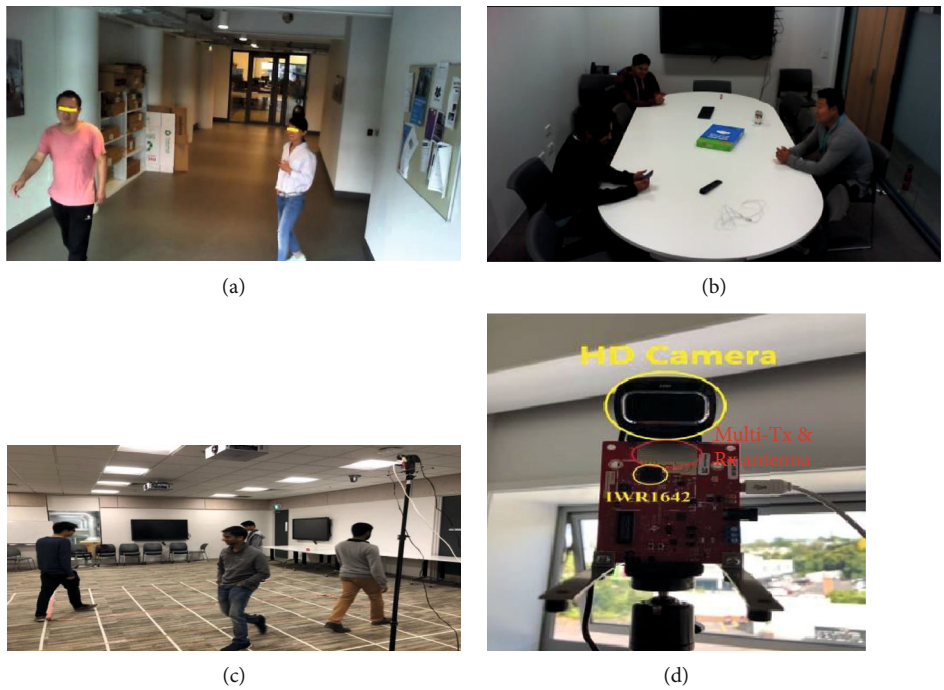
Step 3. Range processing results in local scratch buffers are Enhanced Direct Memory Access (EDMA) to the radar data cube with transpose.

2.4. Clustering and Referencing. The clustering stage is aimed at identifying the number of people in a scene, and since a single centroid is needed to track each person, a referencing process is required.

2.4.1. Clustering. Due to the mmWave sensor field of view, the data's density varies from time and distance against the sensor. For example, the closer the people are to the sensor, the more dense points can be collected. On the other hand, as distance increases, only a few points can be obtained, especially for smaller objects. To demonstrate, Figure 3 shows the different total number of collected points (cluster density) of people located at different distances from the sensor. There are only 31 points (blue and green) of two people around 12 m away from the sensor, while another person is only 3 m away from the sensor, which collected 87 points (blue and green). It showed that a denser cluster represents a person closer to the sensor. In contrast, a person

Steps:
(1) Calculate the distances between each old centroid (objects) and new centroid (observations) in popular coordinate system (e.g., d1, d2, d3, and d4). $d^2 = r_1^2 + r_2^2 - 2r_1 r_2 \cos(\theta_1 - \theta_2)$.
(2) Find the one with the global smallest distance from the total distances (e.g., d1).
(3) Associate the object with the new centroid linked by this distance (e.g., associate Track 2 and new centroid 1).
(4) Repeat steps 2-3 until all unassociated new centroids and objects are associated (e.g., associate Track 1 and new centroid 2).

Algorithm 3: Data association: simplified GNN.



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

Figure 7: Various data collection sites and experimental setup: (a) corridor; (b) meeting room; (c) seminar room; (d) experimental setup.

who is further away is represented by a less dense and more variable cluster.

The black cross stands for the density-based noise points after the clustering stage. These clusters are identified to be the noise of points that are too small to represent people. The density-based noise identifies works by treating each point as a node and then calculating the distance matrix between itself and all the other nodes. The distance between each node is the difference in displacement in the x-direction and the displacement in the y-direction. If a node is within a distance threshold of 0.2 m to the other nodes, those nodes are extracted.

2.4.2. Referencing. After clustering, all detected people are represented by clusters. A reference point on the X and Y plane needs to be found to locate each cluster's position. This reference point will later be used for tracking clusters and extracting trajectory information. The reference point can be the mean center of a cluster and also can be the real center point (both can be called centroid) of a cluster. For people clusters, both can be used as the reference point.

The two density-based clustering and referencing algorithms that we implemented are DBmeans and DBmedoids, respectively. The algorithms can get the centroid location or the centroid point of a cluster with a shallow misclassification rate. The two algorithms are presented as Algorithm 1 and Algorithm 2.

2.5. Tracking

2.5.1. Recursive Kalman Filter. The tracking stage is necessary to locate people as they move through the indoor space and maintain accurate and reliable measurements. In this paper, a recursive estimation method with Kalman filter (RKF) plus a motion model is applied for motion state prediction and estimation of people. Since there are inconsistencies in the rate at which data was lost from the sensor, we decided to recursively calculate the error covariance matrix and Kalman gain in each update stage. We could get a more accurate update compared to a static Kalman gain.

We opted for the RKF, as it has not been researched in mmWave indoor people tracking, and we theorized that we could make it lightweight for real-time application. The

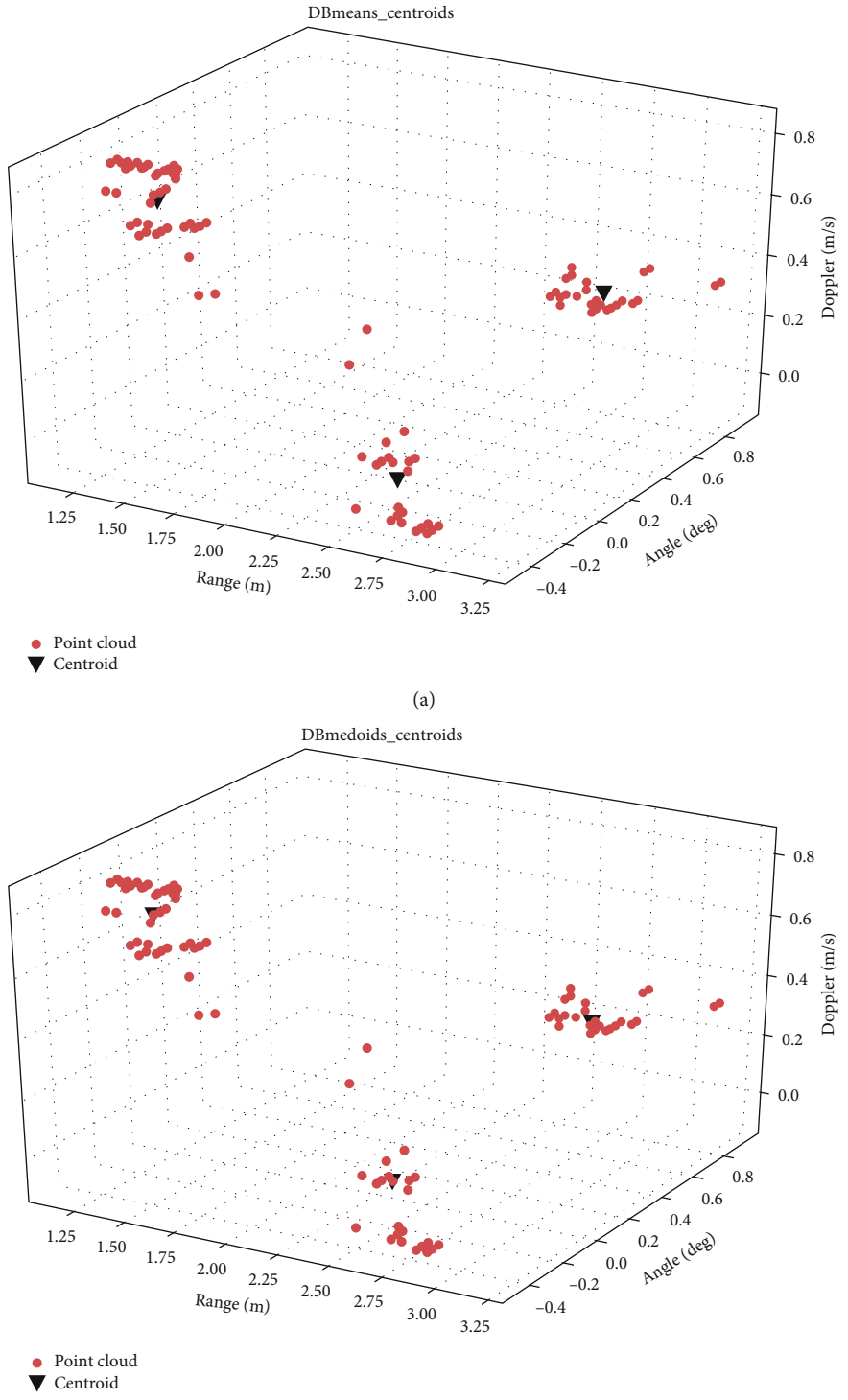
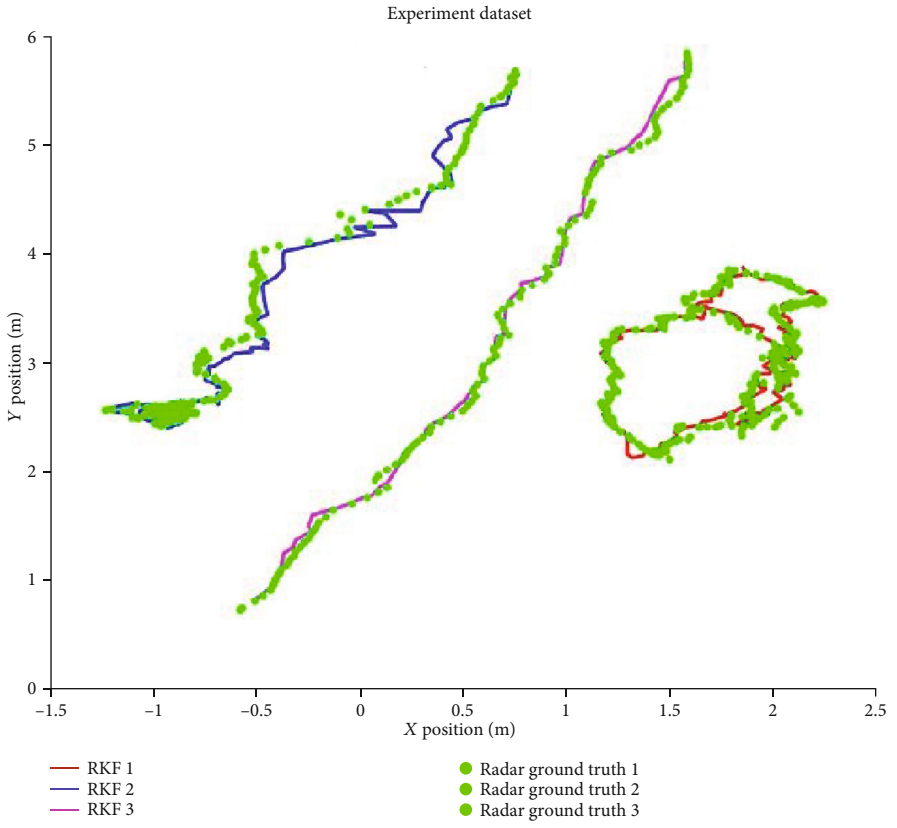


Figure 8: Example frame of the DBmeans and DBmedoids algorithms: (a) DBmeans; (b) DBmedoids.

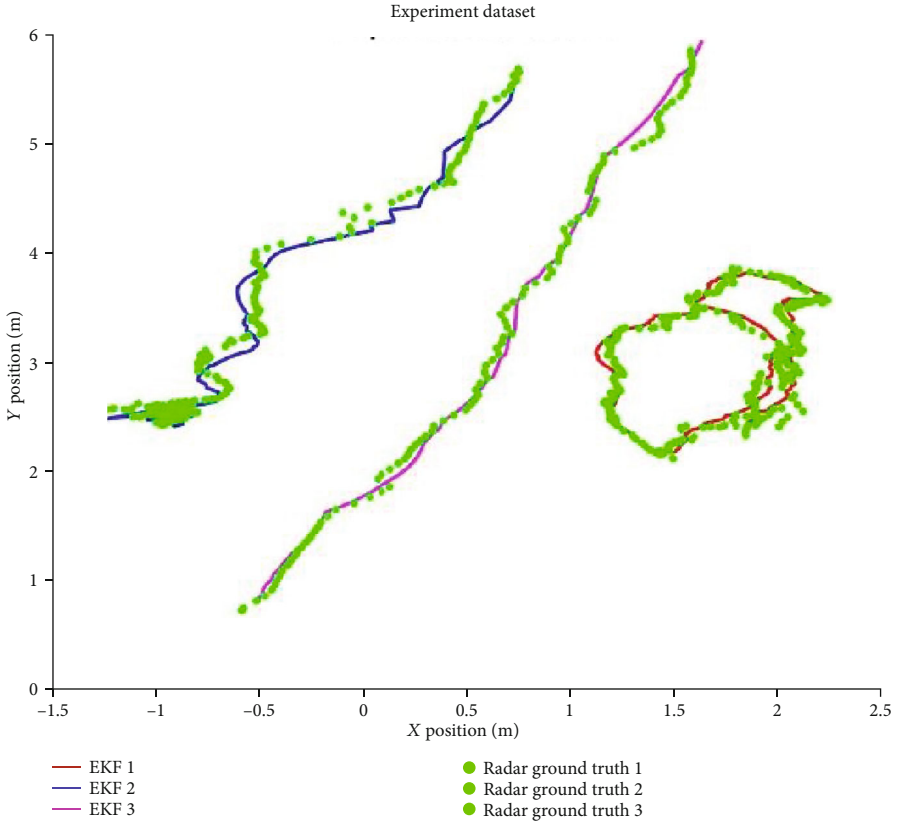
RKF is well-suited for indoor people tracking when using a constant velocity (CV) model, and we also considered an acceleration model by random noise. Moreover, it can improve accuracy by avoiding the EKF's process errors caused by linearization by keeping computation in the polar coordinates, illustrated in Figure 4. Figure 4 illustrates the single reflection point at time k. Multiple reflection points

Table 1: Comparison between the DBmeans and DBmedoids algorithms.

Algorithms	Total frames	Misclustering	Time (ms)	Accuracy
DBmeans	341	52	29.11	84.75%
DBmedoids	341	59	121.67	82.70%



(a)



(b)

Figure 9: RKF vs. EKF: (a) RKF; (b) EKF.

represent real-life radar objects. Each point is represented by range r ($R_{\min} < r < R_{\max}$), angle ($-\theta_{\min} < \theta < \theta_{\max}$), and radial velocity $\dot{r}$ (range rate). To employ RKF, we keep the raw data processing from detection to tracking under the polar system and keep the visualization under the Cartesian coordinates for the best view.

The system state in the polar system at step k can represent as

$$x_k = \left[r\ \dot{r}\ \theta\ \dot{\theta}\right]^T. \tag{1}$$

The motion state model and observation model of people can be built as follows:

$$\begin{aligned} x_k &= \begin{bmatrix} 1 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 1 \end{bmatrix} x_{k-1} + Q, \\ y_k &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} x_k + R, \end{aligned} \tag{2}$$

where Δt is the mmWave sensor sampling time interval and was set to 50 ms. Q and R are the system noise covariance matrix and measurement noise covariance matrix, respectively.

F is a transition matrix,

$$F = \begin{bmatrix} 1 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where H is a measurement matrix,

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{4}$$

An implementation flowchart of the proposed RKF algorithm is summarized in Figure 5. As shown, the update step involves recursively calculating the Kalman gain K, then calculating the current data frame's state x_k^+ and the error covariance matrix P_k^+. Recalculating the Kalman gain and error covariance can give the estimate system more robust and practical flexibility. Moreover, if no measured data are available, the estimated values are used as the updated values. The algorithm is described as follows.

In the initialization step, the mean values x_0^+ and covariance matrix P_0^+ of the states are set up at $k = 0$, where the superscript "+" indicates that the estimate is a posteriori, and P is the error state covariance matrix.

Table 2: Comparison between the RKF and the EKF.

Algorithms	Total frames	Number of people	RMSE	Total time (ms)
RKF	245	3	0.0471	62.5
EKF	245	3	0.0488	281.3

In the prediction step, the state and its covariance matrix at $k-1(x_{k-1'}^+, P_{k-1}^+)$ are projected one step forward to obtain the a priori estimates at $k(x_k^-, P_k^-)$.

In the update step, the actual measurement is compared with predicted measurement based on the a priori estimate. The difference is used to obtain an improved a posteriori estimate as in Figure 6. Symbols z_k and S_k are the measurement vector and innovation covariance, respectively.

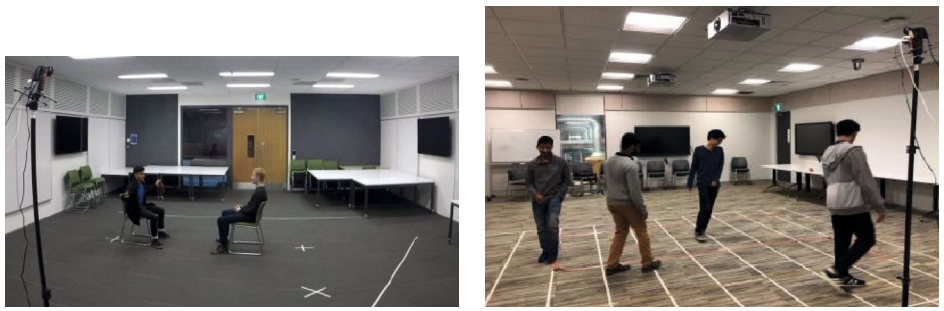
2.5.2. Data Association. Since there could be multiple people at any time, and the Kalman filter can only track a single person at a time; therefore, we implement a lightweight data association approach with a recursive Kalman filter to work on multiple objects. The global nearest neighbor (GNN) data association algorithm used in our system is a simplified version and based on the centroid data after the clustering and referencing step. The simplified GNN diagram is shown in Figure 5, and the algorithm description is shown in Algorithm 3.

After GNN is processed, the associated centroids can be passed through the update step of the RKF to be a multiobject tracker. Each track goes through a life cycle of events. At the maintenance step, we decide to change the state or delete the track that is not used anymore.

3. Experiment and Evaluation

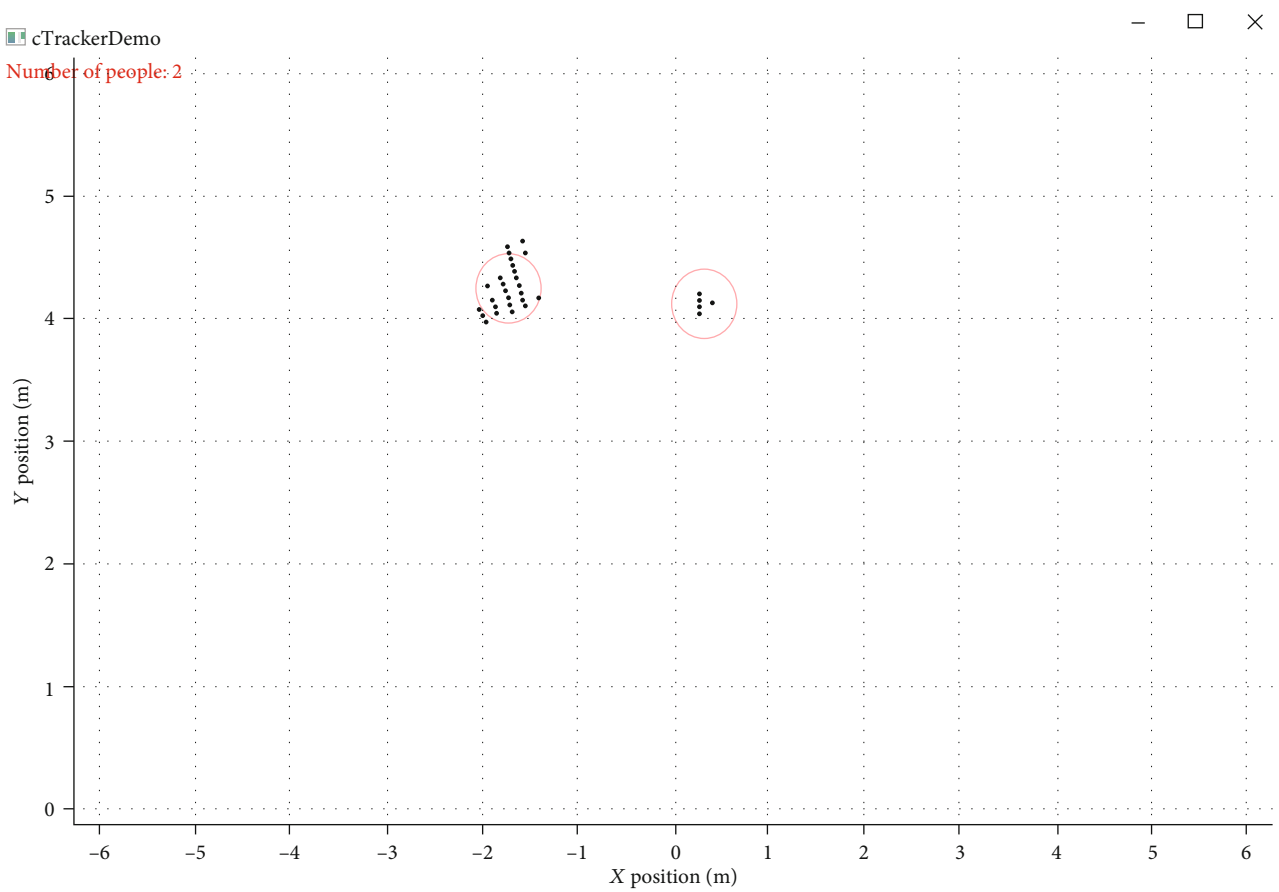
3.1. Experiment Setup. To evaluate our algorithms' performance, we set up experiments at three different data collection sites around the University of Auckland Newmarket campus to model various real indoor scenarios. The mmWave sensor data was captured using the TI IWR1642-BOOST. The IWR1642BOOST radar sensor includes an FMCW transceiver, operating at 76 GHz to 81 GHz (4 GHz available bandwidth) with four receive channels and two transmit channels. It outputs a data frame containing the point cloud, with information for each detected point, including range, azimuth angle, and Doppler velocity. Various settings and modes, such as different ranges, can be selected for using the chirp configuration parameters. There are settings for short-range (10 m), midrange (30 m), and long-range (80 m), albeit at the expense of a narrower field of view. For indoor detection and tracking, we opted for a range of 6 m to maximize the resolution and view field.

For each of the data collection sites, the mmWave sensor was mounted on a tripod and elevated to a height of 1.8 to 2 m. The sensor is placed in the environment so that the field of view covers the range r of 1 to 6 meters and an azimuth θ of -60° to 60°, oriented towards the direction in which people would enter the scene. Additionally, an HD camera was also mounted on top of the millimeter-wave sensor to gather



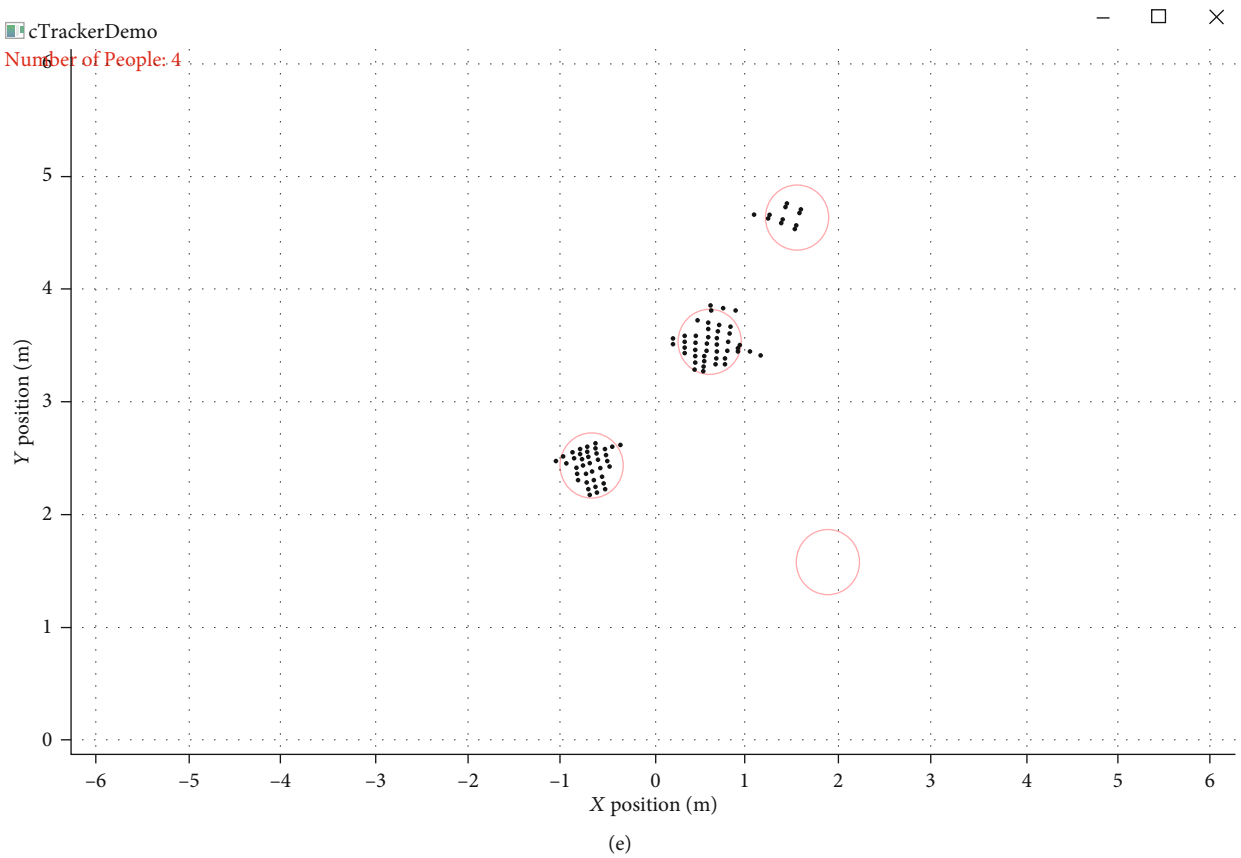
(a) (b)

(c)

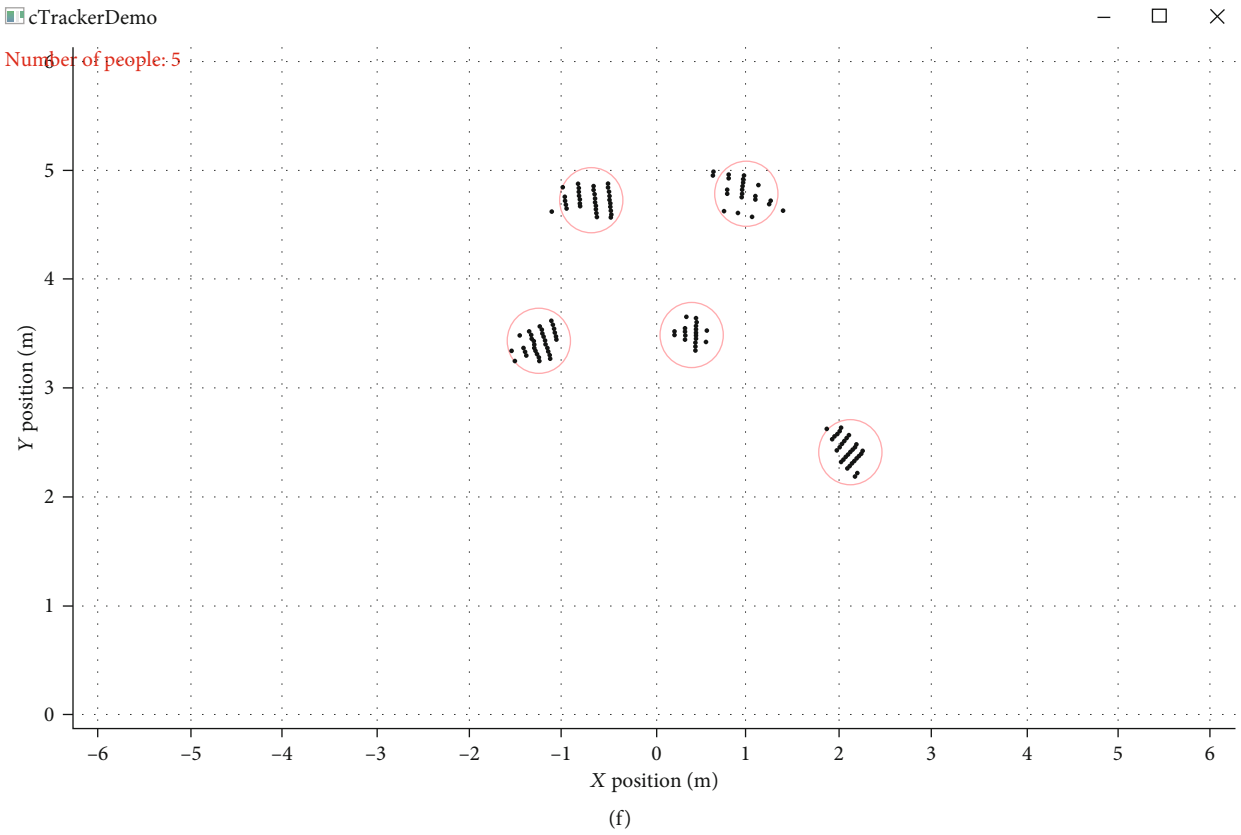


(d)

Figure 10: Continued.



(e)



(f)

Figure 10: cTracker GUI and camera screenshots at the experimental sites: (a) 2 people sitting; (b) 4 people walking; (c) large group; (d) 2 people sitting; (e) 4 people walking; (f) large group.

ground truth information and recording. Figure 7 shows the sensor setup at the different data collection sites. We use data from the three sites to evaluate the methods and algorithms described in the previous sections.

3.2. Evaluation of the Clustering and Referencing. To evaluate the two clustering and referencing algorithms, we tested and compared DBmeans against DBmedoids using a wealth of data obtained from the one experiment site. Experiments were conducted simulating various indoor activities. Data were recorded simultaneously using the TI sensor as well as a video camera, which was used to gather ground truth data. The room was selected to maximize the full range of the sensor. A 6 m by 6 m grid was drawn on the floor to contain the experiment within the sensor range, which allowed us to control when people entered and left the site. The walking activities were selected to test the clustering capabilities of the sensor and to model real indoor scenarios.

Figure 8 shows an example frame of using the DBmeans and DBmedoids algorithms separately. As can be seen, for DBmeans, the centroids are reference locations of each cluster, and for the DBmedoids, the centroids are the real reference points of each cluster. Table 1 also shows the comparison between the DBmeans and the DBmedoids in terms of average misclustering rate and processing time (per frame) using the same total number of data sample frames. In comparison, DBmeans achieves a better average accuracy with 84.75% than DBmedoids with 82.70%. Additionally, DBmeans has a much lower processing time than DBmedoids. Hence, we choose DBmeans as the clustering algorithm of our system.

The density-based clustering algorithm we designed for this task can manage variable cluster densities. Moreover, this algorithm can also handle noise as well as DBSCAN.

3.3. Evaluation of the Tracking. To evaluate the RKF, we compared the tracking accuracy and the processing time of our method to EKF, which TI used.

For the RKF weighting matrix initialization and optimization, we ran through various options and got the best performing combinations. The weighting matrices of the RKF can be initialized as follows:

$$P_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, Q = \begin{pmatrix} 0.2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.2 \end{pmatrix}, R = I. \tag{5}$$

By contrast, the EKF is employed to track the same objects. Figure 9 shows the filter results between RKF and EKF using the experiment data set, and Table 2 shows the comparison of the Root Mean Squared Error (RMSE) and process timing (total frames) between the RKF and EKF.

As can be seen, both EKF and RKF can estimate unmeasurable system states and smooth out the process/measurement noise very well. However, in terms of algorithmic complexity and time consumption, the RKF is much more lightweight than the EKF of TI since the RKF does not need to perform coordinate system conversion and calculate the Jacobian matrix which contribute a lot of additional computational load to the system.

Table 3: Comparison between the number of people and time per frame.

Data set	Total frames	Number of people	Time per frame (ms)
RKF	245	3	62.5
EKF	245	3	281.3

3.4. Evaluation of the Merged Process. To evaluate the merged process for scalability and portability, we merged all the algorithms into a tracking system called centroids-Tracker (cTracker) to pars and present the raw point cloud data in real time. Proof of concept for real-time application on a portable embedded platform was demonstrated using a Raspberry Pi 4. This feature's challenge is to design the algorithm to be lightweight enough, such that the processing time is less than each frame's duration. It was done by minimizing the algorithms' timing complexity and writing the program in python with efficient libraries, however, with the limitation of fewer libraries being available on the Raspberry Pi. Efficient code strategies include using lightweight libraries such as NumPy. It provided us with a very quick run time.

Figure 10 presents part of the objects and the camera ground truth. The black points represent the raw point data returned by the mmWave sensor device at experimental sites, and the colored circles represent the clustered and tracked people. As can be seen, all movements, including the walking/standing movements of people, were tracked and represented.

Apart from the code successfully executing on the Raspberry Pi 4, the embedded application's performance in real time also depends on the time complexity of algorithms. Table 3 shows the average processing time (per frame) from a different number of people with data samples between 1 and 5 people. As expected, an increase in the number of people increases the number of points to be processed. More importantly, run time with five people (a high processing load) is below the 50 ms (the frame rate of the mmWave sensor) constraint ensuring consecutive frames are not missed. Besides, our cTracker can track each person correctly, even with some radar measurement data lost (see Figures 10(b) and 10(e)).

The obvious benefit of the algorithm is the Kalman filter's implementation, as no measurement data input would result in the Kalman filter predicting the missing people until they reappear. Besides, if people's data disappears for long periods, the Kalman filter slowly moves people under their predicted velocity, as predicted using the constant velocity model. The prediction eventually estimates the person as having left the room.

3.5. Comparison with TI System. Compared with the misclustering TI system, Figure 11 shows the average misclustering

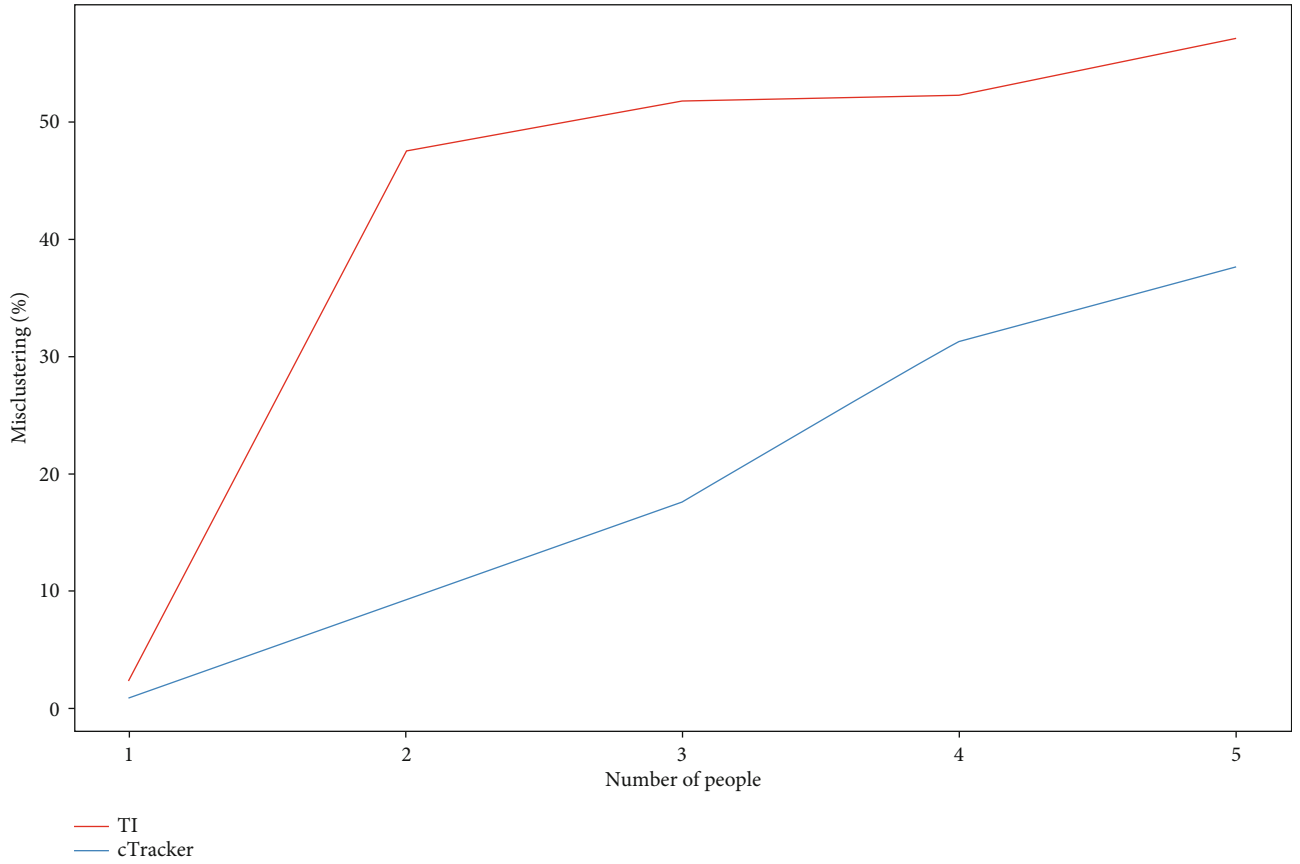


Figure 11: Misclustering_TI_vs_cTracker.

Table 4: Marked ground truth for tracking accuracy comparison.

Location	X (m)	Y (m)
1	-2	3.8
2	1.2	4.2
3	1.5	2.8

rate for different numbers of people (total 12917 frames). As can be seen, our system's misclustering rate is much lower than TI's between 1 and 5 people data sets. However, it also displays that both general trends are increasing as the number of people increases. Since the number of people increases, a higher proportion of objects begin occluding each other, leading to a rise in errors. Additionally, missing data from the sensor is another significant reason for increasing the misclustering rate between both systems.

For tracking accuracy comparison, three data sets were collected from a person walking at the position-known location. We then ran those data sets through both our system and the TI system and calculated the RMSE in the X and Y directions. The location coordinate from the sensor is shown in Table 4. Table 5 shows that our system's average position error was 0.2992 meters in location 1, 0.3271 meters in location 2, and 0.3171 meters in location 3. In comparison, the TI system's average position error was 0.3283 meters, 0.3116 meters, and 0.3343 meters, respectively. The TI system was relatively more accurate only at location 2.

Table 5: RMSE comparison between two systems.

Data set	Total frames	Systems	RMSE X	RMSE Y	RMSE
Location 1	357	TI	0.3030	0.3518	0.3283
Location 1	357	cTracker	0.2630	0.3316	0.2992
Location 2	163	TI	0.2482	0.3641	0.3116
Location 2	163	cTracker	0.2496	0.3894	0.3271
Location 3	126	TI	0.4331	0.1895	0.3343
Location 3	126	cTracker	0.4218	0.1523	0.3171

4. Conclusion

In this paper, the indoor people detection and tracking system is designed based on the proposed data process algorithms. Our methodology processed in the order of static clutter removal, clustering into clusters, and referencing to identify the centroids, then tracking the centroids by using a recursive Kalman filter (RKF). The experiments are set up at three different data collection sites modelling various indoor scenarios. Comparing with the TI system, our system can detect and track each object more accurately. The processing pipeline cycle is under 50 ms (per frame), which can work in real time on an embedded platform such as a Raspberry Pi. Our future work consists of data fusion from multiple mmWave radar sensors to increase the useful field

of view of the system and accuracy. Moreover, we will also use deep learning approaches for tracking and classifying various species objects.

Data Availability

The data used to support the findings of this study can be freely accessed at https://github.com/has-c/Occupancy Detection/tree/master/Data. The mmWave radar sensor product is obtained from https://www.ti.com/tool/IWR1642BOOST#technicaldocuments.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

The authors wish to acknowledge the technical support of the University of Shandong Ying Cai (Jinan, China). This research was funded by the Key Research and Development Project of Shandong Province, China, Grant number 2015GGX101048.

References

[1] E. Yavari, C. Song, V. Lubecke, and O. Boric-Lubecke, "Is there anybody in there?: Intelligent radar occupancy sensors," *IEEE Microwave Magazine*, vol. 15, no. 2, pp. 57–64, 2014.

[2] A. Santra, R. V. Ulaganathan, and T. Finke, "Short-range millimetric-wave radar system for occupancy sensing application," *IEEE Sensors Letters*, vol. 2, no. 3, pp. 1–4, 2018.

[3] G. Monaci and A. V. Pandharipande, "Passive infrared sensor system for position detection," US Patent App. 10/209, 2019, pp. 124.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[6] J. Zhao, H. Xu, H. Liu, J. Wu, Y. Zheng, and D. Wu, "Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors," *Transportation Research Part C: Emerging Technologies*, vol. 100, pp. 68–87, 2019.

[7] X. Guo, B. Liu, C. Shi, H. Liu, Y. Chen, and M. C. Chuah, "WiFi-enabled smart human dynamics monitoring," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, p. 16, Delft, Netherlands, November 2017.

[8] Z. Peng and C. Li, "Portable microwave radar systems for short-range localization and life tracking: a review," *Sensors*, vol. 19, no. 5, p. 1136, 2019.

[9] C. Iovescu and S. Rao, "The fundamentals of millimeter wave sensors," *Texas Instruments, SPYY005*, vol. 1, pp. 1–9, 2017.

[10] S. H. Ryu and H. J. Moon, "Development of an occupancy prediction model using indoor environmental data based on machine learning techniques," *Building and Environment*, vol. 107, pp. 1–9, 2016.

[11] M. K. Masood, C. Jiang, and Y. C. Soh, "A novel feature selection framework with hybrid feature-scaled extreme learning machine (HFS-ELM) for indoor occupancy estimation," *Energy and Buildings*, vol. 158, pp. 1139–1151, 2018.

[12] A. Nessa, B. Adhikari, F. Hussain, and X. N. Fernando, "A survey of machine learning for indoor positioning," *IEEE Access*, vol. 8, pp. 214945–214965, 2020.

[13] Y. Benezeth, H. Laurent, B. Emile, and C. Rosenberger, "Towards a sensor for detecting human presence and characterizing activity," *Energy and Buildings*, vol. 43, no. 2-3, pp. 305–314, 2011.

[14] A. B. Chan, Z. S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: counting people without people models or tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, Anchorage, AK, USA, June 2008.

[15] K. Garcia, "Bringing intelligent autonomy to fine motion detection and people counting with TImmWave sensors," *Texas Instruments*, vol. 1, pp. 1–9, 2018.

[16] X. Wang, L. Kong, F. Kong et al., "Millimeter wave communication: a comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1616–1653, 2018.

[17] T. Wei and X. Zhang, "mTrack: high-precision passive tracking using millimetre wave radios," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pp. 117–129, Paris, France, September 2015.

[18] J. Palacios, G. Bielsa, P. Casari, and J. Widmer, "Single- and multiple-access point indoor localization for millimeter-wave networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 3, pp. 1927–1942, 2019.

[19] M. Livshitz, "Tracking radar targets with multiple reflection points," TI internal document, 2017.

[20] M. Ester, H. P. Kriegel, J. Sander, X. Xu et al., "A density-based algorithm for discovering clusters in large spatial databases with noise," *Knowledge Discovery in Databases*, vol. 96, pp. 226–231, 1996.

Hindawi
Journal of Sensors
Volume 2021, Article ID 6610806, 10 pages
https://doi.org/10.1155/2021/6610806

Research Article

Development and Validation of a Novel Interpretation Algorithm for Enhanced Resolution of Well Logging Signals

Qiong Zhang[1] **and Jean-Baptist Peyaud**[2]

[1]*University of Electronic Science and Technology, China*
[2]*Fluid Minerals Interactions, Australia*

Correspondence should be addressed to Qiong Zhang; zhanqio@uestc.edu.cn

Received 14 December 2020; Revised 18 December 2020; Accepted 23 December 2020; Published 8 January 2021

Academic Editor: Bin Gao

This work presents a novel algorithm that achieves enhanced resolution of well logging signals, e.g., from 1 ft of a pulsed neutron mineralogy tool to 0.04 ft of an imaging tool. The algorithm, denoted as "Digital Core," combines mineralogical and sedimentological information to generate a high-resolution record of the formation mineralogy which can be consequently applied to thin bedded environments. The keystone to the philosophy of this algorithm is that the spectral information recorded by mineralogy tool is a weighted average of the mineralogy of each lithological component in the analyzed volume. Therefore, by using a high-resolution image log to determine the proportion of each lithological component, their composition can be determined from the mineralogy log data. A field case from a well located in South Australia is presented in this work, and the results validate the feasibility of an integrated core-level petrophysical analysis in a cost-effective and timely manner compared to conventional core measurements.

1. Introduction

Oil well logging has been known for many years and provides an oil and gas well driller with information about the particular earth formation being drilled. Accurate and detailed knowledge of earth formations that may contain reservoirs of the hydrocarbons is required for the exploration and production of hydrocarbons [1–3]. Typically, the recording of rock physical properties (logs) is the only information available in a borehole. These include the resistivity of the formation [4], the acoustic properties [4], the porosity (actually the interaction between neutron and hydrogen in the formation) [4], the electronic density of the formation [4], the natural radioactivity, and the gamma emission induced by neutron stimulation. These data are then interpreted in terms of mineralogy of the formation, matrix density, total and effective porosity, fraction of hydrocarbon in the fluids, etc. For example, it is important to know the lithology of the earth formations as a function of depth, particularly in thinly bedded formations. A lithology characterizing a formation may be determined using one or more of several techniques. A common technique is to retrieve core samples from an earth formation and perform intensive analysis of the core sample at the surface. Typically, this is conducted off-site at a specialized facility remote from the well site. While core samples can provide the detailed knowledge petroanalysts and geophysicists desire, obtaining the samples from deep within the earth formation and performing the analysis can be quite time intensive [2].

A neutron logging tool may be used to obtain lithology parameters by measuring radiation resulting from neutron irradiation of the earth formation. The measured radiation is indicative of the reaction of the neutrons with constituents of the formation and thus contains information about the earth formation. As one example, interactions between the neutrons and the formation may result in the emission of gamma rays with energy levels characteristic of the materials with which the neutrons have interacted [2]. Measurements are repeated at several borehole depths along the longitudinal axis of the borehole. Each measurement is associated with a borehole depth at which it is taken. Unfortunately, however, neutron logging generally provides a coarse vertical

resolution, i.e., along the borehole axis, between approximately one and two feet. This resolution is insufficient to locate boundaries of thin beds (e.g., beds less than one foot in thickness). Thus, at the current level of technology, a direct measurement is not possible as these object are commonly far below the vertical resolution of logging tools. For example, a sandstone with 20% clays could correspond to a "dirty" sand with dispersed clay minerals or a "clean" sand with a few argillaceous layers. The properties of the sandstone in terms of flow dynamics, vertical connectivity, porosity, and water content will be very different depending on the actual geological makeup of the formation [3]. This can lead to by-passed pay and/or a misunderstanding of the layer properties. Getting a higher vertical resolution is therefore crucial in complex environments.

In this work, a method is proposed to integrate the sedimentological information from a borehole image and the information from a spectroscopic mineralogy log to improve the vertical resolution of the mineralogy log. The image log provides information about layering: depending on the tool used to acquire the data, a layer is marked by a contrast in resistivity or acoustic properties. Layers are commonly associated to sedimentary beds, but they can be occasionally related to the formation of nodules or the development of diagenetic cements [4]. The vertical resolution of the record is of the order of half an inch. The spectroscopic log provides information about the bulk chemistry of the formation from which the mineralogy is derived. Its vertical resolution is of the order of one foot. The first stage involves comparing the image and spectroscopic logs over the same interval to determine the level of complexity of the logged formation: if the formation displays a layering thinner than the resolution of the spectroscopic mineralogy log, the formation is deemed "thinly bedded" [4]. The algorithm proposed in this work applies to these situations. The borehole is split into intervals ranging from one to several feet in length by clustering similar regions based on the log response of both the mineralogy and the borehole image. Within each interval, lithotype facies, identified as layers of similar properties, are determined based on the image log. In the scope of this work, the term "lithotype" refers to a geological unit characterized by a set of parameters, such as, for example, specific lithology, mineralogical composition, porosity, permeability, grainsize distribution, sedimentological texture, and sedimentological structures. The term "facies" refers to a specific range of the values of these characteristics that characterize a body of rock and allow discriminating it from its surroundings either by way of measurement, observation, or both. For example, natural gamma radiation logs and neutron-induced radiation logs may provide accurate identification of lithotype facies, but with a coarse vertical resolution of approximately two feet in some tools. A high-resolution borehole image log may provide accurate vertical resolution of resistivity and changes in resistivity as small as a few millimeters but with limited ability to identify lithotype facies or minerals.

By employing a high-resolution image log, the volume of each lithological component in the considered interval can be determined. Several classes of high-resolution image logs are available for this purpose: high-resolution electromagnetic imager (e.g., resistivity imager), acoustic imager, and optical image log. From the spectroscopic log, we know the overall composition of the interval, but not the detailed composition of each layer. The mineralogical composition associated with each lithotype is determined by solving a constrained optimization problem [3] that maximizes the likelihood of the determined layer composition. For solving the optimization problem, a constrained sampling algorithm specifically generated for this purpose is utilized and presented in this work.

This work presents in detail the developed mathematical algorithm and methods for combining information obtained from a high-resolution log and low-resolution spectroscopic tool. The targeted application of this method is thin-bedded formations. An example is provided from a well in Australia where spectroscopic mineralogy, a resistivity borehole image, and a core were available. The development of this method allows to increase the resolution of the mineralogy log to the level of the image log: what can be recorded on the borehole image can be discriminated in the mineralogy log. This method will enhance our understanding of complex environments, for instance the distribution of organic matter or the distribution of argillaceous beds in reservoir formations. It allows a better understanding of the reservoir properties and their vertical distributions, a more precise location of hydrocarbon-bearing intervals, and can be used to extend core information more accurately over the logged interval. The higher resolution information could also greatly benefit well completion in cases where stimulation is required or in optimizing the completion design. More specific embodiments may also provide mineralogy logs, matrix density logs, and total porosity logs with a high vertical resolution, as well as allowing calculation of net-to-gross and net pay in thin bed formations.

2. Mathematical Theory and Algorithm Implementation

This section provides an overview of the mathematical theory based on which the algorithm is constructed.

A lithology, e.g., sandstone, limestone, or shale, is characterized by a set of measurable parameters such as grain size distribution and composition [4]. The composition can simply be thought of as being characterized by the relative amount of different minerals contained in the lithology. As an example, sandstone contains more quartz than coal but coal contains more organic matter [4]. However, even among lithologies that are categorized into the same group, the composition can vary. Therefore, it is convenient to characterize the content of a certain mineral in a class of lithologies by a probability density functions. These probability density functions characterize the likelihood of finding a certain mineral with a given concentration within the lithology of interest.

Within this work, we are interested in inferring the mineral compositions of geological layers contained in a logging interval from two separate source of data: (1) from a mineralogical logging tool of coarse resolution; i.e., it gives the average mineralogy over all lithotypes; (2) an image tool that does not provide mineralogical information but has a much finer resolution and can be used to delineate layer

boundaries within the interval. For accomplishing this task, a novel mathematical optimization algorithm is developed to find the most likely combination of lithotypes assigned to the identified layers along with the mineralogical composition of each layer. The solution to the optimization problem must obey several constraints that will be detailed within this section.

The basic idea of the algorithm is to combine information from a facies analysis with the mineralogical information from a mineralogy log derived from spectral data. For each mineral indexed by $m = 1, \cdots, M$, the mineralogy log data provides a mineral volume fraction denoted by I_m integrated over the whole borehole interval. The facies analysis provides the depth of the boundaries between different layers measured from the surface that can be used to compute the volume fractions ξ_j of layers $j = 1, \cdots, J$. For each possible lithotype $l = 1, \cdots, L$, M pdfs denoted by $p_{m,l}(x_{m,l})$ are available, one for each mineral describing the likelihood of finding the mineral volume fraction within this lithology as follows:

$$p_{m,l}(x_{m,l})dx_{m,l} = \text{probability of mineral } m \text{ in lithology } l \text{ to have volume fraction } x_{m,l}. \quad (1)$$

From the mineralogy and image log data, it is unknown which lithotypes are present in an interval. A proposition of lithotypes is defined as an assignment of a lithotype to each layer in the interval, i.e., the map $j \leftarrow l$ for $j = 1, \cdots, J$. There are a total of $L!/(L - J)!$ combinations all of which have to be analyzed. It will later be demonstrated that the vast majority of these combinations is infeasible because they cannot satisfy the constraints. Nevertheless, a smart preselection of relevant lithotypes by experienced users might be necessary if L is large.

In Section 2.1, the probability density functions $p_{m,l}(x_{m,l})$ are discussed with particular focus on where the pertinent data can be obtained from. For using a MCMC method [5] to find the most likely $x_{m,j}$ for all m and j, we need to be able to create uniformly distributed samples over the space of all permissible $x_{m,j}$. To this end, constraints and an efficient sampling algorithm are described in Sections 2.2 and 2.3, respectively.

2.1. Probability Density Functions. The goal of the presented algorithm is to find the most likely combination of mineral volume fractions $x_{m,l}$ that satisfies all constraints detailed in Section 2.2. The likelihood of a combination is given by the joint probability density function $p(\vec{x})$, where for convenience of notation, we collect the $x_{m,l}$ in the vector $\vec{x}$. Within this work, the joint probability function is simply the product of the single mineral probability density functions:

$$p\left(\vec{x}\right) = \prod_{m,l=1}^{M,L} p_{m,l}(x_{m,l}). \quad (2)$$

The single probability density functions $p_{m,l}(x_{m,l})$ are input to the presented algorithm and must be specified by the user. This appears to be an equally or maybe more difficult task than to infer the $x_{m,l}$ manually using the geophysicist's experience and knowledge. For making the developed method practical, it is necessary to create a database of $p_{m,l}(x_{m,l})$ that can be reused for subsequent analysis. We decided to implement a bootstrapped-learning algorithm. At first, the geoscientist analyzing the dataset creates probability density functions following geological insight and ensures that the outcome matches the expectations. Once a sufficiently reliable database of probability density functions has been obtained, the algorithm described in Section 2 can be used. As more log data is analyzed, the process is a two-way street: existing probability density functions can be used to analyze results, and actual results can be used to enhance the probability density functions. Missing probability density functions are successively added.

For creating a useful library of probability density functions, we group sets of probability density functions by similarity with respect to general geological indicators, e.g., lithotypes. These geological indicators are known before logging commences at a particular location. A suitable set of probability density functions is retrieved from a database using the set of indicators for the logging location. Note that depending on the general geological features of a location, some minerals or lithologies may be present in one data set but absent in another. It is important to point out that the set of pdfs associated with the particular set of indicators is extended and possibly enhanced as analysis progresses by adding missing probability density functions and fixing inconsistencies of analysis results with observed mineral compositions.

2.2. The Setup of Constraints. Both equality and inequality constraints apply to the solution of the optimization problem. The obvious inequality constraints simply state that the volume fractions need to be between zero and one:

$$0 < x_{m,j} < 1. \quad (3)$$

However, most pdfs are zero almost everywhere, and therefore, it makes sense to set somewhat tighter bounds on the volume fraction $x_{m,j}$:

$$\alpha_{m,j} < x_{m,j} < \beta_{m,j}. \quad (4)$$

There are two types of equality constraints present in the optimization problem. The first type of constraint incorporates the knowledge of the interval integrated mineralogy into the optimization problem. It states that the sum of mineral contents over all layers weighted by the volume fraction of each layer must equal the mineral content determined by the mineralogy tool:

$$\sum_{j=1}^{J} \xi_j x_{m,j} = I_m. \quad (5)$$

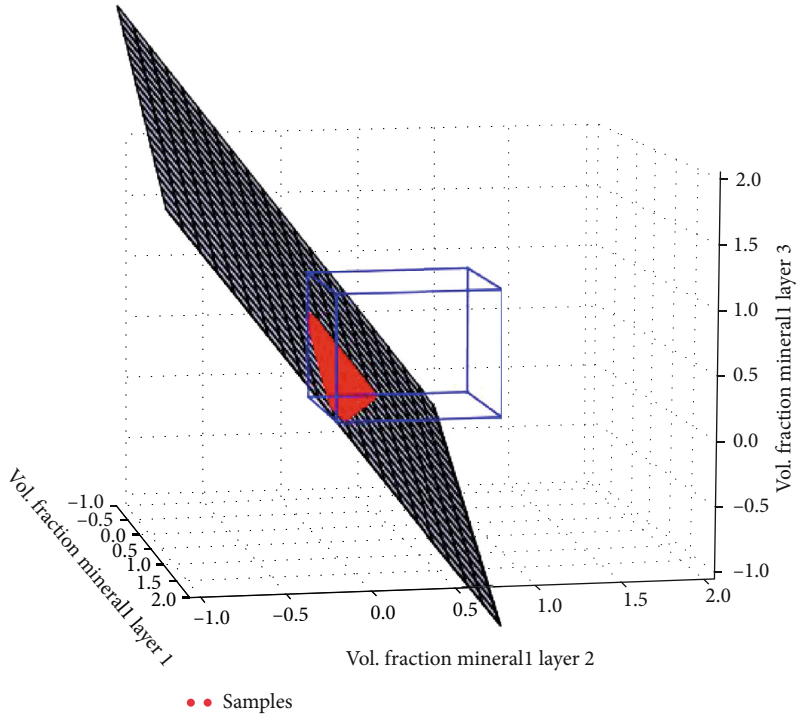


Figure 1

Secondly, the volume fractions of all minerals within each layer must sum to unity:

$$\sum_{m=1}^{M} x_{m,j} = 1. \tag{6}$$

In addition, a lithology l assigned to layer j might have a zero probability to have any traces of mineral m in it and in such cases makes sense to strictly enforce

$$x_{m,j} = 0, \tag{7}$$

to avoid numerical issues in the solution process. There are a total of $M + J$ equality constraints of types given by Equations (5) and (6) and P equality constraints of type Equation (7). The total number of equality constraints is $K = P + M + J$.

We want to create uniformly distributed samples of $x_{m,j}$ that satisfy Equations (5)–(7). To this end, we collect the equality constraints into the matrix E:

$$E\vec{x} = \vec{r}, \tag{8}$$

where $\vec{x}$ collects the $x_{m,j}$, E collects the coefficients of Equations (5)–(7), and $\vec{r}$ collects the right hand sides of Equations (5)–(7).

The matrix E is rank deficient nominally containing more columns than rows (short and wide) and has a nullspace of rank $N = M{\bullet}J - K$. Usually, the matrix is made square by adding N rows consisting of all zeros. As there always exists a vector $\vec{x}$ that satisfies Equation (8), it follows that there are infinitely many solutions to Equation (8) as opposed to none if $\vec{r}$ was not in the range of E.

Basis vectors spanning the null space of E can be determined by using SVD [6]. These orthonormal basis vectors are denoted by $\vec{z}_n, n = 1, \cdots, N$, and are collected in the matrix Z of dimensions $M{\bullet}J \times N$ (it is tall and skinny). The complete set of solutions of Equation (8) can then be written as follows:

$$\vec{x} = \vec{x}_0 + Z\vec{q}, \tag{9}$$

where $\vec{x}_0$ is a particular solution that can, e.g., be computed by using the Moore-Penrose inverse [7]:

$$\vec{x}_0 = E^{-1}\vec{r}. \tag{10}$$

The vector $\vec{q}$ has fewer entries than the vector $\vec{x}$ and denotes a space of reduced dimensionality in which the solution lives. In Figure 1, a three-dimensional space, in which there are three $x_{m,j}$, is depicted that has a single constraint. In this case, the null space is of dimension two permitting samples to be located on a plane. In general settings, the space of permissible samples is a N-dimensional hyperplane in a $M{\bullet}J$ dimensional space.

Using Equation (9), we can use uniform samples of $\vec{q}$ for creating uniform samples of $\vec{x}$ [8]. However, these samples do not necessarily satisfy the inequality constraints, because it is unclear within which limits we need to sample q_n to limit the $x_{m,j}$ to within $\alpha_{m,j}$ and $\beta_{m,j}$ [3].

The inequality constraints Equation (4) define a $M{\bullet}J$-dimensional hyperbox that the valid samples living on the

N-dimensional hyper-plane given by Equation (9) are constrained to. Figure 1 depicts the hyperbox and the hyperplane intersecting it. Valid samples are located on the fraction of the hyperplane that intersects the box and are colored in red.

2.3. Constrained Sampling Algorithm Development. The simplest possible algorithm would be to sample $\vec{q}$ in a large enough space enclosing the hyperbox, e.g., the smallest ellipsoid that fully includes the box, and then use rejection sampling. However, as the dimensionality of the space is much larger than three (usually about 15-20), rejection sampling will be extremely inefficient. This is fairly typical for high-dimensional spaces and is usually referred to as the curse of dimensionality [3].

The algorithm described here is a novel and highly efficient algorithm that does not require rejection and therefore has great capability in addressing high dimension problems as are typical in the described mineralogy analysis.

First, let us assume that we have a seed vector $\vec{x}$ satisfying Equations (4) and (9). A bootstrapping algorithm for obtaining this seed vector is detailed in Section 2.4. Then, we can sample another random vector $\vec{y}$ satisfying the equality constraints Equation (5) through (7) but not the inequality constraints Equation (4). This can be accomplished by first randomly sampling $\vec{q}$ followed by using Equation (9). The difference vector $\vec{d} = \vec{x} - \vec{y}$ lies within the hyperplane defined by the inequality constraints, and all points located on the line drawn through the end points of $\vec{x}$ and $\vec{y}$ are given by

$$\vec{p} = \vec{x} + s\,\vec{d}, \tag{11}$$

where s is the step size that needs to be determined. The limits on the step size are computed by first computing $s_{m,j,\min}$ and $s_{m,j,\max}$:

$$\begin{aligned} s_{m,j,\min} &= \min\left(\frac{\alpha_{m,j} - x_{m,j}}{d_{m,j}}, \frac{\beta_{m,j} - x_{m,j}}{d_{m,j}}\right), \\ s_{m,j,\max} &= \max\left(\frac{\alpha_{m,j} - x_{m,j}}{d_{m,j}}, \frac{\beta_{m,j} - x_{m,j}}{d_{m,j}}\right). \end{aligned} \tag{12}$$

The constraint for s are given by

$$\min\left(s_{m,j,\min}\right) < s < \max\left(s_{m,j,\max}\right). \tag{13}$$

The simplest Algorithm 1 that yields uniformly distributed samples is to perform a random walk as follows:

However, in practice, we found that the following approach performed better given the same execution time:

2.4. Obtaining the Seed Vector. Within this section, an algorithm for obtaining the first seed vector denoted by $\vec{x}$ in Algorithm 2 is introduced. However, a seed vector satisfying inequality and equality constraints does not always exist. It only exists if the hyperplane defined by the equality constraints intersects the box defined by the inequality constraints. Therefore, we first check if a solution exists which is detailed in Section 2.4.1. If a solution exists, obtaining the seed vector is trivial from the previously obtained information. Nevertheless, for the sake of completeness, we describe this procedure in Section 2.4.2.

1. Uniformly sample $\vec{q}$.
2. Compute $\vec{y} = \vec{x} + Z\vec{q}$.
3. Compute $\vec{d} = \vec{x} - \vec{y}$.
4. Uniformly sample s in between $s_{m,j,\min}$ and $s_{m,j,\max}$.
5. Compute $\vec{x}' = \vec{x} + s\,\vec{d}$.
6. Set $\vec{x} = \vec{x}'$ and go to 1.

Algorithm 1: Random walk algorithm.

2.4.1. Check Solution Existence. To facilitate the existence check, we first introduce a coordinate transformation such that limits $\alpha_{m,j}$ and $\beta_{m,j}$ are mapped to zero and one, respectively. Quantities expressed in these new coordinates are denoted by a tilde to distinguish them from the original coordinate system. The coordinate transformation is given by

$$\tilde{x}_{m,j} = \frac{1}{\beta_{m,j} - \alpha_{m,j}} x_{m,j} - \frac{\alpha_{m,j}}{\beta_{m,j} - \alpha_{m,j}}. \tag{14}$$

For compactness of notation, we collect all the coefficients of the first term into the matrix T and the second term of the different into the vector $\vec{b}$ such that we can write

$$\vec{\tilde{x}} = T\vec{x} + \vec{b}. \tag{15}$$

All permissible solution, i.e., all solutions on the hyperplane, can now be expressed in this new coordinate system:

$$\vec{\tilde{x}} = TZ\vec{q} + T\vec{x}_0 + \vec{b}. \tag{16}$$

The transformation transforms the hyperbox to be the unit cube around the midpoint $\vec{l} = [1/2, \cdots, 1/2]$. The unit cube corresponds to the locus of points whose distance from $\vec{l}$ is less than or equal to 1/2 in the maximum norm:

$$\text{Interior of the unit cube}: \left\| \vec{l} - \vec{\tilde{x}} \right\|_\infty \leq \frac{1}{2}. \tag{17}$$

The vector $\vec{\tilde{x}}$ is the locus of all points on the hyperplane as given by Equation (9). Hence, by solving the linear optimization problem,

$$g = \min_{\vec{q}} \left\| \vec{y} - P\vec{q} \right\|_\infty, \tag{18}$$

$$\vec{y} = \vec{l} - T\vec{x}_0 - \vec{b}, \tag{19}$$

$$P = TZ, \tag{20}$$

1. Uniformly sample $\vec{q}$.
2. Compute $\vec{y} = \vec{x} + Z\vec{q}$.
3. Compute $\vec{d} = \vec{x} - \vec{y}$.
4. Perform a line search between $s_{m,j,\min}$ and $s_{m,j,\max}$ to find the maximum $\vec{x}^*$ of the joint probability density $p(\vec{x})$.
5. Set $\vec{x} = \vec{x}^*$ and go to 1.

Algorithm 2: Line search algorithm.

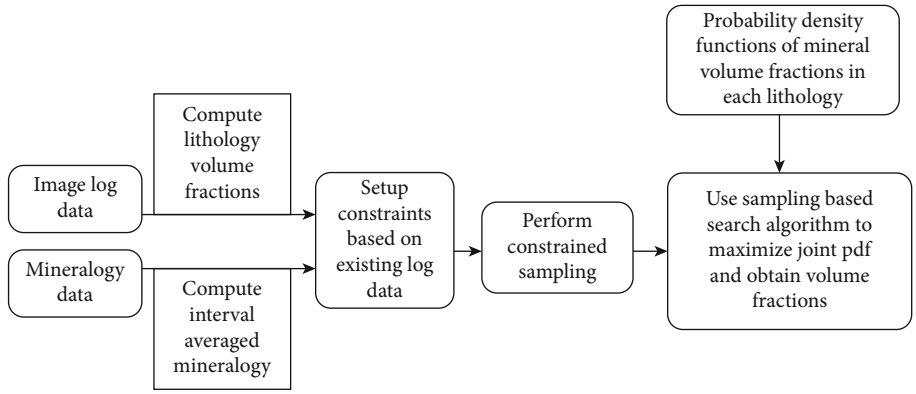


Figure 2

the condition for the hyperplane to intersect the hyperbox simply is follows:

$$g \leq \frac{1}{2}. \tag{21}$$

The solution to the minimization problem Equation (20) is obtained by transforming it into the corresponding linear programming problem [9, 10]:

Minimize t subject to

$$\begin{aligned} \vec{y} - P\vec{q} &\leq t\vec{1}, \\ \vec{y} - P\vec{q} &\geq -t\vec{1}, \end{aligned} \tag{22}$$

where $\vec{1}$ is a vector of the correct length of all ones.

2.4.2. Algorithm for Obtaining a Seed Vector. Once existence of a solution is established, one can simply use the point on the hyperplane with the smallest distance in the infinity norm to the center of the box and transform it back to the original basis. To this end, let us denote the solution as the minimization problem as $\vec{q}^*$. Then, we substitute $\vec{q}^*$ into Equation (16) and apply the inverse transformation:

$$\vec{x} = T^{-1}\left(TZ\vec{q}^* + T\vec{x}_0 + \vec{b} - \vec{b}\right) = \vec{x}_0 + Z\vec{q}^*. \tag{23}$$

While $\vec{x}$ is a valid seed vector for the random walk algorithm, it was found that it could be far away from the maximum. A better initial guess can be obtained by moving it closer to where we expect the maximum. For each one-dimensional pdf used for forming the joint pdf in Section 2.1, we can easily infer where it assumes its maximum value. Let us denote this point as $x_{m,j}^{\max}$ and collect all of these abscissas in the vector $\vec{x}^{\max}$. Then, we estimate the maximum of the joint pdf to be close to $\vec{x}^{\max}$, but we note that it does not satisfy either the equality or inequality constraints. Therefore, we project the vector onto the hyperplane described by Equation (9) and denote the projection as $\vec{x}_\perp^{\max}$. While $\vec{x}_\perp^{\max}$ is guaranteed to satisfy the equality constraints, it could possibly be in violation of the inequality constraints. For approximating the best possible estimate close to $\vec{x}_\perp^{\max}$, we solve the following one-dimensional maximization problem for obtaining the best step size s_{step}:

$$s_{\text{step}} = \max_s p\left(\vec{x} + s\left(\vec{x}_\perp^{\max} - \vec{x}\right)\right) \tag{24}$$

and then update the initial seed for the random walk process as follows:

$$\vec{x} \leftarrow \vec{x} + s_{\text{step}}\left(\vec{x} + s\left(\vec{x}_\perp^{\max} - \vec{x}\right)\right). \tag{25}$$

Figure 2 shows an overview of the mathematical solver flowchart.

2.5. Digital Core Interpretation Model. This algorithm, presented above, has been consequently implemented into an interpretation model denoted as Digital Core, because this algorithm has the potential to replace the conventional,

expensive coring process in well logging and thus provides a comprehensive characterization of formation core.

This interpretation model executes the algorithm based on a multistep workflow as follows:

(1) Determine zones of consistent mineralogy and resistivity response

(2) Determine the volume of each lithology from image logs using facies

(3) Execute the algorithm as described in Section 2.2

(4) Obtain the lithology and the mineralogical compositions to their corresponding facies

(5) Conduct quality control: a moving average at the resolution of the pulsed neutron log. The average should fit with the pulsed neutron measurement

(6) Produce a set of deliverables that typically include:

(i) Lithology corresponding to each volumetric fraction in the analyzed section

(ii) High-resolution mineralogy corresponding to each lithology layer

3. Manufactured Test Case

In this section, the Digital Core model is applied to a manufactured test case and is consequently validated as discussed below.

3.1. Manufacturing the Reference Solution and Preparing Input Data. We selected a three-layer test case and used three lithologies, sandstone, shale, and coal (note: usually, the number of layers and the total number of tested lithologies are not identical). The pdfs for these lithologies are known. The volume fractions for the three layers are selected and listed in Table 1. Along with these volume fractions, we selected that layer 1 was sandstone, layer 2 was shale, and layer 3 was coal.

From the pdfs, the most likely composition for each lithology is selected. Then, we use the volume fraction to compute the averaged mineralogy that typical pulsed neutron geochemical tool measures. This data is detailed in Table 2.

3.2. Numerical Results. To simulate noise, we perturb the mineralogy data, the layer volume fractions (i.e., the image log data), and both the pdfs' abscissa and ordinate values using Gaussian noise with noise levels chosen as the Gaussian's standard deviation being x% of the mean value of the perturbed quantity. We selected x to be 0, 1, 2.5, 5, and 10.

Using the 3 lithologies and 3 layers, there are a total of 6 ways of assigning the lithologies to the different layers. The algorithm always found the correct permutation, i.e., 1: >sandstone, 2: >shale, and 3: >coal, for $x < 10\%$. Moreover, all other permutations were rejected. For $x = 10\%$, out of 5 trials, the VC returned:

Table 1: Volume fraction and lithology in each layer.

Layer	Volume fraction	Lithology
1	0.55	Sandstone
2	0.3	Shale
3	0.15	Coal

Table 2: Mineralogy of each layer and the computed average mineralogy.

	Layer 1	Layer 2	Layer 3	Measured
Organic	0.00000	0.00937	0.88791	0.13600
Kaolinite	0.03107	0.10146	0.02823	0.05176
Quartz	0.90045	0.35264	0.02823	0.60527
Siderite	0.01850	0.03883	0.05046	0.02939
Illite	0.05125	0.49746	0.00677	0.17844

(i) It rejected all permutation once

(ii) It found permutations (1: >sandstone, 2: >shale, and 3: > coal) and (1: >sandstone, 2: >coal, and 3: > shale) to be viable but found that the correct permutation was 50% more likely

(iii) It accepted the correct permutation and rejected all others 2 times

(iv) It accepted the incorrect (1: >sandstone, 2: >coal, and 3: > shale) and rejected all other combinations

Note that 10% standard deviation about the mean could make the volume fractions of layer 2 and layer 3 "switch places" in which case the algorithm cannot distinguish (1: >sandstone, 2: >coal, and 3: > shale) would be the correct answer!

For $x = 0$, 1, 2.5, and 5, selected predicted and reference mineralogies are depicted in Figure 3.

4. Field Data Case Study

A representative field data example is presented in this section and compared to core data to demonstrate the feasibility of the Digital Core interpretation model. This case study proves that the introduced method is readily capable of generating a mineralogical record with high vertical resolution for a formation logged with a pulsed neutron geochemistry tool and a high-resolution borehole imager.

4.1. Field Data Collection. A well from Australia where all log information and a core section of 10 ft was collected. This section was used to test the method with the result presented in Figure 4. The mineralogy was recorded with a pulsed neutron tool and was calibrated to XRD measurements. The borehole image was recorded with a resistivity imaging tool, processed, and flattened, and a high-resolution resistivity curve was generated. At its conventional resolution, the spectroscopic mineralogy shows relatively constant weight fractions for illite (37.5%), kaolinite (10.5%), siderite (about

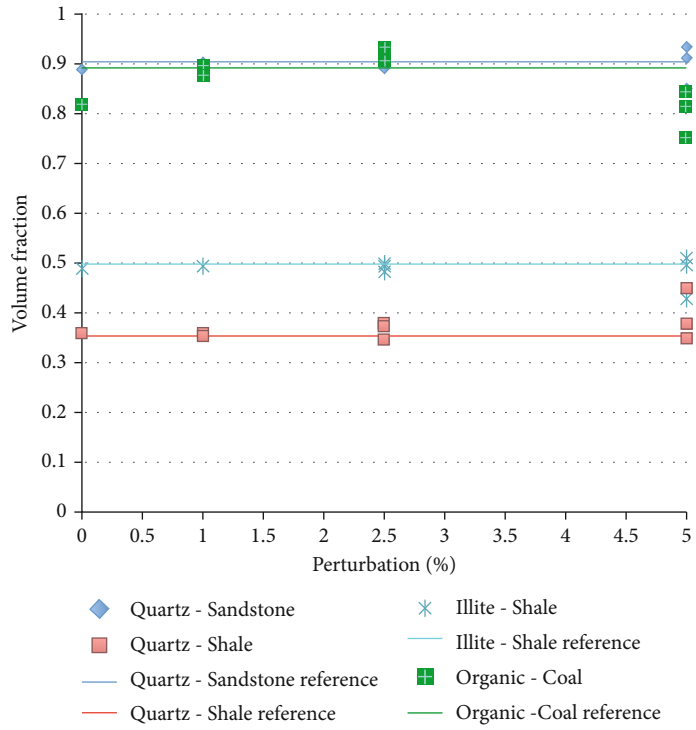


Figure 3

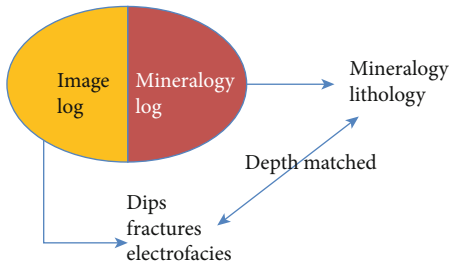


Figure 4

7%), quartz (44%), and organic carbon (about 1%), which are consistent with values generally encountered in this formation for this part of the basin.

The resistivity image log is used to determine facies over the cored interval. It shows a series of thin beds with varying resistivities that were discriminated in three facies: less than 140 ohm·m, assumed to be more argillaceous silts identified as "shale," 140 to 180 ohm·m, expected to contain more quartz and labelled as "silt," and higher than 180 ohm·m, expected to be siderite-cemented silts identified as "cemented shale." These labels indicate which type of formation would have a similar mineralogical composition: for instance, "shale" means that the formation has a mineralogical composition similar to that of a shale. The core shows that the formation consists of a siltstone with variations of mineralogical composition.

4.2. Field Data Processing and Interpretation Analysis. Based on XRD data available for this well, PDFs are generated for each of these lithological components and are consequently used in the algorithm. The algorithm determined two possible solutions, each characterized by the probability (the most likelihood) that this solution could occur: only the solution with higher probability is presented in this communication.

Table 3: Comparison between the weighted average mineralogical composition (WAC) and the average composition determined from the spectroscopic mineralogy log for the 2 solutions determined by the algorithm. 70%: solution with 70% probability; 30%: solution with 30% probability. Mineral compositions expressed as % of the formation weight; EF: electrofacies, in ohm·m.

(a)

70%	Illite	Kaolinite	Organic	Siderite	Quartz	EF
Shale 1	50.0	10.0	2.3	4.0	33.7	<140
Silt	26.6	10.5	0.4	2.4	60.2	<140<180
Shale 2	35.0	10.0	1.5	24.3	29.2	180<
WAC	37.1	10.2	1.3	6.8	44.6	
Log	37.3	10.3	1.2	7.0	44.2	

(b)

30%	Illite	Kaolinite	Organic	Siderite	Quartz	EF
Silt	25.0	10.5	0.1	2.0	62.3	<140
Shale 1	48.7	10.0	2.4	4.0	34.8	<140<180
Shale 2	35.0	10.0	1.2	24.7	29.1	180<
WAC	37.1	10.2	1.3	6.8	44.6	
Log	37.3	10.3	1.2	7.0	44.2	

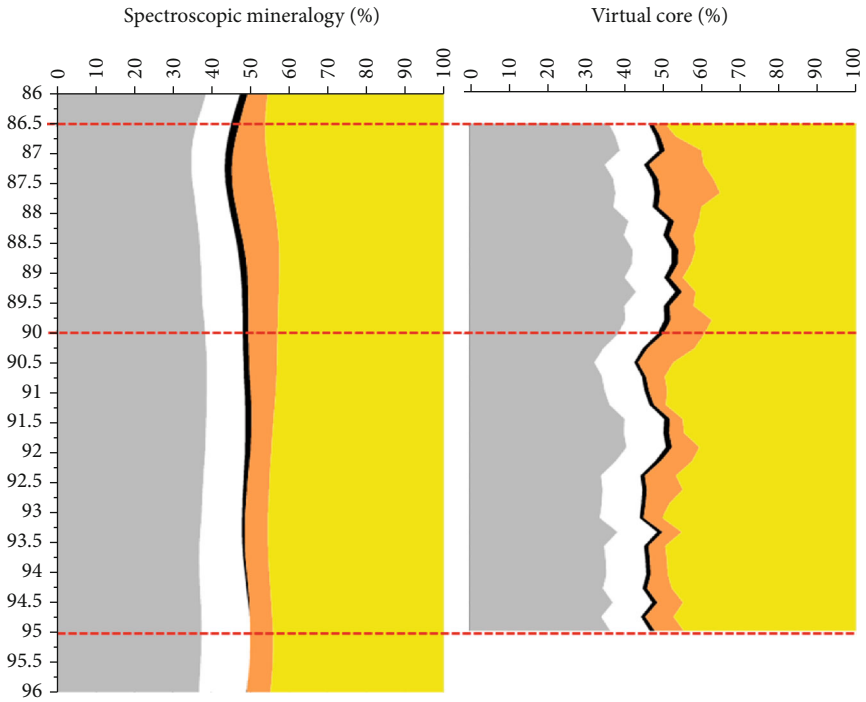


Figure 5

Table 3 shows that the average mineralogical compositions for each solution are consistent with the average mineralogical composition determined from the spectroscopy log over the interval, indicating that both solutions are mathematically acceptable. The solution with the highest probability is presented in Figure 1 along with the core. Aside from depth matching issues, the highly resistive layers observed on the image and associated with siderite-cemented intervals correspond to brownish streaks marking the occurrence of siderite cements on the core. The distribution of the "shale" facies (more argillaceous sediments) also generally corresponds to the darker intervals in the core. The high-resolution mineralogy, after applying a 1 ft mobile average filter, shows a good agreement with the spectroscopic mineralogy recorded by the logging tool (Figure 5).

Some discrepancies are observed in the distribution of "silt" facies in the upper part of the log where an interval with argillaceous composition corresponds to light-colored levels in the core (86-87 ft) (Figure 6). The occurrence of the "shale" can be explained by the dark streak on the image, probably related to data acquisition that affected the determination of the facies. At 87 ft, a siderite-cemented interval could not be located with confidence on the core. This might be caused by the lighter color of the silt, next to that of the siderite-cemented shale in the picture. However, the complex variation of mineralogy between 87.5 and 88.5 ft, as well as below 90 ft, displays a good correspondence to the color variations observed on the core.

In complex environments as the one in the example, where the grain size distribution does not change significantly and the laminations mostly correspond to variations in mineralogy, this result demonstrates that the method is globally working: the high-resolution mineralogy displays a

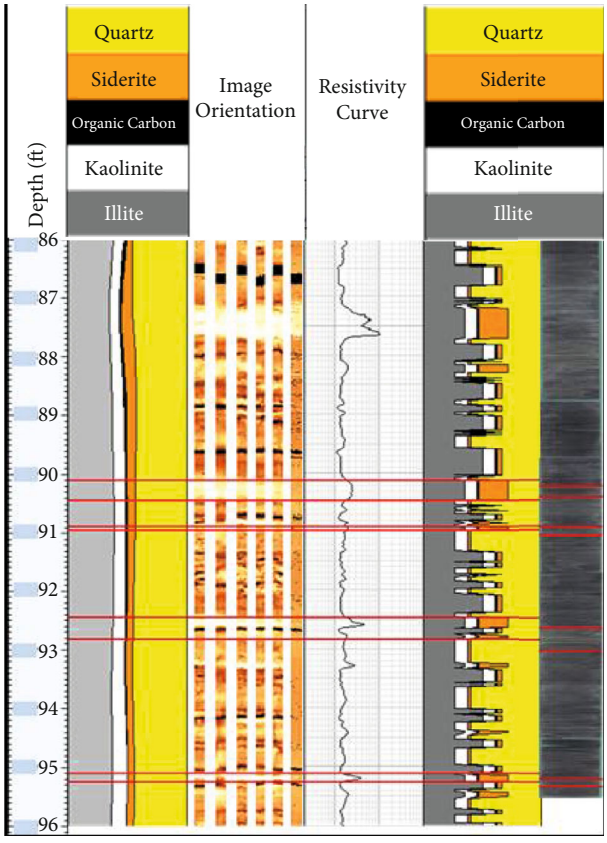


Figure 6

fairly representative account of the variability observed in the core. The level at 86-87 ft shows that the method is sensitive to thresholding while defining the facies. A good knowledge of the mineralogy for the logged formation and the geology of the area would naturally be important as lateral variation can occur: this ensures the PDF can take into account the regional variations of facies. The process itself, including the workflow and the algorithm, appears to be both robust and versatile to account for complex formations such as siltstone, thinly bedded sandstones, and most probably thinly bedded limestone or partially dolomitized limestone.

Further data from other wells have also been explored, and similar performance is achieved from the algorithm. Due to sensitivity issue, this work only presents one field data case as discussed above in this section.

5. Conclusion

The presented method enables a direct quantification of mineralogy at the scale of the lamination in thinly bedded formations and thus allows for determining organic content, matrix density, and total porosity at high resolution, resulting in a more accurate calculation of net pay and net-to-gross ratio. The method provides great advantage over conventional mineralogy log interpretation by revealing the full vertical variability of a formation that would otherwise appear very homogeneous, revealing by-passed potentially targets. The interpretation methodology, once fully implemented and validated for the mineralogy log, could also be applied to other nuclear logs such as density and porosity logs, as well as NMR and acoustic logs. Furthermore, a future work will combine machine learning into various aspects of the Digital Core, e.g., facies recognition and log prediction.

Data Availability

Underlying data is available upon request.

Disclosure

The initial work has been presented in 2016 SEG Conference with an abstract submitted to the 2016 SEG Technical Program.

Conflicts of Interest

The authors declare that there are no potential conflicts of interest, such as financial interests, affiliations, or personal interests or beliefs, that could be perceived to affect the objectivity or neutrality of the manuscript.

Acknowledgments

The authors thank Alberto Mezzatesta and Lena Thrane for their support and input towards this work. This work is funded by research funds from the University of Electronic Science and Technology of China.

References

[1] Q. Zhang, F. Mendez, J. Longo, S. Gade, and J. Peyaud, "Development of reduced order modelling techniques for downhole logging sensor design," in *SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition*, Bali, Indonesia, Oct. 2015.

[2] Q. Zhang, "Development of a surrogate model for elemental analysis using a natural gamma ray spectroscopy tool," *Applied Radiation and Isotopes*, vol. 104, pp. 5–14, 2015.

[3] D. P. de Farias and B. van Roy, "On constraint sampling in the linear programming approach to approximate dynamic programming," *Mathematics of Operations Research*, vol. 29, no. 3, pp. 462–478, 2004.

[4] J. B. Peyaud, M. Pagel, J. Cabrera, and H. Pitsch, "Mineralogical, chemical and isotopic perturbations induced in shale by fluid circulation in a fault at the Tournemire experimental site (Aveyron, France)," *Journal of Geochemical Exploration*, vol. 90, no. 1–2, pp. 9–23, 2006.

[5] P. Diaconis, "The Markov Chain Monte Carlo Revolution," *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 179–205, 2009.

[6] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins, 3rd ed edition, 1996.

[7] R. Penrose, "A generalized inverse for matrices," *Proceedings of the Cambridge Philosophical Society*, vol. 51, no. 3, pp. 406–413, 1955.

[8] K. Van den Meersche, K. Soetaert, and D. V. Oevelen, "xsample(): AnRFunction for sampling linear inverse problems," *Journal of Statistical Software*, vol. 30, no. Code Snippet 1, 2009.

[9] E. Stiefel, "Note on Jordan elimination, linear programming and Tchebycheff approximation," *Numerische Mathematik*, vol. 2, no. 1, pp. 1–17, 1960.

[10] G. B. Dantzig and M. N. Thapa, *Linear Programming 1: Introduction*, Springer Verlag, 1997.

Hindawi
Journal of Sensors
Volume 2020, Article ID 6622047, 11 pages
https://doi.org/10.1155/2020/6622047

Research Article

Variable Aperture Method of Ultrasonic Annular Array for the Detection of Addictive Manufacturing Titanium Alloy

Wenchao Li[1], **Junjie Chang**[1], **Wentao Li**[2] **and Xiaoyun Long**[3]

[1]*Key Lab of Nondestructive Testing, Ministry of Education, Nanchang Hangkong University, Nanchang 330063, China*
[2]*School of Mechanical and Electronical Engineering, Lanzhou University of Technology, Lanzhou 730050, China*
[3]*College of Information Engineering, Nanchang University, Nanchang 330031, China*

Correspondence should be addressed to Junjie Chang; changjunjie_nhu@126.com

Received 18 November 2020; Revised 14 December 2020; Accepted 16 December 2020; Published 31 December 2020

Academic Editor: Bin Gao

The ultrasonic annular array transducer usually has a stronger focusing acoustic field than the linear array and matrix transducer with the same number of array elements, and is more suitable for the detection of large thickness and high attenuation components. However, due to the special arrangement of array elements, the focusing beam cannot be deflected and has a large near field, which limits its application in practical detection. The element parameters of annular array transducer are often designed and analyzed according to the 2-D acoustic field model of a linear array transducer. Therefore, the 3-D acoustic field distribution model of the annular array transducer is established, and the influence of the element parameters on its spatial acoustic field focusing characteristics is analyzed. The design criteria of the array element division mode and element size are proposed, which can avoid the generation of high-energy side lobe and grating lobe, and have good axial acoustic field. Then, the influence of excitation aperture on the energy and size of focal spot at different depths is discussed. The dynamic focusing method with variable aperture of annular array is established, and the C-scan detection experiment is carried out on the addictive manufacturing titanium alloy specimen. The detection results show that the variable aperture method has better central amplitude consistency and imaging accuracy for different depth defects, and has better near surface detection ability than the fixed aperture method.

1. Introduction

Ultrasonic array technology is a kind of multichannel ultrasonic testing technology which arranges several piezoelectric wafers into an array according to a certain combination mode. By controlling the excitation sequence and delay time of piezoelectric wafers, the deflection and focusing of synthetic acoustic beam can be realized. At present, ultrasonic array transducers commonly used in industry can be divided into 1-D linear array, 1.5-D annular array, 2-D matrix array, etc. according to the element arrangement mode [1]. At present, ultrasonic the linear array transducer and ultrasonic matrix transducer are the most widely used in theoretical research and industrial application. The ultrasonic matrix array transducer can realize 3-D imaging [2], but the acoustic beam control algorithm is complex, and the manufacturing process and equipment hardware costs are high [3–5]. The ultrasonic linear array transducer is easier to manufacture, and its transmitting and receiving delay control method is relatively simple, which is mostly used in practical applications. However, the C-scan detection results of ultrasonic linear array transducer are affected by its focal spot asymmetry, and the shape quantitative error of defects in ultrasonic linear array C-scan results is large [6]. The single element size of ultrasonic annular array transducer is larger, and it can achieve stronger focusing energy than ultrasonic linear array transducer with fewer array elements [7]. Meanwhile, its focal spot is completely symmetrical along the radial direction. It is an effective method to solve the problem of low signal-to-noise ratio (SNR) and large defect distortion in C-scan testing results of large thickness and high attenuation materials. However, due to the fact that the acoustic beam of ultrasonic annular array can only focus along the central axis, the beam cannot be deflected and the near field is large,

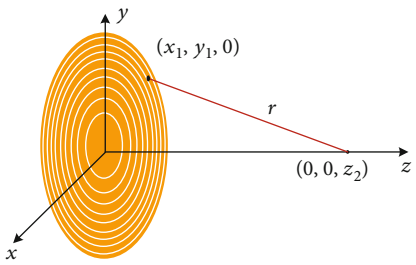


Figure 1: Calculation diagram of acoustic field of annular transducer.

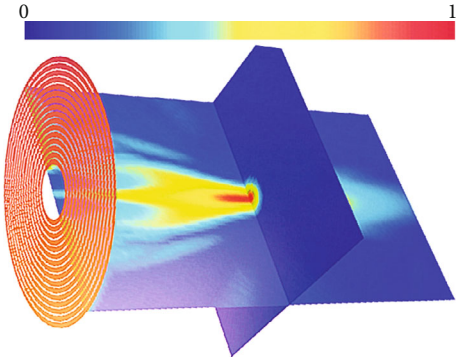


Figure 2: The focusing acoustic field of ultrasonic annular array transducer.

so there is little research and application in the industrial field, and the design and verification method of the parameters and focusing algorithm of the ultrasonic annular array transducer are insufficient.

Firstly, based on the 3-D acoustic field calculation theory, the influence of the parameters of the ultrasonic annular array transducer on the focusing acoustic field characteristics is analyzed, and the design criteria of the detection parameters suitable for the annular array transducer are proposed. Then, the influence of focusing depth and excitation aperture size on the focal spot size and acoustic energy of annular array transducer is analyzed, and a variable aperture dynamic focusing method of annular array is established. The C-scan experimental results of additive titanium alloy show that the variable aperture dynamic focusing method can improve the detection ability and sensitivity of different depth defects. It provides theoretical guidance for the application of ultrasonic annular array transducer in the detection of large thickness and high attenuation components.

2. Influence of Transducer Parameters on Acoustic Field

2.1. Focusing Acoustic Field of Annular Array. Different from the acoustic field calculation of ultrasonic linear array transducer, only calculating the 2-D cross-sectional acoustic field along the array direction cannot accurately reflect the real focusing acoustic field of annular array transducer [8]. It is necessary to calculate the superimposed acoustic field in 3-D space after the circular array element is discretized at a certain angle along the circumference. Firstly, the physical state relationship between any two points in space is established according to the transformed wave equation [9]:

$$\int_{S_T} \left(\overline{P_2} \frac{\partial \overline{P_1}}{\partial n} - \overline{P_1} \frac{\partial \overline{P_2}}{\partial n} \right) dS_T = \int_V \left(\overline{P_1} \overline{f_2} - \overline{P_2} \overline{f_1} \right) dV. \quad (1)$$

Here, P_1 and P_2 are the acoustic pressures of two points in the medium at a certain time, f is the body force, S_T is the element area of the transducer, and V is the outside of the whole boundary of the array element. To calculate the acoustic field distribution in space, Green's function $\bar{G}(x; y, \omega)$ in 3-D space should be solved firstly:

$$\bar{G}(x; y, \omega) = \frac{\exp(-jkr)}{4\pi r}, \quad (2)$$

where k is the wavenumber and r is the distance from a coordinate point to a fixed point in 3-D space.

As shown in Figure 1, it is expressed as the distance from any point of transducer to any point on the central axis, i.e., $r = \sqrt{x_1^2 + y_1^2 + z_2^2}$.

Assuming that $\delta(x - y)$ is the shock response function and $\bar{G}(x; y, \omega)$ is the acoustic solution of the function as a body force f, the two solutions of the wave equation can be assumed as follows:

$$\begin{cases} \bar{P}_1(x, \omega) = \bar{P}(x, \omega), \\ \bar{f}_1(x, \omega) = 0, \end{cases} \quad \begin{cases} \bar{P}_2(x, \omega) = \bar{G}(x; y, \omega), \\ \bar{f}_2(x, \omega) = \delta(x - y). \end{cases} \quad (3)$$

Take Equation (3) into Equation (1), and then express it as follows:

$$\bar{P}(y, \omega) = \int_S \left[\bar{G}(y; x, \omega) \frac{\partial \bar{P}(y, \omega)}{\partial n} - \bar{P}(y, \omega) \frac{\partial \bar{G}(y; x, \omega)}{\partial n} \right] dS. \quad (4)$$

Combined with the deformation formula $\bar{P} = -j\omega\rho_0\bar{\varphi}$ of Newton's second law and Equation (2), for any point in the radiation space, the total acoustic pressure of the synthetic beam can be obtained by adding the acoustic pressure of N elements with the following equation.

$$p(x, t) = \sum_{i=1}^{N} \sqrt{\frac{c}{t}} \otimes \int_{S_T} \frac{V_n(t - (r/c))}{4\pi r} dS_T(t), \quad (5)$$

where V_n is the excitation signal of the array element, x and t represent the spatial position and time, respectively, c is the

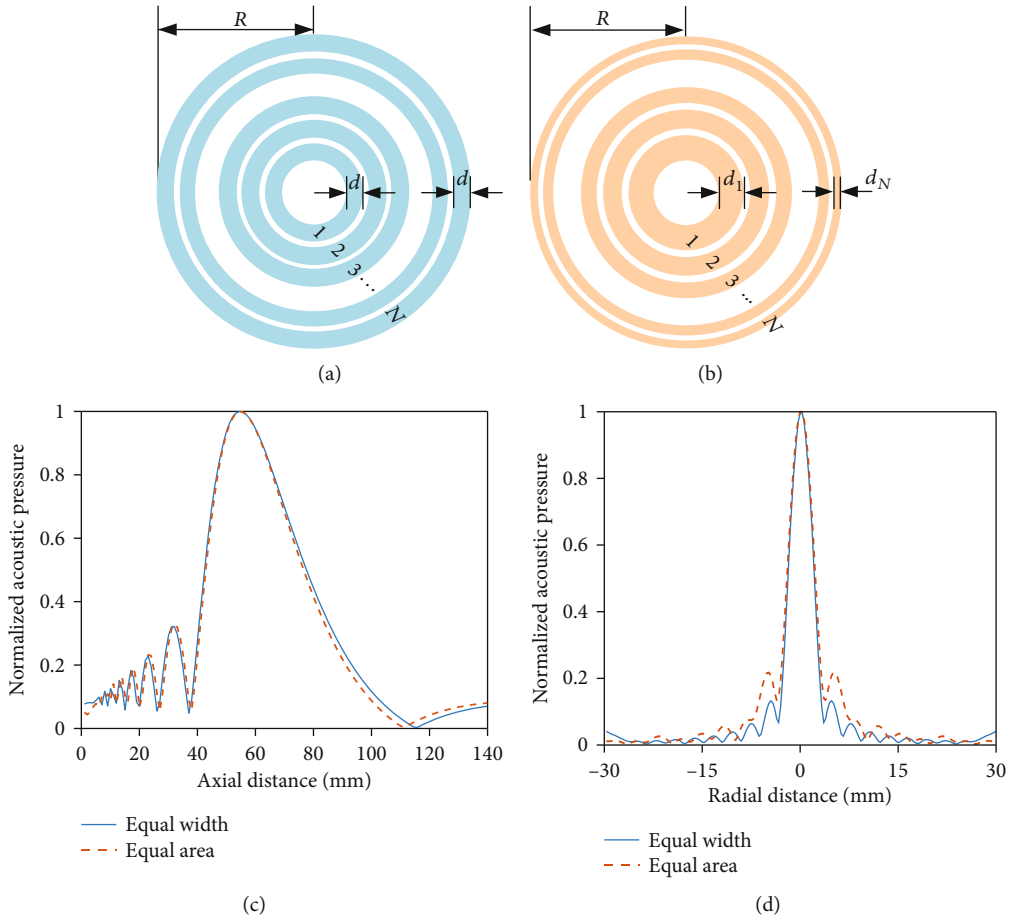


Figure 3: Acoustic pressure distribution of two types of annular array elements. (a) The element with equal width. (b) The element with equal area. (c) Axial acoustic field distribution. (d) Radial acoustic field distribution.

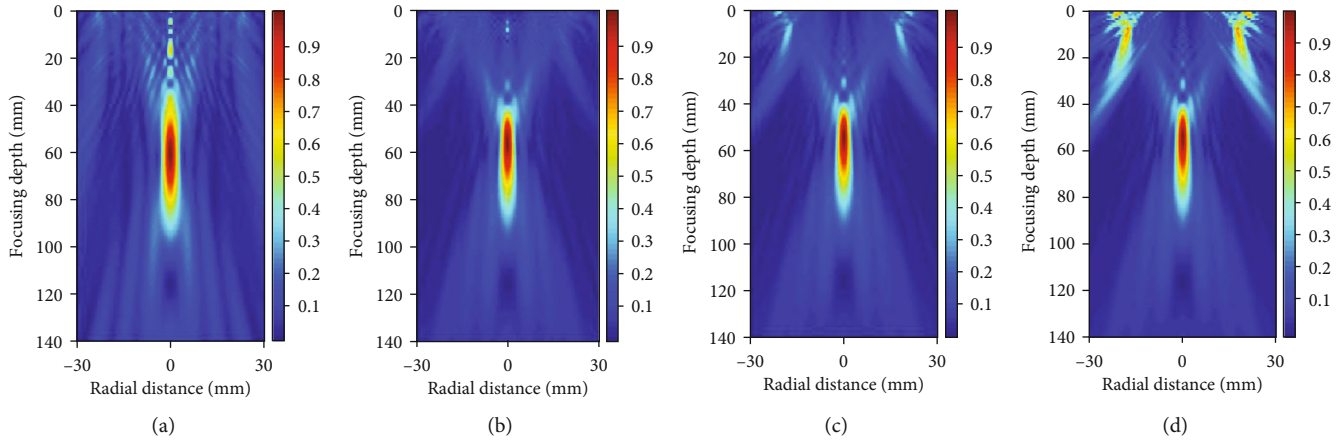


Figure 4: Focusing acoustic field distribution of annular array with different array element sizes. (a) $N = 4$, $d = 2.6\,\text{mm} \approx 4\lambda$. (b) $N = 8$, $d = 1.3\,\text{mm} \approx 2\lambda$. (c) $N = 16$, $d = 0.65\,\text{mm} \approx \lambda$. (d) $N = 32$, $d = 0.325\,\text{mm} \approx 0.5\lambda$.

sound velocity of the medium, and S_T is the array element's area [10, 11].

The spatial focusing acoustic field of the ultrasonic annular array transducer is shown in Figure 2.

2.2. Selection and Design Method of Array Element Parameters. When the total element area is fixed, the annular array transducer element can be divided into two types: the element with equal width and the element with equal area,

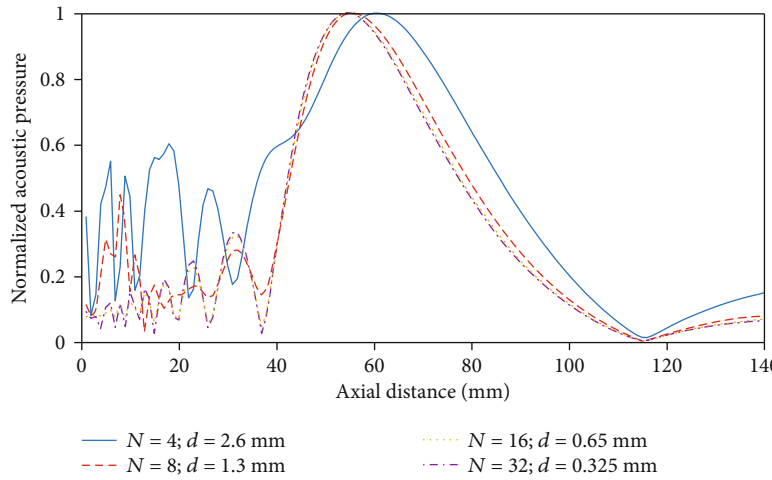


Figure 5: Axial acoustic pressure of annular array with different array element sizes.

Table 1: Simulation parameters of ultrasonic annular array.

No.	Element pitch (mm)	Focusing depth (mm)	Resolution (mm)	Focusing method
1	0.65	20-120	0.1	Fixed aperture (16)
2	0.65	20-120	0.1	Fixed F/D (~4.4)
3	0.65	20-120	0.1	Fixed F/N_f (~0.5)

as shown in Figure 3(a) and Figure 3(b). The area of the element increases with the outer element from the inner element under the partition form of equal element width. The width of the outer element will decrease exponentially with the increase of the number of elements when the area of each element is equal. Figure 3(c) and Figure 3(d) show the acoustic pressure distribution at the depth of 60 mm of the annular array transducer with 12 elements of the equal width and equal area array elements. It can be seen that the difference of axial acoustic field distribution between the two array element partition methods is very small, and the radial focal spot size is basically the same. However, when the number of array elements required is large, the width difference between the outer element and the inner element will be too large, leading to the difficulty of micromachining.

The number and width of the elements can be directly affected by the size of the elements when the total area of the circular array element is fixed. The influence of different element sizes on the focusing acoustic field of annular array transducer is analyzed to establish the design criteria of element parameters. Figure 4 shows the 2-D cross-sectional simulation results of 3-D focusing acoustic field distribution of annular array with different array element partition sizes.

Due to the equal area of the total array elements, the focal spot sizes are close to each other under different element sizes [12]. However, different from the design criteria for the elements of ultrasonic linear array transducer, the ultrasonic annular array transducer still has strong main lobe energy at 60 mm focal depth when there are only $N = 4$ elements. The influence of element width on the side lobe and grating lobe of focusing acoustic field is also different from that of ultrasonic linear array transducer: When the element width d is close to half wavelength, as shown in Figure 4(d), there are strong grating lobes on both sides of the transducer. However, the side lobe with higher energy will be produced when the element width $d \approx 4\lambda$, as shown in Figure 4(a). Figure 4(b) and Figure 4(c) show that when the element width of the ultrasonic annular array transducer is between the wavelength and twice the wavelength, that is, when $\lambda \leq d \leq 2\lambda$, the side lobe and grating lobe with higher energy can be avoided at the same time. In addition, the axial acoustic field distribution characteristics under different array element sizes are further analyzed to better realize the full depth range detection, as shown in Figure 5.

The energy and size of the main lobe of the focusing beam are mainly determined by the element number when the total array area is fixed. According to the results of the axial acoustic pressure, the focal spot sizes of the three results are consistent when the number of elements $N \geq 8$. For the axial acoustic pressure in the nonfocusing region, the peak value of side lobe has reached 50% of the peak value of the main lobe in the initial 20 mm depth when $d = 2.6$ mm $\approx 4\lambda$ and $d = 1.3$ mm $\approx 2\lambda$. When $d = 0.65$ mm $\approx \lambda$ and $d = 0.325$ mm $\approx 0.5\lambda$, the acoustic pressure fluctuation in the near-field region is relatively stable. In conclusion, when the element width $d \approx \lambda$, the ultrasonic annular array transducer has the best overall focusing acoustic field characteristics, which can avoid the generation of high-energy side lobe and grating lobe, and has better axial acoustic field.

3. Variable Aperture Focusing Detection Method

3.1. Theory Analysis. The distribution of the focusing acoustic field of the ultrasonic annular array transducer is like that of the single crystal disc with different focal length lenses, that is, it satisfies the ultrasonic focusing theory of the disk

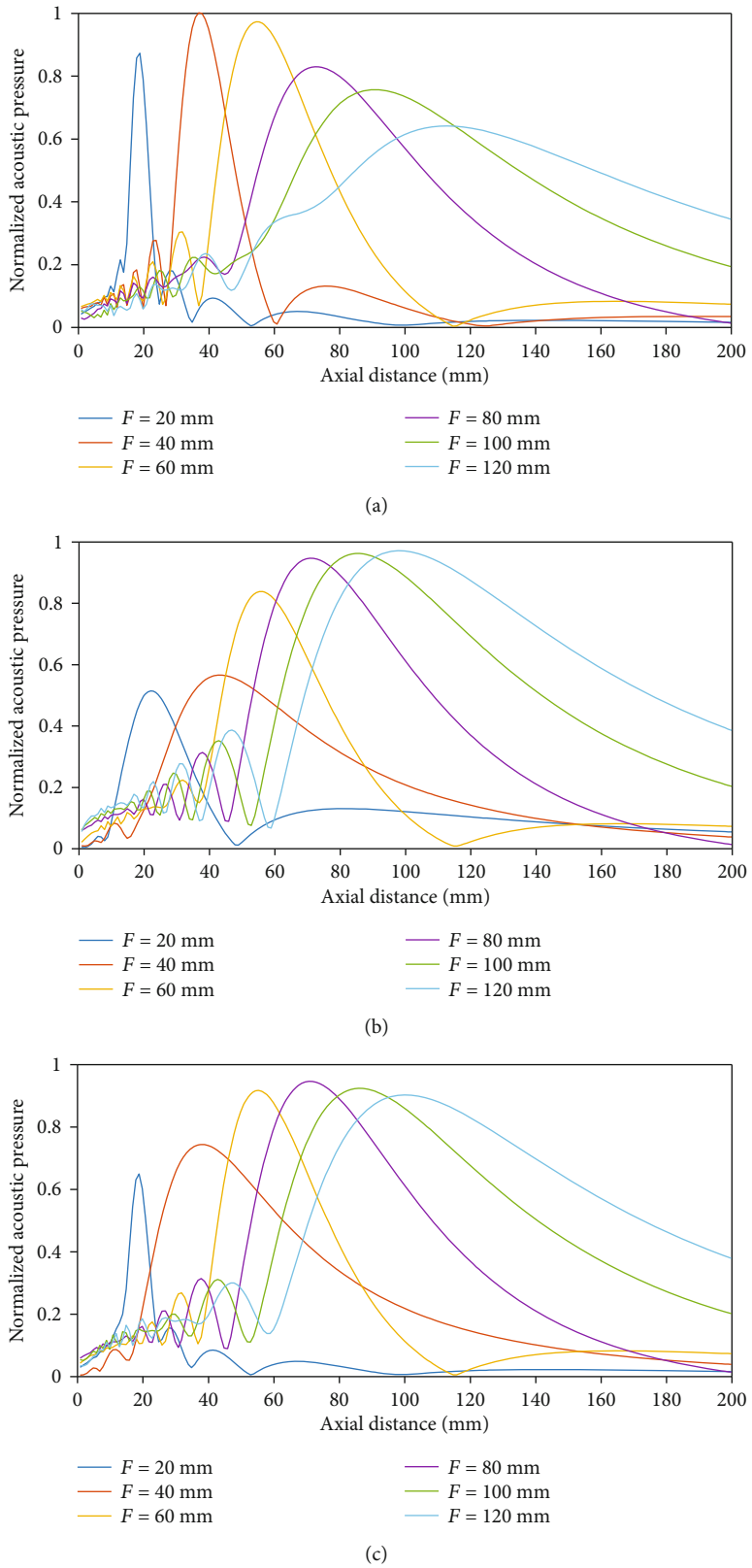


FIGURE 6: Axial acoustic field distribution of annular array with different aperture excitation methods. (a) Fixed aperture method. (b) Variable aperture method for fixed F/D. (c) Variable aperture method for fixed F/N_f.

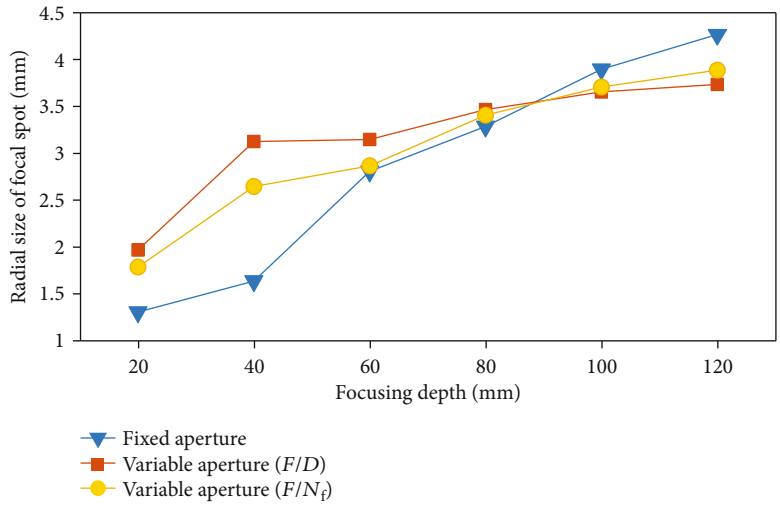


Figure 7: Variation of radial size of focal spot with different focusing depths.

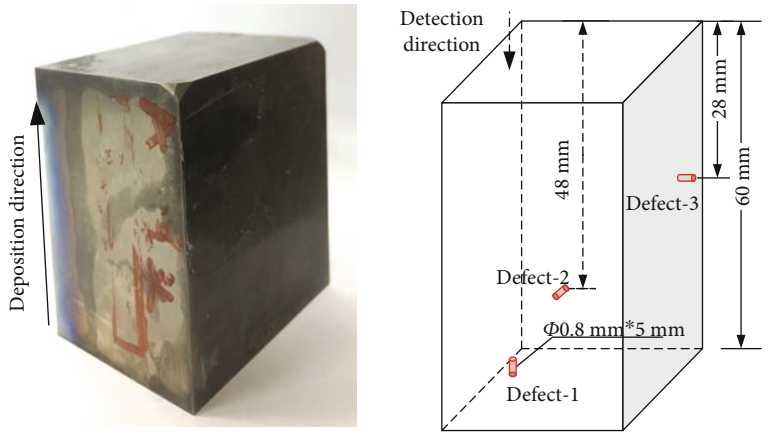


Figure 8: Test specimen of additive manufacturing with prefabricated defects.

acoustic field [13]. The width of the focusing beam is measured by -6 dB method as the focal spot diameter, and the focal spot diameter of the focusing beam can be expressed as follows:

$$\Phi = \Phi_{-6\,\mathrm{dB}} \approx \frac{\lambda F}{D}, \tag{6}$$

where λ is the wavelength, F is the focal length, and D is the diameter of the disc.

Therefore, when the wavelength is constant, the similar focal spot diameter can be obtained in different depth regions by fixing the ratio of focal length to element diameter of annular array transducer. However, this fixed focal spot diameter variable aperture method will lead to changes in the near-field length and focusing energy. Therefore, the variable aperture method can also be established by fixing the focusing intensity and the near-field ratio [14]. According to the near-field formula and the intensity of focusing acoustic field, the approximate characterization methods are as follows:

$$\begin{cases} N_\mathrm{f} = \dfrac{D^2}{4\lambda}, \\ I \approx \left[\dfrac{\pi}{2} \times \left(\dfrac{F}{N_\mathrm{f}}\right)\right]^2, \end{cases} \tag{7}$$

where N_f is the near-field length of the transducer and I is the relative intensity of the focusing acoustic field.

It can be seen that the intensity of focusing acoustic field at different depths is relatively consistent when the ratio of focal length to near-field length is fixed. Therefore, two different variable apertures focusing detection methods can be established, namely, fixed F/D and F/N_f.

3.2. Analysis of Focusing Acoustic Field Characteristics. In the simulation model, the focusing acoustic field is distributed along the axial direction in the depth range of 0-120 mm, and a total of 6 focusing points are set at the interval of 20 mm. The excitation elements of the fixed aperture method are all set as 16 elements. In order to ensure the dynamic

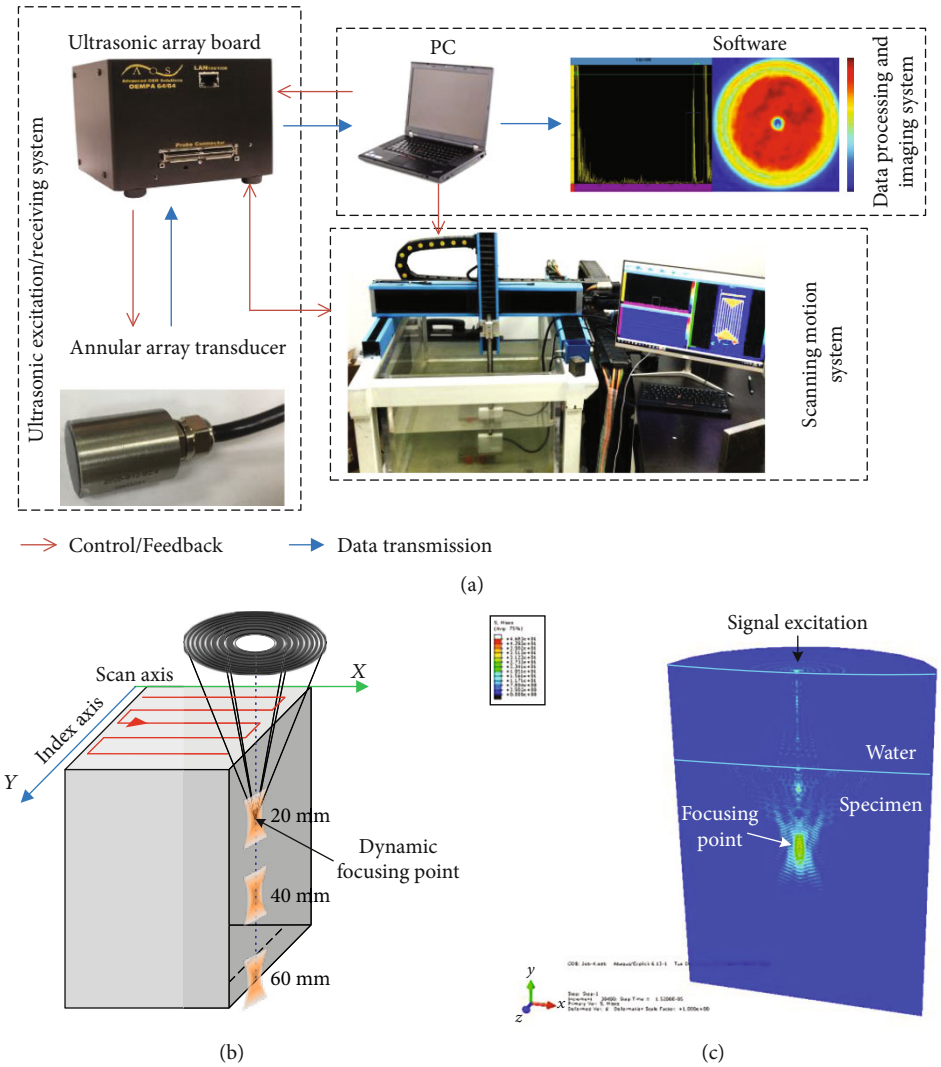


Figure 9: Experimental setup of ultrasonic annular array detection system. (a) Main components of detection system. (b) Schematic of the dynamic focusing method using the annular array. (c) The 3D FEM model of acoustic beam propagation of annular array in double-layer medium.

focusing effect, the minimum excitation elements of the two variable aperture methods are not less than 4. By changing the number of excitation array elements, the scale factor of the two variable aperture methods is kept unchanged as far as possible. The specific simulation parameters are shown in Table 1.

The axial acoustic field distribution of the three focusing methods at different depths is shown in Figure 6. It is shown in Figure 6(a) that the focusing method of annular array with fixed aperture has higher focusing beam energy in the focusing depth of 20-80 mm. However, with the increase of the focusing depth, the axial size of the focal spot increases obviously, and the focusing energy decreases rapidly, with the maximum difference of nearly 40%. In addition, the focal spot length of the near surface is small under this detection method, and when the dynamic focusing points are spaced at a large distance, the nonfocusing region (such as 30 mm in Figure 6(a)) may become a detection blind zone. The simulation results of the fixed F/D variable aperture method are shown in Figure 6(b). Different from the fixed aperture focusing method, the axial size difference of the focal spot at a larger focusing depth (>40 mm) is smaller. However, the energy of focal spot field is lower at a smaller focusing depth, and the difference between the maximum and minimum amplitude of each focusing center is about 47%. Figure 6(c) shows the simulation results of the fixed F/N_f variable aperture method. It can be found that the difference of energy and length of focal spot is the smallest at different focusing points. The maximum difference of central amplitude of all focusing center is about 31%, and the difference of central amplitude is less than 5% in the detection range of 60-120 mm, which can ensure more consistent detection sensitivity for different depth defects of the component. In addition, the acoustic pressure of the variable aperture method for fixed F/N_f is increased obviously in the nonfocusing region with the depth of 30 mm, which is about 3 times of the fixed aperture method and 2 times of the variable aperture method for fixed F/D. Therefore, the variable

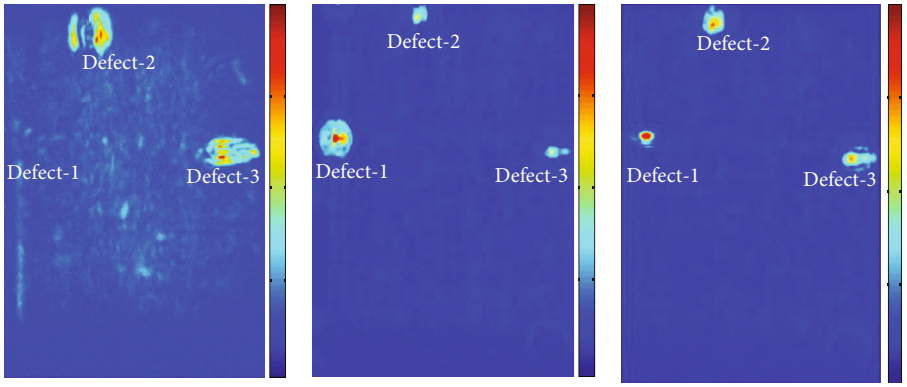


Figure 10: C-scan results of the additive manufacturing specimen of dynamic focusing method. (a) Fixed aperture method of linear array (10 MHz, 64 elements). (b) Fixed aperture method of annular array. (c) Variable aperture method of annular array.

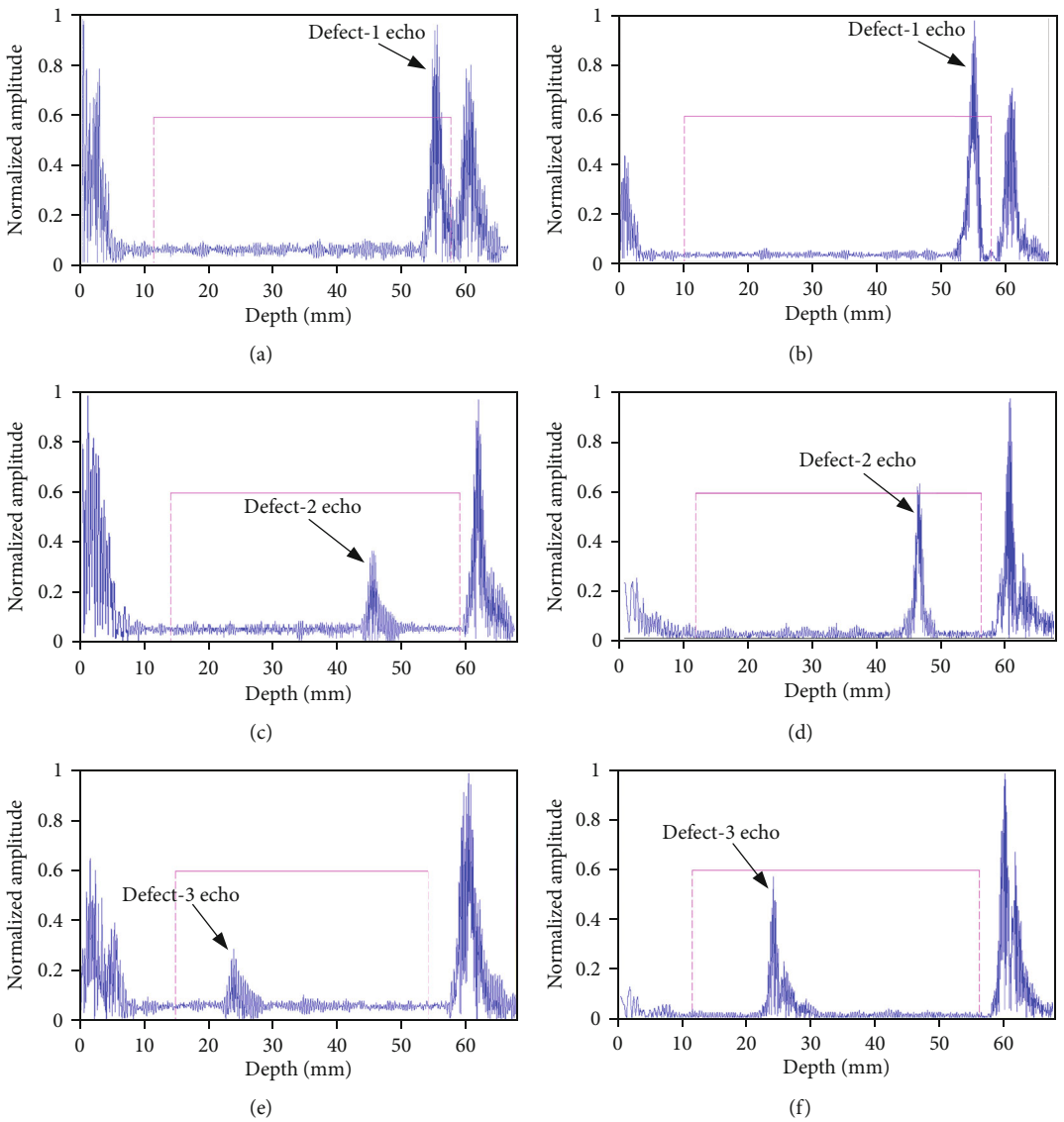


Figure 11: Echo signal of the defects using annular array. (a, c, e) Echo signals of defect-1, defect-2, and defect-3 using the fixed aperture method, respectively. (b, d, f) Echo signals of defect-1, defect-2, and defect-3 using the variable aperture method, respectively.

aperture method for fixed F/N_f can avoid the problem of missing detection caused by the large spacing of the focusing points.

The axial distribution characteristics of acoustic field will affect the detection sensitivity and maximum detection depth, while the radial distribution characteristics of acoustic field will directly affect the quantitative accuracy of C-scan inspection results [15, 16]. The -6 dB method is used to measure the radial size of the focal spot of the fixed aperture method and the two variable aperture methods. The variation of the radial size of the focal spot with the focusing depth in the range of 20-120 mm is shown in Figure 7.

The measurement results show that the radial size of the focal spot changes mostly with the increase of the focusing depth under the fixed aperture focusing method, and the difference between the maximum value and the minimum value in the range of 20-120 mm is about 3.3 times. The fluctuation of focal spot size of the two variable aperture focusing methods is better than that of the fixed aperture focusing method. The change range of the focal spot size of the fixed F/D variable aperture focusing method is the smallest, the difference between the maximum and the minimum is about 1.9 times, and the change range of focal spot size is less than 0.7 mm when the depth is more than 40 mm. The simulation results show that the axial and radial acoustic field distribution characteristics of the variable aperture focusing method are better than those of the fixed aperture focusing method. For different depth focusing points, the fixed F/D variable aperture method has the best performance in the consistency of focal spot size, and the fixed F/N_f variable aperture method is more excellent in the consistency of focusing energy intensity and detection sensitivity.

4. Dynamic Focusing C-Scan Experiment of Annular Array

4.1. Specimen and Experimental Setup. The detection of titanium alloy specimen is prepared by additive manufacturing method, which has high attenuation and anisotropy properties [17]. Due to the particularity of manufacturing process, the acoustic wave will produce distortion and strong attenuation when propagating inside the specimen [18, 19]. Therefore, the focal spot symmetry and focusing acoustic field energy of transducer have great influence on the defect imaging accuracy and SNR, resulting in the poor detection results of additive manufacturing titanium alloy using ultrasonic linear array transducer and single-crystal ultrasonic transducer. The thickness of the specimen along the deposition direction is 60 mm, and there are flat bottom holes with a diameter of 0.8 mm on three adjacent surfaces to verify the design method of annular array transducer and the variable aperture focusing method, as shown in Figure 8.

An ultrasonic array water immersion C-scan automatic detection system was established to evaluate the defects of additive manufacturing titanium alloy specimen [20], as shown in Figure 9(a). The ultrasonic annular array transducer with a center frequency of 10 MHz and 16 elements is used for the detection experiment. The 64/128 channel ultrasonic array board produced by American AOS company is used as the excitation/receiving hardware of acoustic wave detection. As shown in Figure 9(b), the single detection region along the axis direction is discretized into dynamic focusing points with a certain interval. The dynamic focusing points are set at 20 mm, 40 mm, and 60 mm intervals, which are consistent with the parameters in acoustic field simulation analysis. According to the simulation results of axial acoustic field, the defect-2 and defect-3 are located in the nonfocusing region, and the bottom defect-1 is located in the focusing region of 60 mm. Finally, the 3D finite element simulation model of acoustic beam propagation of annular array in double-layer medium is established, as shown in Figure 9(c). The correctness of the focusing delay time of the annular array transducer in the water coupling automatic scanning is verified, and the propagation of the focusing acoustic beam in the specimen is analyzed.

4.2. Detection Results and Discussion. The detection results of the commonly used linear array are shown in Figure 10(a). Due to the high attenuation and anisotropy of the additive manufacturing titanium alloy, the defect-1 at the bottom of the specimen cannot be detected by the 64 elements of linear array transducer. In addition, the noncircular symmetric array element distribution structure of linear array will lead to more serious imaging distortion. The C-scan detection results of the annular array are shown in Figure 10(b) and Figure 10(c) by using the dynamic focusing method of fixed aperture and fixed F/N_f variable aperture, respectively [21]. The two focusing methods of annular array can detect all the three defects in the specimen with different depths. However, although the depth of defect-2 and defect-3 is smaller, the center amplitude of bottom defect-1 with fixed aperture focusing method is much larger than that of near surface defect-2 and defect-3, while the center amplitude of defects in variable aperture focusing method is more consistent. In addition, since the variable aperture method is more consistent in the focal spot size at different depths, the size difference between defects in Figure 10(c) is smaller than that of the C-scan results of fixed aperture.

Figure 11 is the echo signals of defects in Figure 10(b) and Figure 10(c), respectively. The detection results show that the SNR and near-field characteristics of the variable aperture method are better than the fixed aperture method. Among them, the improvement of detection results for defect-2 and defect-3 in the nonfocusing region is more obvious than defect-1 in the focusing region with variable aperture method. Although the excitation aperture of the variable aperture method is less than that of the fixed aperture method at this depth, the defect center amplitude of the variable aperture method is higher and the overall noise level is relatively low. For example, the SNR of defect-3 with variable aperture method is 18.1 dB, while that of fixed aperture method is only 9.6 dB. In addition, it can be seen from the A-type signal that the length and amplitude of the near-field signal using the annular array variable aperture method is far better than that of the fixed aperture method. The near-field signal length of the fixed aperture method is 7.5 mm, and the maximum amplitude is about 65% of the full screen; the near-field length of the variable aperture method is

3.2 mm, and the maximum amplitude is only 14% of the full screen.

5. Summary and Conclusions

The relationship between the annular array element parameters and the 3-D acoustic field distribution is analyzed, and the parameter design criteria of ultrasonic annular array transducer are established. Different from the design criteria of linear array transducer element size, the annular array transducer can obtain better acoustic field distribution characteristics when the element width is close to the detection wavelength, and the stronger main lobe energy can be achieved with 4 array elements. In addition, when the element width is $\lambda \le d \le 2\lambda$, the high-energy side lobe and grating lobe can be avoided at the same time; when the array element number is $N \ge 8$, and the element width is $0.5\,\lambda \le d \le \lambda$, the ultrasonic annular array transducer has better axial acoustic field.

The variable dynamic focusing method of ultrasonic annular array transducer is established; compared with the fixed aperture method, the consistency of the focusing point energy and the focusing spot size is higher, and the detection sensitivity of the nonfocusing region is improved. The detection results of additive manufacturing titanium alloy specimen show that the difference of center amplitude and imaging size of different depth defects is smaller by using the variable aperture method. In addition, the SNR of the nonfocusing region defect reaches $\Phi 0.8\,\text{mm} - 18.1\,\text{dB}$ and has better near-field characteristics. The proposed method has a good application prospect in the detection of large thickness and high attenuation materials in the whole depth range. The distortion of defect imaging caused by the anisotropy of additive manufacturing titanium alloy will be studied in the next step.

Data Availability

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

Conflicts of Interest

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

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) No. 11464030.

References

[1] L. W. Schmerr, *Fundamentals of Ultrasonic Phased Arrays*, vol. 215 of Solid Mechanics and Its Applications, Springer International Publishing, Cham, 2015.

[2] Y. X. Dai, S. G. Yan, and B. X. Zhang, "Ultrosonic beam steering behavior of linear phased arrays in solid," in *2019 14th Symposium on Piezoelectrcity, Acoustic Waves and Device Applications (SPAWDA)*, Shijiazhuang, China, China, 2019.

[3] C. J. L. Lane, A. K. Dunhill, B. W. Drinkwater, and P. D. Wilcox, "The inspection of anisotropic single-crystal components using a 2-D ultrasonic array," *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, vol. 57, no. 12, pp. 2742–2752, 2010.

[4] R. J. Housden, A. H. Gee, G. M. Treece, and R. W. Prager, "Ultrasonic imaging of 3D displacement vectors using a simulated 2D array and beamsteering," *Ultrasonics*, vol. 53, no. 2, pp. 615–621, 2013.

[5] J. G. McKee, R. L. T. Bevan, P. D. Wilcox, and R. E. Malkin, "Volumetric imaging through a doubly-curved surface using a 2D phased array," *NDT & E International*, vol. 113, article 102260, 2020.

[6] Z. Zhou, "Development of ultrasonic phased array immersion C-scan automatic detection system," *Journal of Mechanical Engineering*, vol. 53, no. 12, pp. 28–34, 2017.

[7] J. Woo and Y. Roh, "Design and fabrication of an annular array high intensity focused ultrasound transducer with an optimal electrode pattern," *Sensors and Actuators A: Physical*, vol. 290, pp. 156–161, 2019.

[8] X. Guan, J. Zhang, E. M. Rasselkorde, W. A. Abbasi, and S. Kevin Zhou, "Material damage diagnosis and characterization for turbine rotors using three-dimensional adaptive ultrasonic NDE data reconstruction techniques," *Ultrasonics*, vol. 54, no. 2, pp. 516–525, 2014.

[9] P. D. Wilcox, C. Holmes, and B. W. Drinkwater, "Advanced reflector characterization with ultrasonic phased arrays in NDE applications," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 54, no. 8, 2007.

[10] S. R. Kolkoori, M. U. Rahman, P. K. Chinta, M. Ktreutzbruck, M. Rethmeier, and J. Prager, "Ultrasonic field profile evaluation in acoustically inhomogeneous anisotropic materials using 2D ray tracing model: numerical and experimental comparison," *Ultrasonics*, vol. 53, no. 2, pp. 396–411, 2013.

[11] O. Martinez, L. G. Ullate, and F. Montero, "Analysis of the ultrasonic field radiated by segmented annular arrays," *Journal of Computational Acoustics*, vol. 9, no. 3, pp. 757–772, 2011.

[12] L. Feng and X. Qian, "Enhanced sizing for surface cracks in welded tubular joints using ultrasonic phased array and image processing," *NDT & E International*, vol. 116, article 102334, 2020.

[13] X. Lei, H. Wirdelius, and A. Rosell, "Experimental validation of a phased array probe model in ultrasonic inspection," *Ultrasonics*, vol. 108, article 106217, 2020.

[14] C. Li, D. Pain, P. D. Wilcox, and B. W. Drinkwater, "Imaging composite material using ultrasonic arrays," *NDT & E International*, vol. 53, pp. 8–17, 2013.

[15] Y. Wu, D. Guo, K. Que, B. Chen, and J. Cheng, "Annular phased array dynamic focusing method for large target ultrasonic testing," *Ferroelectrics*, vol. 459, no. 1, pp. 14–23, 2014.

[16] W. T. Li, Z. G. Zhou, and Y. Li, "Inspection of butt welds for complex surface parts using ultrasonic phased array," *Ultrasonics*, vol. 96, pp. 75–82, 2019.

[17] S. E. Zeltmann, N. Gupta, N. G. Tsoutsos, M. Maniatakos, J. Rajendran, and R. Karri, "Manufacturing and security challenges in 3D printing," *JOM*, vol. 68, no. 7, pp. 1872–1881, 2016.

[18] P. Howard, P. Klaassen, N. Kurkcu, and J. Barshinge, "Phased array ultrasonic inspection of titanium forgings," *Review of Progress in Quantitative Nondestructive Evaluation*, vol. 894, no. 1, pp. 854–861, 2007.

[19] X. Wang, W. Li, Y. Li et al., "Phased array ultrasonic testing of micro-flaws in additive manufactured titanium block," *Materials Research Express*, vol. 7, no. 1, article 016572, 2020.

[20] Q. Y. Jiang, X. R. Gao, C. Y. Peng, and J. L. Li, "Application of water immersion ultrasonic phased array technology in wheel rim inspection," *Advanced Materials Research*, vol. 468-471, pp. 733–737, 2012.

[21] J. Oh, "Phase delay quantization error analysis at a focal plane for an ultrasonic annular arrays imaging system," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E90-A, no. 5, pp. 1105-1106, 2007.