[4] Liu K, Liang Y, 2021, Enhancement of Underwater Optical Images Based on Background Light Estimation and Improved Adaptive Transmission Fusion. Optics Express, 29(18): 28307–28328. https://doi.org/10.1364/OE.428626

[5] Song H, Wang R, 2021, Underwater Image Enhancement Based on Multi-Scale Fusion and Global Stretching of Dual-Model. Mathematics, 9(6): 595. https://doi.org/10.3390/math9060595

[6] Zhang W, Wang Y, Li C, 2022, Underwater Image Enhancement by Attenuated Color Channel Correction and Detail Preserved Contrast Enhancement. IEEE Journal of Oceanic Engineering, 47(3): 718–735. https://doi.org/10.1109/JOE.2022.3140563

[7] Yan H, Zhang Z, Xu J, et al., 2023, UW-CycleGAN: Model-Driven CycleGAN for Underwater Image Restoration. IEEE Transactions on Geoscience and Remote Sensing, 61: 1–17. https://doi.org/10.1109/TGRS.2023.3315772

[8] Wang C, Xu H, Jiang G, et al., 2024, Underwater Monocular Depth Estimation Based on Physical-Guided Transformer. IEEE Transactions on Geoscience and Remote Sensing, 62: 1–16. https://doi.org/10.1109/TGRS.2024.3373904

[9] Ren T, Xu H, Jiang G, et al., 2022, Reinforced Swin-Convs Transformer for Simultaneous Underwater Sensing Scene Image Enhancement and Super-Resolution. IEEE Transactions on Geoscience and Remote Sensing, 60: 1–16. https://doi.org/10.1109/TGRS.2022.3205061

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Leveraging Green AI Technology to Build Sustainable Schools: A Conceptual Model Based on AI Agents

Yiting Qiu[1,2*], Yihan Lu[1], Guoqing Xia[3], Md Munir Hayet Khan[2], Deshinta Arrova Dewi[4]

[1]Zhejiang Technical Institute of Economics, Hangzhou 310018, Zhejiang, China
[2]Faculty of Innovation and Technology Program, Engineering and Quantity Surveying, INTI International University, Nilai 71800, Negeri Sembilan, Malaysia
[3]Department of General Practice, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou310016, Zhejiang, China
[4]Faculty of Data Science and Information Technology, INTI International University, Nilai 71800, Negeri Sembilan, Malaysia

Corresponding author: Yiting Qiu, barbaraqiu@dingtalk.com

Abstract: The integration of Green Artificial Intelligence (AI) technologies into educational systems offers a promising avenue to enhance operational efficiency while addressing sustainability challenges. Through a rigorous three-phase methodology combining literature review, AI agent development, and participatory workshop-based case analysis, this paper highlights the pivotal role of AI agents, as applications of Green AI technologies, in driving transformative outcomes within schools. By directly improving self-learning efficiency and reducing learning costs for students, enhancing management and service efficiency, reducing labor costs for schools, as well as minimizing resource dependence for both teachers and students, AI agents create a foundation for sustainable operations. These direct effects generate positive spillover effects, cascading into broader outcomes, including innovation performance, economic efficiency, and environmental sustainability, aligning with the United Nations Sustainable Development Goals (SDGs). By presenting a comprehensive conceptual model, this study demonstrates the pathways through which Green AI contributes to sustainable development in education and emphasizes its critical role in bridging technological innovation with sustainability. This framework provides significant theoretical insights for further empirical research while offering actionable strategies for policymakers and educators to harness Green AI for building sustainable schools with a student-centered approach.

Keywords: Green AI technology; AI agents; Conceptual model; Spillover effects; Sustainable development

Online publication: April 2, 2025

1. Introduction

The pursuit of sustainability in education has become a pressing global issue, particularly aligned with the United

Nations Sustainable Development Goals (SDGs) [1]. Building sustainable schools requires more than policy support and fund allocation; it demands the integration of innovative, technology-driven approaches. Green Artificial Intelligence (Green AI) has emerged as a vital enabler of sustainable practices by optimizing resource usage and reducing environmental impacts [2]. While Green AI has demonstrated significant potential in sectors like healthcare and industrial applications, its applications in educational sustainability remain inadequately explored. This research gap underscores the urgency of a systematic framework to explore its potential contributions.

To address this, this paper presents a conceptual model that examines Green AI's impact on developing sustainable schools and uses AI agents as an example to represent an innovative application of Green AI technologies [3]. These AI agents are equipped with diverse functionalities, including personalized learning support, multimodal intelligent communication, process automation, and knowledge generalization through advanced large models [4]. By improving operational efficiency, reducing costs, and minimizing reliance on traditional resources, AI agents actively drive sustainable education systems and establish student-centered educational practices [5]. This study fully demonstrates how Green AI technologies enable sustainable school development through direct and indirect pathways, thereby aligning with global sustainability goals.

2. Research questions

This paper aims to address the following research questions:

RQ1: What are the direct pathways through which Green AI impacts sustainability metrics such as innovation performance, environmental performance, and economic performance in schools?

RQ2: What are the indirect pathways through which Green AI affects sustainability metrics in schools?

RQ3: How can these direct and indirect impact pathways be integrated into a conceptual model to support the construction of sustainable schools?

By exploring these research questions, the study establishes a conceptual framework that serves as a theoretical and practical guide for leveraging Green AI in education. It highlights the synergy between technology-driven innovation and student-centered educational philosophies, offering a comprehensive perspective on how schools can achieve sustainable development goals.

3. Methodology

To explore the role of Green AI technology in building sustainable campuses, this study adopts a methodological approach combining "literature review", "AI agent development", and "case analysis", supported by successful applications of AI technology in campus scenarios. Using a participatory workshop model, the study aims to construct a scientifically rigorous and practically valuable framework by integrating theoretical and practical perspectives.

4. Literature review

The literature review focuses on three core themes: (1) the concept, advantages, and characteristics of Green AI technology and AI agents; (2) the integration and application of technological innovation and education within the framework of sustainable development; and (3) the potential of artificial intelligence technologies in educational paradigms, resource allocation, and management [1, 3, 6]. This systematic review synthesizes the latest research

and practices, emphasizing the value of Green AI in enhancing resource efficiency, optimizing campus resource management, and improving environmental performance [3]. It also explores how technology-driven, student-centered educational philosophies can catalyze the sustainability process on campus [2]. The findings provide a solid theoretical foundation for subsequent AI agent development, case analysis, and the design of participatory workshops by identifying both technological gaps and opportunities for innovation [7].

5. AI agent development

To translate theoretical assumptions into practical applications, this study developed a series of AI agents tailored to diverse campus scenarios. These agents were designed to address specific challenges such as resource management, academic advising, digital platform navigation, and moral education. The AI-enabled tools drew on principles of Green AI technology to optimize energy efficiency and minimize computational waste during operations. This phase also emphasized the usability of AI tools to ensure both students and teachers could easily integrate them into their daily academic and administrative activities [8]. Over 20 AI agents were deployed, collectively serving more than 2,300 users, providing both valuable data and an experiential foundation for further analysis.

6. Case analysis: participatory workshop-based approach

Building on the results of the literature review and AI agent development, this study employed participatory case analysis to validate theoretical assumptions against practical results. Through workshops that engaged users directly, the value of AI agents in campus sustainability was thoroughly examined. Discussions gathered perspectives on the synergies among green technology, performance, and sustainable development. Students focused on how AI agents could optimize learning experiences, improve learning efficiency, and reduce learning costs. Teachers, on the other hand, explored the potential of AI tools to foster educational innovation and resource optimization from a managerial perspective while promoting waste reduction and fostering green campus environments. This multi-layered participatory approach constructed a comprehensive analytical framework linking practical application, technological development, and user feedback. Case data and workshop findings were synthesized to evaluate the feasibility and future potential of Green AI in real campus scenarios, offering both theoretical insights and practical guidelines for fostering sustainable, technology-driven, and student-centered campuses. This three-phase methodology—spanning literature review, AI agent development, and participatory case analysis—outlines the pathways by which Green AI technology can contribute to educational systems and sustainable development frameworks. It also lays the groundwork for further exploration of green educational technologies.

7. Conceptual model design and discussion

7.1. Green AI technique, efficiency, and innovation performance

H1: The application of AI agents directly improves learning efficiency of students and promotes innovation performance, leading to sustainable development.

AI agents, as a core application of Green AI technologies, significantly enhance students' learning efficiency through various practical functions and mechanisms. Their value is reflected in three key aspects. First, AI agents

provide personalized and adaptive learning experiences through human-AI interaction, acting as round-the-clock virtual learning mentors [9]. For instance, they offer step-by-step explanations of complex problems and generate tailored exercises to reinforce understanding and encourage critical thinking. This student-centered approach not only improves comprehension and problem-solving skills but also compensates for the time and resource limitations inherent in traditional education systems [3].

Second, AI agents leverage internal knowledge bases and the generalization capabilities of large language models to deliver precise academic guidance and foster diverse intellectual exploration [4]. They can quickly address specific student queries with accurate responses while simultaneously facilitating brainstorming and divergent thinking. This guided learning approach cultivates creativity and critical reasoning, enabling students to thrive in dynamic, multidimensional learning environments [4].

Lastly, the accessibility of AI agents significantly reduces barriers to quality education in under-resourced regions. With minimal requirements such as electricity and internet access, AI agents provide instant access to knowledge and learning methods. This accessibility is particularly impactful in addressing disparities in education, as it bridges the digital divide and empowers students in underserved areas to realize their learning potential [10]. By democratizing access to resources, AI agents also contribute to greater educational equity.

Thus, through personalized academic support, intellectual stimulation, and resource accessibility, AI agents directly enhance students' learning efficiency across multiple dimensions, leveraging the strengths of Green AI to address diverse educational challenges.

H2: The application of AI agents directly improves management and service efficiency and promotes innovation performance, leading to sustainable development.

The development and application of AI agents help consolidate the teaching experience of educators and the operational expertise of administrative staff. This significantly enhances schools' management and service efficiency. First, AI agents can serve as digital extensions of teachers, providing targeted academic support on demand. They ensure students have access to virtual mentors 24/7 to resolve their queries anytime, anywhere [4]. Second, AI agents integrate knowledge bases to address questions related to academic standards and procedural requirements. They reduce repetitive and mundane administrative tasks, giving teachers and staff more time and energy to focus on innovative projects.

Third, AI agents are powered by continuously upgraded large language models [4]. These models support rich multimodal interactions and improve the precision and breadth of responses. This makes up for knowledge gaps that human teachers or administrators may have. Furthermore, by automating repetitive workflows and enabling scalable solutions, AI agents foster sustainable innovation within schools. They promote smarter resource allocation and long-term operational efficiency [10]. In summary, the effective use of AI agents greatly enhances school management and services, making them more efficient, intelligent, and sustainable.

7.2. Green AI technique, resource dependence and environmental performance

H3: The application of AI agents directly reduces management and service in school and promotes environment performance, leading to sustainable development

The application of AI agents significantly reduces resource reliance and enhances environmental performance by reshaping workflows, centralizing information, and streamlining education and management support [11]. First, AI agents promote paperless and digital-first operations, ensuring that tasks like academic grading, assignments,

exercises, creating mind maps, team collaboration, and communication are entirely digitized. This comprehensive transformation drastically reduces the use of physical resources such as paper, printing supplies, and office consumables, thereby cutting down on waste and aligning institutions with sustainability goals [11].

Second, they enhance the efficiency of information usage by centralizing data storage and providing personalized, on-demand access to academic and organizational resources. By eliminating the need for redundant physical or digital copies, AI agents prevent unnecessary duplication while meeting diverse user needs with precision. Third, AI agents elevate operational efficiency through scalable, reusable virtual services that replace traditional methods of manual academic advising or repetitive administrative processes. Such AI-powered solutions deploy seamlessly across different contexts without increasing the strain on physical infrastructure or escalating material consumption [12]. This approach not only improves service delivery but also fosters sustainable systems.

In summary, the integration of AI agents establishes a robust framework for reducing resource demands, optimizing environmental performance, and supporting the development of eco-friendly, sustainable institutions.

7.3. Green AI technique, costs and economic performance

H4: The application of AI Agents directly reduces learning costs of students and promotes economic performance, leading to sustainable development

AI agents effectively reduce learning costs while promoting sustainable development through technological innovation and resource efficiency [5]. Firstly, AI teaching assistants and similar tools largely replace the need for private tutors. For example, AI systems like Doubao, which support multilingual learning through voice interaction, enable self-study, reducing both tutoring and transportation costs. Secondly, AI agents provide multimodal digital learning resources, such as interactive courseware, exercises, and assessments, replacing expensive physical textbooks and workbooks [13]. These tools allow students to learn and practice in a more flexible and adaptive way, eliminating unnecessary spending while streamlining study processes. Additionally, AI agents utilize cloud service platforms to store and manage digital resources, significantly reducing knowledge management and storage costs [14]. By advancing digital-first, cost-efficient, and eco-friendly educational practices, AI agents not only alleviate financial burdens but also contribute to sustainable development by minimizing the environmental impact associated with traditional learning resources.

H5: The application of AI Agents directly reduces labor costs of school and promotes economic performance, leading to sustainable development

AI agents, through intelligent and automated technologies, help schools significantly reduce labor costs while enabling a lightweight and low-cost operation model that aligns with the principles of a sustainable green campus. Firstly, AI agents can replace routine and repetitive tasks, thereby reducing the need for low-skill, low-creativity labor and significantly cutting associated personnel budgets [15]. This allows schools to allocate more resources to higher-value areas, such as innovative teaching methods and student development. Secondly, AI agent-driven automation optimizes task allocation and improves execution efficiency, shortening the time and resources required for tasks and further reducing labor demands [15]. Additionally, digital teams composed of multiple AI agents can autonomously collaborate to complete complex tasks, achieving seamless and highly efficient workflows. This intelligent collaboration not only enhances productivity but also substantially reduces staffing costs. By promoting a lightweight operational structure and minimizing unnecessary resource consumption, AI agents contribute to the

vision of a truly sustainable green campus, balancing cost-efficiency with environmental responsibility.

7.4. Green AI technique, student-oriented education and sustainable development

H6: The application of AI agents promotes sustainable development of education through student-oriented transformation

AI agents have emerged as timely and essential technological enablers of student-oriented transformation, playing a critical role in advancing the sustainable development of education. First, AI agents leverage big data and natural language processing to deliver personalized, adaptive, and tailored learning experiences, breaking away from traditional, one-size-fits-all teaching methods [3, 4, 15]. By dynamically addressing individual student needs, these technologies optimize learning outcomes and foster a more student-centered approach to education. Second, AI agents promote accessibility and inclusivity by offering tools such as real-time language translation, adaptive interfaces, and assistive technologies [13, 15]. These innovations ensure that students with diverse abilities, needs, and languages can equitably access learning resources, creating a fair and supportive educational system. Lastly, AI agents enhance future readiness by improving AI literacy and equipping students with essential skills such as prompt engineering, responsible AI usage, and basic AI development [15–18].

These competencies enable students to succeed in an AI-driven world, positioning them for success in emerging industries and job markets. By fostering innovation, inclusivity, and future readiness, AI agents enable a student-oriented transformation that advances the sustainable development of education in an ever-evolving global context.

7.5. Green AI technique, technological innovation and sustainable development

H7: The application of AI agents promotes sustainable development of education through technological innovation

AI agents play a pivotal role in driving the sustainable development of education by advancing technological innovation in three key areas. First, AI agents revolutionize the development of educational tools by automating repetitive tasks, analyzing diverse learner data, and modeling pedagogical strategies[1, 5, 19]. This accelerates the creation of adaptive learning platforms, intelligent tutoring systems, and personalized feedback solutions, significantly reducing development cycles and enabling swift adaptation to diverse educational needs [15]. Second, AI agents enhance creativity in teaching and learning by facilitating the design of interactive, immersive, and gamified learning experiences. Their ability to dynamically interact with both educators and learners allows for real-time adjustments in content delivery and problem-solving, fostering engagement and innovation within the educational ecosystem [2].

Furthermore, AI agents actively support students in hands-on exploration of fields such as robotics, machine learning, and digital technology, helping them develop practical skills and innovative thinking to address real-world challenges [3, 15]. Lastly, AI agents provide adaptive and accessible technologies, ensuring diverse learners, including those with special educational needs, can engage with advanced tools to develop technological literacy and problem-solving capabilities for an AI-driven future [5, 12]. Together, these contributions position AI agents as transformative enablers of technological advancement, ensuring the education sector evolves sustainably to meet the demands of an ever-changing world.

Therefore, based on the research questions and hypotheses, this study proposes a conceptual framework (**Figure 1**) that encompasses both direct and indirect pathways to reveal how Green AI technology contributes to sustainable development in education. First, Green AI technology affects learning efficiency, management

and service efficiency, resource dependence, learning costs, and labor costs through direct pathways (H1–H5), leading to improvements in innovation performance, environmental performance, and economic performance. Additionally, student-oriented educational transformation and technological innovation, driven by Green AI technology (H6, H7), serve as indirect pathways to support and empower the transformation of educational systems. Together, these pathways form an integrated model that links Green AI technology, technological innovation, educational reform, and sustainability goals, providing theoretical foundations and practical guidance for building sustainable schools.

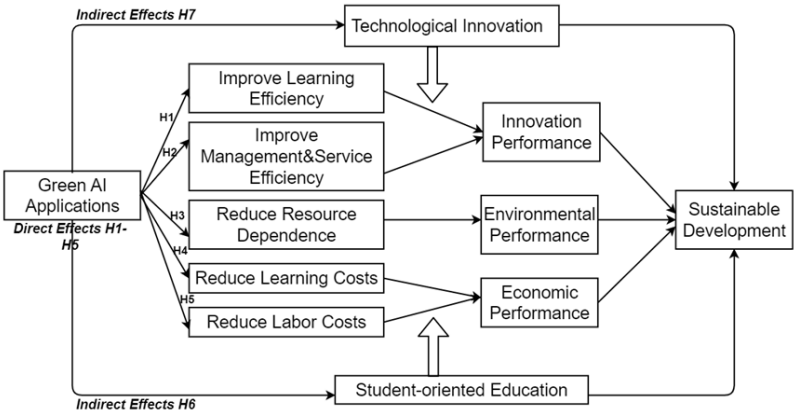


Figure 1. Conceptual model of green AI technique for promoting sustainable development in education

7.6. Conclusion and Future Work

This study presents a comprehensive conceptual framework that elucidates the transformative role of Green AI technology, particularly through AI agents, in fostering sustainable development within educational institutions. Through a rigorous three-phase methodology combining literature review, AI agent development, and participatory workshop-based case analysis, our research demonstrates that AI agents contribute to sustainability metrics via both direct and indirect pathways. The direct pathways encompass enhanced learning and management efficiency, reduced resource dependence, and optimized cost structures, while indirect pathways operate through student-oriented transformation and technological innovation [12, 20]. Notably, the positive spillover effects extend beyond the educational domain, creating cascading benefits for innovation performance, environmental sustainability, and economic efficiency across the broader educational ecosystem [5, 21]. Future research endeavors should prioritize: empirical validation of the proposed framework through large-scale quantitative studies; investigation of implementation barriers across diverse educational contexts; development of standardized sustainability impact metrics; examination of long-term effects on student outcomes; and assessment of scalability across different educational levels. This research not only bridges the theoretical gap in understanding Green AI's role in educational sustainability but also provides practical insights for educational institutions pursuing sustainable development goals.

Funding

2024 Academic Research of Zhejiang Technical Institute of Economics: "Spillover Effects of Multimodal AI Agents on Green School Development" (Project No.: X2024038); 2024-2025 Research and Creative Project, Department of Culture and Tourism: "The Application of Digital Information Technology in Safety Early

Warning and Supervision of Cultural Relics in Zhejiang, China" (Project No.: 2024KYY045); 2024 General Research Project of Zhejiang Provincial Department of Education: "Empirical Research on Low-Carbon Economy Driving the Development of New Quality Productivity: A Case Study of Zhejiang Province" (Project No.: Y202456145)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Ziemba EW, Doung CD, Ejdys J et al., Leveragind Artificial Intelligence to Meet the Sustainable Development Goals. Journal of Economics and Management, 46(1): 508–583

[2] Alzoubi Y, Mishra A, 2024, Green Artificial Intelligence Initiatives: Potentials and Challenges. Journal of Cleaner Production, 468: 143090.

[3] Osondu J, 2025, Red AI vs. Green AI in Education: How Educational Institutions and Students Can Lead Environmentally Sustainable Artificial Intelligence Practices, thesis, Ohio University.

[4] Xi Z, Chen W, Guo X, et al., 2025, The Rise and Potential of Large Language Model-Based Agents: A Survey. Science China Information Sciences, 68(2): 1–44.

[5] Luo S, Liu J et al., 2024, Enterprise Service-Oriented Transformation and Sustainable Development Driven by Digital Technology. Scientific Reports, 2024(14): 1–18.

[6] Mikalef P, Lemmer K, Schaefer C, et al., 2023, Examining How AI Capabilities Can Foster Organizational Performance in Public Organizations. Government Information Quarterly, 40(2): 101797.

[7] Celik I, 2023, Exploring the Determinants of Artificial Intelligence (AI) Literacy: Digital Divide, Computational Thinking, Cognitive Absorption. Telematics and Informatics, 83: 102026.

[8] Chiu TKF, 2024, Future Research Recommendations for Transforming Higher Education With Generative AI. Computers and Education: Artificial Intelligence, 6: 100197.

[9] Dennis AR, Lakhiwal A, Sachdeva A, 2023, AI Agents as Team Members: Effects on Satisfaction, Conflict, Trustworthiness, and Willingness to Work With. Journal of Management Information Systems, 40(2): 307–337.

[10] Li Y, Wen H, Li X, et al., 2024, Personal LLM Agents: Insights and Survey About the Capability, Efficiency and Security. arXiv:2401.05459: 1– 62.

[11] Jhurani J, 2025, Revolutionizing Enterprise Resource Planning: The Impact of Artificial Intelligence on Efficiency and Decision-Making for Corporate Strategies. International Journal of Computer Engineering and Technology, 13(2): 156–165.

[12] Xu Z, Pan R, 2024, Effects of Intelligent Manufacturing on the High-Quality Development of Manufacturing Industry: The Mediating Role of Green Technology Innovation. Scientific Reports, 14: 1–14.

[13] Qiu Y, Chen Q, Ng PSJ, 2023, Research on the Spillover Effects of Digital Transformation on the Sustainable Growth of Green Schools. PBES, 6(6): 16–23.

[14] Qiu Y, Yuan CS, Yie LW, 2024, Creating a "Ready-to-Use" AI Agent for Navigating Digital Platform to Enhance Collaborative Efficiency. INTI Journal, 2024: 1–11.

[15] Qiu Y, Khan MH, Shuqing Z, SiYuan C, Choonkit C, 2024, Enhancing Sustainability in Academic Guidance: Develop an AI-Driven Agent for Education 5.0. INTI Journal, 2024: 1–10.

[16] Korte SM, Cheung WM, Maasilta M, et al., 2024, Enhancing Artificial Intelligence Literacy Through Cross-Cultural Online Workshops. Computers and Education Open, 6(4): 100164.

[17] Stolpe K, Hallström J, 2024, Artificial Intelligence Literacy for Technology Education. Computers and Education Open, 6: 100159.

[18] Qiu Y, Li L, Wang X, 2024, Quantitative Analysis Digital Literacy of Secondary and Higher Vocational School Students in the Digital Economy Background – Preliminary Empirical Research Based on 181 Samples From Zhejiang, China. INTI Journal, 2024: 1–7.

[19] Usman M, Khan R, Moinuddin M, 2024, Assessing the Impact of Artificial Intelligence Adoption on Organizational Performance in the Manufacturing Sector. Revista Española de Documentación Científica, 18(2): 95–116.

[20] Chen D, Wang, S, 2024, Digital Transformation, Innovation Capabilities, and Servitization as Drivers of ESG Performance in Manufacturing SMEs. Scientific Reports, 14(1): 24516.

[21] Arshad M, Qadir A, Rafique M, 2024, Enhancing Organizational Sustainable Innovation Performance Through Organizational Readiness for Big Data Analytics. Humanities and Social Sciences Communications, 11(1): 1–15.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

A Brief Discussion on Data Encryption and Decryption Technology and Its Applications

Zhihong Jin*

The 22nd Research Institute of China Electronics Technology Group Corporation,Xinxiang 453000, Henan Province, China

***Corresponding author:** Zhihong Jin, 81984088@163.com

Abstract: With the rapid development of information technology, data security issues have received increasing attention. Data encryption and decryption technology, as a key means of ensuring data security, plays an important role in multiple fields such as communication security, data storage, and data recovery. This article explores the fundamental principles and interrelationships of data encryption and decryption, examines the strengths, weaknesses, and applicability of symmetric, asymmetric, and hybrid encryption algorithms, and introduces key application scenarios for data encryption and decryption technology. It examines the challenges and corresponding countermeasures related to encryption algorithm security, key management, and encryption-decryption performance. Finally, it analyzes the development trends and future prospects of data encryption and decryption technology. This article provides a systematic understanding of data encryption and decryption techniques, which has good reference value for software designers.

Keywords: Data encryption; Data decryption; Communication security; Data storage encryption; Key management

Online publication: March 28, 2025

1. Introduction

In today's digital age, strengthening the security management of computer data information and enhancing data encryption technology have profound significance for ensuring the security of data information in the entire society [1]. Data encryption and decryption technology has received widespread attention and research worldwide. In developed countries such as the United States, data encryption technology started early and has become relatively mature, widely used in multiple fields. On the basis of researching existing database encryption technologies, some scholars have designed and implemented an encryption system for specific information systems using the Java platform. These studies have made important contributions to the development of data encryption technology. This article explores the principles, applications, and challenges of data encryption and decryption technology in depth and proposes effective response strategies. It is expected to make beneficial contributions to the development of data security, with theoretical guidance and international application value.

2. Overview of data encryption and decryption technologies

2.1. Principles of data encryption technology

The basic principle of data encryption technology is to convert raw plaintext data into difficult to understand ciphertext data through specific algorithms and keys. The purpose of data encryption is to prevent unauthorized data access and data leakage. Data encryption technology can be divided into two categories based on whether the encryption and decryption keys used are the same: symmetric encryption algorithms and asymmetric encryption algorithms. The encryption process involves complex mathematical operations and key extension to ensure the confidentiality and integrity of data [2]. Chaos encryption, as an emerging encryption method, utilizes the characteristics of chaotic systems to generate complex chaotic sequences for encrypting and scrambling data. This method has good security performance and brings new ideas and methods to the field of data encryption.

2.2. Principles of data decryption technology

Data decryption technology is the inverse operation of data encryption, which is the process of restoring ciphertext data to original plaintext data through specific algorithms and the same or associated keys. The encryption process is a process of restoring data through complex mathematical calculations and the core of decryption technology lies in the management of keys and the selection of algorithms. In practical applications, data decryption technology is often combined with various security protocols and mechanisms to ensure the integrity and authenticity of data. In digital signature technology, decryption techniques are used to verify the validity of signatures, thereby confirming the source and unaltered state of data.

2.3. The relationship between data encryption and decryption

Data encryption and decryption are two core components of data security protection. Without the decryption process, encrypted data will become unusable and lose its original value. Similarly, without encryption, the security of data cannot be guaranteed, and there is a risk of theft or tampering. Therefore, data encryption and decryption are two interdependent and mutually reinforcing links, playing an indispensable role in the field of data security [3].

In practical applications, data encryption and decryption techniques need to be selected according to specific scenarios. With the continuous development of technology, data encryption and decryption methods are also constantly innovating and optimizing. Currently, the data encryption and decryption methods of collaborative cognitive models combine the basic principles of data encryption and decryption with those of collaborative cognitive models, providing a simple, secure, and reliable method for data encryption and decryption.

3. Classification of data encryption and decryption technologies

3.1. Symmetric encryption algorithm

Symmetric encryption algorithm uses the same key in the encryption and decryption process, with fast encryption and decryption calculation speed, which can meet the requirements of large-scale data processing or high real-time applications.

The Advanced Encryption Standard (AES) algorithm is the most typical symmetric encryption algorithm. The AES algorithm offers multiple key length options, including 128-bit, 192-bit, and 256-bit, to meet the needs of different security levels. As the key length increases, the security of the algorithm also improves accordingly, enabling AES to provide strong protection in various security-sensitive applications. The AES algorithm is

lightweight and highly flexible, making it easy to deploy and implement on different platforms and environments.

The most prominent problem faced by symmetric encryption algorithms is key management. Due to the use of the same key for encryption and decryption, once the key is leaked, the security of the entire encryption system will be seriously threatened. Therefore, in practical applications, how to securely and effectively manage and distribute keys has become a major challenge in the implementation of symmetric encryption algorithms. At the same time, secure key exchange protocols and regular key replacement can be used to increase the security of system data exchange.

In addition to AES, the Data Encryption Standard (DES) algorithm, as an earlier symmetric encryption algorithm, can still have certain application value in scenarios that require high performance but less strict security requirements. In addition, there are some symmetric encryption algorithms for specific application scenarios, such as stream encryption algorithms, block encryption algorithms, etc., which play important roles in their respective application fields.

3.2. Asymmetric encryption algorithm

Asymmetric encryption algorithms use a pair of public and private key mechanisms, with the public key used for data encryption and the private key used for data decryption, bringing revolutionary changes to the field of data security. This algorithm not only provides high-strength security but also solves the problem of key distribution in symmetric encryption algorithms. In asymmetric encryption algorithms, the public key is public and anyone can use it for data encryption, but decryption requires the use of the corresponding private key, which greatly enhances the security of the data.

The RSA algorithm, as a representative of asymmetric encryption algorithms, has attracted much attention since its inception. It is based on the mathematical problem of large number decomposition, and generates a public key and a private key by multiplying two large prime numbers, ensuring its security. In the RSA algorithm, the encryption process uses a public key, while the decryption process uses a private key. By using this method, even if the data is illegally intercepted during transmission, the interceptor cannot easily decrypt it, thus protecting the security of the data.

In addition to RSA algorithm, elliptic curve cryptography (ECC) algorithm is also a highly regarded asymmetric encryption algorithm in recent years. Compared to RSA, ECC requires shorter key lengths while providing the same level of security, which means faster and more efficient encryption and decryption. The ECC algorithm is based on the mathematical theory of elliptic curves and achieves secure encryption and decryption of data by selecting points on the elliptic curve as public and private keys.

The advantages of asymmetric encryption algorithms are high-strength security and flexible key management mechanisms, but the disadvantages are relatively slow encryption and decryption computation speed and complex private key storage. Therefore, in practical applications, asymmetric encryption algorithms are often combined with other encryption techniques to fully leverage their respective advantages and ensure data security.

3.3. Hybrid encryption algorithm

Hybrid encryption algorithm is a technology that combines the advantages of symmetric and asymmetric encryption algorithms, aiming to find the best balance between encryption efficiency and security. In this algorithm, asymmetric encryption algorithm is not directly used to encrypt large amounts of data but is used to encrypt the key of symmetric encryption algorithm.

The workflow of hybrid encryption algorithm is as follows: Firstly, the sender and receiver will each generate a pair of public and private keys for asymmetric encryption algorithm. The sender will use the receiver's public key to encrypt the key of the symmetric encryption algorithm, so even if the data is intercepted during transmission, the key of the symmetric encryption algorithm cannot be directly obtained. Then, the sender uses the key of this symmetric encryption algorithm to encrypt the actual data to be transmitted. Finally, after receiving the encrypted data and key, the recipient will use their own private key to decrypt the key and then use the borrowed key to decrypt the actual data to be transmitted.

The hybrid encryption algorithm combines the efficiency of symmetric encryption algorithms with the security of asymmetric encryption algorithms. Moreover, since the key of symmetric encryption algorithm is randomly generated and a new key is changed every time encryption is performed, even if a certain encryption is cracked, it will not affect the security of other data.

Common hybrid encryption algorithms include encryption algorithms in TLS/SSL protocols. These algorithms have been widely applied in network communication, providing strong guarantees for the secure transmission of data. In scenarios such as web browsing, email, and online shopping, hybrid encryption algorithms silently protect our data security.

4. Typical applications of data encryption and decryption technology

4.1. Application of data encryption in communication security

In terms of network communication, in addition to the application of HTTPS protocol and VPN technology, data encryption is also widely used in tools such as email and instant messaging to ensure that email content and chat records are not stolen or tampered with by third parties, thereby protecting users' privacy and information security. Data encryption is also commonly used to protect the security of network devices and servers such as firewalls, preventing hacker intrusion and data leakage.

In terms of mobile communication, data encryption is applied to sensitive information of mobile phone users such as bank transfers, shopping payments, etc., to prevent information from being easily stolen. In satellite communication, the use of data encryption technology can ensure that data in satellite communication is not stolen or tampered with, thereby ensuring the security and reliability of communication.

In short, the application of data encryption in communication security is comprehensive, involving various communication methods and scenarios. By using data encryption technology, users' privacy and information security can be effectively protected, preventing data leakage and theft.

4.2. Application of data encryption in data storage

In terms of database encryption, using symmetric encryption algorithms such as AES can encrypt sensitive data in the database, ensuring that even if the database files are illegally obtained, attackers will have difficulty directly reading the plaintext data.

Data encryption in cloud storage has also been a research hotspot in recent years. Through encryption technology, users can encrypt their data before uploading it and then store the encrypted data in the cloud. This can effectively protect cloud data from being accessed by illegal malicious users.

With the continuous development of technology, some emerging data encryption technologies are gradually being applied in the field of data storage. For example, attribute based encryption algorithms (ABE) can control

data access permissions based on the attributes of the data, thereby achieving finer grained data protection. Homomorphic encryption algorithms allow for computation and processing of data without decryption, providing a new solution for data privacy protection in scenarios such as cloud computing.

4.3. Application of data decryption in data recovery

In the process of data backup, encryption technology is usually used to process the backup data in order to ensure its security. After data backup, even if the backup data is illegally obtained, the data content cannot be directly read. At the same time, when data is lost or damaged due to hardware failures, software errors, human error, and other reasons, we need to rely on data decryption technology to recover the loss as much as possible.

The application of data decryption in data recovery is not limited to traditional local data recovery scenarios. In the era of cloud computing and big data, more and more data is being stored in the cloud or on remote servers. In this case, data backup and recovery often require crossing different physical locations and network environments. Therefore, data decryption technology also needs to adapt to this distributed and heterogeneous environment to ensure the security, integrity, and availability of data.

5. Challenges and countermeasures faced by data encryption and decryption technologies

5.1. Security issues of encryption algorithms

In recent years, with the rapid improvement of computing power and the deepening of cryptographic research, some classic encryption algorithms such as DES have been proven to have security risks and have even been successfully cracked, posing huge challenges to data encryption and decryption technology.

To address this challenge, cryptographers and computer scientists are constantly innovating and developing more secure and efficient encryption algorithms. For example, the AES algorithm, as an alternative to DES, provides higher security and encryption strength. Meanwhile, asymmetric encryption algorithms such as RSA and ECC also provide higher security for data encryption due to their unique public-private key mechanisms.

Relying solely on updating encryption algorithms is not enough to completely solve security issues. We need to establish a continuous encryption evaluation mechanism to conduct regular security assessments and vulnerability scans of encryption algorithms, promptly identify and address potential security risks, and ensure the security of encryption algorithms.

5.2. Key management issues

Key management is crucial in data encryption and decryption. Once the key is leaked or illegally obtained, the security of the entire encryption system will be seriously threatened.

The generation of keys must follow strict security standards, which typically involve using strong random number generators to generate sufficiently complex and unpredictable keys. The storage of keys also requires careful handling. Keys need to be stored in a secure device environment to prevent unauthorized access and ensure their physical security. The distribution of keys uses secure communication protocols such as SSL/TLS to encrypt the transmission keys. The key needs to be updated regularly, and leaked keys should be promptly revoked and replaced with new keys.

In addition to technical measures, strengthen the training and management of key management personnel, fully understand the importance of key management and related operating procedures, to ensure that management

personnel will not inadvertently leak keys or engage in improper operations.

5.3. Encryption and decryption performance issues

The performance issues of encryption and decryption are not only related to the efficiency of data processing and transmission, especially when dealing with large-scale data, which can seriously affect user experience and may also have a negative impact on the operational efficiency of enterprises. The following measures can be taken to optimize the performance of encryption and decryption calculations. Firstly, the efficiency of encryption and decryption operations can be improved by utilizing specialized hardware accelerators such as encryption cards or TPUs.

Secondly, optimizing encryption algorithms is also key to improving performance. On the one hand, algorithm-level optimization can be used to adopt more efficient mathematical operation methods, optimize lookup tables, etc., reducing the computational load during encryption and decryption processes. On the other hand, suitable encryption algorithms can be selected based on the application scenario to reduce unnecessary computation.

Finally, adopting a hierarchical encryption strategy to reduce computational overhead. High strength encryption is used for important data, while lower strength encryption is used for non-sensitive data. This can ensure data security while reducing the overall computational cost of encryption and decryption.

6. Development trends and prospects

Data encryption and decryption technology, as key technologies for information security, are systematically elaborated in this article. With the continuous advancement of information technology, data encryption and decryption techniques are also evolving, showing some new development trends:

(1) With the rise of quantum computing and the application of computational complexity theory, traditional encryption algorithms are facing unprecedented challenges and encryption algorithms will develop towards higher security and greater complexity.

(2) With the development of artificial intelligence and blockchain technology, key management will achieve higher levels of automation and intelligence.

(3) With the popularization of technologies such as cloud computing, big data, and the Internet of Things, the processing speed and data throughput of data encryption and decryption have become key performance indicators.

(4) Data encryption and decryption technology is expected to deeply integrate with cutting-edge technologies such as artificial intelligence, blockchain, and cloud computing.

(5) Governments and industry organizations around the world are increasingly concerned about data security issues, and data encryption and decryption technologies will inevitably move towards standardization and legalization.

Through continuous research and innovation, we look forward to developing more efficient, secure, and intelligent data encryption and decryption technology solutions, providing a more solid guarantee for information security.

Disclosure statement

The author declares no conflict of interest.

References

[1] Ping H, 2022, Network Information Security Data Protection Based on Data Encryption Technology. Wireless Personal Communications, 126(3): 2719–2729.

[2] Liu GX, 2022, The Application of Data Encryption Technology in Computer Network Communication Security. Mobile Information Systems, 2022(5): 1–10.

[3] Logofatu PC, Udrea C, Garoi F, 2024, Physical Encryption-Compression and Decryption-Decompression of Data Using the Fourier Transform. Romanian Reports in Physics, 76(1): 1–12.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Algorithm and Application in Vehicle Routing Problem: A Review

Zhenyu Chen*

Business School, Shandong University of Technology, Zibo 255000, China

Corresponding author: Zhenyu Chen, czy0807111@163.com

Abstract: This paper systematically reviews the latest research developments in Vehicle Routing Problems (VRP). It examines classical VRP models and their classifications across different dimensions, including load capacity, operational characteristics, optimization objectives, vehicle types, and time constraints. Based on literature retrieval results from the Web of Science database, the paper analyzes the current state and trends in VRP research, providing detailed explanations of VRP models and algorithms applied to various scenarios in recent years. Additionally, the article discusses limitations in existing research and provides perspectives on future development trends in VRP research. This review offers researchers in the VRP field a comprehensive overview while identifying future research directions.

Keywords: Vehicle routing problem; VRP; Delivery route optimization; Logistics planning

Online publication: April 2, 2025

1. Introduction

The Vehicle Routing Problem (VRP) has been a central challenge in operations research and logistics for over six decades, its origins tracing back to Dantzig and Ramser's seminal work in 1959 [1]. While the initial formulations focused on simple cost minimization, the explosive growth of e-commerce, urban logistics, and the increasing complexity of supply chains have transformed VRP into a far more multifaceted and critical area of study. No longer are we solely concerned with minimizing distance; modern VRPs demand the simultaneous optimization of multiple, often conflicting, objectives. These include considerations of time windows, diverse vehicle types and capacities, environmental impact (carbon emissions, fuel efficiency), dynamic demand fluctuations, real-time traffic conditions, and even fairness in workload distribution. The integration of big data analytics and advances in artificial intelligence further amplify both the opportunities and challenges inherent in contemporary VRP research [2].

This review critically examines the evolution of VRP research, offering a seasoned perspective on both its theoretical advancements and practical implementations. We will analyze the diverse classifications of VRP models based on load capacity, operational characteristics, objective functions, vehicle heterogeneity, time

constraints, and other crucial factors. Drawing on extensive literature analysis, including a review of recent publications from the Web of Science database, we will highlight key breakthroughs in algorithm design and their application in diverse real-world settings. Furthermore, we will identify persistent challenges, limitations of current methodologies, and promising avenues for future research that are crucial to addressing the growing complexity and sophistication demanded by the modern logistics landscape. This paper aims to provide both established researchers and newcomers with a comprehensive and insightful understanding of the current state-of-the-art and future directions in VRP.

2. Vehicle routing problem

VRP, a cornerstone of operations research and transportation science, formally emerged in 1959 with Dantzig and Ramser's seminal work on optimizing fuel tanker routes. This initial formulation, focusing on minimizing transportation costs from a single depot to multiple delivery points, has since evolved dramatically. The problem's inherent complexity and its crucial role in optimizing logistics efficiency have spurred decades of research, leading to a rich landscape of VRP variants. These variants arise from incorporating a multitude of real-world constraints and objectives beyond simple cost minimization, reflecting the nuanced challenges faced in modern logistics and supply chain management. Understanding these variations, categorized along several key dimensions as detailed below, is essential for effective problem modeling and algorithm design. The classical VRP model describes a scenario where a single distribution center serves multiple user nodes, with the core objective of minimizing transportation costs through route optimization. As research has progressed, scholars have systematically categorized the VRP across various dimensions:

Table 1. Classification of VRP

Dimension	Classification	Description	Example
Load Capacity	Non-full load	Vehicles may not be fully loaded on each route	Delivering small packages across a city; not every vehicle is full
	Full load	Vehicles must be fully loaded before departure (or close to full)	Transporting large, bulky goods where each vehicle requires a full load
Operational Characteristics	Pure loading	Only loading operations at customer locations	Picking up goods from multiple suppliers
	Pure unloading	Only unloading operations at customer locations	Delivering goods to multiple customers from a central warehouse
	Mixed loading-unloading	Both loading and unloading operations occur at customer locations	Picking up goods from some locations and delivering to others
Optimization Objective	Single-objective	Minimizing a single objective (e.g., total distance, cost, time)	Minimizing the total travel distance for all vehicles
	Multi-objective	Optimizing multiple conflicting objectives (e.g., cost & time, cost & emissions)	Minimizing total cost while minimizing total delivery time and CO^2 emissions
Vehicle Homogeneity	Homogeneous vehicles	All vehicles have the same characteristics (capacity, speed, cost)	A fleet of identical delivery vans
	Heterogeneous vehicles	Vehicles have different characteristics (capacity, speed, cost)	A fleet with different sized trucks and vans

Table 1 (Continued)

Dimension	Classification	Description	Example
Time Constraint	With time windows	Deliveries must be made within specified time windows	Delivering perishable goods with strict delivery timeframes
	Without time windows	No time constraints on deliveries	Delivering non-perishable goods with flexible delivery times
Demand Divisibility	Split delivery	Demand at a customer can be split among multiple vehicles	Delivering a large order in multiple trips
	Non-split delivery	Demand at a customer must be fulfilled by a single vehicle	Delivering a furniture set that must be transported together
Priority Constraint	With priority restrictions	Some customers have higher priority than others	Delivering urgent medical supplies before regular orders
	Without priority restrictions	All customers have equal priority	Delivering standard packages to customers
Route Closure	Open routes	Vehicles do not need to return to the depot	A one-way delivery route where vehicles are left at the final delivery point
	Closed routes	Vehicles must return to the depot after completing their routes.	Standard delivery routes where vehicles return to the warehouse.

Notably, practical applications often require consideration of multiple constraint combinations, resulting in more complex VRP variants. These theoretical classifications provide an important conceptual framework for subsequent research while reflecting the rich connotations and research value of VRP problems.

3. Research progress in vehicle routing problem

A comprehensive literature search was conducted in the Web of Science (WOS) database using the keyword "Vehicle routing problem" for full-text articles. As of January 2025, a total of 10,154 papers were retrieved. According to the citation report provided by WOS, these papers have been cited 196,168 times, with an average citation frequency of 19.32 times per paper. **Figure 1** shows the number of papers published and their corresponding citations from 2015 to January 2025. It is evident that since 2015, the quantity of VRP-related academic achievements has steadily increased with a rapid growth rate, while the number of citations has also shown year-on-year growth.

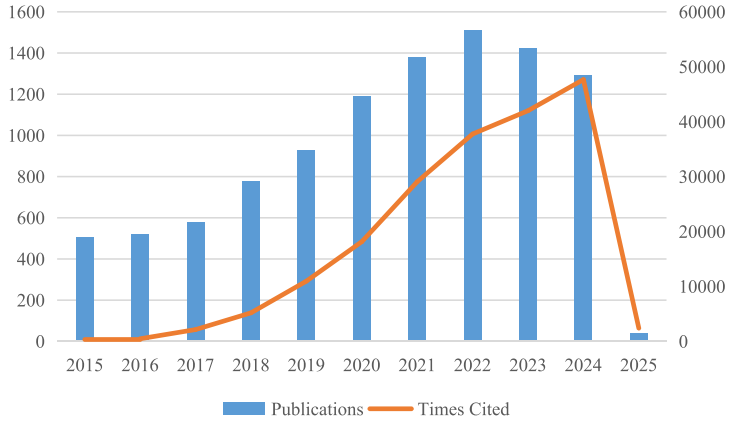


Figure 1. Number of papers and corresponding citation counts from 2015 to 2025

Based on **Figure 2**, through reviewing the relevant literature, we have observed that VRP problems have increasingly attracted attention from researchers in the computer science field, beyond the traditional domains of operations research and transportation. This paper systematically reviews the major research achievements in recent years from two aspects: VRP applications and model algorithm research.

Figure 2. Number of VRP-related papers in various research fields

3.1. Applications

The applications of VRP are expanding rapidly, driven by the increasing complexity of modern logistics and the availability of advanced computational tools. The following examples, drawn from recent literature, highlight both the breadth of VRP applications and the evolving sophistication of solution methodologies. However, it is essential to examine the limitations often present in reported findings critically. Many studies focus on specific scenarios, utilizing carefully selected datasets which may not fully capture the complexities of real-world operations. Furthermore, the benchmark problems often employed do not always provide a rigorous comparison across diverse algorithms and conditions. A more robust and standardized evaluation framework is crucial for future progress.

(1) Sustainable cold chain E-commerce fulfillment: Tsang *et al*. presented a noteworthy attempt to integrate order packaging optimization with vehicle routing in cold chain e-commerce [3]. Their JOSOPMDP model aims for sustainability by minimizing packaging materials. While addressing a significant contemporary issue, the model's real-world applicability hinges on the accuracy of its demand and packaging cost estimations, which are often challenging to obtain 1accurately.

(2) Medical waste transportation optimization: Anityasari *et al.* apply the Periodic Vehicle Routing Problem (PVRP) to medical waste collection in Surabaya [4]. Their work demonstrates the practical benefits of optimized routing for waste management. However, the generalizability of their findings to other contexts necessitates careful consideration of local infrastructure and regulations.

(3) Multi-objective time window VRP for E-commerce: Guo *et al.* utilize an improved Intelligent Water Drops algorithm for a Multi-objective Time Window Vehicle Routing Problem (MTWVRP) applied to

Suning.com's operations [5]. While showcasing algorithmic improvement, the study's reliance on a single case study limits the generalizability of performance claims. Robustness testing across a broader range of problem instances would strengthen these conclusions.

(4) Dynamic VRP for autonomous vehicles in agriculture: Andersen *et al.* tackle the challenging Dynamic Vehicle Routing Problem (DVRP) with dynamic nodes, focusing on autonomous vehicles in agriculture [6]. The findings on optimal detectability levels are valuable, but the context-specific nature of their problem demands careful consideration of transferability to other applications of DVRP.

(5) Capacitated VRP with fuzzy stochastic demand: Singh *et al.* addressed uncertainty in demand using fuzzy stochastic variables within a Capacitated Vehicle Routing Problem (CVRP) [7]. Their approach offers improved realism, yet the computational complexity of handling fuzzy uncertainty needs further investigation for larger-scale real-world applications.

(6) VRP with constraints in car-sharing: Feng and Xiao developed a Vehicle Routing Problem with Constraints (VRPC) for car-sharing, focusing on minimizing operational costs and maximizing user experience [8]. Their hybrid algorithm demonstrates improvement but lacks extensive comparative analysis against a wider range of state-of-the-art algorithms.

(7) Cold chain logistics VRP with travel time prediction: Bai *et al.* integrates data fusion technology for travel time prediction into a cold chain logistics VRP [9]. The use of real-time traffic data is a significant advance, yet the accuracy and robustness of the prediction model in diverse traffic conditions require further validation.

(8) Low-carbon VRP with time windows: Lou *et al.* incorporated carbon emission considerations into a VRP with time windows [10]. The inclusion of high-granularity time-dependent speeds is commendable, but the computational burden associated with high resolution needs further analysis, especially for large-scale deployments.

(9) Electric vehicle-drone coordinated delivery: Ma *et al.* address the novel challenge of coordinated delivery using electric vehicles and drones, considering carbon emissions [11]. This work highlights the growing interest in integrating different transportation modes, however, the practicality of this integrated approach depends on various operational factors, including drone regulations, infrastructure, and battery technology advancements, warranting further investigation.

In conclusion, while these studies offer valuable insights into diverse VRP applications, a critical assessment reveals a need for more rigorous validation methodologies, broader comparative analyses, and a deeper understanding of the limitations inherent in transferring research findings to practical implementation. Future research should prioritize the development of robust, generalizable solutions that address the full spectrum of real-world complexities.

3.2. Model algorithm research

This section reviews recent advancements in model algorithms for solving various VRPs. The research highlights the application of metaheuristics, including hybrid approaches that combine the strengths of different optimization techniques. Specific algorithms discussed include Improved Multi-directional Local Search (IMDLS), Student Psychology-Based Optimization combined with Ant Colony Optimization (SPBO-ACO), Dynamic Population Island Genetic Algorithm with Hybrid Genetic Search (DPIGA-HGS), and an Improved hybrid Fish-Ant Colony Optimization algorithm (IFACO). Additionally, the development of standardized validation methodologies for

algorithm comparison is examined.

Feng *et al.* established the VRPTWRB model and conducted numerical studies on the rational selection of workload resources and fairness functions [12]. An Improved Multi-directional Local Search (IMDLS) algorithm was proposed to solve the model and approximate the Pareto frontier. The IMDLS algorithm restricts the archive size and adaptively determines the number of current solutions and search directions. Experimental results demonstrated that when considering time window constraints, using duration to evaluate workload resources is more appropriate than using distance; more sophisticated fairness functions can effectively identify high-quality non-dominated solutions with good fairness. The IMDLS algorithm outperformed existing multi-directional local search algorithms in terms of both efficiency and solution quality.

Li *et al.* proposed an integrated route planning method for autonomous vehicle delivery systems to improve the efficiency of urban "last-mile" delivery [13]. Experimental results demonstrated that this method significantly improved delivery efficiency and reduced total travel distance and time in real urban delivery scenarios.

Wei *et al.* introduced a hybrid metaheuristic algorithm (SPBO-ACO) combining Student Psychology-Based Optimization and Ant Colony Optimization for solving the Multi-Depot Vehicle Routing Problem with Time Windows for Electric Vehicles (MDVRPTW-EV) [14]. The algorithm integrates ACO's global search capability with SPBO's local search efficiency, employing strategies such as path length classification, strong-weak perturbation, and learning operators to enhance search efficiency and solution quality. Results demonstrated that the algorithm exhibits high scalability and stability and significantly reduced travel distances of electric loaders in industrial applications, proving its practicality.

Rezaei *et al.* proposed a novel hybrid metaheuristic algorithm called Dynamic Population Island Genetic Algorithm with Hybrid Genetic Search (DPIGA-HGS), combining the advantages of Dynamic Population Island Genetic Algorithm (DPIGA) and improved Hybrid Genetic Search (HGS) to solve the Capacitated Vehicle Routing Problem (CVRP) [15]. DPIGA is a specialized variant of island genetic algorithms that allows islands to lose their populations over time. Experimental results demonstrated that DPIGA-HGS outperformed existing state-of-the-art algorithms on multiple benchmark instances (including Uchoa, CMT, Golden, and LoggiBUD), achieving higher solution quality, finding more Best Known Solutions (BKS), and reducing both average and maximum solution gaps compared to BKS. The paper also included parameter tuning and analysis of the impact of the proposed multi-step restart mechanism.

Jastrzab *et al.* introduced a standardized validation methodology for vehicle routing algorithms, addressing the lack of unified and widely adopted algorithm validation methods in existing research [16]. The methodology consists of three main modules: a benchmark generator, a solver, and a post-processing module. The paper demonstrated the flexibility and effectiveness of this approach through experiments on the NP-hard Pickup and Delivery Problem with Time Windows (PDPTW), providing comprehensive performance comparison and analysis of different algorithms.

Hosseini *et al.* investigated the Green Cold Vehicle Routing Problem (GC-VRP), considering traffic congestion and variable speed. The author proposed a mixed-integer nonlinear programming model that incorporates variable vehicle speeds and the impact of traffic congestion during different time periods on travel times and fuel consumption [17]. A hybrid solution method combining Benders decomposition and Binary Particle Swarm Optimization (BPSO) algorithm was developed. Finally, numerical experiments and sensitivity analyses were conducted to validate the algorithm's effectiveness and the model's rationality.

Zhang *et al.* proposed an Improved hybrid Fish-Ant Colony Optimization algorithm (IFACO) to solve the

Vehicle Routing Problem with Time Windows (VRPTW) [18]. The algorithm integrates Artificial Fish Swarm Algorithm (AFSA) with Ant Colony Optimization (ACO). The paper designed three neighborhood search strategies (2-opt exchange, crossover, and insertion) to further enhance solution quality. Experimental results demonstrated that IFACO possesses strong search capabilities and good convergence when solving VRPTW problems of different scales, effectively avoiding local optima while outperforming several existing algorithms in terms of solution accuracy and efficiency.

Several studies employ metaheuristic algorithms to tackle the complexities of VRPs. Genetic algorithms (GAs) remain a popular choice, as demonstrated by Rezaei *et al.* DPIGA-HGS algorithm, which incorporates a dynamic population island strategy and hybrid genetic search to improve solution quality and efficiency on benchmark instances. Lou *et al.* combine a Hybrid Genetic Algorithm with Adaptive Variable Neighborhood Search (HGA-AVNS) to address the VRP with time windows and carbon emission considerations. Ant Colony Optimization (ACO), known for its exploration capabilities, is utilized in Wei *et al.* SPBO-ACO hybrid algorithm, which incorporates Student Psychology-Based Optimization to enhance local search. The comparative effectiveness of these and other GA-based approaches should be investigated further using standardized benchmarks (as suggested by Jastrzab *et al.*).

4. Academic perspectives and personal analysis

The academic research on VRP primarily diverges into two main directions: practical application studies and model-algorithm research. One group of scholars emphasizes the importance of fundamental theory, arguing that VRP requires more solid theoretical support through the enhancement of mathematical models and algorithmic research. Another group of researchers focuses more on the feasibility and applicability of algorithms in real-world scenarios, advocating for the development of customized solutions that address practical needs.

In this analysis, the integration of theoretical model innovation and practical application is crucial for VRP research. On one hand, theoretical research needs to propose more efficient algorithms for complex scenarios (such as multi-objective, multi-constraint problems). On the other hand, the true value of VRP research can only be realized by combining industry requirements to design routing planning solutions that feature both robustness and real-time capabilities.

Overall, vehicle routing problem research shows a trend toward closer alignment with practical applications, adoption of more sophisticated modeling approaches, and development of more efficient algorithms. This development emphasizes both theoretical innovation and practical value, providing strong support for modern logistics management.

5. Existing challenges

Despite extensive research and numerous algorithms developed for Vehicle Routing Problems, several challenges remain unresolved. Firstly, existing studies often take an overly idealistic approach, primarily emphasizing single objectives such as cost minimization and route length optimization while neglecting conflicts between customer satisfaction and cost minimization. These conflicts arise from real-world factors such as travel delays and varying customer demands. Secondly, while research on single-constraint VRP is relatively mature, its application scope is limited, constraining practical implementation. Lastly, widely used heuristic algorithms generally suffer from limitations such as insufficient local search capabilities, slow convergence rates, and susceptibility to local optima,

which restrict their effectiveness in solving complex VRP problems.

6. Future development trends

Based on ongoing technological advancements and evolving demands, this paper identifies the following trends in VRP research:

(1) Dynamic and real-time optimization: Future research will increasingly focus on routing problems in dynamically changing scenarios, such as real-time traffic conditions and order modifications. The development of real-time optimization algorithms, leveraging big data and Internet of Things technologies, will become a dominant trend.

(2) Intelligence and automation: Artificial intelligence, particularly deep learning and reinforcement learning, will play an increasingly crucial role in VRP. Intelligent algorithms, by simulating human decision-making processes, can plan routes more efficiently in large-scale complex networks.

(3) Green solutions and sustainable development: Against the backdrop of growing global environmental concerns, research will increasingly focus on green delivery problems, including electric vehicle route optimization and carbon emission reduction.

(4) Interdisciplinary research: VRP research will integrate with fields such as biology and sociology, for example, designing new algorithms through biological evolution simulation or optimizing distribution network structures through social network analysis.

7. Conclusion

This paper reviews the current state and development trends of VRP. Some researchers have achieved notable progress in model development and algorithm applications, while others have made significant contributions to theoretical framework construction and practical implementations. Looking ahead, VRP research will evolve towards dynamic optimization, intelligence, sustainability, and interdisciplinary integration. It is recommended that researchers, while exploring new algorithms, should emphasize their practical application potential to provide more scientific support for logistics transportation and urban development.

Disclosure statement

The author declares no conflict of interest.

References

[1] Dantzig GB, Ramser JH, 1959, The Truck Dispatching Problem. Management Science, 6(1): 80–91.

[2] Schiffer M, Schneider M, Walther G, et al., 2019, Vehicle Routing and Location Routing with Intermediate Stops: A Review. Transportation Science, 53(2): 319–343.

[3] Tsang YP, Ma H, Tan KH, et al., 2024, A Joint Sustainable Order-Packing Vehicle Routing Optimisation for the Cold Chain E-Fulfilment. Annals of Operations Research (2024): 1–24.

[4] Anityasari M, Rinardi HC, Warmadewanthi IDAA, 2024, Analysing Medical Waste Transportation Using Periodic Vehicle Routing Problem for Surabaya Public Health Facilities. Journal of Material Cycles and Waste Management,

(2024): 1–18.

[5] Guo Z, Karimi HR, Jiang B, et al., 2024, Enhanced Intelligent Water Drops With Genetic Algorithm for Multi-Objective Mixed Time Window Vehicle Routing. Neural Computing and Applications, 2024: 1–15

[6] Andersen T, Belward S, Sankupellay M, et al., 2023, Reoptimisation Strategies for Dynamic Vehicle Routing Problems With Proximity-Dependent Nodes. TOP, 32: 1–21.

[7] Singh VP, Kirti S, Debjani C, 2023, Solving Capacitated Vehicle Routing Problem With Demands as Fuzzy Random Variable. Soft Computing: A Fusion of Foundations, Methodologies, and Applications, 27(21): 16019–16039.

[8] Feng Z, Xiao R, 2023, Spatiotemporal Distance Embedded Hybrid Ant Colony Algorithm for a Kind of Vehicle Routing Problem With Constraints. Frontiers of Information Technology & Electronic Engineering, 24: 1062–1079.

[9] Bai Q, Yuan Y, Fu X, et al., 2024, Vehicle Routing Problem for Cold Chain Logistics Based on Data Fusion Technology to Predict Travel Time. Operational Research, 24(4): 55.

[10] Lou P, Zhou Z, Zeng Y, et al., 2024, Vehicle Routing Problem With Time Windows and Carbon Emissions: A Case Study in Logistics Distribution. Environmental Science and Pollution Research, 29: 31.

[11] Ma J, Ma X, Li C, Li T, 2024, Vehicle-Drone Collaborative Distribution Path Planning Based on Neural Architecture Search Under the Influence of Carbon Emissions. Discover Computing, 27(1): 1–26.

[12] Feng B, Wei L, 2022, An Improved Multi-Directional Local Search Algorithm for Vehicle Routing Problem With Time Windows and Route Balance. Applied Intelligence, 53(10): 11786–11798.

[13] Li T, He Z, Wu Y, 2022, An Integrated Route Planning Approach for Driverless Vehicle Delivery System. PeerJ Computer Science, 8(4): e1170.

[14] Wei X, Niu C, Zhao L, et al., 2025, Combination of Ant Colony and Student Psychology-Based Optimization for the Multi-Depot Electric Vehicle Routing Problem With Time Windows. Cluster Computing, 28: 99.

[15] Rezaei B, Guimaraes FG, Enayatifar R, et al., 2024, Exploring Dynamic Population Island Genetic Algorithm for Solving the Capacitated Vehicle Routing Problem. Memetic Computing, 16(2): 179–202.

[16] Jastrzab T, Myller M, Tulczyjew L, et al., 2024, Standardized Validation of Vehicle Routing Algorithms. Applied Intelligence, 2024: 1–30.

[17] Hosseini M, Rahmani A, 2024, The Green Cold Vehicle Routing Problem With Traffic Congestion and Variable Speed. Energy Systems, 2024: 1–37.

[18] Zhang J, Zhang J, Qin Z, et al., 2022, Vehicle Routing Problems With Time Windows Based on the Improved Hybrid Fish Swarm-Ant Colony Algorithm. International Journal on Interactive Design and Manufacturing (IJIDeM), 2022.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Urban Hydrological Extraction Based on Otsu Threshold: A Case Study in Zhejiang Province, China

Lijun Hu[1], Xunyuan Zheng[1], Yizhong Qi[1], Liang Liu[1], Hongyu Sun[1], Jiahui Chen[2,3], Ling Peng[2,3], Yinghui Han[3]*

[1]Quzhou Guangming Electric Power Design Co., LTD., Quzhou 324000, China
[2]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China
[3]College of Resource and Environment, University of Chinese Academy of Sciences, Beijing 101408, China

***Corresponding author:** Yinghui Han, hanyinghui@ucas.ac.cn

Abstract: This work uses advanced remote sensing to precisely extract hydrological information, supporting transmission network planning. High-resolution water body mapping lets designers optimize routes to avoid ecologically sensitive areas, achieving environmental protection, cost efficiency, and enhanced operational safety. The methodology provides a scalable, replicable framework for intelligent obstacle avoidance in power grid development, applicable to other regions and sectors with similar planning needs.

Keywords: Water extraction; Power transmission system; Intelligent decision

Online publication: April 2, 2025

1. Introduction

In power grid line planning, avoiding obstacles like water bodies (rivers, lakes, reservoirs, wetlands) is crucial due to environmental regulations and the need to minimize ecological disruption [1]. Therefore, intelligently extracting water body information using remote sensing imagery is vital [2]. This study focuses on Zhejiang Province, China, which has a diverse hydrological system. We used the Google Earth Engine (GEE) platform and the Modified Normalized Difference Water Index (MNDWI) method to create an accurate water body map [3, 4]. This map will aid power grid planning by reducing construction costs and environmental impact [5]. The methodology can be applied to other regions facing similar challenges.

2. Methods

2.1. Water body information extraction platform

The Google Earth Engine (GEE) platform, with its global open-source satellite imagery and cloud computing

power, greatly speeds up large-scale remote sensing data analysis. In this study, the GEE platform was used to efficiently obtain and preprocess 2022 Sentinel-2 satellite imagery of Zhejiang Province [6]. The Modified Normalized Difference Water Index (MNDWI) was also computed on the platform, which enhanced computational efficiency and streamlined the analysis.

2.2. Sentinel-2 satellite data

The Sentinel-2 satellite, comprising 2A and 2B, has a 10 - day revisit period per satellite, but together, they can reduce it to 5 days. It carries a multispectral imager (MSI) covering 13 spectral bands , with ground resolutions of 10 m, 20 m, and 60 m [6]. The main spectral bands are shown in **Table 1**.

Table 1. Main bands of sentinel-2

Band	Description	Spatial resolution	Wavelength
B2	Blue	10m	496.6 nm(S2A) / 492.1 nm(S2B)
B3	Green	10m	560 nm(S2A) / 559 nm(S2B)
B4	Red	10m	664.5 nm(S2A) / 665 nm(S2B)
B8	NIR	10m	835.1 nm(S2A) / 833 nm(S2B)
B11	SWIR1	20m	1613.7 nm(S2A) / 1610.4nm(S2B)
B12	SWIR2	20m	2202.4 nm(S2A) / 2185.7 nm(S2B)

2.3. Water body index method

The NDWI method uses the ratio of band differences to extract water body information, highlighting the contrast between water bodies (with strong absorption in the near-infrared band) and other land features (with enhanced reflectance in the near-infrared band). The MNDWI method replaces the near-infrared band in the NDWI with the short-wave infrared band (SWIR), enhancing the distinction between water bodies and buildings and reducing confusion in urban areas [7]. The MNDWI method was used in this study, and its calculation formula is as follows:

$$MNDWI = \frac{\rho_{Green} - \rho_{SWIR}}{\rho_{Green} + \rho_{SWIR}} \quad (1)$$

In this equation, $_{Green}$ and $_{SWIR}$ are the reflectance values in bands B3 and B12 of Sentinel-2 imagery, respectively.

2.4. Otsu's thresholding method

The complexity of the Earth's surface and spectral variability can lead to incorrect water body extractions or omissions. To address this, the Otsu method, an adaptive thresholding technique, is used. It divides image pixels into background and target pixels, using variance as a measure of grayscale distribution uniformity. A larger variance indicates greater differences between image parts [8]. By maximizing between-class variance, the algorithm determines the threshold that minimizes misclassification probability, subsequently enabling robust image binarization. Within this framework, consider an image with grayscale intensities spanning the interval[0, K], where n_i denotes the pixels count at grayscale level i and N representing the total number of pixels in the

image and its corresponding occurrence probability is defined as follows:

$$P_i = \frac{n_i}{N}, i = 0,1,2,\ldots K, \sum_{i=0}^{K} P_i = 1 \tag{2}$$

Divide the pixels in the image into two categories, A and B, based on a grayscale threshold value t. Category A consists of pixels with grayscale values between 0 and t, while category B consists of pixels with grayscale values between $t + 1$ and K. The probabilities of categories A and B are as follows:

$$\omega_0 = \sum_{i=0}^{t} P_i, \omega_1 = \sum_{i=t+1}^{K} P_i = 1 - \omega_0 \tag{3}$$

The mean grayscale values of categories A and B are as follows:

$$\mu_0 = \sum_{i=0}^{t} \frac{iP_i}{\omega_0}, \mu_1 = \sum_{i=t+1}^{K} \frac{iP_i}{\omega_1} \tag{4}$$

The overall mean grayscale value of the entire image is:

$$\mu = \omega_0\mu_0 + \omega_1\mu_1 \tag{5}$$

The between-class variance is defined as:

$$\sigma^2 = \omega_0(\mu_0 - \mu)^2 + \omega_1(\mu_1 - \mu)^2 \tag{6}$$

Let t vary in the range [0, K] with a step size of 1, and the optimal threshold value corresponds to the value of t that maximizes the between-class variance.

2.5. Water body extraction accuracy verification

We choose the confusion matrix based on Python language to evaluate the classification results. The binary confusion matrix consists of TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative) [9]. F1-score is calculated using the constructed confusion matrix. Precision represents the proportion of true water bodies among the extracted water bodies, while recall represents the proportion of extracted water bodies among all true water bodies. The calculation formulas are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F_1 score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

3. Results

3.1. Water area analysis

Applying the methodology described in this study, the total water surface area in Zhejiang Province is estimated to be approximately 4500.67 square kilometers. The areas of water bodies in cities of Zhejiang Province are shown in **Table 2**.

Table 2. Water area of each city in Zhejiang Province (Unit: square kilometers)

Region	Huzhou	Ningbo	Taizhou	Shaoxing	Wenzhou	Jiaxing	Hangzhou	Jinhua	Lishui	Quzhou	Total
Area	466.3902	699.4834	324.3846	404.7413	301.6418	790.6482	920.3597	236.2061	186.887	169.9326	4500.67

In this paper, we select one region in Zhejiang Province to verify the accuracy of water body extraction, using visual interpretation results as the evaluation standard for model accuracy, and calculating the F1-score. In this demonstration area, the F1-score value of water extraction was 0.87.

3.2. Research analysis and implications for power grid line planning

The extraction of water bodies using remote sensing techniques, as demonstrated in this study, serves as a critical foundation for power grid line planning. The accurate identification and mapping of water bodies are essential for several reasons:

3.2.1. Minimizing environmental impact

Water bodies, especially those in ecologically sensitive areas, are subject to strict environmental regulations. By accurately identifying water bodies, planners can design power grid routes that avoid these areas, thereby minimizing ecological disruption and ensuring compliance with environmental laws. For example, avoiding wetlands and coastal areas can prevent habitat destruction and protect biodiversity.

3.2.2. Optimizing construction costs and logistics

Water bodies pose significant challenges for the construction of power lines. Crossing rivers, lakes, or wetlands often require specialized infrastructure such as bridges, elevated lines, or underwater cables, which can be costly and technically complex. By identifying water bodies in advance, planners can optimize routes to avoid such areas, thereby reducing construction costs and simplifying logistics.

3.2.3. Enhancing operational safety and maintenance

Water bodies can impact the operational safety of power lines. For instance, proximity to water bodies may increase the risk of flooding, which can damage power infrastructure and disrupt service. By identifying water bodies and planning routes to avoid them, power grid operators can enhance the resilience of their infrastructure against natural disasters and reduce the frequency of maintenance interventions.

3.2.4. Supporting sustainable development goals

The integration of water body information into power grid planning aligns with broader sustainability goals. By minimizing the impact on water bodies and surrounding ecosystems, power grid projects can contribute to the preservation of natural resources and the promotion of sustainable development practices.

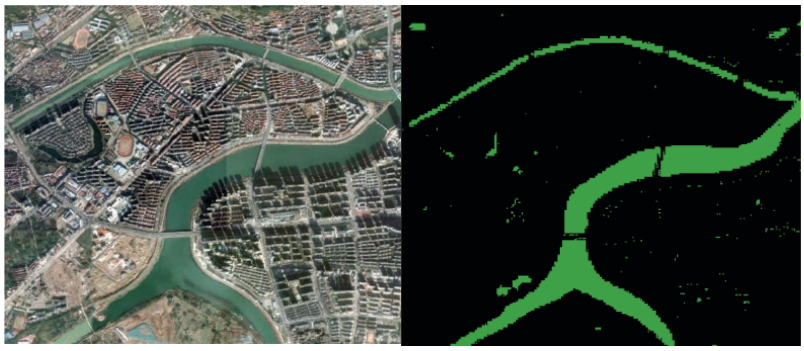


Figure 1. Water extracted by the threshold method

3.3 Case study analysis: power grid line planning in Zhejiang Province

In this study, the water body map was overlaid with proposed power grid routes in Zhejiang Province, leading to several important findings. In Huzhou, the original route intersected small lakes and wetlands; by leveraging the water body data, planners successfully adjusted the route to avoid these areas, thereby reducing both infrastructure demands and environmental impact. In Ningbo, the initial plan involved crossing a major river, but with the aid of water body mapping, an alternative route was identified that eliminated the need for a river-crossing bridge and shortened the power line, significantly lowering construction costs. Similarly, in Hangzhou, the extraction of water bodies revealed the presence of protected wetlands, enabling planners to reroute the grid to avoid these sensitive areas, thus ensuring compliance with environmental regulations and preventing potential legal disputes.

4. Conclusions

Advanced remote sensing techniques, like those in this study, provide a strong framework for better power grid line planning. By identifying and mapping water features, planners can design routes that avoid ecologically sensitive areas, achieving environmental preservation, cost efficiency, and enhanced safety. The methodology offers a scalable solution for intelligent obstacle avoidance in power grid development and has potential for use in other regions. Future research may involve improving extraction algorithms, using multi-temporal data analysis and integrating additional environmental data to enhance the precision and sustainability of power grid planning.

Funding

The State Grid Independent Research and Development Project (Project No.: CY1124SHF02), National Natural Science Foundation of China (Project No.: 52320105003), the Fundamental Research Funds for the Central Universities (Project No.: E3ET1803)

Disclosure statement

The authors declare no conflict of interest.

Author contributions

Fund acquisition: Lijun Hu
Transmission guidance: Xuanyuan Zheng
Line design experience: Yizhong Qi
Resource provision: Liang Liu
Project management: Hongliang Sun
Algorithm implementation: Jiahui Chen
Supervision: Ling Peng
Visualization and review: Yinghui Han

Data availability

The dataset supports the findings of the study are available from the corresponding author upon request.

References

[1] National Bureau of Statistics, 2022, Statistical Communiqué of the People's Republic of China on National Economic and Social Development in 2021, visited on October 23, 2024, https://www.stats.gov.cn/english/PressRelease/202202/t20220227_1827963.html

[2] Qianzhan Intelligence Co Ltd, 2022, Report of Market Prospective and Investment Strategy Planning on China Distributed Energy Industry (2022–2027). The Ministry of Commerce of the People's Republic of China, Beijing.

[3] World Bank Group, 2019, Where Sun Meets Water: Floating Solar Market Report, ESMAP, visited on October 23, 2024 , https://www.esmap.org/where_sun_meets_water_floating_solar_market_report

[4] Chen D, 2017, New Opportunities, Developments and Challenges of Chinese Photovoltaic Power Plants. Electronic Engineering & Product World, 24(5): 3–5.

[5] Qian H, Jiang L, Liang Q, et al., 2021, The Effect of "Fishery-PV Integration" Module Shading Rate on Pond Ecology and Grass Carp Growth, 48(6): 1–12

[6] Zhang M, Wang Y, 2020, Global Floating PV Industry Development Status and Market Prospect Analysis. Solar Energy, 2020(7): 19–24.

[7] Song X, Bei Y, Yuan B, et al., 2022, The Influence of Floating Photovoltaic Power Station on Key Indicators of Water Ecological Environment. Water Resources Protection, 2022(5): 1–9. http://kns.cnki.net/kcms/detail/32.1356.TV.20210901.1039.002.html

[8] Sun J, 2017, Application Technology and Solution of Floating Photovoltaic Power Station. JIE NENG YU HUAN BAO, 2017(2): 48–51.

[9] Xiong Z, Liu J, Shen M, et al., 2020, Benefit Analysis of Integrated Project of Fishing and Lighting—Taking TW New Energy Provincial Fishery Boutique Park as an Example. Tropical Agricultural Engineering, 2020, 44(3): 38–42.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

An Optimization Method for Reducing Losses in Distribution Networks Based on Tabu Search Algorithm

Jiaqian Zhao[1]*, Xiufang Gu[1], Xiaoyu Wei[1], Mingyu Bao[2]

[1]School of Electric Power, Inner Mongolia University of Technology, Hohhot 010051, Inner Mongolia, China
[2]Inner Mongolia Ultra-High Voltage Power Supply Company, Hohhot 010080, Inner Mongolia, China

Corresponding author: Jiaqian Zhao, zhaojjaqian1998@163.com

Abstract: With the continuous growth of power demand and the diversification of power consumption structure, the loss of distribution network has gradually become the focus of attention. Given the problems of single loss reduction measure, lack of economy, and practicality in existing research, this paper proposes an optimization method of distribution network loss reduction based on tabu search algorithm and optimizes the combination and parameter configuration of loss reduction measure. The optimization model is developed with the goal of maximizing comprehensive benefits, incorporating both economic and environmental factors, and accounting for investment costs, including the loss of power reduction. Additionally, the model ensures that constraint conditions such as power flow equations, voltage deviations, and line transmission capacities are satisfied. The solution is obtained through a tabu search algorithm, which is well-suited for solving nonlinear problems with multiple constraints. Combined with the example of 10kV25 node construction, the simulation results show that the method can significantly reduce the network loss on the basis of ensuring the economy and environmental protection of the system, which provides a theoretical basis for distribution network planning.

Keywords: Distribution network; Loss reduction measures; Economy; Optimization model; Tabu search algorithm

Online publication: April 2, 2025

1. Introduction

Under the background of the continuous growth of electricity consumption in the whole society, the proportion of tertiary industry and residential electricity consumption in the electricity consumption structure is increasing, and the influence of distribution network in the power system is also increasing. According to statistics, the power loss caused by 10kV distribution network accounts for about 50% of the power loss of the entire system [1]. And with the development of smart grid, the addition of distributed power sources also makes the loss distribution of distribution network more complicated. Research on the loss reduction of distribution network

can produce good economic and environmental benefits.

At present, distribution network loss reduction studies tend to select loss reduction measures by studying the loss reduction effect of a single measure in the distribution network, ignoring the loss change distribution law and interaction of different loss reduction measures in the network, and have not considered other better measures that may indirectly affect the heavy loss area. The existing research shows that the use of energy-saving wires can reduce the resistance value of the line, to reduce the power loss of the transmission line. At the same time, combined with the analysis of different regions, the application of energy-saving wires has significant economic and social benefits [2]. The particle swarm clustering algorithm with adaptive inertia weight can cluster similar load curve users and classify them for phase optimization adjustment, which can significantly reduce the unbalance of three-phase load and effectively reduce the line loss [3].

One of the important tasks of power enterprise operation is to ensure economy and the rational use of loss reduction measures can improve the economy of power system operation. However, the existing research on loss reduction often fails to comprehensively consider the economic benefits generated by the combination of economic and practical operation of each loss reduction measure. The existing research comprehensively considers the economic operation of the distribution network and the loss reduction measures of the distribution network. By analyzing the economic operation conditions of distribution transformers, the key factors affecting the economic operation of distribution transformers are identified and a reasonable distribution transformer is selected to reduce the loss [4]. According to the loss reduction scheme set formed by different loss reduction measures, the optimal decision model is constructed, which takes into account the comprehensive planning cost and the economic benefit of the lost power, and the optimization algorithm is used to determine the combination scheme of loss reduction measures with the best comprehensive benefit [5]. Existing research is based on the power grid energy efficiency index corresponding to the proposed loss reduction measures. It uses the sum of the total input of loss reduction measures and the loss cost resulting from electricity loss as the objective function. In addition, the research takes into account various constraints, such as voltage deviation, branch transmission capacity, the number of measures, and the total input capital. Finally, through the enumeration method in the optimization algorithm, the measure combination scheme with the best overall benefit is obtained [6].

Optimization algorithm is a way to solve the problem of optimal combination of measures. The combination optimization of loss reduction measures for distribution network can be reduced to discrete and nonlinear programming problems. The available comprehensive optimization algorithms include genetic algorithm, particle swarm algorithm, tabu search algorithm and so on [7]. In the existing research, tabu search algorithm is used in reactive power optimization of power system, and the feasibility and effectiveness of this algorithm are verified by practical cases [8]. At the same time, genetic algorithm is applied to the power supply planning of distribution network and the reactive power reduction optimization of distribution network [9]. In the distribution network model with distributed power access, the optimization model of loss reduction based on genetic algorithm is constructed to solve the line loss of distribution network, and the corresponding loss reduction measures are given [10].

This paper analyzes the comprehensive benefits of different combinations of loss reduction measures, taking into account the construction cost of each measure and the economic benefits of loss reduction. It establishes an optimization model for distribution network loss reduction, with the ultimate goal of maximizing the comprehensive benefits of loss reduction. The model is solved using the tabu search algorithm. Through the combination and parameter optimization of different measures, The comprehensive loss reduction scheme of

distribution network with the best economic benefits is obtained.

2. Establishment of distribution network loss reduction optimization model

Distribution network planning for loss reduction is a multi-objective optimization problem. The decision optimization model should fully consider the comprehensive benefits brought by the implementation of each loss reduction measure to the current distribution network. On this basis, the optimal optimization objective function of comprehensive benefits and the distribution network loss reduction optimization model under different constraints are established. The combined optimization model is as follows:

$$\begin{aligned} L_1 \quad & \min F(\lambda,\xi) \\ s.t. \quad & G(\lambda,\xi) \le 0 \\ & H(\lambda,\xi) = 0 \end{aligned} \tag{1}$$

In the formula, (1) F- optimization model objective function; (2) G- inequality constraint and; (3) H- equality constraint.

2.1. Optimize the model objective function

Under the current development goal of a low-carbon society, the construction costs, economic benefits of power saving, and environmental benefits of various measures at different transformation levels should be comprehensively considered. In the distribution network loss reduction optimization model, the objective function includes the acquisition costs and other construction costs related to equipment installation and replacement. The comprehensive benefits of distribution network loss reduction mainly account for the direct power benefits generated by reducing electricity loss, as well as the indirect low-carbon environmental benefits resulting from the reduced electricity loss [11]. Therefore, it is proposed that the objective function of distribution network planning loss reduction model is:

$$max\, F = F_1 + F_2 - (F_T + F_L + F_C) \tag{2}$$

In the formula, F_1– the economic benefits directly generated by reducing the loss of electricity, 10,000 yuan/year; F_2– The low carbon environmental benefit indirectly generated by the reduction of power loss, 10,000 yuan/year; F_T– Comprehensive investment cost of replacing different types of transformers, 10,000 yuan/year; F_L– The comprehensive investment cost of replacing different models of distribution lines, 10,000 yuan/year; F_C– ifferent capacity reactive power compensation device comprehensive investment cost, 10,000 yuan/year.

In the objective function of the distribution network optimization model, the calculation methods of different benefits and costs are as follows:

The economic benefits directly generated by reducing the loss of power. F_1 can be obtained by reducing the loss of power multiplied by the corresponding electricity price:

$$F_1 = \left(\sum_{n=1}^{n_{all}} \Delta A_{Tn} + \sum_{m=1}^{m_{all}} \Delta A_{Lm}\right)\sigma \times 10^{-4} \tag{3}$$

In the formula, the power loss reduced by the ΔA_{Tn} , ΔA_{Ln} – n transformer and the m line in one year, kWh; n_{all}, m_{all} –The number of transformers and lines in the distribution network; σ– Local electricity price, yuan.

The low-carbon environmental benefits indirectly generated by the reduction of power loss F_2 can be calculated according to the price per ton of carbon dioxide emission rights in the current carbon trading market and the carbon dioxide emissions reduced by the reduction of power loss:

$$F_2 = \left(\sum_{n=1}^{n_{all}} \Delta A_{Tn} + \sum_{m=1}^{m_{all}} \Delta A_{Lm}\right) \times \beta \times \frac{\rho}{1000} \times 10^{-4} \quad (4)$$

In the formula, β– carbon dioxide emissions reduced by saving per unit of electricity, kg;

ρ– carbon trading market price per ton of carbon dioxide emission rights, yuan.

Annual comprehensive investment cost of reactive power compensation device F_C, according to the distribution network planning to reduce the input of compensation device capacity, unit capacity cost, other installation and construction investment and economic service life:

$$F_C = \frac{1}{N_C}(f_{C1} \cdot Q_C + f_{C2}) \quad (5)$$

From this equation, f_{Cl} – reactive power compensation device unit capacity cost, ten thousand yuan; Q_C – Reactive power compensation capacity, kVar; f_{C2} –Other construction expenses of reactive power compensation device, ten thousand yuan; N_C – Economic service life of the reactive power compensation device: years.

The annual comprehensive investment cost of distribution line replacement F_L, according to the acquisition cost required to replace the large diameter distribution line, other construction costs and the economic service life of the line:

$$F_L = \frac{1}{N_L}(f_{L1} + f_{L2})L \quad (6)$$

Based on Equation 6, f_{Ll} – Purchase cost per unit length of different types of distribution lines, ten thousand yuan; f_{L2} – Other construction costs per unit length of distribution line, ten thousand yuan; L – Distribution line replacement length, km; N_L – Distribution line economic service life, years.

Distribution transformer replacement annual comprehensive investment cost, F_T, according to the replacement of new energy-saving transformer equipment acquisition cost, other construction costs and economic service life:

$$F_T = \frac{1}{N_T}(f_{T1} + f_{T2}) \quad (7)$$

Based on Equation 7, f_{Tl} – different types of transformer acquisition cost, ten thousand yuan; f_{T2} – Distribution transformer other construction costs, ten thousand yuan; N_T – Distribution transformer economic applicable life, years.

2.2. Optimization model constraints

In the distribution network loss reduction optimization model, the establishment of constraint conditions mainly includes the following parts:

(1) Power flow equation constraints

$$\begin{cases} P_i = P_{Li} + U_i \sum_j^N U_j (G_{ij}\, cos\, \delta_{ij} + B_{ij}\, sin\, \delta_{ij}) \\ Q_i = Q_{Li} + U_i \sum_j^N U_j (G_{ij}\, sin\, \delta_{ij} - B_{ij}\, cos\, \delta_{ij}) \end{cases} \quad (8)$$

From this equation, node $P_i, Q_i - i$ input active power, reactive power, kW, kVar; U_i, U_j –The voltage amplitude of nodes i and j, kV; P_{Li}, Q_{Li} – Active and reactive power absorbed by node i load, kW, kVar; $G_{ij}, B_{ij}, \delta_{ij}$–Conductance, susceptance, and voltage phase Angle difference of branches between nodes i and j.

(2) Constraint of voltage deviation

In the current power system, the voltage deviation requirement for the 10kV distribution network is $-10\% \le \Delta U\% \le +5\%$, and the voltage deviation requirement for the 0.4kV distribution network is $|\Delta U\%| \le +7\%$, where the voltage deviation calculation formula is:

$$\Delta U\% = \frac{U_i - U_N}{U_N} \times 100\% \quad (9)$$

In the formula, U_i, U_N node operating voltage and rated voltage value, kV.

(3) Line transmission capacity constraints

The actual transmission capacity of the line should not exceed its maximum allowed transmission capacity, transmission capacity is generally expressed by the transmission current:

$$I_j \le I_{j.max} \quad (10)$$

In the formula, I_j, $I_{j,max}$ The actual current flowing through the branch and the maximum current allowed through, A.

3. Solution of distribution network loss reduction optimization model

3.1. Distribution network loss reduction optimization model solving algorithm

For the solution of distribution network optimization model, its process can be regarded as a combinatorial optimization problem. The methods commonly used to solve combinatorial optimization problems in current research include: genetic algorithm, particle swarm optimization, enumeration, and tabu search algorithm [12–13]. Tabu Search algorithm (Tabu Search) is a kind of optimization algorithm based on the global scope proposed by scholar F. Lover in the 1970s. The basic idea of the algorithm is to search for the extended neighborhood and constantly marking the local optimal solution found in the search process to avoid the problem of roundabout search at the local extreme value and obtain the global optimal solution through continuous optimization [14]. The algorithm is suitable for nonlinear, multi-constraint, and discrete multi-variable global optimization problems, mainly composed of the following seven parameters: tabu object, tabu length, contempt criterion, neighborhood function, evaluation function, memory frequency information, and termination criterion.

3.2. Solution steps of distribution network loss reduction optimization model

The tabu search algorithm is used to solve the optimization model and realize the combination and parameter optimization of the distribution network planning loss reduction scheme. The implementation steps are as follows:

(1) Read the distribution network line information, load data, distribution transformer information and other network parameters, operation data parameters and equations, inequality constraint parameters;

(2) According to the preliminary collection of loss reduction measures, generate a combined loss reduction scheme set of different loss reduction measures, and establish a neighborhood structure;

(3) According to the set of combined loss reduction schemes, input algorithm parameters, place tabu table $TL = \varnothing$, calculate the objective function of the initial scheme, and take the scheme $X^{best} = X^0$ as the current optimal scheme;

(4) Generate a neighborhood scheme according to the changes of parameters in the neighborhood structure, pay attention to only considering the change of a single parameter when the parameters change, and do

not take the scheme in the tabu table as the neighborhood scheme;

(5) To judge whether the constraint conditions in each neighborhood scheme are satisfied, subtract the penalty function of maximum value from the objective function of the neighborhood scheme that does not meet the constraint conditions;

(6) Calculate the objective function value of the neighboring scheme and compare it with the current optimal solution. If it is better than the current optimal solution, take the neighboring scheme as the current optimal scheme $X^{best}=X^{0}$;

(7) Determine whether the termination conditions are met. If the termination requirements are met, output the current optimal scheme as the distribution network loss reduction measures optimization specific implementation scheme; if not, add one iteration number and return to step (4) to continue the iterative calculation.

4. Analysis of numerical examples

In this paper, the 25-node radial distribution network of 10kV as shown in **Figure 1** is used as an example, and four loss reduction measures including 23-node reactive power compensation, L-1 trunk line replacement wire model, 12-node reactive power compensation, and 23-node replacement distribution transformer are selected to form the primary measure set of the comprehensive loss reduction scheme of the distribution network. In the combination optimization of loss reduction measures, there are 15 kinds of combination measure loss reduction schemes randomly generated in the set of primary measures. The specific selection of measures for each scheme is shown in **Table 1**.

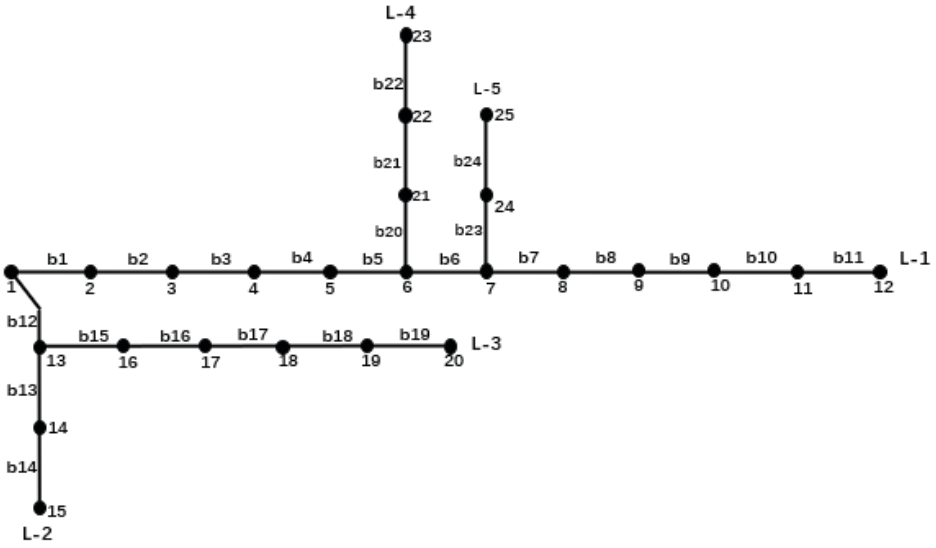


Figure 1. 25-node rural distribution network model

Table 1. Loss reduction measures portfolio options

Loss reduction measures portfolio options	Loss mitigation measures			
	23-node reactive power compensation	L-1 trunk replacement wire model	12-node reactive power compensation	23-node replacement of distribution transformer
Option 1	√			
Option 2		√		
Option 3			√	
Option 4				√
Scheme 5	√	√		
Option 6	√		√	
Scheme 7	√			√
Scheme 8		√	√	
Scheme 9		√		√
Option 10			√	√
Option 11	√	√	√	
Scheme 12	√	√		√
Scheme 13	√		√	√
Scheme 14		√	√	√
Scheme 15	√	√	√	√

In the parameter optimization of loss reduction measures, there are mainly the following options:

(1) In view of the replacement of the main line L-1 wire model loss reduction plan, considering the requirements of future distribution network development transmission capacity, LGJ-150 model is no longer selected on the basis of the original LGJ-120 wire of the branch road, and directly choose to replace the model LGJ-185 or LGJ-240 transmission wire to ensure the future load change requirements;

(2) For 12 nodes, 23 nodes to install reactive power compensation device loss reduction program, the study found that the higher the compensation power factor is not the higher the benefit, in order to make the best comprehensive benefit usually control the power factor of the distribution network at about 0.95, while the current distribution network reactive power compensation usually requires a power factor greater than 0.9 [16]. Therefore, when reactive power compensation is selected, the load power factor reaches 0.9 or 0.95 compensation capacity;

(3) For the replacement of 23-load node distribution transformer loss reduction scheme, on the basis of the original selection to replace the current more used S11 or S13 energy-saving new transformer parameters optimization.

In consideration of parameter optimization, there are hundreds of comprehensive loss reduction schemes for the distribution network, and manual sorting with fewer schemes will produce a large amount of calculation. It is necessary to combine the parameter cost of each measure and use tabu search algorithm to optimize the combined parameter calculation. The optimization price of each measure parameter is shown in **Table 2**.

Table 2. Cost optimization of loss reduction measures parameters

Measures	Voltage rating (kV)	Replacement parameters	Engineering comprehensive cost	Units	Other construction expenses	Units
Replacement of transmission wires	10	LGJ-185	1.700	RMB /km	1	Million yuan /km
		LGJ-240	2.400			
Replacement distribution transformer	10	S11-630	5.570	RMB/set	0.3	Ten thousand yuan/set
		S13-630	7.039			
Reactive power compensation mounting	0.38	-	0.011	RMB /kVar	0.94	Yuan/place

In the objective function calculation of the optimization model, a power purchase price of 0.7 yuan/kWh is used, while the transmission line and distribution transformer costs are based on an economic service life of 15 years. Additionally, the costs of the installed reactive power compensation device are calculated based on an economic service life of 10 years, considering the comprehensive costs of various measures and other construction expenses. The tabu search algorithm is used to obtain the annual economic benefits generated by different combination of measures and parameter optimization in different loss reduction schemes. The tabu step size is 4, and no change in the optimal objective function in 10 iterations is taken as the termination criterion. The specific measures for the top three comprehensive economic benefits in the optimization model are shown in **Table 3**.

Table 3. Optimization results of distribution network loss reduction scheme model

Scheme No.	Specific measures and parameter optimization results	Annual comprehensive benefit / 10,000 yuan
1	23-node reactive power compensation to cos = 0.95, 12-node reactive power compensation to cos = 0.95, 23-node load distribution transformer replaced to S13-630.	3.311
2	23-node reactive power compensation to cos = 0.95, 12-node reactive power compensation to cos = 0.95, 23-node load distribution transformer replaced to S11-630.	3.196
3	23-node reactive power compensation to cos = 0.95, 12-node reactive compensation to cos = 0.95.	3.167

According to the results obtained by the optimization algorithm, the specific loss reduction scheme and optimization parameters for the loss reduction planning of the distribution network are obtained. From the analysis in **Table 3**, it can be found that the more loss reduction measures taken, the better comprehensive benefits will not be produced. When the power supply department carries out practical application, the possibility of implementing measures should be considered comprehensively in combination with economic benefits.

5. Conclusion

This paper starts with the combination and parameter optimization of loss reduction measures in distribution

network, establishes an objective function combining the comprehensive cost and the comprehensive benefit of the implementation of each loss reduction measure in distribution network, and optimizes the combination and parameter optimization of the measures concentrated on primary measures through optimization algorithm under certain constraints. Considering that there are many comprehensive measures schemes, tabu search algorithm is used to solve the optimization model. The combination of loss reduction measures and parameter scheme with the best comprehensive benefits are obtained. The following conclusions can be drawn through the analysis of numerical examples:

(1) Under the comprehensive consideration of various operating constraints, the optimized scheme can significantly reduce network loss and improve the economy and environmental protection of the system.

(2) Tabu search algorithm can effectively avoid the defect that the traditional optimization algorithm is easy to fall into the local optimal, and shows a good global optimization ability in the loss reduction optimization of distribution network.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Yang Y, 2018, Comprehensive Analysis and Research on Power Loss in Urban Medium and Low Voltage Distribution Network. Electric Times, 2018(9): 72–73.

[2] Yan L U, Co H P, Ltd, 2015, Measures of Reducing Line Loss through the Optimization of Chengmai Power Grid Structure. China Electric Power (Technology Edition), 2015(1): 41–43.

[3] Zhou R, Wang J, Hou X, Wang M, Qiu X, 2015, Phase Optimization Adjustment Method for Three-phase Feeder Loss Model. Journal of Electric Power Systems and Automation, 27(4): 1–6.

[4] Wang X, 2018, Research on Economic Operation and Loss Reduction Measures of Distribution Network, thesis, Heilongjiang: Northeast Agricultural University, DOI:10.7666/d.Y3516823.

[5] Li T, 2010, Research on Energy Saving and Loss Reduction of 10kV Distribution Network, thesis, South China University of Technology.

[6] Wang Y, Zheng T, Shi Y, 2019, A Combined Optimization Model for Distribution Network Loss Reduction Considering Low-Carbon Benefits. Distributed Energy Resources, 4(1): 13–16.

[7] Rouhi F, Effatnejad R, 2015, Unit Commitment in Power System by Combination of Dynamic Programming (DP), Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). Indian Journal of Science and Technology, 8(2): 134.

[8] Wang H, Xiong X-G, Wu Y-W, 2002, Reactive Power Optimization of Power System Based on Improved Tabu Search Algorithm. Power Grid Technology, 26(1): 15–18.

[9] Zhang P, Liu Y, 1999, Simplified Dynamic Programming Method for Voltage Control and Reactive Power Optimization of Distribution Systems. Journal of Electric Power Systems and Automation, 11(4): 49–54.

[10] Ying L, Liu M, Deng L, Sun J, 2017, Review on Loss Reduction of Distribution Network. Power System Protection and Control, 45(19): 162–169.

[11] Tang H, Wang X, Xie G, Feng M, 2019, Combinatorial Optimization Model of Distribution Network Loss Reduction

Scheme Considering Low Carbon Benefit. Journal of Electric Power Systems and Automation, 32(2): 113–118.

[12] Golberg D E, 1989, Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley, 1989(102): 36.

[13] Tungadio D H, Numbi B P, Siti M W, et al, 2015, Particle Swarm Optimization for Power System State Estimation. Neurocomputing, 148: 175–180.

[14] Sun Y, Shen T, Liu C, Sun Z, Zhang Z, 2017, Loss Reduction Decision of Distribution Network Based on Grey Relational Degree Algorithm. Sichuan Electric Power Technology, 40(4): 24–2847.

[15] HHe W, 2008, Reactive Power Optimization of Regional Power Network Based on Tabu Search Algorithm, thesis, Shaanxi: Xi'an University of Science and Technology, DOI:10.7666/d.y1322167.

[16] Ren G, 2021, Research on Reactive Power Compensation Method of Distribution Network with Distributed Power Source. Electronic Technology and Software Engineering, 2021(19): 212–213.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

The Application of Machine Vision in Defect Detection Systems

Peihang Zhong*, Jiawei Lin, Muling Wang

Guangdong Technology College, Zhaoqing 526100, Guangdong, China

Corresponding author: Peihang Zhong, zhangry010129@163.com

Abstract: With the rapid development of computer vision technology, artificial intelligence algorithms, and high-performance computing platforms, machine vision technology has gradually shown its great potential in automated production lines, especially in defect detection. Machine vision technology can be applied in many industries such as semiconductor, automobile manufacturing, aerospace, food, and drugs, which can significantly improve detection efficiency and accuracy, reduce labor costs, improve product quality, enhance market competitiveness, and provide strong support for the arrival of Industry 4.0 era. In this article, the concept, advantages, and disadvantages of machine vision and the algorithm framework of machine vision in the defect detection system are briefly described, aiming to promote the rapid development of industry and strengthen China's industry.

Keywords: Machine vision; Defect detection system; Image preprocessing

Online publication: April 2, 2025

1. Introduction

The *Guiding Opinions on Accelerating Scene Innovation to Promote High-Quality Economic Development with High-Level Application of Artificial Intelligence* clearly points out that with a smarter city and a more caring society as the guidance, it will continue to explore opportunities for artificial intelligence application scenarios in urban management, traffic management, ecological environmental protection, medical health, education, elderly care, and other fields, and carry out application demonstrations of intelligent society scenarios. In the field of urban management, urban brain, urban IoT perception, invisible government data availability, digital procurement, and other scenarios are explored. In the field of traffic governance, scenarios such as traffic brain, smart roads, smart parking, self-driving travel, smart ports, and smart waterways will be explored. In the field of ecological and environmental protection, we will focus on scenarios such as intelligent environmental monitoring and autonomous drone inspection. In the field of smart communities, scenarios such as future communities, unmanned delivery, community e-commerce, and digital restaurants will be explored. In the

medical field, scenarios such as intelligent diagnosis assisted by medical images, decision support assisted by clinical diagnosis and treatment, medical robots, Internet hospitals, intelligent medical equipment management, smart hospitals, and intelligent public health services are actively explored. In the field of education, online classes, virtual classes, virtual simulation training, virtual teaching and research rooms, new teaching materials, teaching resource construction, and smart campuses are actively explored. The elderly care sector is actively exploring scenarios such as home intelligent monitoring and smart wearable device applications. In the rural sector, we are actively exploring scenarios such as smart rural governance, digital rural housing, and online government services. Machine vision, as a part of artificial intelligence, should give play to its detection ability to better detect defects, so as to better promote the development of industry and achieve industrial progress [1].

2. Overview of machine vision

2.1. The concept of machine vision

Machine vision is a rapidly developing branch of the field of artificial intelligence, which mainly refers to the use of machines instead of human eyes to make measurements and judgments [2]. Machine vision is a combination of image processing, mechanical engineering, optical imaging, sensor technology, analog and digital video technology, as well as computer hardware and software technology (including image enhancement and analysis algorithms, image cards, etc.), and other fields of knowledge. It also includes image capture, light source system, image digitization, digital image processing, intelligent judgment decision and mechanical control execution, and other modules, to achieve automated detection and decision.

2.2. The advantages of machine vision

2.2.1. Higher production efficiency and accuracy

On the production line, compared with the traditional manual inspection, the machine vision system can not only continue to work but also will not affect the detection efficiency due to fatigue or inattention. On the contrary, it will detect faster, have a stronger ability to process information, and can quickly process and analyze the data to achieve automatic integration and real-time feedback of information. It also improves the ability of enterprises to optimize the production process [3]. For example, in the automobile manufacturing workshop, the factory can use machine vision to detect the paint film applied to the body, check whether the car has scratches, and whether there is a dent or uneven color problem; if there is any problem, the factory can immediately make decisions and changes.

2.2.2. Strong adaptability and flexibility

Machine vision systems can adapt to high or low temperatures, humidity, dust, and other complex environments, where people cannot work [4]. For example, in the process of growing food, farmers can use machine vision technology to monitor the growth of crops, weed distribution, and disease and insect pests; according to this precise application and irrigation, they can reduce the overuse of pesticides, improve the yield and quality of crops, and can even customize and optimize according to the needs of different farmers to meet different needs.

2.3. The shortcomings of machine vision

2.3.1. Limited comprehension and reasoning ability

Although it can process and analyze a large amount of image data, it is far inferior to humans in terms of in-

depth understanding and reasoning of images, because it relies on pre-set algorithms and models for image recognition and analysis, and these models are not able to analyze complex images. Recognition of abstract concepts in images and understanding of the relationship between the context of the scene and emotional knowledge is difficult to process, so machine vision is difficult to completely replace human beings in some application scenarios requiring high intelligence and flexibility [5].

2.3.2. Poor adaptation to complex scenes and backgrounds

Machine vision system mainly relies on optical imaging technology for image capture and processing, so it is easy to be affected by lighting changes, occlusions, noise, and other factors, which will lead to the performance decline of machine vision systems [6]. In the environment of low light or large light changes, the machine vision system may not be able to accurately capture and identify the target object in the image. In the case of a complex background or the presence of multiple similar target objects, there may also be misrecognition or missed recognition [7].

3. The algorithm framework of machine vision in the defect detection system

3.1. Image preprocessing

Due to the imperfection of the imaging system, interference of the transmission medium, or improper operation in the image processing process, these images exist in the form of Gaussian noise and salt and pepper noise (isolated pixels or pixel blocks. Although they have nothing to do with the information of the image itself, it has a great impact on the image processing steps of the following edge extraction and feature recognition. How to suppress these noises in the defect detection system, improve the image quality, and ensure the accuracy and efficiency of the detection data has become the key we need to discuss [8]. The detection system can first use the median filter, which will sort the pixel values of the neighborhood, and then select the middle value as the output, effectively removing the isolated noise points while maintaining the edge and details of the image. Then, using the Gaussian filter in the way of weighted summation, give different weights to the neighborhood pixels, the weight size is determined by the Gaussian function, so that not only can smooth the noise but also can better retain the edge information of the image, reduce the influence of Gaussian noise, enhance the overall contrast of the image. After such pretreatment steps, the image will improve the quality of the image, highlight the defect features, and provide a solid foundation for the subsequent feature extraction and classification recognition [9].

3.2. Region of interest extraction

Threshold segmentation and edge detection are the two most commonly used methods of ROI extraction image segmentation technology, so we will elaborate on the application of these two methods in ROI extraction, defect recognition, and classification in machine vision defect detection systems through specific examples.

First, the application of threshold segmentation in ROI extraction. In the defect detection system, threshold segmentation is an algorithm that divides the image into areas with properties significantly different from the background or other normal areas by setting one or more thresholds based on the image's grayscale value or color and other attributes. Taking semiconductor chip defect detection as an example, defects on the surface of the chip (such as scratches, stains, omissions, etc.) usually appear as grayscale differences from the surrounding normal area. First of all, the detection system can use image processing software or algorithms to process the chip image, such as denoising and enhancing contrast, which can better improve the accuracy of

threshold segmentation [10]. Secondly, according to the gray histogram of the pre-processed image, one or more appropriate thresholds can be selected for segmentation, and the defective area on the surface of the chip can be divided from the normal area. At this time, according to the location, shape, and size of the flaw area, the smallest rectangular box or irregular polygon containing the flaw can be manually or automatically outlined as ROI, which will greatly reduce the amount of computation and improve the processing speed [11].

Second, the application of edge detection in ROI extraction. Edge detection is based on the mutation of pixel values in the image to identify the edge of the object to determine ROI. In textile defect detection, the detection system extracts the flaw edge contour of the textile such as holes, stains, and color difference, and preprocesses the textile image using image processing algorithms such as filter denoising and edge enhancement. After that, Canny operator, Sobel operator can be used as the classic edge detection operator, which will identify the edge by calculating the image gradient to extract the pre-processed image. After edge extraction, the edge points that constitute the contour of the edge of the defect can be obtained. The obtained edge points can not only be connected to the cultivated edges by morphologic operations such as expansion and corrosion, but also the complete contour of the defect can be extracted by a contour tracking algorithm, or a rectangular box or irregular polygon containing the smallest defect can be taken as ROI in order to ensure the accuracy of subsequent detection [12].

Third, defect identification and classification. In order to identify the defects of the color, texture, and shape of the metal surface, the detection system can use the image processing algorithm to extract the color features (such as RGB value, HSV value, etc.), texture features (such as grayscale co-occurrence matrix, local binary mode, etc.), shape features (such as area, circumference, aspect ratio, circularity, etc.) within the ROI; according to the needs of defect recognition, select a representative subset for subsequent processing. Then select a suitable classifier from a vector machine (SVM), decision tree, random forest, neural network (such as convolutional neural network CNN) to identify classification defects, and finally output detection results [13].

Threshold segmentation and edge detection are suitable for different types of defect detection tasks respectively, and can be combined with technical means such as feature extraction, feature selection, and classifier design to achieve accurate identification and classification of defects and improve the accuracy and efficiency of detection [14].

3.3. Defect recognition and classification

3.3.1. Defect identification

In PCB defect detection systems, machine vision technology always uses high-precision cameras and image sensors to capture common problems such as poor welding, broken wires, missing components, and short circuits. Therefore, the machine vision system will enhance and preprocess the acquired images by image smoothing, brightness, and contrast adjustment steps before flaw recognition, so that it can better show the detailed characteristics of the PCB surface, and provide reliable input data for subsequent defect recognition. The machine vision system will also be based on the threshold segmentation algorithm can separate the foreground and background in the image, better distinguish the defect area, understand the content of the image, and improve the accuracy of defect recognition [15].

3.3.2. Defect classification

On the one hand, the statistical classification method is to identify and classify defects by analyzing defect

features in historical data and using cluster analysis and discriminant analysis. For example, the K-means clustering algorithm needs a large amount of historical data as a training model, divides defects into different categories according to the size, shape, and other characteristics of defects, establishes an accurate statistical model, and realizes the accurate classification of defects. On the other hand, decision trees, support vector machines (SVM), and convolutional neural networks (CNN) can also learn the visual features of defects through training, and realize automatic identification of defects. For example, for PCB defect detection, CNN can be used to train the model, so that it can distinguish between normal areas and different types of defects, label the type and location of defects, and achieve better classification of defects.

3.3.3. Algorithm design

Image enhancement and pre-processing algorithms (denoising algorithm, image smoothing algorithm, brightness and contrast adjustment algorithm) can improve the accuracy of detection, and feature extraction and design classification algorithms are the keys to achieving defect detection and classification. As the name suggests, the denoising algorithm can remove the noise in the image and improve the image quality. Image smoothing algorithm is able to reduce the detail information in the image, so that the defects are more obvious. Brightness and contrast adjustment algorithm is to adjust the brightness and contrast of the image to make the defects of the image more prominent. The feature extraction algorithm can extract the texture, shape, and color similar to the defect. Classification algorithm means to classify defects according to the extracted features. In daily life, the appropriate classification algorithm can be selected according to the specific detection requirements and data characteristics, so as to improve the accuracy of classification.

3.3.4. Practical application cases

PCB defect detection system built by electronic manufacturing company using machine vision technology can identify common defects such as poor welding, wire breakage, missing components, and short circuits, and classify and record these defects to improve detection efficiency, reduce labor costs and time, ensure the stability and unity of product quality, so as to reduce unqualified products. It also improves production efficiency and provides strong support for the sustainable development of enterprises. Machine vision technology in the defect detection system can use a high-precision camera and image sensor to capture the fine features of the product, the use of image processing and analysis technology for high-precision product detection, in order to play an important role in machine vision technology.

4. Conclusion

Machine vision technology through simulation and beyond the limits of human vision for defect detection has brought a lot of changes and can solve the limitations of traditional manual detection methods in terms of efficiency, accuracy, and stability. More industrial manufacturing can provide efficient, accurate, reliable defect detection solutions, in order to promote industrial manufacturing to higher quality and efficiency, and more intelligent direction. In this paper, the application of machine vision in the defect detection system is briefly summarized, aiming to provide references and a basis for researchers studying this direction, contributing their strength to the high-quality development of industrial manufacturing, realizing the rapid development of industry, leading the industrial automation and intelligence level to a new level.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Li X, 2024, Research on Fabric Defect Detection System based on Deep Learning, dissertation, Heilongjiang University.

[2] Tao N, 2024, Research on Deep Learning-based Towel Defect Visual Detection Algorithm and its Application, dissertation, Hebei University.

[3] Huo Y, 2024, Design of Flaw Detection System for Glass Production Line, dissertation, North University of China.

[4] Hu B, 2023, Research and System Implementation of Intelligent Flaw Detection Method for Cloth, dissertation, Jishou University.

[5] Jiang K, 2023, Flaw Detection System of Wine Box Product Label based on deep Learning, dissertation, Nanjing University of Information Science and Technology.

[6] Xiao J, Guo H, Wang N, 2019, Design of Towel Defect Detection System based on Convolutional Neural Network. Computer Integrated Manufacturing Systems, 30(11): 3977–3983.

[7] Zhu H, 2021, Research on Fabric Defect Detection System based on Machine Vision, dissertation, South China University of Technology.

[8] Guan Z, 2021, Research on Defect Detection Technology of Automobile Instrument based on Machine Vision and Deep Learning, dissertation, Jilin University.

[9] Li L, 2021, Flaw Detection System of Glass Panel based on Industrial Vision, dissertation, Taiyuan University of Technology.

[10] Tang C, 2021, Development of Wheel Hub Defect Detection System based on Machine Vision, dissertation, Zhejiang Normal University.

[11] Cheng Y, 2021, Research on Yarn Defect Detection System based on Fusion Sensing, dissertation, Fujian University of Technology.

[12] Yang Y, 2021, Flaw Detection of Warp Knitting Fabric based on Digital Image Processing, dissertation, China University of Geosciences (Beijing).

[13] Zhu Q, 2021, Research and Application of Wine Bottle Defect Detection System based on Machine Vision, dissertation, Qingdao University of Science and Technology.

[14] Hou Y, 2021, Design of TFT Screen Defect Detection System based on Deep Learning, dissertation, Beijing University of Posts and Telecommunications.

[15] Liu Y, 2021, Research on Defect Detection of Industrial Steel based on Deep Learning, dissertation, Anhui University of Civil Engineering and Architecture.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Web Visualization Application of Large Mesh Models Based on Simplification Algorithms

Shengtai Shi*

Cranfield University, Cranfield MK43 0AL, United Kingdom

***Corresponding author*:** Shengtai Shi, shengtai.shi63@gmail.com

Abstract: This paper studies polygon simplification algorithms for 3D models, focuses on the optimization algorithm of quadratic error metric (QEM), explores the impacts of different methods on the simplification of different models, and develops a web-based visualization application. Metrics such as the Hausdorff distance are used to evaluate the balance between the degree of simplification and the retention of model details.

Keywords: QEM algorithm; Mesh simplification; WebGL rendering; Java web development

Online publication: April 3, 2025

1. Introduction

In the field of industrial design, 3D modeling enables designers to simulate and inspect product designs in detail before manufacturing. This improves the accuracy of designs, helps to detect design flaws in advance, and reduces subsequent costs [1]. High-precision 3D models can also be used in engineering design for precise load and stress analysis and to verify the integrity of structures. Although elaborate models can bring more accurate simulations, the performance degradation, high memory consumption, reduced interactivity, and increased costs caused by a large number of polygons cannot be ignored. This paper aims to balance the simplification rate and the retention of model details according to actual needs while maintaining visual quality. The simplification algorithm is combined with web applications, and the rationality of the final product is evaluated through data processing, calculation, and visualization.

1.1. Theoretical basis

Vertex clustering: The mesh is evenly divided into multiple partitions, and all vertices within a partition are clustered into a single point, usually the centroid [2,3]. However, uniform cell division sometimes cannot fit the original mesh well. In some cases, a vertex tree system is used to divide more strictly to conform to the original mesh [4].

Vertex deletion: It is an incremental reduction method. After deleting a vertex, triangles are regenerated to patch the resulting holes [5], aiming to gradually reduce the complexity.

Based on traditional simplification methods, there are also some innovative solutions:

(1) Mesh reconstruction and resurfacing: New vertices are evenly distributed on a coarser mesh to approximate the original topology and replace the original mesh [6].

(2) Symmetry-aware algorithms [7].

1.2. Quadratic error metric method

The quadratic error metric method studies the impact of geometric errors caused by vertex movement by defining a quadratic error matrix for each vertex and describing the deviation of the vertex from the plane it lies on, thereby guiding the simplification work [8].

Representing the plane equation with a matrix, we can get:

$$P = [abcd]^T \tag{1}$$

In the formula, a, b, and c are the components of the plane normal vector, determining the direction of the plane in 3D space, and d is the offset of the plane, representing the distance of the plane from the origin in the direction of the normal vector.

For a vertex, its initial error matrix is a zero-matrix. In a 3D model, this vertex may be connected to multiple triangular faces. By summing the contribution values of all the faces passing through this vertex, we can calculate the required error matrix Q_i, its form is as follows:

$$Q_i = \sum_{F \in \mathcal{F}(v_i)} \mathbf{p}\mathbf{p}^T \tag{2}$$

In the formula,$F(v_i)$ represents all the faces passing through the vertex v_i.

To simplify the mesh, it is necessary to identify and select all possible vertex pairs for merging. The position of the merged vertex needs to be calculated to ensure the minimum error, and this metric can be represented by the error matrix. The algorithm traverses all vertex pairs, calculates the error of each merging pair, and then selects the pair with the minimum error for merging. During this process, common priority queues such as Min-Heap can be used to improve the efficiency of selecting vertex pairs. After selecting the vertex pair and determining the merging position, the mesh structure is updated, the redundant vertex is deleted, all the faces related to it are reconnected, and the error matrix of the new vertex is calculated again [9].

1.3. Hausdorff distance

The Hausdorff distance is commonly used in fields such as geometry, image processing, and computer vision. It defines the similarity or dissimilarity between two point sets and is a measure of the maximum distance between two point sets [10].

The shortest distance from any point a_i in set A to any point in set B is d_i. The longest di among these distances is selected as the Hausdorff distance between sets A and B [11].

$$\mathrm{h(A,B)} = \max_{a \in A} \left\{ \min_{b \in B} \{d(a,b)\} \right\} \tag{3}$$

The Hausdorff distance is also applicable when A and B are polygons instead of discrete point sets [12]. The conventional Euclidean shortest distance only applies to one vertex of each polygon, while the Hausdorff distance takes into account all other vertices of the polygon. The Hausdorff distance is sensitive to the distance between the positions of two polygons but not sensitive to the shortest distance between the positions of polygons [13]. In this paper, it is used to show the local maximum mismatch between the two models.

2. Research plan

2.1. Simplification algorithm

The algorithm separately defines the processing of triangles in 3D space and stores the information of their vertices in a pointer array. By defining sine and cosine values, a series of rotations are performed on the triangle so that one of its vertices falls at the origin of the coordinate system and the other two vertices are located on the corresponding coordinate axes. This process flattens the triangle in 3D space to simplify subsequent geometric calculations and can also determine whether the target is a valid triangle.

The simplification process is divided into three main steps: loading the mesh, simplifying the mesh, and updating the mesh. By default, the quadratic error metric is used as a guide to calculate the error of each edge, then the optimal position of the vertex is calculated, and a new vertex is generated. If two vertices have the same boundary attributes, they will be merged. A variant merging method that does not generate new vertices is derived from this method. It selects the vertex with a smaller error from the original two vertices as the merged position. On this basis, after calculating the errors, the errors are sorted, and the edges with the smallest errors are processed first, which improves the efficiency and accuracy when the simplification rate is fixed. Four variants of the quadratic error metric can be obtained according to whether the errors are sorted and whether new vertex positions are generated.

Shorter edges often have less impact on the appearance of the model, which can also be used as one of the criteria for merging. Calculate and save the Euclidean distances of the three sides of each triangle, and save the index of the shortest side to a new triangle array by comparison to obtain the shortest-edge algorithm.

The last type of method is the Melax threshold method. It uses Melax's error-calculation method to combine the length and curvature of the edge. The calculation method of the edge length is as described above, and the curvature is obtained by calculating the minimum angle between the normal of each triangle adjacent to the two vertices of the edge. Finally, the calculated edge length and curvature are multiplied as the error [14].

In the mesh update, the first iteration identifies and marks all the points on the boundary to prevent damage to the model. If the current iteration is not the first iteration, the triangles marked for deletion are first removed, the remaining triangles are moved forward in the array, then the references to vertices and triangles are initialized, the starting position of the triangle references for each vertex in the array is calculated based on the reference count, and the reference information of each triangle vertex is written. In addition, to make the model simplification more stable, some auxiliary functions are added to the algorithm, such as checking whether the removal of an edge will cause the triangle to flip.

To quantify the error, it is necessary to calculate the bounding box of the model and the length of its diagonal, then derive the error coefficient according to the simplification level, and finally divide the diagonal length by this error coefficient to obtain the scaled error range, which is used to control the accuracy of error calculation. Any distance between a vertex and a triangle that is less than this scaled error range is considered

to be within an acceptable error range [15]. The Hausdorff distance calculated using the vertex set of one model and the triangle-face set of another model in the common bounding box of the two models can reflect the large deviations in the local area.

2.2. Web visualization application

The main content area contains the structure and layout of the entire page and initializes the application when the page is loaded. It contains two canvas areas and a model information display area. The models are rendered through WebGL and can be interacted with separately. The model information mainly includes the number of vertices and triangles and the file size to visually compare the differences before and after simplification from a data perspective.

The head of the web page is responsible for linking and referencing external resources and initializing default settings. Meta tags are used to adjust the display character encoding and viewport configuration to ensure that the web page works properly on different types of mobile devices. Structurally, the rendering and interaction functions are achieved by embedding the WebGL renderer and JS functions in the head section so that they can be called on the web page. The fragment and vertex shader information required for each of the two models is also stored in this section. The body section only considers the layout of the web page and makes the user interface more reasonable and friendly by adding elements and components.

The JS code section covers the main functions such as managing user interactions, buffering and loading models, event handling, and interface updates. It processes a series of logical operations and feeds back the results to the interface for viewing. For example, it reads the file content and parses it into JSON data, loads and compiles shaders, creates WebGL program objects, and links them to the corresponding WebGL context.

3. Result output and analysis

When the simplification rate is low, the overall appearance and local details of the model are well-preserved, with almost no obvious differences. When the simplification rate is increased to 50%, the edges of the model become sharper. This change is particularly evident on surfaces with large curvatures, such as the edges of the rabbit's ears, which change from smooth curves to distinct straight-line segments, while details such as dents are still well-preserved as shown in **Figure 1**.

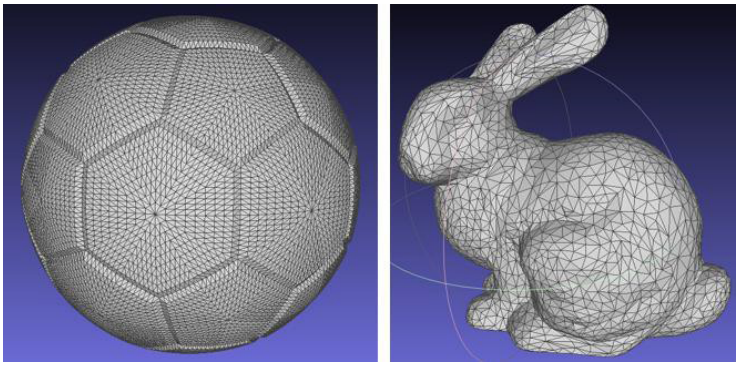


Figure 1. Grid comparison football – rabbit

Under high-level simplification, the appearance of the model becomes further blurred, and some

deformation occurs. The number of line segments forming the curves is further reduced, resulting in a further decrease in smoothness. In addition, most geometric details, such as the eyes, ears, and dents on various parts of the body, will be lost. For simple models, due to the small number of triangles themselves, the remaining triangles after high-level simplification are not enough to support all the detailed features. To ensure the overall appearance, the algorithm preferentially discards these details. Based on observations, the simplification results often visually affect the surfaces with large curvatures first, causing obvious defects. To analyze this feature, a sphere model composed of uniform continuous surfaces was selected for comparative research.

For a sphere model with a continuous and uniform surface, even under high-level simplification, the visual performance remains good. The mesh shows that such graphics often have a high degree of symmetry and consistency, and the numerical error distribution of the uniform mesh is more consistent, which helps to improve stability and overall accuracy. The topological errors caused by the removal of triangles are also evenly distributed throughout the model, so the local and overall appearance changes are also mitigated. In contrast, the mesh triangles of the rabbit model vary greatly. The triangles in the gentle slope area on the back are relatively uniform. Similarly, in the three simplification levels, this part is the least visually affected, while areas with significant triangle changes, such as the ears and concave surfaces, are the most deformed parts in the simplified model, which is consistent with the above-derived and observed phenomena.

In addition to the influence of the model structure and mesh characteristics on the visualization of the simplified model, when the original model is fine-grained enough, the number of remaining triangles after simplification will be larger, which may make the simplified model more similar to the original model. Therefore, a Buddha statue model with both high-frequency details and large, flat surfaces was selected for subsequent experiments as shown in **Table 1**.

Table 1. Simplified data of Buddha statues

	Original model	**20% simplification**	**50% simplification**	**90% simplification**
Number of vertices	149970	120006	75036	15000
Number of triangles	100000	80000	50000	10000
File size	3062.54 kb	2436.33 kb	1497.79 kb	277.46 kb

In actual results, QEM performs similarly when processing complex and simple models. It does not improve the visual effect due to the increase in the number of triangles and exposes the problem of poor handling of high-frequency details under high-level simplification. Although it is very effective in maintaining the overall shape, it often erases almost all small details. For example, bumps and dents blend in with the adjacent large planes and are difficult to distinguish. QEM does not distinguish between areas with rich and sparse details. Therefore, even considering error minimization, vertex merging will still lead to loss of details.

Analysis of the data shows that the Hausdorff error increases with the increase in the degree of simplification, which means that the gap between the output model and the original model gradually widens. When the degree of simplification is not large (less than 60%), the error increase is relatively gentle. When the degree of simplification is high (more than 60%), the error will increase significantly. Although the file size is compressed, the visual effect may become unacceptable. The positive impact of the uniformity of the model mesh on the simplification results is also well-reflected in the line chart. The error change of the football model is more controllable. Without considering model specificity, QEM has a superior simplification

performance. Sorting the triangles can slightly improve the calculation efficiency, and the selection of whether to generate a new vertex merging position can be made according to requirements. To balance the convenience of transmission and model quality, in most cases, a simplification rate of no more than 60% is appropriate. In cases where there are fewer details or only the appearance of the model is emphasized, a higher simplification rate can be selected as needed.

4. Review and summary

This paper analyzes the ideas and processes of model simplification around the QEM algorithm and compares it with other common methods to show their respective applicable scenarios, advantages, and disadvantages. To make model simplification more user-friendly, a series of model processing processes from simplification, transformation, to visualization are proposed. Through experiments, the developed algorithm can simplify well according to the set compression ratio.

In the future, it is expected to further integrate the three processing processes into the web end to achieve high-level integration and automation. Based on the characteristics of various algorithms and their performances at different simplification degrees, users can be provided with references when choosing the simplification ratio.

Disclosure statement

The author declares no conflict of interest.

References

[1] Durk J, 2019, Model Simplification in Manufacturing Simulation – Review and Framework. *Computers & Industrial Engineering,* 127: 1056–1067.

[2] Arvo J, Euranto A, Jarvenpaa L, et al., 2015, 3D Mesh Simplification: A Survey of Algorithms and CAD Model Simplification Tests, University of Turku, Technology Research Center, Finland.

[3] Rossignac J, Borrel P, 1993, Multi-resolution 3D Approximations for Rendering Complex Scenes, Springer Berlin Heidelberg, 455–465.

[4] Erikson C, Luebke D, 1997, View-dependent Simplification of Arbitrary Polygonal Environments. Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH ’97), New York, NY, USA, 199–208.

[5] Schroeder W, 1997, A Topology Modifying Progressive Decimation Algorithm, *IEEE Visualization ’97*, Phoenix Arizona USA, 205–212.

[6] Voutchkov I, Keane A, Shahpar S, et al., 2018, (Re-)Meshing Using Interpolative Mapping and Control Point Optimization. Journal of Computational Design and Engineering, 5(3): 305–318.

[7] Podolak J, Funkhouser T, Golovinskiy A, 2009, Symmetry-aware Mesh Processing, Mathematics of Surfaces XIII*, Springer Berlin Heidelberg*, 170–188.

[8] Heckbert P, Garland M, 1997, View-dependent Simplification of Arbitrary Polygonal Environments. Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH ’97), New York, NY, USA, 209–216.

[9] Hugues H, 1996, Progressive meshes. Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '96), Association for Computing Machinery, New York, NY, USA, 99–108.

[10] Klanderman G, Rucklidge W, Huttenlocher D, 1993, Comparing Images Using the Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 15(9): 850–863.

[11] Munkres J, 1999, *Topology*, 2nd ed, Upper Saddle River, NJ, USA, Prentice Hall.

[12] Rucklidge W, 1997, Efficiently Locating Objects Using the Hausdorff Distance. *International Journal of Computer Vision,* 24(3): 251–270.

[13] Vito D, Valery S, 1999, Distance-based Functions for Image Comparison. *Pattern Recognition Letters*, 20(3): 207–214.

[14] Stan M, 1998, A Simple, Fast, and Effective Polygon Reduction Algorithm. *Game Developer Magazine*, 5(11): 44–49.

[15] Taha A, Hanbury A, 2015, An Efficient Algorithm for Calculating the Exact Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11): 2153–2163.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Research on the Development Strategies of Real-time Data Analysis and Decision-support Systems

Wei Tang*

Sichuan Provincial Architectural Design and Research Institute Co., Ltd., Chengdu 610000, Sichuan, China

***Corresponding author*: Wei Tang, 13980526616@163.com

Abstract: With the advent of the big data era, real-time data analysis and decision-support systems have been recognized as essential tools for enhancing enterprise competitiveness and optimizing the decision-making process. This study aims to explore the development strategies of real-time data analysis and decision-support systems, and analyze their application status and future development trends in various industries. The article first reviews the basic concepts and importance of real-time data analysis and decision-support systems, and then discusses in detail the key technical aspects such as system architecture, data collection and processing, analysis methods, and visualization techniques.

Keywords: Real-time data analysis; Decision-support system; Big data; System architecture; Data processing; Visualization technology

Online publication: April 3, 2025

1. Introduction

In the current era of rapid development of information technology, data has become an indispensable and important asset for enterprises and organizations. The application of real-time data analysis and decision-support systems enables enterprises to make informed decisions quickly in a volatile market environment, thereby enhancing their competitiveness and improving operational efficiency [1]. Decision-support systems can not only integrate a large amount of real-time data but also provide solid decision-making support through complex data analysis models. However, building an efficient real-time data analysis and decision-support system is no easy task. It requires careful planning and design in many aspects, including system architecture, data collection and processing, analysis methods, and visualization techniques.

2. Basic principles and concepts of real-time data analysis and decision-support systems

The real-time data analysis and decision-support system (DSS) is an important application of modern

information technology in the field of enterprise decision-making. The basic working principle of this system covers a series of continuous processes, including data collection, transmission, processing, analysis, and presentation.

Real-time data analysis involves collecting data from various sources, such as sensors, network logs, and social media. The operation of the data collection system depends on strong high-concurrency processing capabilities and stable and reliable transmission channels, which are the keys to ensuring the timeliness and integrity of the collected data. Data transmission usually relies on high-speed networks and message queue systems, such as Kafka, to ensure that data can quickly reach the central processing system [2].

In the field of real-time data analysis, data processing plays a crucial role. For example, stream processing frameworks like Apache Flink and Apache Storm are characterized by their ability to process data immediately when it arrives, rather than using the traditional batch-processing method. This link includes steps such as data cleaning, format conversion, data aggregation, and in-depth analysis, aiming to ensure the accuracy of the data and meet the requirements of subsequent analysis.

In this process, techniques from statistics, machine learning, and data mining are used to conduct in-depth analysis of real-time generated data. Models are built to predict future trends, revealing potential valuable information and patterns. The analysis results are usually presented to decision-makers through visualization tools, such as Tableau and Power BI, to provide an intuitive understanding and application of the data. The DSS uses the results of real-time data analysis to assist decision-makers in making appropriate decisions when facing semi-structured or unstructured decision-making situations [3]. The DSS consists of components such as a database, a model base, a knowledge base, and a user interface, and its functions cover aspects such as data management, model management, and user interaction. **Figure 1** shows the structure of an intelligent command decision-support system using artificial intelligence.

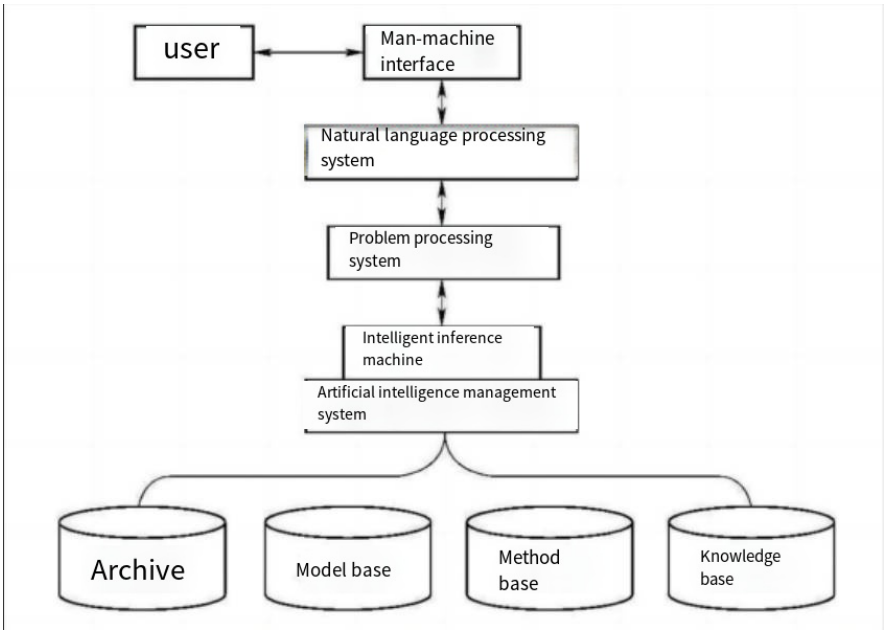


Figure 1. Structure of the intelligent command decision-support system using artificial intelligence

3. Development strategies of real-time data analysis and decision-support systems

3.1. System architecture design

3.1.1. Data collection layer

The data collection layer is responsible for collecting data in real-time from multiple sources. To ensure that the data remains up-to-date within a specific period and covers all the necessary content, an efficient information collection strategy must be established. Common data collection techniques include various methods, such as using sensor networks, log collectors, API interfaces, and data stream platforms (such as Apache Kafka).

3.1.2. Data transmission layer

The data transmission layer undertakes the crucial task of ensuring that the collected data can be quickly and accurately delivered to the central processing system. In the field of computer network communication, common protocols include TCP/IP, HTTP, etc. In addition, message queue technology, such as RabbitMQ, is also widely used in data transmission [4]. To improve the efficiency of data transmission, data compression methods and batch-transmission techniques can be adopted.

3.1.3. Data processing layer

The data processing layer is responsible for performing real-time processing and operations on the data transmitted to this layer. For example, Apache Flink and Apache Storm are both stream-processing frameworks that play a vital role in realizing real-time data processing. Data processing includes several links, such as data cleaning, format conversion, aggregation, and analysis.

3.1.4. Data storage layer

In the process of real-time data processing, both temporary and persistent storage technologies must be used. As a medium for temporarily storing data during the data-processing stage, in-memory databases (such as Redis) and NoSQL databases (such as MongoDB) are usually used. These tools and technologies aim to improve data access speed and ensure data consistency [5]. For the long-term storage and analysis of historical data, distributed file systems (such as HDFS) and relational databases (such as MySQL) are usually adopted.

3.1.5. Analysis layer

In the field of data analysis, it is crucial to conduct in-depth analysis and model building on the processed data by applying various algorithms and models. The range of analysis tools includes machine-learning libraries (e.g., TensorFlow, Scikit-learn) and statistical analysis tools (e.g., R, SAS), which are commonly used in this field.

3.2. Data collection and processing strategies

3.2.1. Data collection strategy

During the data collection process, it is particularly important to select appropriate data sources and collection methods. When choosing data sources, factors such as data real-time update, accuracy, and relevance to the research objectives must be considered. Data collected from sensor networks, Internet of Things devices, as well as information recorded by log systems and API interfaces, are common data sources. To ensure the efficiency of the data collection process, parallel processing, and distributed collection strategies can be implemented [6].

3.2.2. Data processing strategy

In the data-processing strategy, ensuring the real-time update and accuracy of the data is of great significance. Data cleaning is a key operation carried out at the initial stage of the data-processing flow. Its purpose is to eliminate irrelevant information in the data set and correct incorrect data. In the data pre-processing process, common processing methods include filling in missing data, detecting and identifying outliers, and identifying and removing duplicate records. Data in different formats are adjusted to the same format for subsequent processing.

3.3. Analysis methods and algorithms

3.3.1. Statistical analysis methods

As a key means of real-time data analysis, statistical analysis plays an indispensable role in in-depth data research. In the field of data analysis, common analysis methods include basic data summarization (descriptive statistics), inferring the population from samples (inferential statistics), exploring the relationships between variables (regression analysis), and studying the patterns of data changes over time (time-series analysis). These methods form the basic framework for researching data and information and provide a scientific basis for decision-making [7].

3.3.2. Machine-learning algorithms

Machine-learning technology is the core of dealing with real-time data analysis problems. Supervised learning algorithms, such as linear regression, decision trees, and support vector machines; unsupervised learning algorithms, such as cluster analysis and principal component analysis; and reinforcement learning, which is a learning method that trains models through reward and punishment mechanisms. Machine-learning algorithms can automatically obtain models from data and make predictions and classifications based on them [8].

3.3.3. Data mining techniques

Data mining involves the in-depth exploration of massive data to discover valuable information and patterns. In the field of data mining, common techniques include discovering association rules, data classification, data clustering, and extracting frequent patterns. Through data-mining techniques, hidden potential patterns and relationships in the data can be discovered, providing data support for the decision-making process [9].

3.3.4. Real-time analysis techniques

Real-time analysis techniques are the core elements for realizing real-time data analysis and play a crucial role. For example, stream-processing frameworks such as Apache Flink and Apache Storm, as well as real-time analysis platforms like Spark Streaming, can perform immediate processing and in-depth analysis on continuous data streams. Real-time analysis techniques require high-concurrency processing and low-latency response from the system.

3.4. Visualization techniques

Visualization techniques are key tools that present the results of data analysis and processing in a graphical way, making it more intuitive to explore and understand information. Through graphical representation, geographical mapping, and interface display, data and analysis conclusions can be more clearly perceived and grasped by users [10].

3.4.1. Selection of visualization tools

When presenting data, choosing the appropriate visualization method is a decisive factor in determining the information presentation effect and efficiency. In the field of data visualization, FineBI, Tableau, Power BI, and D3.js are common and widely used tools. These tools have excellent data processing and visualization capabilities and can adapt to diverse application scenarios.

3.4.2. Principles of visualization design

When conducting visualization design, simplicity, intuitiveness, and understandability must be emphasized. When choosing a chart, the characteristics of the data and the display objectives should be considered. For example, for time-series data, a line chart is a suitable display method. For categorical data, bar charts or pie charts can effectively visualize the data. During the page design process, appropriate attention should be paid to color matching and layout structure to avoid visual confusion caused by too many graphical elements and to ensure that the information presentation does not exceed the audience's reception capacity [11].

3.5. System integration and testing

3.5.1. System integration strategy

System integration involves integrating multiple independent modules into a unified whole. During the system integration process, the interaction interfaces between modules should be closely monitored, and seamless data flow should be ensured. In the practice of system integration, common methods include hierarchical integration, phased incremental integration, and continuous integration strategies [12].

3.5.2. Testing strategy

Ensuring system quality depends on the key step of testing. The testing strategy covers multiple levels, including unit testing, integration testing, system testing, and acceptance testing, aiming to comprehensively evaluate software quality. Unit testing is the process of testing individual modules to verify the accuracy of their functions. Integration testing is to comprehensively test the integrated modules to ensure that the interfaces and data flow meet the expected standards. System testing is to comprehensively test a complete system to ensure that all functions operate normally and meet the established performance standards [13].

3.5.3. Performance testing

In the field of real-time data analysis and the development of DSS, performance testing plays a key role. The inspection of system performance mainly covers the evaluation of the system's load, stress, and stability. Load testing is a method for evaluating the performance of a system under various expected load conditions. Stress testing is to apply pressure to the system to evaluate its stability under extreme conditions. Stability testing aims to evaluate the performance of the system during long-term continuous operation [14].

3.6. Security and privacy protection

3.6.1. Data security

In the development of real-time data analysis and DSS, data security is an important aspect that cannot be ignored. During data transmission, encryption technologies such as SSL/TLS should be used to effectively prevent data from being intercepted and modified. During data storage, strict access-control mechanisms must be implemented, and data should be encrypted to resist unauthorized access attempts.

3.6.2. Privacy protection

Privacy protection includes maintaining the security of users' personal information to ensure that it is not accessed, used, disclosed, or tampered with without authorization. During software development, relevant laws and regulations, such as the General Data Protection Regulation (GDPR), must be complied with to ensure data security and compliance. To achieve the confidentiality of personal information, protection measures include anonymizing data, applying privacy-protecting algorithms, and setting strict system access permissions [15].

4. Conclusion

The development strategies of real-time data analysis and DSS are the keys to ensuring the efficiency, reliability, and scalability of the system in practical applications. By discussing in detail aspects such as system architecture design, data collection, and processing, analysis methods and algorithms, visualization techniques, system integration and testing, security and privacy protection, and user-experience design, this study provides comprehensive guidance and references for the development of high-performance real-time data analysis and DSS. Against the backdrop of the rapid development of big data and the Internet of Things, the demand for real-time data analysis and DSS is constantly increasing. In the future, with the in-depth integration of artificial intelligence and machine-learning technologies, real-time data analysis and DSS will become more intelligent and efficient. The application of edge computing will further reduce the latency of data processing and improve the system response speed. Innovations in data visualization techniques will make data presentation more intuitive and understandable, helping decision-makers understand and utilize data more quickly and effectively.

Disclosure statement

The author declares no conflict of interest.

References

[1] Yuan L, 2023, The Development of Sports Data Analysis and Sports Decision-Support Systems. Sports Goods & Science and Technology, (23): 142–144.

[2] Tang X, Xu Y, 2023, Research on the Correction Strategy for Calculation Deviations in Decision-support Systems-taking the Recruitment Support System as an Example. Journal of Intelligence, 42(6): 87–95.

[3] Liu S, 2023, Research on Medical Information Management and Decision-Support Systems Based on Big Data Analysis. Science and Informatization, (24): 160–162.

[4] Ai G, 2024, Research on the Maintenance Decision-Support System for Coal Mine Tunneling Equipment Based on Big Data. China Plant Engineering, (07): 166–168.

[5] Chang X, 2024, Communication Data Analysis and Real-Time Decision-support in the Internet of Things Environment. Telecommunications Power Technology, 41(1): 142–144.

[6] Jing H, 2024, Real-time Data Analysis and Decision-Support of Artificial Intelligence in Grid Digitalization. China Strategic Emerging Industry, (02): 48–50.

[7] Wu Z, 2024, The Application of Intelligence in Equipment Management. Mechanical Engineering & Automation, (03): 222–223 + 226.

[8] Du L, Chen D, 2024, Research and Analysis of Water Conservancy Engineering Survey Based on the Smart Big Data Platform. Stone, (05): 101–103.

[9] Yue M, 2025, Construction of the Whole-Process Cost Dynamic Monitoring and Decision-Support System for Enterprises. Modern Enterprise, (01): 45–47.

[10] Wang P, 2025, Design of a Procurement Decision-Support System Based on Big Data Analysis. Automation Application, 66(01): 199–201.

[11] Wu D, 2024, Intelligent Monitoring and Real-Time Data Analysis Methods for Chemical Equipment. Chemical Engineering & Equipment, (12): 107–109.

[12] Wang H, 2024, The Application of Computer Systems in Emergency Rescue Decision-Support Systems. Applications of IC, 41(5): 266–267.

[13] Zhang Q, Gao L, Zhang Y, 2024, Design of Big Data Analysis and Intelligent Decision-support Systems in Power System Communication. Telecommunications Power Technology, 41(6): 69–71.

[14] Wang H, 2024, The Application of Computer Systems in Emergency Rescue Decision-support Systems. Applications of IC, 41(5): 266–267.

[15] Qin Z, 2024, Research on the Assembly Quality Control System for Automobile Engines Based on Big Data Analysis. Farm Machinery Using & Maintenance, (6): 24–26.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Face Expression Recognition on Uncertainty-Based Robust Sample Selection Strategy

Yuqi Wang, Wei Jiang*

School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, China

***Corresponding author:** Wei Jiang, jiangwei@ncwu.edu.cn

Abstract: In the task of Facial Expression Recognition (FER), data uncertainty has been a critical factor affecting performance, typically arising from the ambiguity of facial expressions, low-quality images, and the subjectivity of annotators. Tracking the training history reveals that misclassified samples often exhibit high confidence and excessive uncertainty in the early stages of training. To address this issue, we propose an uncertainty-based robust sample selection strategy, which combines confidence error with RandAugment to improve image diversity, effectively reducing overfitting caused by uncertain samples during deep learning model training. To validate the effectiveness of the proposed method, extensive experiments were conducted on FER public benchmarks. The accuracy obtained were 89.08% on RAF-DB, 63.12% on AffectNet, and 88.73% on FERPlus.

Keywords: Facial expression recognition; Uncertainty; Sample selection strategy.

Online publication: April 3, 2025

1. Introduction

Emotions are expressed through various non-verbal means, with facial expressions being the most intuitive and widespread form of emotional communication. In recent years, numerous deep learning-based facial expression recognition (FER) methods have been proposed with promising performance. However, there are annotation discrepancies between different labelers for similar expressions, which leads to uncertainty [1-4]. Typically, uncertainty can lead to insufficient feature learning by the model, particularly when dealing with complex expressions where key facial features are not adequately captured. Studies have shown that in the final classification results of a facial expression recognition task trained with the ResNet-18 model, misclassified samples are often uncertain ambiguous expression images, accompanied by high confidence scores. The model gradually memorizes the facial features of noisy samples and overfits as the training epochs progress, leading to overly confident misclassifications.

To address the aforementioned challenges of uncertainty, this paper proposes a sample selection framework

based on the ResNet-18 model. The framework utilizes an uncertainty determination mechanism to select samples that positively contribute to model training. Specifically, the training data is first divided into a clean set and an uncertain set. To enhance training stability, the sample exclusion strategy prevents uncertain samples from participating in the training process and temporarily reserves them for reassessment in the next epoch. Additionally, this framework incorporates RandAugment [5] during data preprocessing, utilizing a richer feature space to improve the model's ability to distinguish between different facial expression categories, thereby increasing intra-class compactness and inter-class separability. The main contributions of this paper are as follows:

We propose an efficient sample exclusion framework to mitigate the effect of uncertain samples on model training, allowing the model to learn clean facial expression features. We conduct comprehensive experiments on both real-world and synthetic noise Facial Expression Recognition (FER) datasets to validate the strong robustness of our method.

2. Related work

2.1. Facial expression recognition

The uncertainty stemming from inter-class similarity and annotation ambiguity in facial expressions makes it challenging to accurately recognize emotions. Recent methods [1-4] leverage learnable uncertainty to enhance model robustness. SCN [1] suggests quantifying the uncertainty of each sample and ranking them to suppress uncertain samples through re-labeling. DMUE [2] introduces a multi-branch learning framework to explore latent distributions and decay the weights of uncertain samples using confidence scores. RUL [3] quantifies the relative uncertainty of images and uses it for the weight of facial features mix. IPA2LT [4] is the first to address label inconsistency in FER datasets, proposing to assign pseudo-labels to each image to obtain a latent distribution.

2.2. Learning with noisy labels

Learning with noisy labels is currently mainly classified into two categories: adjusting the loss function [6,7] and clean sample selection methods [8,9]. Adjusting the loss function [6,7] focuses on estimating the noise transition matrix, inferring the probability of different class data points being corrupted using clean samples, and modifying the loss of each sample to minimize the adverse effects of label noise. Sample selection research [8,9] currently focuses on methods based on probability distribution. These methods typically utilize confidence to address noisy labels, where confidence reflects the model's certainty about each sample prediction.

3. Proposed method

3.1. Overview of proposed method

This paper designs a training framework based on confidence error to prevent deep networks from overfitting uncertain facial images. We observe that the model tends to memorize noisy facial features during training, which leads to a gradual increase in the classification accuracy of uncertain labels after an early fluctuation. Inspired by [9], we compute the confidence error of training samples. To prevent the model from memorizing uncertain facial images, samples with confidence errors above the fixed threshold are excluded from training and retained for re-evaluation in the next epoch. In addition, we employ RandAugment during the data preprocessing stage to expand the feature space of facial expressions.

3.2. Data augmentation module

To mitigate the deterioration of training caused by uncertainty, we introduce RandAugment [5]. RandAugment expands the training dataset by randomly selecting transformations for each sample in a mini-batch and increases the diversity of images. In FER tasks, facial expression images may be blurry or occluded and the model may fail to learn useful knowledge for distinguishing uncertain samples. Therefore, we choose to use RandAugment to introduce perturbations to simulate the real-world data distribution, making the model sensitive and adaptive to uncertainty.

3.3. Confidence error

Confidence error [9] is defined as the difference between the predicted label and the original label of a sample, serving as a sieving strategy to distinguish clean samples from noisy samples. During the training of CNN for Facial Expression Recognition (FER), the neural network model $F(x_i, \theta) \in R^{m \times k}$ is a k-class classifier with trainable parameters θ. The probability computed by the softmax activation function for each class represents the confidence of the sample in that class. Assuming a classification task on a training dataset $D = \{(x_i, y_i) | x_i \in X, y_i \in Y\}_{i=1}^{n}$, where n is the number of samples, X and Y denote the training sample and label spaces respectively.

Given a set of facial expression images $D_1 = (x_i, \tilde{y}_i)$, we apply RandAugment to the input images and then feed them into the neural network model to obtain the predicted probability for each class, denoted as $P = F(x_i, \theta)$. The model confidence $P^{(l)} = F(x_i, \theta)^{(l)}$ is generated through the original labels $l \in \{1, \ldots, k\}$. Subsequently, the predicted confidence is considered as the current maximum probability: $P^{arg} = \text{argmax}\big(F(x_i, \theta)\big)$. From these two confidences, we can extract the confidence error of facial expression images:

$$E_P(D_1) = P^{arg} - P^{(l)} \quad (1)$$

3.4. Uncertainty judgment module

To select uncertain data samples, this paper employs a proven effective sample exclusion method that uses cross-entropy to exclude samples exceeding a fixed threshold from training. Furthermore, these excluded samples can be re-evaluated instead of being deleted in the next epoch. For the multi-class facial expression task, we denote our loss as Sample Exclusion Loss (L_{SEL}), formulated as follows:

$$L_{\text{SEL}} = \sum_{b=1}^{m} \mathbf{1}\,(E_P(D_1) \leq \delta) H(P, l) \quad (2)$$

Where $H(P,l)$is defined as the cross-entropy of the probability distribution , with as the fixed threshold. We calculate the loss for clean samples based on the formula. Importantly, samples exceeding the threshold do not participate in the current model training round, but their confidence error will be recalculated for judgment in the next epoch.

4. Experiments

4.1. Implementation details

This experiment utilizes the RAF-DB [10], FERPlus [11] and AffectNet [12] datasets, implemented with the PyTorch framework and executed on three GTX 1080 Ti GPUs. By default, we employed a ResNet-18 [13] pre-trained on MS-Celeb-1M [14] and trained it end-to-end as the backbone network. Facial images were resized to 224×224 pixels for fair comparison. To improve the effectiveness of the sieving strategy, we applied horizontal flipping

with a probability of 0.5, Random Erasing [15], and RandAugment [5] to the images. The batch size was set to 1024 during training. The initial learning rate was 0.001 and training ended at epoch 60. Additionally, we used the Adam optimizer with a weight decay of 0.0001 to expedite convergence. To decrease the learning rate after each epoch, the ExponentialLR learning rate scheduler was set with a gamma of 0.9.

4.2. Evaluation on noise for datasets

To quantitatively analyze noisy labels, we explore the robustness of our method across three parameters as shown in **Table 1**.

Table 1. Evaluation of the sample selection framework on synthetic uncertainty FER datasets

Noise (%)	Method	RAF-DB (%)	AffectNet-7 (%)
10	SCN [1]	82.14	58.56
	DMUE [2]	83.19	61.21
	RUL [3]	86.22	60.54
	Ours	88.17	61.24
20	SCN [1]	79.79	57.21
	DMUE [2]	80.31	58.66
	RUL [3]	84.34	58.36
	Ours	87.09	60.84
30	SCN [1]	77.46	54.84
	DMUE [2]	79.41	56.88
	RUL [3]	82.06	56.65
	Ours	85.10	59.63

Levels on the RAF-DB and AffectNet datasets. Specifically, we select 10%, 20%, and 30% of the training data in each category and randomly change their labels to assign them labels from other categories. For a fair comparison, we choose ResNet-18 as the backbone network and compare its performance with other state-of-the-art FER uncertainty quantification methods based on ResNet-18. As shown in **Table 1**, our method outperforms superior performance compared to SCN and other state-of-the-art FER uncertainty methods.

5. Conclusions

This paper proposes a simple framework based on confidence error to prevent deep networks from overfitting uncertain facial images. The sample selection framework employs an uncertainty Judgment module to filter out samples above the fixed threshold and retain them for training, which prevents the model from overfitting uncertain facial expression images. Additionally, this paper introduces RandAugment to simulate real-world data distribution, making the model sensitive and adaptive to uncertainty. Experimental results on multi-dimensional synthetic and real-world FER datasets demonstrate the robustness of this framework. Furthermore, compared to other uncertainty training methods, the proposed sample selection framework achieves state-of-the-art performance.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Wang K, Peng X, Yang J, et al., 2020, Suppressing Uncertainties for Large-scale Facial Expression Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6897–6906.

[2] She J, Hu Y, Shi H, et al., 2021, Dive Into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6248–6257.

[3] Zhang Y, Wang C, Deng W, 2021, Relative Uncertainty Learning for Facial Expression Recognition. Advances in Neural Information Processing Systems, 34: 17616–17627.

[4] Zeng J, Shan S, Chen X, 2018, Facial Expression Recognition with Inconsistently Annotated Datasets. Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 222–237.

[5] Cubuk ED, Zoph B, Shlens J, et al., 2020, Randaugment: Practical Automated Data Augmentation with a Reduced Search Space. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 702–703.

[6] Yao Y, Liu T, Gong M, et al., 2021, Instance-dependent Label-noise Learning Under a Structural Causal Model. Advances in Neural Information Processing Systems, 34: 4409–4420.

[7] Yao Y, Liu T, Han B, et al., 2020, Dual t: Reducing Estimation Error for Transition Matrix in Label-noise Learning. Advances in Neural Information Processing Systems, 33: 7260–7271.

[8] Nguyen D, Mummadi C, Ngo T, et al., 2019, Self: Learning to Filter Noisy Labels with Self-ensembling. arXiv: 1910.01842. https://doi.org/10.48550/arXiv.1910.01842.

[9] Torkzadehmahani R, Nasirigerdeh R, Rueckert D, et al., 2022, Label Noise-robust Learning using a Confidence-based Sieving Strategy. arXiv: 2210.05330. https://doi.org/10.48550/arXiv.2210.05330.

[10] Li S, Deng W, Du J, 2017, Reliable Crowdsourcing and Deep Locality-preserving Learning for Expression Recognition in the Wild. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2852–2861.

[11] Barsoum E, Zhang C, Ferrer C, et al., 2016, Training Deep Networks for Facial Expression Recognition with Crowd-sourced Label Distribution. Proceedings of the Proceedings of the 18th ACM International Conference on Multimodal Interaction, 279–283.

[12] Mollahosseini A, Hasani B, Mahoor M, 2017, Affectnet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Transactions on Affective Computing, 10(1): 18–31.

[13] He K, Zhang X, Ren S, et al., 2016, Deep Residual Learning for Image Recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 770–778.

[14] Guo Y, Zhang L, Hu Y, et al., 2016, Ms-celeb-1m: A dataset and Benchmark for Large-scale Face Recognition. Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14, 2016, Springer International Publishing, 87–102.

[15] Zhong Z, Zheng L, Kang G, et al., 2020, Random Erasing Data Augmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07): 13001–13008.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Study on Preparation and Forming of TiC Steel-Bonded Cemented Carbide Paste for Direct Writing Printing

Zhi Wang[1,2]*, Bing Xu[1], Jiawei Yuan[1], Xiang Jie Cheng[1], Ting Pu Yue[1], Mei Ye Zhang[1], Sheng Hui Zhou[1]

[1]School of Materials Science and Engineering, Shijiazhuang Tiedao University, Shijiazhuang 050043, China
[2]Engineering Research Center of Metamaterials and MicroDevices of Hebei Province, Shijiazhuang 050043, China

***Corresponding author:** Zhi Wang, zhiwang_stdu@126.com

Abstract: TiC steel-bound cemented carbide body was prepared by direct writing printing. The effects of powder content (89.28, 89.49, 89.69, 89.88, and 90.07 wt%) and dispersant content (0.017, 0.034, 0.051, and 0.068 wt%) on the slurry and printing body were studied. The experimental results show that with the increase of powder content, the viscosity of the slurry gradually increases, the settlement rate gradually decreases, and the size and linewidth of the blank body gradually decreases. When the powder content is 89.69 wt%, the sedimentation stability and extrusion stability of the slurry are the best, and the density of the blank body is the highest, which is 3.8275 g/cm^3, which is suitable for direct writing printing. The addition of dispersant reduced the viscosity of the slurry; With the increase of dispersant content, the surface line width and size of the printed body gradually increased. When the dispersant content is 0.034 wt%, the extrusion stability of the slurry is the best, and the density of the body is the highest, which is 3.8901 g/cm^3.

Keywords: Direct writing printing; Steel-bonded cemented carbide; Powder content; Dispersant

Online publication: April 10, 2025

1. Introduction

TiC base steel-bonded cemented carbide is a kind of metal matrix composite material with TiC as the main ceramic phase and Fe as the binder, which combines the excellent properties of metal and ceramics, with high hardness, high melting point, good wear resistance, chemical inertness, and thermal stability [1]. Widely used in mold, tool, oil drilling, and other fields [2].

At present, the preparation process of steel-bonded cemented carbide materials is still based on the traditional powder metallurgy process are pressing and sintering, and the high hardness and high wear resistance lead to the mechanical processing of steel-bonded cemented carbide materials is very difficult. This process makes the preparation process complex, difficult to prepare special shape parts, and also the loss of

raw materials, increasing the production cost of parts. The preparation of some parts will also be limited by processing technology and tools. Therefore, the steel-cemented carbide product development process is simple and cost-effective. As a result, low-cost near-net forming technology has become the focus of the field of steel-cemented carbide research content.

3D printing is based on the "discrete-stacking" principle of a forming technology, also known as additive manufacturing, which is a non-traditional advanced processing and manufacturing method, no mold, free interface, in principle can achieve the forming of any complex structure [3]. The near-net forming characteristics of 3D printing also reduce the waste [4] of materials in the subsequent processing process. At present, the most widely used types of 3D printing are light curing (SLA) [5], melt deposition molding (FDM) [6,7], direct writing printing (DIW) [8,9], and laser selection sintering molding (SLS) [10-12]. Among them, direct writing printing is a kind of non-die-forming technology that extrudes the rheological paste through a specific extrusion device. It has the advantages of simple equipment, low investment, and fine and complex three-dimensional structures that can be prepared at room temperature [13]. At present, there is research on the preparation of TiC steel-bonded cemented carbide by direct writing printing. Therefore, this paper takes the direct writing 3D printer as the extrusion tool to study the influence of powder and dispersant content on slurry viscosity, settlement stability, extrusion stability, and printing billet and clarify its mechanism, aiming to provide a theoretical basis for the preparation of TiC steel-bonded cemented carbide and related materials.

2. Experiment

2.1. Experimental materials

High-purity titanium carbide powder with an average particle size of 5 μm (Hebei Yanyu Metal Materials Co., LTD.), spherical iron powder with an average particle size of 1 μm (Shanghai Xiangtian Nanomaterials Co., LTD.), alloy powder with a particle size of 300 mesh (Hebei Yanyu Metal Materials Co., LTD.), carbon black with a particle size of 500 nm and 5 μm Mo powder are used as raw materials. Oleic acid is used as a dispersant.

2.2. Preparation of slurry

The raw materials were weighed according to the mixture ratio of powder as shown in **Table 1**, and then put into the roller ball mill for ball grinding. The ball mill speed was 240 rpm/min and the ball milling time was four hours to obtain the steel-bonded cemented carbide mixture powder.

Table 1. Ratio of mixed powder (wt%)

Ingredients	High-purity TiC powder	Mo	C	Alloy powder	Spherical iron powder
Content	35.00	2.10	0.5	17.1	45.3

Mix the solvent and binder in a certain proportion to get the premix, then weight X (89.28, 89.49, 89.69, 89.88, 90.07 wt%, X = the mass of the mixed powder)/(the mass of the mixed liquid + the mass of the mixed powder), add the mixed powder in batches to the premix and stir well with a mixer. The oleic acid of different content Y (0.017, 0.034, 0.051, 0.068, 0.085 wt%, Y = the mass of oleic acid/the mass of the mixed powder) is

added to the slurry, and the slurry can be printed by stirring evenly with the blender.

2.3. Direct write print molding

At room temperature, select a needle with an inner diameter of 0.33 mm, pressure is 0.05 MPa, printing speed is 20 mm/s, layer thickness is 0.475 mm, line width is 0.5 mm, and after stacking layers, the blank body of direct writing is obtained, and the size of the blank body is set to 30 mm × 10 mm × 2.85 mm.

2.4. Characterization method

The viscosity of the slurry was characterized by NDJ-8S digital display viscometer. The parameters of the viscometer were rotor 4 and speed 6 rpm. A 3D printer was used to record the extrusion quality of the slurry every 10 seconds and calculate the extrusion deviation of the slurry, the formula is: where is the average mass of the slurry extrusion per unit time, is the quality of the slurry extrusion per unit time, the greater the extrusion deviation, the worse the extrusion stability of the slurry as shown below:

$$\sqrt{\frac{1}{n-i}\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (1)$$

Static settling experiment is used to measure the settling stability of the slurry, the settling rate is, where H_0 is the height of decline over time, and H is the initial height. The shape of the blank was characterized by SEM. The density of the billet was characterized by the Archimedean drainage method.

$$(H_0 / H) \times 100\% \quad (2)$$

3. Results and discussion

3.1. Rheological properties of slurry

3.1.1. Influence of powder content on rheological properties of slurry

(1) The viscosity of the slurry with different powder content, the viscosity of the slurry with powder content of 89.28, 89.49, 89.69, 89.88, and 90.07 wt% is 17382.2, 19614.7, 20275.23, 22522.37 and 24764.26 mPa·s, respectively. That is, with the increase of powder content, the viscosity of the slurry gradually increases. When the powder content increases, the number of powder particles in the slurry increases, resulting in an increase in the number of collisions between powder particles, and the distance between particles decreases. When the slurry flows, the resistance it receives will increase with the increase in the powder content. The larger the powder content, the larger the specific surface area of the powder, the larger the contact area between the powder and the premix, the greater the friction resistance generated by the movement between the powder and the premix, resulting in the larger shear viscosity of the slurry [14].

(2) The sedimentation rate of the slurry with different powder content, and the sedimentation rate of the slurry with different powder content gradually increases with the increase of time, which is caused by the gravity of the slurry itself (**Figure 1**). With the increase of powder content, the settlement rate is reduced, because the powder content increases, the Van der Waals force between particles gradually increases, resulting in a lower settlement rate of the slurry with high powder content, that is, the greater the viscosity of the slurry, the lower the settlement rate [15].

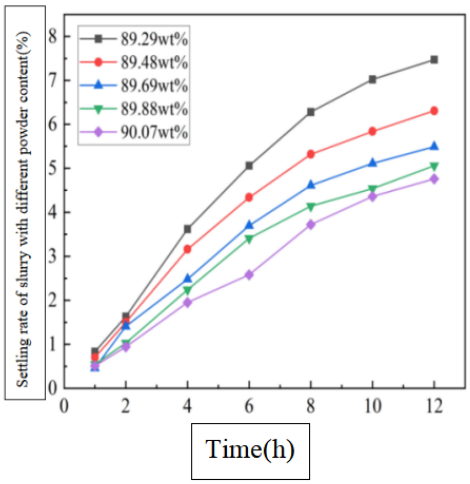


Figure 1. The sedimentation rate of the slurry with different powder content against time

(3) Extrusion standard deviation of slurry with different powder content is shown in **Figure 2**. From the figure, with the increase of powder content, the extrusion standard deviation decreases first and then increases. When the powder content is 89.69 wt%, the extrusion standard deviation is the smallest, and the extrusion standard deviation is 6.29%. When the powder content is lower than 89.69 wt%, the excess premix exists in the slurry, the settlement rate increases and the premix may be unevenly distributed during the extrusion process, resulting in a high extrusion standard deviation. When the powder content is 89.69 wt%, the powder dispersion is more uniform, so the extrusion deviation of the slurry is minimal when the powder content is 89.69 wt%. When the powder content is higher than 89.69 wt%, the powder is present in the slurry, and the powder may flocculate or agglomerate, resulting in a larger extrusion standard deviation. In short, when the powder content is too high or too low, the powder or premix distribution is not uniform, and the extrusion stability of the slurry is poor.

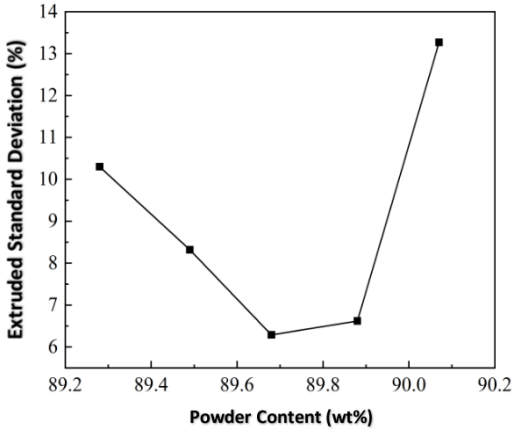


Figure 2. Extrusion standard deviation of slurry with different powder content

3.1.2. Influence of dispersant on rheology of slurry

(1) To obtain a uniform slurry, it is necessary to have a good dispersion of the mixed powder in the slurry. However, when a variety of powders are mixed, due to the interaction between the particles, there is

inevitably a tendency of agglomeration between the particles. A common way to reduce the viscosity of the slurry is to add a dispersant to the slurry because the raw material used in this paper does not have an electric charge, therefore, this paper uses a non-ionic dispersant oleic acid as an additive. Add different amounts of oleic acid to the viscosity of the slurry. The addition of oleic acid reduced the viscosity of the slurry, and the viscosity of the slurry gradually decreased with the increase of oleic acid content, and the viscosity was 20179.45, 20054.39, 19284.50, 18952.17 and 18212.33 mPa·s, respectively shown as **Figure 3**. This is due to the hydrophilic carboxyl (COOH) and lipophilic groups in its molecular structure. The hydrophilic end of oleic acid molecules will be attached to the surface of powder particles, resulting in repulsion between particles and separation from each other. The particles are gradually wetted and become lipophilic, which can be well suspended in organic solvents.

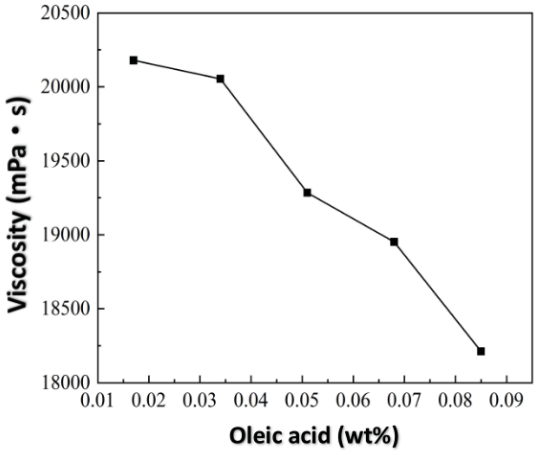


Figure 3. Viscosity of the slurry against oleic acid content

(2) Add different content of oleic acid slurry sedimentation rate. With the increase of time, the sedimentation rate of the slurry with different oleic acid content gradually increases, which is due to the gravitational action of the slurry itself. With the increase of oleic acid content, the sedimentation rate of the slurry decreases first and then increases. When 0.034 wt% oleic acid is added, the sedimentation rate of the slurry is the lowest, which is 4.69%. When oleic acid is not added or the oleic acid content is lower than 0.034 wt% (0.017 wt%), the surface of the particles is not completely wet, and the particles cannot be effectively modified, resulting in the increased sedimentation rate. When the oleic acid content is properly increased, it is conducive to improving the coverage rate of the particle surface, easy to form a network structure inside the slurry, and forms an organic protective film on the particle surface to prevent particles from colliding with each other, and the adsorption of the particle surface reaches saturation. The concentration of oleic acid is too high, the free dispersant molecules will exist in the skeleton between the particles, the resistance between the particles of the powder is reduced, and the interaction force is reduced, resulting in an increase in the settlement rate.

(3) When the powder content is 89.69 wt%, the standard deviation of extrusion of oleic acid slurry with different content is obtained. With the increase of oleic acid content, the extrusion standard deviation of the slurry first decreased and then increased, while the extrusion standard deviation of the slurry without oleic acid was 6.29%. The standard deviation of 0.034 wt% oleic acid is the smallest, which

is 5.82%, and the extrusion stability of the slurry is the best. When 0.017 wt% oleic acid was added, the extrusion standard deviation of the slurry increased, which may be because the addition of a small amount of oleic acid made the slurry incomplete dispersion, and the interaction between particles led to the instability of the slurry extrusion. When oleic acid is added greater than 0.034 wt%, the fluidity of the slurry becomes larger, and excessive and free dispersant molecules will exist between the particles, resulting in uneven dispersion of the slurry at the top and bottom, and poor extrusion stability during the extrusion process.

3.2. Influence of powder content on printing billet

3.2.1. The influence of powder content on the surface morphology of the printed billet

The surface morphology of the slurry-forming billet with different powder content is shown in **Figure 4**. As can be seen from the figure, the linewidths of the body corresponding to different powder contents (89.28, 89.49, 89.69, 89.88, 90.07 wt%) are 0.525, 0.515, 0.505, 0.49 and 0.48 mm, respectively. With the increase of powder content, the linewidths of the body surface gradually decrease, and the surface lines of the body become clearer. This is due to increase of powder content, the viscosity of the slurry increases, and the extrusion amount of the slurry decreases. When the printing speed is the same, the slurry stroke is the same, and the line width of the billet decreases gradually. With the increase of powder content, the fluidity of the slurry is poor, the "groove" formed in the printing process is difficult to fill, and the lines on the surface of the billet are clearer. When the powder content is 89.69 wt%, the line width is 0.505 mm, the error between the line width and the set value is minimal, and the filling property of the blank body is better at this time.

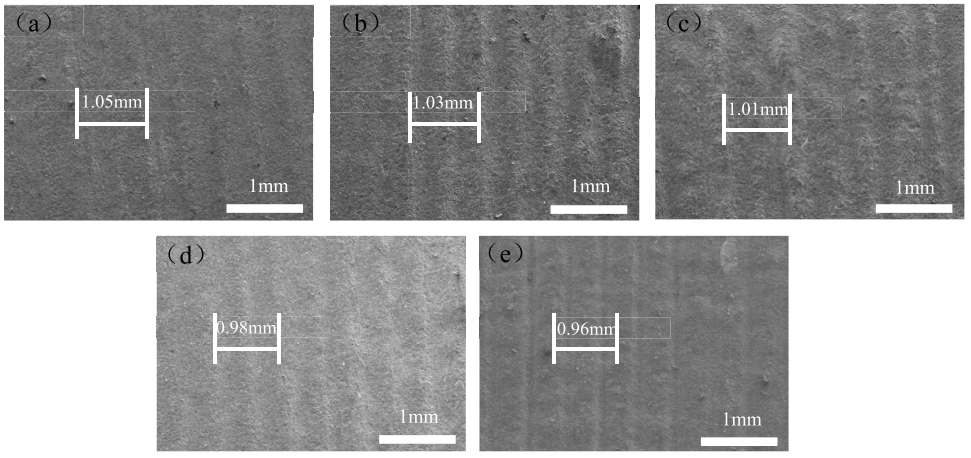


Figure 4. The apparent morphology of the paste printing body with different powder content. (a) 89.28 wt%; (b) 89.49 wt%; (c) 89.69 wt%; (d) 89.88 wt%; (e) 90.07 wt%

According to the figure, the side topography of the paste printing body with different powder content. It can be seen from the figure that with the increase in powder content, the side morphology of the blank body changes from uneven to flat to convex (concave). This is because the slurry with low powder content (89.28, 89.49 wt%) has low viscosity, large fluidity, and poor extrusion stability of the slurry. When the slurry is extruded for a long time, the slurry flows more to the side of the billet, which looks "high," and conversely, it is "low," resulting in uneven side morphology of the billet. And high powder content (89.88, 90.07 wt%) of

the slurry viscosity is larger, when the extrusion amount is more, the slurry fluidity is poor, forming a pile-like convex, and vice versa to form a depression.

3.2.2. Influence of powder content on the fracture morphology and density of print billet

Figures 5 and **6** show the section morphology and density of the printed billet under different powder content. As can be seen from the figure, when the powder content is low (89.28 wt% and 89.49 wt%), the blank body has large pores (**Figure 5a, 5b**). This is because on the one hand, when the powder content is low, the spacing between particles is large, and more pores are formed after the liquid phase in the premix evaporates. On the other hand, when the powder content is low, the slurry viscosity is low, the fluidity is large, the layer spacing is the same, the filling between layers is poor, and the billet has large pores. When the powder content is high (89.88 wt% and 90.07 wt%), there are more pores in the blank body (**Figure 5d, 5e**). It may be because the spacing between particles decreases when the powder content is high, the slurry may tend to agglomerate or flocculate, and the slurry is not evenly dispersed, so there are more pores. The body with a powder content of 89.69 wt% had fewer pores (**Figure 5c**). As can be seen from **Figure 6**, the density of the body is the highest when the powder content is 89.69 wt%, and the density is 3.8275 g/cm^3, which corresponds to **Figure 5**.

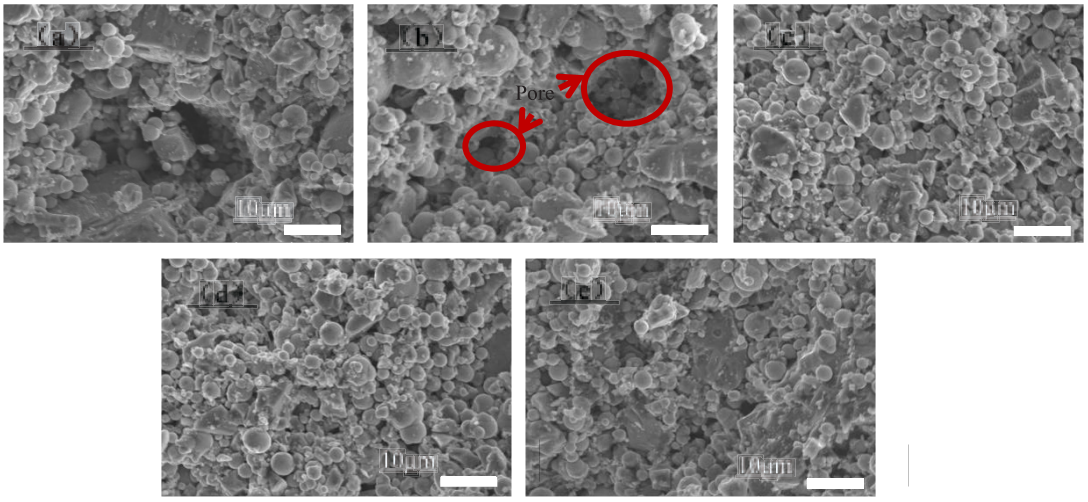


Figure 5. Section morphology of printed billet under different powder content. (a) 89.28 wt%; (b) 89.49 wt%; (c) 89.69 wt%; (d) 89.88 wt%; (e) 90.07 wt%

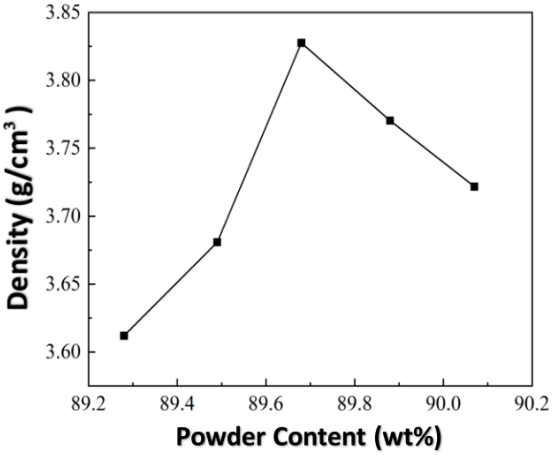


Figure 6. Density of printed billet under different powder content

3.2.3. Influence of oleic acid content on the surface morphology of printed billet

With the increase of oleic acid content, the body length, width, and height of the blank gradually increase, and the change of the size of the paste with different oleic acid content corresponds to the viscosity of the slurry with the same oleic acid content. The lower the viscosity of the slurry, the more slurry will be extruder per unit time. When the printing speed is the same and the stroke is the same, the body size will gradually increase. **Figure 2** shows the side topography of the paste printing billet with different oleic acid content. As can be seen from the figure, with the increase of oleic acid content, the profile of the sideline first changes from uneven to flat and then becomes convex. This is because when the oleic acid content is low (0.017 wt%) or high (0.051, 0.068, 0.085 wt%), the extrusion deviation of the slurry is large and the fluidity is high, resulting in the uneven profile of the printing billet. When the oleic acid content is 0.034 wt%, the slurry printing molding effect is the best.

3.2.4. Influence of oleic acid content on the fracture morphology and density of printing blank

Section morphology and density of different content oleic acid printing billet. With the increase of oleic acid content, the porosity of the body decreases first and then increases, and the porosity is the least when the oleic acid content is 0.034 wt%. With the increase of oleic acid content, the body density first increased and then decreased. When the oleic acid content is 0.034 wt%, the density is the highest, and the density is 3.8901 g/cm^3. This is because the addition of oleic acid overcomes the van der Waals force between particles, making the slurry disperse more evenly and have a higher density. When the oleic acid content is low (0 wt%, 0.017 wt%), oleic acid cannot disperse the powder sufficiently, resulting in its low density. When the oleic acid content is high (0.051 wt%, 0.068 wt%, 0.085 wt%), due to the larger fluidity of the slurry, the poor filling between layers, excessive dispersant will exist between particles, and the spacing between particles of the powder will be large, resulting in its low density shown in **Figure 7.**

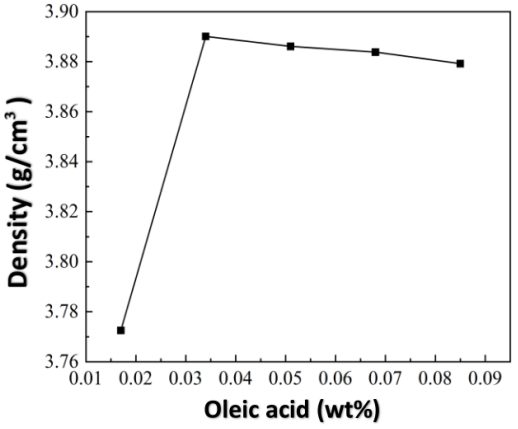


Figure 7. The density of printed billets with different oleic acid content

4. Conclusion

In this paper, the pulp with different powder content and oleic acid content was prepared and its rheological property was tested. The shape, size, and density of the blank printed with different slurries were studied, and the following conclusions were drawn:

(1) With the increase of powder content, the viscosity of the slurry increases and the sedimentation rate decreases. When the powder content is 89.69 wt%, the sedimentation rate and standard extrusion deviation are 5.49% and 6.29% respectively, and the rheological property of the slurry is the best. The line width and size of the printed billet gradually decrease the side topography from uneven to flat to convex (concave), and the density of the billet is the highest, 3.8275 g/cm^3.

(2) With the increase of oleic acid content, the viscosity of the slurry showed a decreasing trend. When 0.034 wt% oleic acid was added, the sedimentation rate and standard extrusion deviation were 4.69% and 5.82%, respectively. The rheological property of the slurry was the best. The line width and size of the printing body gradually increased, from uneven to flat and then convex, the density of the body is the highest, the density is 3.8901 g/cm^3, and the forming effect of the body is the best.

Funding

(1) "TiC Steel Composite 3D Gel Printing Molding and its Densification Mechanism" Hebei Province Natural Science Foundation funded project – Youth Science Fund Project (E2021210094)

(2) "TiC Steel-bonded Cemented Carbide 3D Gel Printing Forming and Densification Mechanism" Sichuan Powder Metallurgy Engineering – Technology Research Center open project (SC-FMYJ2020-07)

(3) "Design Preparation of MAC Composite Powder and Research on the Mechanism of Strengthening Steel-bonded Cemented Carbide" Hebei Provincial Department of Education Natural Science – Youth Fund project (QN2024021)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Zheng J, Zhao Z, Liu X, et al., 2018, Effect of Nano-composite Inhibitor on the Composition and Microstructure of WC-TiC-Co Cemented Carbide Prepared by Microwave Method. Powder Metallurgy Industry, 28(04): 21–25.

[2] Li G, Chen W, Sun L, et al., 2017, Microstructure and Properties of TiC Steel-bonded Carbide Used Fe/Mo Pre-alloyed Powder as Binder. Materials Science Forum, 898: 1459–1467.

[3] Chen Z, Li Z, Liu C, et al., 2018, 3D Printing of Ceramics: A review. Journal of the European Ceramic Society, 39(4): 661–687.

[4] Guo Y, 2008, Metal Cutting Liquid Pollution and Green Cutting Technology. Sci-Tech Information Development and Economy, 18(36): 84–85.

[5] Alexander S, Egmont R, Tinus S, et al., 2023, SLA-printed K-band Waveguide Components using Tollens' Reaction Silver Plating. IEEE Transactions on Components, 13(2): 230–239.

[6] Mustafa K, Volkan_K, Tugce T, et al., 2023, Three-dimensional Printability of Bismuth Alloys with Low Melting Temperatures. Journal of Manufacturing Processes, 92: 238–246.

[7] Ferro P, Fabrizi A, Berto F, et al., 2023, Creating IN718-high Carbon Steel Bimetallic Parts by Fused Deposition Modeling and Sintering. Procedia Structural Integrity, 47: 535–544.

[8] Giovanni P, Chiara G, Paolo C, et al., 2016, Direct Ink Writing of Micrometric SiOC Ceramic Structures using a

Preceramic Polymer. Journal of the European Ceramic Society, 36: 1589–1894.

[9] Liu K, Lei Q, Zhang Y, et al., 2023, Additive Manufacturing of Continuous Carbon Fiber-reinforced Silicon Carbide Ceramic Composites. International Journal of Applied Ceramic Technology, 20(6): 3455–3469.

[10] Rios J, Zambrano R, Taborda J, et al., 2023, Process Parameters Effect and Porosity Reduction on AlSi10Mg Parts Manufactured by Selective Laser Melting. The International Journal of Advanced Manufacturing Technology, 129(7): 3341–3351.

[11] Cao J, Wang Pei, Liu Z, et al., 2022, Research Progress on Powder-based Laser Additive Manufacturing Technology of Ceramics. Journal of Inorganic Materials, 37(3): 241–254.

[12] Zhang Y, Zhang Y, Chen Y, et al., 2022, Effect of Microwave on Mechanical Properties of Laser-sintered Carbon Nanotube-polymer Composites. Materials Science and Technology, 38(15): 1239–1243.

[13] Liu Y, Cheng Y, Ma D, et al., 2022, Continuous Carbon Fiber Reinforced ZrB2-SiC Composites Fabricated by Direct Ink Writing Combined with Low-temperature Hot-pressing. Journal of the European Ceramic Society, 42(9): 3699–3707.

[14] Zhang X, Guo Z, Chen C, et al., 2018, Additive Manufacturing of WC-20Co Components by 3D Gel-printing. International Journal of Refractory Metals and Hard Materials, 70: 215–223.

[15] Ruoyu C, Adam B, Joshua R, et al., 2022, Additive Manufacturing of Complexly Shaped SiC with High Density via Extrusion-based Technique-Effects of Slurry Thixotropic Behavior and 3D Printing Parameters. Ceramics International, 48(19): 28444–28454.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

A Collaborative Protection Mechanism for System-on-Chip Functional Safety and Information Security in Autonomous Driving

Zhongyi Xu*, Lei Xin, Zhongbai Huang, Deguang Wei

Shandong Vocational and Technical University of International Studies, Rizhao 276800, Shandong, China

***Corresponding author:** Zhongyi Xu, shuanqiangbr8@163.com

Abstract: This article takes the current autonomous driving technology as the research background and studies the collaborative protection mechanism between its system-on-chip (SoC) functional safety and information security. It includes an introduction to the functions and information security of autonomous driving SoCs, as well as the main design strategies for the collaborative prevention and control mechanism of SoC functional safety and information security in autonomous driving. The research shows that in the field of autonomous driving, there is a close connection between the functional safety of SoCs and their information security. In the design of the safety collaborative protection mechanism, the overall collaborative protection architecture, SoC functional safety protection mechanism, information security protection mechanism, the workflow of the collaborative protection mechanism, and its strategies are all key design elements. It is hoped that this analysis can provide some references for the collaborative protection of SoC functional safety and information security in the field of autonomous driving, so as to improve the safety of autonomous driving technology and meet its practical application requirements.

Keywords: Autonomous driving; SoC functional safety; Information security; Collaborative protection mechanism; Collaborative protection architecture

Online publication: April 3, 2025

1. Introduction

In the current field of autonomous driving, the functions of the system-on-chip (SoC) are a core component, and its security directly affects the safe use of autonomous driving technology. In addition, information security also directly impacts the safety of autonomous driving. Therefore, in the practical application research of modern autonomous driving technology, the collaborative protection mechanism between SoC functional safety and information security has become the focus of attention for researchers in this field. In specific research, researchers need to clarify the functional safety of autonomous driving SoCs and their information security and

design a collaborative protection mechanism with reasonable strategies to achieve the collaborative guarantee of the security of both, providing strong support for the good application and development of autonomous driving technology.

2. Introduction to the functions and information security of autonomous driving SoCs

2.1. Functions of autonomous driving SoCs

System-on-chip is a core integration platform in the modern autonomous driving field and an integrated platform for various functions and capabilities in autonomous vehicles. In the current autonomous driving field, the main components and functions of SoCs include the following aspects. First is the computing and processing function, which involves functions such as sensor data processing, complex algorithm operations, multitasking processing, and scheduling in autonomous driving. Second is the communication and interface function, including sensor interfaces, vehicle bus communication, and external communication interfaces in autonomous driving. Third is the safety and reliability function, covering hardware safety mechanisms, fault detection and tolerance, security authentication, and encryption in autonomous driving. Fourth is the graphics processing and display function, including high-definition image rendering and 3D environment modeling display in autonomous driving. Fifth is the power management function, which includes power consumption control, power management, and power distribution in autonomous driving. In practical applications, secure SoC functions can strongly support the sensor data analysis and obstacle position judgment of autonomous vehicles, ensuring the timely issuance of braking, evasion, and other commands, and providing a good guarantee for the safety of autonomous driving.

2.2. Introduction to information security in autonomous driving

In the field of autonomous driving, information security means that all kinds of information and systems during the autonomous driving process of vehicles are not subject to security attacks, and there is no risk of data tampering, loss, or leakage. Only in this way can it guarantee the safe driving of autonomous vehicles and effectively prevent the leakage of important or private information of passengers. In the current autonomous driving field, the importance of information security is mainly reflected in the following aspects. First, the autonomous driving system highly depends on sensors, electronic control units, and software algorithms. If the information in these systems is lost, tampered with, or interfered with, the autonomous vehicle is likely to get out of control, and in severe cases, even lead to safety accidents [1]. Second, autonomous vehicles often collect a large amount of user data during operation, including driving habits and location information. If the security of this information cannot be effectively guaranteed, important or private user information is likely to be leaked, causing unnecessary disturbances to their normal work and life and even economic losses. Third, in the practical application of autonomous driving technology, if information security problems occur frequently, users' trust in this technology will gradually decrease, which will pose many obstacles to the subsequent application and development of this technology and the automation transformation and development of the automotive driving field.

2.3. Correlation between SoC functional safety and information security in autonomous driving

In the current autonomous driving field, SoC functional safety and information security are not independent of

each other but are interdependent and mutually influential. There is a close correlation between them, and both play a crucial role in the operation of the autonomous driving system. First, there is some redundant design in the SoC functions of autonomous driving, that is, some spare parts identical or similar to the operating hardware are equipped. When a component fails, the spare component can promptly detect the fault and replace it to continue normal operation. In this case, if a piece of hardware is hacked and the data or permissions in it are tampered with, the redundant component can promptly discover the fault and replace it to continue normal operation [2,3]. This not only ensures the security of autonomous driving information but also reduces the probability of unnecessary safety accidents. Second, the information security protection technology in autonomous driving can effectively prevent various data from being attacked during storage or transmission, thus ensuring the integrity and security of data. In this mode, the realization of SoC functions in autonomous driving will be more secure, providing a good guarantee for the safe and stable operation of autonomous vehicles. It can be seen that in the field of autonomous driving, the connection between SoC functional safety and information security is very close. Only by ensuring the security of both SoC functions and information can the safety of autonomous vehicle driving be ensured. Based on this, the collaborative protection mechanism for SoC functional safety and information security has become a key focus and research area in the field of autonomous driving in recent years [4].

3. Main design strategies for the collaborative protection mechanism of SoC functional safety and information security in autonomous driving

3.1. Design of the overall collaborative protection architecture

In the collaborative protection mechanism for SoC functional safety and information security in autonomous driving, the rational design of the overall collaborative protection architecture is of great significance. According to the collaborative protection requirements for SoC safety and information security in the autonomous driving field, this architecture mainly includes the following components.

The first is the SoC functional safety protection module. This module is the key defense line to ensure the healthy and normal operation of SoC functions and the safe and stable operation of SoC hardware. The most critical safety protection links include fault detection, tolerance, and safe state switching. Through the collaborative operation of these links, the probability of SoC hardware failures can be effectively reduced, avoiding autonomous driving risks caused by SoC failures and making the operation state of autonomous vehicles safer and more reliable [5-7].

The second is the information security protection module. This module is the key defense line to prevent autonomous driving systems from being attacked by the network and ensure data security. The most critical security protection technologies include data encryption technology, identity authentication technology, access control technology, and intrusion detection technology. Through the coordinated cooperation of these technologies, various information security risks can be effectively prevented, guaranteeing the data security of the autonomous driving system, preventing the adverse effects of data security risks on system operation, and maintaining the safe and stable operation state of the overall system.

The third is the collaborative protection mechanism implementation module. This module is the key link to ensure the collaborative protection of SoC functions and information security in autonomous driving [8]. It can establish a close connection between SoC function and information security protection, enabling all links

to maintain coordinated cooperation and orderly operation. The three most critical collaborative operation processes are safety monitoring, collaborative decision-making, and response execution. Through the effective control and realization of these three processes, the orderly operation of the overall collaborative protection mechanism can be ensured, meeting the collaborative protection requirements for SoC functional safety and information security in autonomous driving.

3.2. Design of the SoC functional safety protection mechanism

As an important protection module to ensure the functional safety of autonomous driving SoCs, this protection mechanism mainly consists of three functions: fault detection, tolerance, and safe state switching.

In this safety protection mechanism, fault detection is the primary link. Its core operation goal is to detect various potential hardware abnormalities or software errors in the SoC system during operation in a timely and accurate manner. There are many detection methods for hardware abnormalities. First is the real-time monitoring technology of hardware circuits, that is, using sensors to monitor the current, voltage, temperature, etc. of hardware circuits in real time. If the values exceed the specified range, the system will immediately issue an alarm and activate the corresponding protection mechanism [9,10]. Second is the hardware redundancy-based fault detection and hardware switching. On the basis of the operation of the main hardware, pre-installed identical or similar hardware is added as a backup. Once the main hardware fails or malfunctions, the backup hardware will immediately replace the main hardware to continue maintaining the safe and stable operation of the overall system, providing support for subsequent fault detection, operation, and maintenance of the main hardware. There are also many detection methods for software errors. For example, through static analysis methods, the logic, semantics, and syntax of software code can be analyzed to check for potential problems, including memory leaks, null pointer references, etc. Through dynamic testing methods, the software operation scenarios can be simulated, and the software performance and function can be comprehensively tested in the simulated scenarios to verify whether they meet the design requirements. Through unit testing, integration testing, and system testing methods, the functions and operation effects of the software under various conditions can be comprehensively verified to discover software errors in a timely manner [11]. To ensure the system operation effect under software error conditions, the system also has a software redundancy design, that is, the backup system replaces the main system to run, reserving sufficient time for the repair of software errors in the main system.

The fault-tolerance protection mechanism is a key component of this module. Its main goal is to implement fault-tolerance processing on the detected faults through effective means to keep the key functions of the system operating normally. In addition to hardware and software redundancy designs, the system also needs to introduce error-correcting code technology supported by automation and intelligence. That is, the system data is verified through CRC (Cyclic Redundancy Check). If data errors are found, the error-correcting code can automatically correct the wrong data to ensure the accuracy of system operation parameters and prevent operation errors [12].

Safe state switching is the last line of defense for SoC functional safety. Its main goal is to automatically switch the SoC function to a safe state when a fault or abnormality occurs, ensuring the safe and stable operation of the overall system. In the specific design, this function needs to be realized with the help of intelligent fault assessment and automated control logic. Methods such as fault tree analysis are used to assess the SoC function faults or abnormalities, and the assessment results are fed back to the control logic. The control logic then

executes the corresponding safe state switching operations according to the assessment results. In this way, the rapid isolation of SoC system hardware and software faults can be achieved, and the operation state of the overall system can always be kept safe and stable.

3.3. Design of the information security protection mechanism

As the key defense mechanism for the information security of the SoC system, this module's main supporting technologies include data authentication and encryption technology, identity authentication technology, access control technology, and intrusion detection technology.

Data encryption technology is the basic technology for information security protection. Its main goal is to encrypt the data stored or transmitted in the system, converting readable plaintext into unreadable ciphertext. Only authorized users can decrypt and read the data. With the support of SSL (Secure Socket Layer) and TLS (Transport Layer Security), the data transmitted between autonomous vehicles and cloud servers, other vehicles, and infrastructure is encrypted. When data starts to be transmitted, SSL/TLS will encrypt it to prevent attacks during data transmission. With the support of AES (Advanced Encryption Standard), various data encryption algorithms can be adopted according to the actual situation to encrypt all sensitive or important data stored in the system, including passengers' location information and personal information, to prevent illegal use [13,14].

Identity authentication technology is a key means to judge the legitimacy of system access. With the help of this technology, information such as usernames, passwords, faces, or fingerprints can be identified. Only users who pass the identification are allowed to access the system, and those who fail cannot.

Permission control technology is a security protection technology superimposed on the basis of identity authentication technology. Its basic goal is to identify the access permissions of users to the system according to their identity information. For example, ordinary users can only view data within their permission range, while administrative users can manage system configurations, update software, or set security policies. Users cannot operate functions outside their permission range to prevent malicious access and tampering with system resources.

Intrusion detection technology is an important technology for the system to resist external attacks. Its basic goal is to monitor the system operation state and network traffic in real time to detect potential network attack behaviors in a timely manner and issue alarms. In the specific design, this function should be realized with the support of machine learning technology and big data technology. The machine learning model is trained with massive data to enable it to accurately identify and block various network attack behaviors in a timely manner, ensuring system security.

3.4. Design of the workflow and strategies of the collaborative protection mechanism

In the operation of the collaborative protection mechanism, its basic workflow includes three steps: safety monitoring, collaborative decision-making, and response execution. Safety monitoring is the first step in discovering SoC function or information security problems [15]. After detecting various security problems through the above-mentioned safety protection technical measures, this module will continue to use the intelligent algorithm model to accurately predict the types of security threats, their severity, and their impact scope. According to the security prediction results, the intelligent algorithm model will, with the support of big data technology, quickly and effectively decide on subsequent security protection and governance measures and send the corresponding security control instructions to the response execution module. After receiving the

system instructions, the response execution module will immediately isolate the faulty hardware or software according to the instructions or immediately block the corresponding network security attack behaviors. In this way, effective collaborative protection of SoC functions and information security in autonomous driving can be achieved, maximizing the security of the autonomous driving SoC system and avoiding various unnecessary security problems or accidents.

4. Conclusion

In summary, in the practical application of autonomous driving technology, the collaborative protection of SoC functional safety and information security is of great importance. Based on this, researchers need to establish a scientific, reasonable, and complete collaborative protection mechanism according to their internal connections and security protection requirements to ensure the safe and stable operation of the overall system and provide a good guarantee for the safety of autonomous vehicles.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Chen M, Yang Z, Qiu J, et al., 2025, Analysis of the Information Security Platform for the Commercialization of Autonomous Driving. Network Security Technology & Application, 2025(01): 111–113.

[2] Duan W, Wang G, 2024, Analysis of Traffic Safety Management for Autonomous Driving from the Perspective of Security. Auto & Safety, 2024(08): 83–93.

[3] Wang M, Tu H, Xue D, et al., 2024, Optimization of Autonomous Driving Adaptive Cruise Control Based on Safety Risk Prediction. Journal of Tongji University (Natural Science), 2024(04): 512–519.

[4] Zhang X, Chen H, Yang S, et al., 2024, Research on Cybersecurity Policies and Standardization for Autonomous Driving. China Information Security, 2024(02): 26–29.

[5] Wang Y, 2022, Development of On-Board Controllers for Autonomous Vehicles, dissertation, Tianjin University of Technology and Education.

[6] Mao X, Shang S, Cui H, 2018, Research on the Analysis and Countermeasures of Safety-Related Factors for Autonomous Vehicles. Shanghai Auto, 2018(1): 5.

[7] Wang K, Dong Z, Yang F, et al., 2023, Key Technologies and Applications of Vehicle-to-Everything Cooperative Autonomous Driving Based on C-V2X. Telecommunications Science, 39(3): 16.

[8] Xu X, 2024, Research on the Functional Safety Strategy for Multisource Vehicle Body Information in Autonomous Driving. Journal of Information Security Research, 10(11): 1020–1026.

[9] Feng J, 2023, Research and Application of 3D Object Detection Algorithms for Autonomous Driving Scenarios, dissertation, Taiyuan University of Technology.

[10] Han J, 2006, Research on the Attack-Defense Technology of Information Security Chips, dissertation, Fudan University.

[11] Zhang A, Qiao G, 2006, Development of High-Performance Information Security SoC Chips. China Integrated Circuit, 2006(01): 29–31.

[12] Zhang L, Chang C, Dong J, 2012, External Program Secure Access Architecture Based on SoC Chips, CN202102449U.

[13] Wen S, 2009, Design and Implementation of a Dedicated Security Chip Based on SoC, dissertation, The PLA Information Engineering University.

[14] Fang J, Wu P, Ai Y, 2023, Implementation of the Off-line Loading System for Secure SoC Chips. Journal of Engineering Technology (Citation Edition), 2023(4): 4.

[15] Yang T, 2022, Research on the Functional Safety of Vehicle Brake-by-Wire Systems for Autonomous Driving, dissertation, Jilin University.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Analysis of Internet of Things Intrusion Detection Technology Based on Deep Learning

Huijuan Zheng, Yongzhou Wang*

Chongqing University of Mobile Communication, Chongqing 401420, China

***Corresponding author:** Yongzhou Wang, 18983847509@163.com

Abstract: With the rapid development of modern information technology, the Internet of Things (IoT) has been integrated into various fields such as social life, industrial production, education, and medical care. Through the connection of various physical devices, sensors, and machines, it realizes information intercommunication and remote control among devices, significantly enhancing the convenience and efficiency of work and life. However, the rapid development of the IoT has also brought serious security problems. IoT devices have limited resources and a complex network environment, making them one of the important targets of network intrusion attacks. Therefore, from the perspective of deep learning, this paper deeply analyzes the characteristics and key points of IoT intrusion detection, summarizes the application advantages of deep learning in IoT intrusion detection, and proposes application strategies of typical deep learning models in IoT intrusion detection so as to improve the security of the IoT architecture and guarantee people's convenient lives.

Keywords: Deep learning; Internet of Things; Intrusion detection technology

Online publication: April 3, 2025

1. Introduction

In the modern network environment, intrusion detection technology is a basic means to improve network security, especially with important application value in the Internet of Things (IoT) architecture. Intrusion detection systems have the functions of real-time monitoring and analyzing network traffic data and system logs. Thus, they can timely discover intrusion behaviors and immediately respond and feedback, so as to prevent potential threats and detect potential security vulnerabilities, abnormal behaviors, and intrusion behaviors. With the development of artificial intelligence technology, deep learning has demonstrated more powerful feature learning and pattern recognition capabilities, becoming an important means to further optimize intrusion detection technology. With the support of deep learning algorithms, intrusion detection systems can autonomously learn and judge network traffic characteristics and system behavior patterns, greatly improving the accuracy of intrusion detection and adapting to the increasingly complex network security environment.

2. Overview of related concepts and theories

2.1. Internet of Things

The Internet of Things is an interaction architecture between the real world and the network environment composed of servers and computer networks. It is mainly composed of sensors, smart devices, networks and their protocols, and distributed applications, etc., so as to achieve the purpose of monitoring the real world and operating through devices [1]. For example, common networked cameras, intelligent agricultural equipment, and network-monitored industrial robots have been gradually applied in various fields of modern society. Specifically, the IoT is mainly constructed by the perception layer, the network layer, and the application layer. The perception layer mainly collects data through various sensors. The network layer stipulates network protocols to ensure the interactive transmission of data information. The application layer mainly uses application programs, software, and other devices to control underlying data or hardware devices, playing their functions and services [2].

Since the IoT is directly connected to devices, it has also become an important target of network intrusion. Common types of IoT intrusions include denial-of-service attacks, distributed denial-of-service attacks, SQL injection, man-in-the-middle attacks, cross-site scripting attacks, unauthorized access, brute-force cracking, etc. [3].

2.2. Intrusion detection

Intrusion detection systems are mainly used to monitor and identify potential network security factors so as to prevent networks from being invaded and achieve the effect of security protection. The system mainly uses the methods of data collection and analysis to judge whether there are abnormal situations in data traffic, detect potential or hidden attack means, and record and report the monitoring content [4].

Intrusion detection technologies can be divided into several types based on data sources and detection methods. In the classification of data sources, host-based systems mainly detect the system processes, file data, and system call behavior logs of the host system software itself to complete the detection [5]. Network-based systems monitor device traffic information to real-time monitor whether there is abnormal external traffic intrusion. In intrusion detection technologies based on detection methods, they can be divided into two situations: misuse and anomaly. The former detects intrusion behaviors by matching data with signatures, and the latter detects by monitoring abnormal data values.

2.3. Deep learning

Deep learning imitates the neural network structure of the human brain and endows machines with the ability to learn. It is mainly manifested in the layer-by-layer extraction and analysis of data features and then solves problems such as image and speech recognition, natural language processing, and intelligent recommendation. Deep learning has significant advantages in dealing with tasks with large amounts of data and complex content, so it shows good application value in the field of intrusion detection technology. In intrusion detection systems, common deep learning models mainly include convolutional neural networks, recurrent neural networks, and deep neural networks [6].

3. Advantages of applying deep learning in IoT intrusion detection

3.1. Automated feature extraction

The ability of automatic feature extraction is the primary advantage of the application of deep learning in

intrusion detection technology. The traffic data in the IoT environment is large-scale and complex. Traditional intrusion detection systems mainly rely on manually extracted features for matching and analysis, with low work efficiency and poor information capture effects. Deep learning models such as convolutional neural networks and recurrent neural networks have the ability to automatically extract features and conduct matching analysis, which can greatly improve the work efficiency and accuracy of traditional intrusion detection technology [7]. Taking the convolutional neural network as an example, it mainly performs sliding operations on data through the convolutional layer and convolutional kernels and can automatically extract local features, thereby grasping the feature information of data packets, port numbers, etc., and their combination relationships, so as to quickly achieve the effects of port scanning and data recognition.

3.2. Efficient pattern recognition

Pattern recognition ability is a key advantage of the application of deep learning in intrusion detection technology. The data traffic in the IoT system not only has high-dimensional features but also shows non-linear characteristics. Traditional detection methods cannot effectively identify attack patterns for such data. Deep learning models can deeply analyze the complex features and internal relationships of data, thus more accurately and efficiently identifying more complex attack patterns [8]. Taking the application of the deep neural network model as an example, it can learn and master the attack traffic characteristics of distributed denial-of-service attacks and then quickly extract and analyze abnormal traffic or specific data packets in intrusion detection to judge whether there is a DDoS attack.

3.3. Outstanding adaptability and scalability

Adaptability and scalability are also specific advantages of the application of deep learning in intrusion detection technology. On the one hand, the IoT environment is complex and changeable. Especially with the update and upgrade of different technologies, its intrusion means and methods are also constantly changing. Deep learning models can continuously adapt to new environments through continuous learning and model updates and can more effectively deal with new intrusion methods [9]. On the other hand, deep learning models are scalable. They can not only increase the number of network layers and neurons to improve their own detection efficiency and capabilities but also combine and link through distributed computing frameworks to allocate training tasks to multiple computing nodes, thus shortening their learning time.

4. Applications of typical deep learning models in IoT intrusion detection

4.1. Application of convolutional neural network (CNN)

In IoT intrusion detection, the extraction of traffic spatial features is a key link, and the convolutional neural network shows unique advantages in this regard. The convolutional neural network model mainly consists of several parts, such as the convolutional layer, the pooling layer, and the fully connected layer. The convolutional layer mainly uses convolutional kernels to perform convolutional operations on the detected data to automatically extract data features. The pooling layer samples and analyzes the data output by the convolutional layer, reducing the data dimension and calculation amount while ensuring that the feature information can be retained [10]. The fully connected layer mainly integrates the data features output by the pooling layer to classify the feature elements.

Specifically, the intrusion detection of DDoS attacks in the IoT based on the convolutional neural network

can be mainly divided into the following steps. First, relevant researchers should collect a large amount of network data of the IoT system, ensuring that the traffic data contains normal data and DDoS attack traffic. In data preprocessing, the original data needs to be cleaned by removing noise data and outliers and normalizing it to the range of 0–1 [11]. Second, transform the format of the processed data and input it into the convolutional neural network model, ensuring that the traffic data presents a two-dimensional matrix structure, with columns representing differences in feature dimensions and rows representing time steps. Third, complete the construction of the convolutional neural network model, where the convolutional layer and the pooling layer need to be constructed in multiple ways. In the convolutional layer, different convolutional kernels such as 3×3 and 5×5 should be set respectively to ensure the extraction of data features at different scales. Through the convolutional stacking effect, high-level and abstract feature elements can be gradually extracted. The pooling layer should adhere to the maximum pooling operation. For example, if the pooling window is set to 2×2 and the step size is 2, it can not only reduce the resolution of the feature map but also decrease the relevant calculation amount [12]. In the fully connected layer, the ReLU activation function can be used for optimization to enhance its non-linear expression ability. At the same time, classification can be completed based on the softmax function to calculate the probability of sample data and analyze whether it belongs to DDoS attack traffic.

In practical applications, the recognition accuracy of the convolutional neural network model in DDoS attacks reaches over 95%. It has unique advantages in the extraction of IoT traffic spatial features and can significantly reduce the false negative and false positive rates, providing important support for the development of IoT intrusion detection technology.

4.2. Application of Recurrent Neural Network (RNN)

In IoT intrusion detection, the recurrent neural network and its variants mainly show their functional characteristics in the detection of time-series data, especially being good at analyzing and capturing the time-series characteristics of continuous data streams. However, traditional recurrent neural network models face great difficulties in processing long-sequence data and may even encounter problems such as gradient disappearance or gradient explosion. In response to this, researchers have introduced different variant models such as the long short-term memory (LSTM) and the gated recurrent unit (GRU).

The long short-term memory network introduces three “gates” in the traditional recurrent neural network, namely the input gate, the forget gate, and the output gate, to control the flow of information, which can effectively solve the long-sequence problem [13]. The input gate can control the retention degree of input information, the forget gate is used to determine the information data that can be discarded, and the output gate controls the output content as the hidden state. The gated recurrent unit is a simplified variant based on the long short-term memory network. It combines the input and forget gates into an update gate, still having the original functions and effects. At the same time, it further simplifies the original output gate and memory unit, thus improving the calculation efficiency.

For example, in a smart home intrusion detection system, the system usually covers multiple IoT devices such as cameras, door locks, and home appliances. The operating state data and network traffic information of such devices have distinct time-series characteristics. Therefore, the data information and system logs of the devices can be collected in chronological order and used as training data. On this basis, first, data pre-processing is required. Error and duplicate data are removed through data cleaning, and the data is integrated

into the same range through normalization. Second, "time-step" features should be established in chronological order, clarifying the device state information and traffic change features at each time step, and then inputting them into the recurrent neural network model [14]. Finally, the LSTM model should be used for analysis and detection. On the one hand, a diversified hidden layer should be established, and learning and training should be completed through multiple LSTM units. On the other hand, pattern recognition is required. The traffic features in the normal operating state should be determined, and the traffic change rules in abnormal behaviors should be clarified. An alarm should be issued in a timely manner when an abnormality is detected.

Another example is in the context of the industrial IoT. The gated recurrent unit model can further detect the operating state of industrial equipment. In the industrial production process, some industrial equipment needs to run continuously for a long time, resulting in the continuous generation of time-series data such as equipment temperature, pressure, and rotation speed. The gated recurrent unit model can take its equipment parameters as input data to master and learn the data traffic rules during normal operation. When the equipment fails or is attacked externally, the model can immediately respond based on abnormal parameters, issue an early warning, avoid production accidents, and achieve the goal and effect of improving industrial production safety.

4.3. Application of Deep Neural Network (DNN)

In IoT intrusion detection, the deep neural network is mainly applied to large-scale data classification and detection. The deep neural network model generally consists of multiple hidden layers, with the ability to abstract and extract features layer by layer for data analysis, so as to complete the learning of more advanced feature representations and show a higher-level data classification and detection ability. Therefore, in the intrusion detection system, when facing the impact of large-scale network traffic data, the deep neural network can quickly learn the normal and abnormal traffic features, thus making scientific judgments on network data and device state information and achieving the effect of quickly and accurately identifying intrusion behaviors. Its advantage is mainly reflected in its non-linear fitting ability, especially being good at processing data with complex relationships. It is also one of the important technologies that endows the intrusion detection system with the ability to adapt to a changeable network environment.

Taking an enterprise IoT system as an example, its intrusion detection system needs to connect a large number of hardware devices within the enterprise and cover various work contents such as production, office work, and monitoring. Therefore, the generated traffic data is huge in scale and complex in content. To solve this problem, network data containing multiple intrusion traffic such as DDoS attacks, port scans, and malware propagation can be collected as learning and training data [15]. In the data preprocessing link, operations such as cleaning, noise reduction, and feature extraction are required. Both invalid and noisy data should be removed, and core data features such as port numbers, source IP addresses, destination IP addresses, data packet types, and traffic rules should be extracted and then used as input data. In model training, based on a multi-hidden-layer model, the model is continuously optimized through the backpropagation algorithm, and its weights and biases are adjusted to minimize the error. After multiple iterative trainings, the model can master the basic features of multiple intrusion behaviors, providing a reliable guarantee for the security of the enterprise IoT system.

5. Conclusion

In summary, in the context of the rapid development of information technology, improving the security level

of the IoT network environment is a key issue in current research. Facing increasingly changing and upgrading network security problems, IoT intrusion detection technology should also be continuously improved and optimized. Therefore, it is necessary to further give play to the application value and advantages of deep learning models. With the assistance of convolutional neural networks, recurrent neural networks, and deep neural networks, the protection level of intrusion detection systems can be continuously improved, creating a good, stable, healthy, and safe usage environment for the modern IoT environment and ensuring the smooth progress of people's work and life.

Funding

This article is the research result of the 2022 Municipal Education Commission Science and Technology Research Plan Project "Research on the Technology of Detecting Double-Surface Cracks in Concrete Lining of Highway Tunnels Based on Image Blast" (KJQN02202403); the first batch of school-level classroom teaching reform projects "Principles Applications of Embedded Systems" (23JG2166); the school-level reform research project "Continuous Results-Oriented Practice Research Based on BOPPPS Teaching Model—Taking the 'Programming Fundamentals' Course as an Example" (22JG332).

Disclosure statement

The authors declare no conflict of interest.

References

[1] Ge Y, 2024, Research on Internet of Things Intrusion Detection Technology Based on Deep Learning, dissertation, Beijing University of Posts and Telecommunications.

[2] Zhai R, 2024, Research on Open-Set Intrusion Detection Technology in Industrial Internet of Things, dissertation, Donghua University.

[3] Lu H, 2024, Research on Data Processing Technology in Internet of Things-Oriented Intrusion Detection Systems, dissertation, People's Public Security University of China.

[4] Che X, 2024, Research on Adversarial Attacks and Defense Technologies for Internet of Things Intrusion Detection, dissertation, Jilin University.

[5] Chen X, 2024, Research on Intrusion Detection Technology for Intelligent Connected Vehicles Based on Deep Learning, dissertation, University of Electronic Science and Technology of China.

[6] Zhang Y, 2024, Research on Internet of Things Intrusion Detection Technology Based on deep learning Algorithms, dissertation, Tianjin University of Technology.

[7] Feng G, Jiang S, Hu X, et al., 2024, New Progress in Research on Internet of Things-Oriented Intrusion Detection Technology. Netinfo Security, 24(02): 167–178.

[8] Du J, 2023, Research on Network Intrusion Detection Technology Based on Federated Learning, dissertation, Xijing University.

[9] Chen X, 2023, Research on Network Intrusion Detection Technology for Internet of Things. Network Security and Informatization, (10): 148–150.

[10] Xie S, 2023, Research on Internet of Things Intrusion Detection Technology Based on Combined Neural Networks,

dissertation, North China University of Technology.

[11] He F, 2023, Research on Home Internet of Things Intrusion Detection Technology Based on Machine Learning, dissertation, Southeast University.

[12] Liu Y, 2023, Research on Power Internet of Things Intrusion Detection Technology Based on Deep Learning, dissertation, Tianjin University of Science and Technology.

[13] Wang K, 2023, Research on Intrusion Detection Technology for Internet of Vehicles Based on Behavior Analysis, dissertation, Henan University of Science and Technology.

[14] Cui A, 2022, Research on Network Intrusion Detection Methods Based on Machine Learning, dissertation, Lanzhou University of Technology.

[15] Zhang X, 2021, Research on Key Technologies of Intrusion Detection in Internet of Things Environment, dissertation, Zhejiang Gongshang University.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

A Study on the Modeling and Design of Sigma-Delta Modulator for High Precision ADC

Haodong Guo, Longyu Li, Xia Zhang*

Shandong Provincial Key Laboratory of Optical Communications Science and Technology, School of Physics Science and Information Engineering, Liaocheng University, Liaocheng 252000, Shandong, China

***Corresponding author:** Xia Zhang, wenerzhang2002@163.com

Abstract: With the continuous improvement of signal processing accuracy requirements in modern electronic systems, the demand for high-precision analog-to-digital converters (ADCs) is increasing. Sigma-Delta modulator, as the most important component of high-precision ADC, is widely used in high-quality audio, high-precision instrument measurement, and other fields due to its advantages of high precision, strong noise resistance, and low hardware cost. This article designs a discrete structure third-order four-bit high-precision Sigma-Delta modulator through modeling, with an oversampling rate set to 512. Under ideal conditions, the simulation results show that the SDNR reaches 152.7db and the ENOB is 25.24bits. After introducing non-ideal noise, the system performance has decreased. The simulation results show that the SDNR is as high as 124.5db and the ENOB is 20.39bits. This indicates that the design can achieve high-precision conversion and provide assistance for further research in the future.

Keywords: Analog-to-digital converter; Sigma-Delta modulator; High precision; Modeling design

Online publication: April 3, 2025

1. Introduction

With the rapid development of modern electronic technology, high-precision analog-to-digital converters (ADCs) are increasingly widely used in key fields such as audio processing, wireless communication, and precision measurement [1]. High-precision ADCs not only have the advantages of high accuracy, good linearity, and low quantization noise but also have strong anti-interference ability and easy integration with digital circuits [2]. Through the strategy of exchanging speed for accuracy, they can achieve high-resolution analog-to-digital conversion at a lower complexity than analog circuits. Although high-precision ADCs have many advantages, their design also faces many challenges [3]. Especially in the pursuit of higher precision and lower power consumption, it is necessary to comprehensively consider the effects of modulator structure, circuit implementation, and non-ideal factors [4]. As the most important component of Sigma-Delta ADCs, the design of Sigma-Delta modulators directly affects the effectiveness and reliability of the system [5]. Therefore, researching

and designing high-precision modulators plays an important role in high-precision ADCs. Taking into account accuracy, power consumption, error, and linearity, this paper designs a Sigma-Delta modulator that achieves high-quality factor through system modeling and simulation [6].

2. Ideal modulator design

2.1. Modulator design process

In the modeling and design process of high-precision Sigma-Delta modulators, the first step is to determine the design specifications based on the application, select the appropriate system topology structure according to the requirements applicable to the audio field, and complete the system-level modeling. Then, each coefficient is determined based on the calculation, and reasonable scaling is completed according to the circuit design requirements [7]. It is necessary to determine whether the system is stable based on the zero pole diagram, integrator current, and other conditions. When the system is stable, an ideal model simulation can be performed to check whether the simulation indicators meet the actual requirements [8]. After the system simulation is completed, non-ideal factors are analyzed and introduced, and key noise parameters are set to supplement the overall model for simulation verification. The impact of non-ideal factors on the system is examined to provide further guidance for achieving circuit-level design [9].

2.2. Topology structure selection

The main performance parameters of the modulator are signal-to-noise distortion ratio and dynamic range. The following formula is used to calculate the signal-to-noise distortion ratio of the modulator:

$$\mathrm{SQNR} = 6.02N + 1.76 + (20L + 10)\log(\mathrm{OSR}) - 10\log\left(\frac{\pi^{2L}}{2L+1}\right) \tag{1}$$

Among them, N is the quantization bit number of the modulator, L is the modulator order, and OSR is the oversampling multiple [10]. According to formula (1), the Sigma-Delta modulator is mainly related to the integrator order, quantization bit number, and oversampling rate. Next, calculate the dynamic range of the modulator:

$$\mathrm{DR(dB)} = 10\lg\left(\frac{P_{sig,out,max}}{IBN}\right) \approx 10\lg\left[\frac{3}{2}\left(2^N - 1\right)^2 \frac{(2L+1)\mathrm{OSR}^{(2L+1)}}{\pi^{2L}}\right] \tag{2}$$

Among them, the output power of the P_ (sig, out, max) signal and IBN are the in band noise power. As shown in Equation (2) above, the higher the order of the modulator, the stronger its ability to suppress quantization noise within the bandwidth, the higher the signal-to-noise ratio, and the higher the accuracy. It can be seen that selecting a high-order structure significantly improves the performance of the modulator, effectively reducing noise and power consumption. However, considering that high-order structures can cause poor system stability, a third-order structure is selected and the system stability is ensured by setting an appropriate transfer function [11].

As the number of quantization bits increases, the multi-bit quantizer structure becomes more stable, and the performance of the modulator approaches that of an ideal modulator [12]. Therefore, in this design, a four-bit quantizer structure is ultimately chosen to improve the performance of the modulator. In a feedback modulator,

the input terminals of each integrator are affected by the feedback DAC, while in a feedforward modulator, only the input terminal of the first integrator is affected by the DAC [13]. Considering the requirement for high precision and not too high speed, the feedforward CIFF structure was ultimately chosen. The final modulator model is established as shown in **Figure 1**:

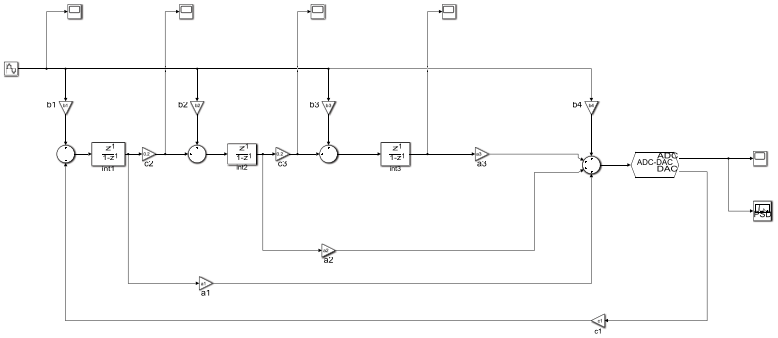


Figure 1. Structure model of third-order four-bit CIFF modulator

2.3. Modulator coefficient optimization and system simulation

After completing Simulink modeling, use functions such as scale in MATLAB toolbox to scale the modulator coefficients, taking into account signal dynamic range, noise shaping efficiency, and system stability. Through theoretical analysis and tool assistance, achieve proportional scaling before and after coefficients. After continuous optimization of various coefficients, the final optimized coefficients were obtained and compared with the initial coefficients as shown in **Table 1**.

Table 1. Comparison of modulator coefficient optimization before and after

Coefficient	a1	a2	a3	b1	b2	b3	b4	c1	c2	c3	g
Initial value	1.8183	2.2305	1.2085	0.2988	0	0	1	0.2988	0.3750	0.1994	0
Final value	3.5	5	3	0.2	0	0	1	0.2	0.2	0.2	0

After obtaining the preliminary optimized modulator coefficients, the zero pole distribution of the system was calculated using MATLAB tools. **Figure 2** illustrates the zero pole distribution of the system.

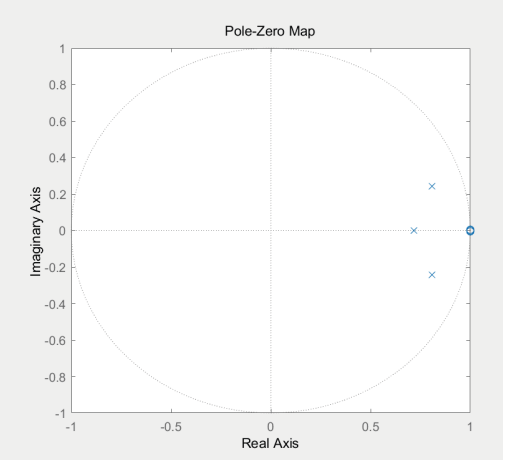


Figure 2. Distribution of zero and pole points in the system

From **Figure 2**, it can be seen that in the zero pole diagram of the noise transfer function, the zeros of the transfer function are separated from each other and all are at z = 1, and all the poles are located within the unit circle. Therefore, it can be concluded that the system remains stable. Next, the optimized coefficients are used to simulate the system. Under the condition of setting the input as a sine signal, the output spectrum of the ideal third-order four-bit CIFF modulator Simulink model is shown in **Figure 3**.

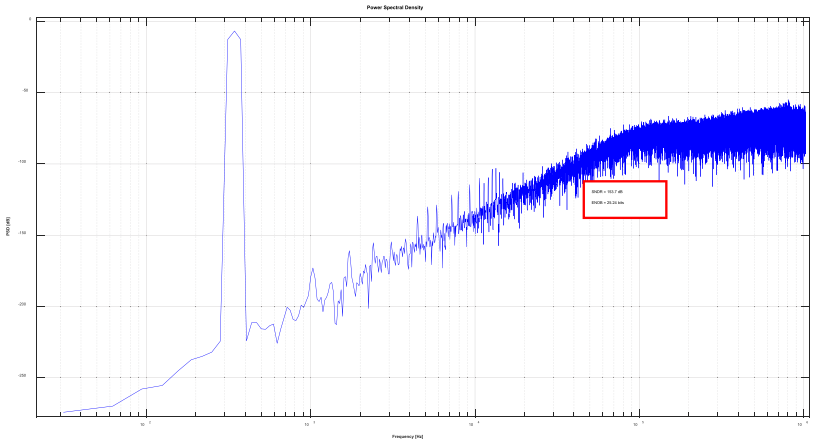


Figure 3. Simulation diagram of ideal modulator

From **Figure 3**, it can be seen that under ideal conditions, the SNDR of the modulator is 153.7dB, with an effective bit count of approximately 25.24bits, which meets the design requirements. In addition, the stability of the output signal can be observed based on the oscilloscope on each integrator. **Figure 4** shows the output results of each level of integrator. It can be seen that the outputs of all levels of integrators are within 0.6V, and there is no overload phenomenon, so the circuit current remains stable, which further proves the stability of the system.

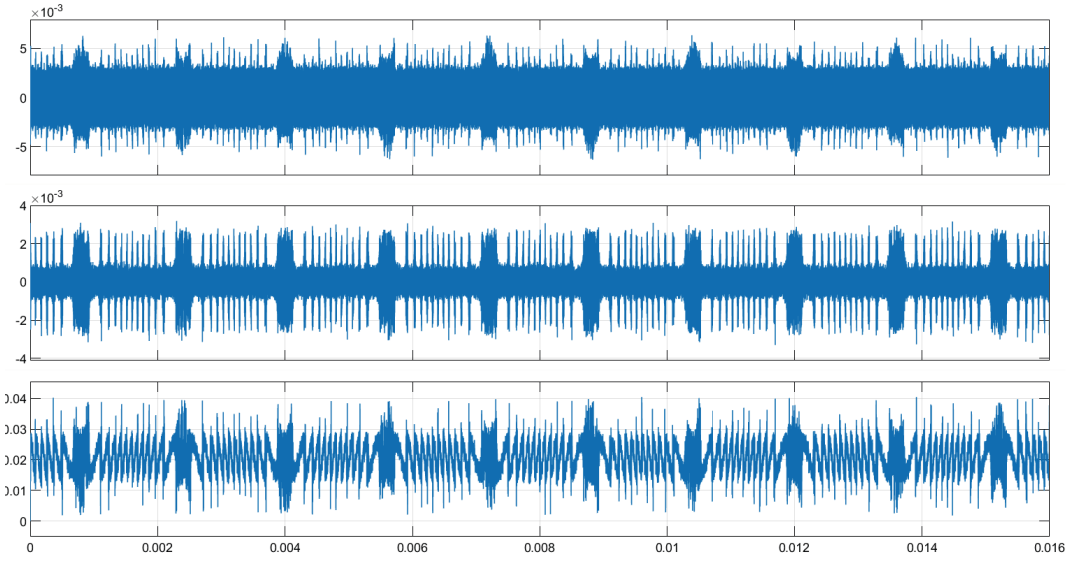


Figure 4. Output of integrators at all levels

3. Complete topology structure

Non-ideal factors and nonlinearity in circuits can reduce the performance of modulators, and these non-ideal

factors affect the quality factor of modulators through their own properties, the influence of specific circuits, and their impact on the noise transfer function. At the input of the modulator, non-ideal factors can be modeled as additive noise sources, and this structure mainly considers independent topology thermal noise. The thermal noise of the switch is mainly related to the sampling capacitor, and the impact of kT/C thermal noise on the in band noise IBN is as follows:

$$\frac{IBN_{KT}}{C} \approx \frac{4KT}{C_S OSR} \tag{3}$$

Among them, K represents Boltzmann constant, T represents absolute temperature, C-S represents sampling capacitance, and OSR is oversampling rate. Research has found that the in band noise generated by kT/C is similar to the in band noise contribution of rational quantization noise. Therefore, a non-ideal model incorporating kT/C noise is shown in **Figure 5**.

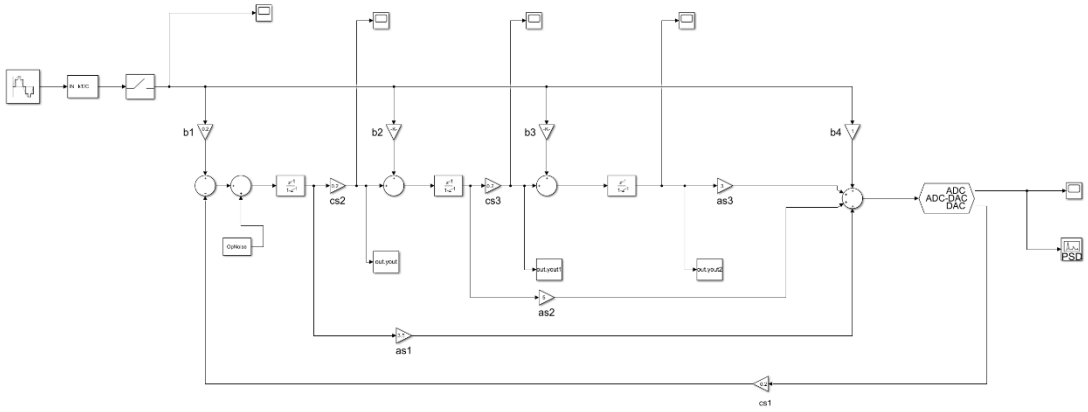


Figure 5. Introduction of non-ideal factor modulator model

According to the design requirements, taking into account the trade-off between actual circuit design and system performance, a clock jitter of 0.5ns, a sampling capacitor of 6 pf, an operational amplifier gain of 60dB, an operational amplifier voltage swing rate of 10V/us, and a sine input signal swing amplitude of 0.8V were ultimately selected. Set the above factors in the system and simulate the modulator as a whole. The simulation results are shown in **Figure 6**:

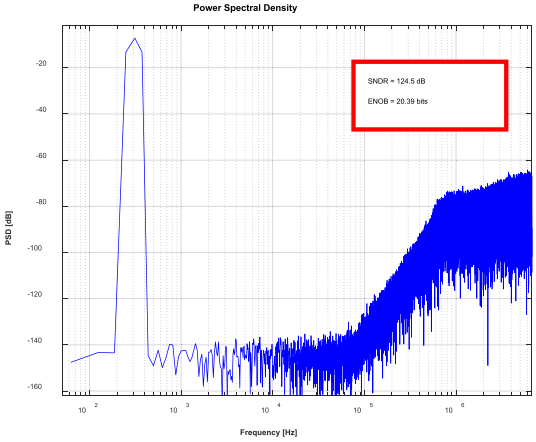


Figure 6. Overall performance simulation results of modulator

According to the simulation results in **Figure 6**, the SNDR with the addition of a non-ideal factor modulator is reduced to 124.5dB, and the effective bit count is reduced to 20.39 bits. This indicates that the influence of non-ideal factors cannot be ignored in practical design, but there is a margin left in the design process to achieve the expected design goals.

4. Conclusion

This article designed a high-precision Sigma-Delta modulator system, which adopts a third-order four-bit quantized CIFF structure with an oversampling rate of 256 and improves system performance by adjusting coefficients. After incorporating non-ideal factors, although the overall performance of the system decreased, the simulation results met expectations. The final simulation results showed that SNDR could reach 124.5dB and ENOB could reach 20.39bits, which meets the expected design goals and is helpful for further circuit design. It can be applied to the field of audio signal conversion.

Disclosure statement

The authors declare no conflict of interest.

References

[1] He B, Guo X, Jia H, et al., 2025, A 100-MHz Bandwidth Continuous-Time Sigma-Delta ADC with 1 V Supply in 28 nm CMOS. Microelectronics Journal, 158: 106597.

[2] Li W, Tian D, Zhu H, et al., 2025, A Programmable Gain Amplifier Featuring a High Power Supply Rejection Ratio for a 20-Bit Sigma-Delta ADC. Electronics, 14(4): 720.

[3] Kohandel R, Yavari M, 2025, A Systematic Time-Based approach for Analysis and Compensation of Excess Loop Delay in Continuous-Time Sigma-Delta Modulators Exceeding One Clock Cycle. Results in Engineering, 25: 103912.

[4] Ye M, Liu Z, Zhao Y, 2024, Design of a Sigma-Delta Analog-to-Digital Converter Cascade Decimation Filter. Electronics, 13(11): 2090.

[5] Brandon M, 2024, Orbital Angular Momentum Small-X Evolution: Exact Results in the Large-Nc Limit. Journal of High Energy Physics, 2024(4): 55.

[6] Wang J, Wang G, Li K, et al., 2024, Erratum: A 1.8 V 115.52 dB Third-Order Discrete-Time Sigma-Delta Modulator Using Nested Chopper Technology. Journal of Circuits, Systems and Computers, 33(09): 2492001.

[7] Khachi CS, Wanas KA, 2024, Two Families of m-fold Symmetric Bi-univalent Functions Involving a Linear Combination of Bazilevic Starlike and Convex Functions. Earthline Journal of Mathematical Sciences, 14(3): 405–419.

[8] Dong S, Ning S, Yuan M, et al., 2024, A Low-Power Sigma-Delta Modulator Based on High-Order Op-Amp Sharing Technique for Speech Communication. AEUE - International Journal of Electronics and Communications, 176: 155116.

[9] Wang J, Wang G, Li K, et al., 2024, A 1.8 V 115.52 dB Third-Order Discrete-Time Sigma-Delta Modulator Using Nested Chopper Technology. Journal of Circuits, Systems and Computers, 33(07): 2450126.

[10] Chen K, Chen M, Cheng L, et al., 2022, A 124 dB Dynamic Range Sigma-Delta Modulator Applied to Non-Invasive

EEG Acquisition Using Chopper-Modulated Input-Scaling-Down Technique. Science China Information Sciences, 65(4): 140402.

[11] Ocampo-Hidalgo JJ, Castillo JA, Molinar-Solis JE, 2022, Processing Electrocardiographic Signals Using a Custom Designed Sigma-Delta Modulator. Journal of Circuits, Systems and Computers, 31(03): 2250040.

[12] Nikolaos G, Alkis PH, 2022, Functional Verification of a Sigma-Delta ADC Real Number Model. International Journal of Electronics, 109(1): 119–134.

[13] Toshihiro I, Shoji M, Yoshiya K, 2022, Sigma-Delta Beamformer DOA Estimation for Distributed Array Radar. IEICE Transactions on Communications, E105.B(12): 1589–1599.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Manifold-Optimized Error-State Kalman Filter for Robust Pose Estimation in Unmanned Aerial Vehicles

Bolin Jia[1], Zongwen Bai[1]*, Yiqun Gao[1], Dong Wang[1], Meili Zhou[1], Peiqi Gao[1], Pei Zhang[1], Zhang Yang[2]

[1]School of Physics and Electronic Information, Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data, Yan'an University, Yan'an 716000, Shaanxi, China
[2]Zichang Vegetable Center, Yan'an 716000, Shaanxi, China

***Corresponding author:** Zongwen Bai, ydbzw@yau.edu.cn

Abstract: This paper presents a manifold-optimized Error-State Kalman Filter (ESKF) framework for unmanned aerial vehicle (UAV) pose estimation, integrating Inertial Measurement Unit (IMU) data with GPS or LiDAR to enhance estimation accuracy and robustness. We employ a manifold-based optimization approach, leveraging exponential and logarithmic mappings to transform rotation vectors into rotation matrices. The proposed ESKF framework ensures state variables remain near the origin, effectively mitigating singularity issues and enhancing numerical stability. Additionally, due to the small magnitude of state variables, second-order terms can be neglected, simplifying Jacobian matrix computation and improving computational efficiency. Furthermore, we introduce a novel Kalman filter gain computation strategy that dynamically adapts to low-dimensional and high-dimensional observation equations, enabling efficient processing across different sensor modalities. Specifically, for resource-constrained UAV platforms, this method significantly reduces computational cost, making it highly suitable for real-time UAV applications.

Keywords: UAV pose estimation; Error-State Kalman Filter; Manifold; GPS; LiDAR

Online publication: April 3, 2025

1. Introduction

Accurate pose estimation, which determines both position and orientation, is essential for unmanned aerial vehicles (UAVs) in tasks such as localization, navigation, path planning [1], and other autonomous operations [2]. In UAV applications, precise pose estimation is particularly critical for autonomous flight control, remote sensing target detection, and precise positioning of detected objects.

Traditional pose estimation methods rely on a single sensor, such as Global Positioning System (GPS), inertial measurement units (IMUs) [3], or cameras, each with inherent advantages and limitations. GPS

provides absolute positioning but lacks orientation information and performs poorly in indoor or GPS-denied environments. IMUs, while capable of high-frequency motion sensing, suffer from drift errors due to cumulative integration noise. Vision-based methods, such as camera-based Simultaneous Localization and Mapping (SLAM), depend on good lighting conditions and visual feature availability, which are not always guaranteed in UAV operations.

To overcome these limitations, sensor fusion techniques have been widely explored to integrate IMU data with GPS or LiDAR, thereby improving pose estimation accuracy and robustness. This paper presents a manifold-optimized Error-State Kalman Filter (ESKF) framework designed specifically for real-time UAV applications, fusing IMU data with GPS or LiDAR to achieve accurate state estimation. Unlike traditional approaches that use rotation matrices or quaternions, which introduce redundant degrees of freedom, this work adopts a manifold-based optimization approach that leverages exponential and logarithmic mappings to transform rotation vectors into rotation matrices, thereby achieving a minimal three-degree-of-freedom representation.

The main contributions of this work are as follows: (1) Development of a real-time sensor fusion framework based on ESKF [4], enabling accurate and robust UAV pose estimation by integrating IMU with GPS or LiDAR data. (2) Adoption of a manifold-based representation and error-state for rotational increments, addressing gimbal lock issues in Euler angles and eliminating redundancy in rotation matrices and quaternions, thereby improving computational stability. (3) Introduction of a novel Kalman gain computation strategy, which dynamically adjusts for low-dimensional and high-dimensional observation equations, ensuring computational efficiency. Specifically, for resource-constrained UAV platforms, this method significantly reduces computational cost, making it highly suitable for real-time UAV applications.

2. Related work

Pose estimation is a fundamental task in autonomous UAV cruising, enabling precise positioning, flight control, and target tracking. However, compared to ground-based systems, UAVs face a series of unique challenges, including limited onboard computational resources, GPS signal degradation, and rapid dynamic motion, making real-time state estimation more challenging than ground systems. To enhance the accuracy and robustness of UAV pose estimation, multi-sensor fusion techniques have been widely adopted. These fusion approaches effectively compensate for the limitations of individual sensors and improve UAV adaptability in complex mission scenarios. IMU-GPS fusion provides global positioning information, while IMU-LiDAR [5] fusion performs well in environments where GPS signals are limited or unavailable. Additionally, IMU-vision [6] fusion methods have played a significant role in autonomous UAV navigation, especially in complex environments with low illumination or no GPS availability.

Optimization-based methods have been widely adopted in SLAM systems, such as VINS-Mono [7] and VINS-Motion [8]. These methods employ IMU pre-integration [6] to fuse IMU data with other sensor inputs, resulting in high-precision pose estimation. However, due to their high computational demands, these methods are challenging to implement in real-time applications, especially in resource-constrained environments such as UAVs or small robots.

Filtering-based methods offer a computationally efficient alternative to optimization-based approaches. However, handling rotational states in UAV pose estimation remains a key challenge. Classical Kalman filters

assume a Gaussian distribution of system states, requiring the state space to be Euclidean. However, rotation matrices and quaternions do not satisfy these conditions due to their nonlinear nature. To address this issue, Euler angles or axis-angle representations are sometimes used, but both suffer from gimbal lock or singularities. When using rotation matrices or quaternions, nonlinear extensions such as Extended Kalman Filters (EKF) [9] or Unscented Kalman Filters (UKF) [10] are typically employed. However, rotation matrices require 9 degrees of freedom, while quaternions require 4 degrees of freedom, introducing redundancy that negatively affects computational efficiency. This issue is particularly relevant for real-time UAV applications, where computational resources are limited, and efficient state estimation is required.

Given these challenges, this paper proposes a fusion method that integrates IMU, GPS, or LiDAR data, using the ESKF as the fusion algorithm for UAV pose estimation**.** This method optimizes the representation of rotational increments, allowing them to be expressed using a minimal three-degree-of-freedom parameter set. By utilizing the error-state formulation, it avoids gimbal lock and singularity issues that arise when describing rotation with three degrees of freedom, thereby improving the accuracy of pose estimation. Additionally, a new Kalman gain calculation formula is introduced, enabling different formulas to be applied when dealing with low-dimensional and high-dimensional observation signals, ensuring computational efficiency.

3. Error-State Kalman Filter algorithm composition

3.1. Problems faced in rotation calculations

In classical Kalman filter algorithms, it is generally assumed that the initial system values, system noise, and observation noise are mutually independent and follow a Gaussian distribution. This assumption requires the vector space defined by the state variables to be closed under addition and scalar multiplication. However, in the process of IMU pose estimation, the rotation quantities are typically defined as rotation matrices or quaternions, with quaternions subject to the constraint: $||q|| = I$ and rotation matrices subject to the constraint [11]: $RR^T = I$. Obviously, neither of these representations is closed under addition.

When using Euler angles to solve for IMU attitudes, although there are no constraints and both addition and multiplication operations are closed, the classical Kalman filter algorithm can be directly applied. However, as the number of iterations increases, singularities and gimbal lock problems can occur. It can be easily verified that with increasing iterations, the rotation quantities quickly move away from the origin, leading to these issues.

To address these problems, the manifold space error-state Kalman filter algorithm is introduced. Its advantages are: In handling rotation quantities, the error-state Kalman filter can use rotation vectors to directly describe rotations. The error state is a small quantity that always remains near the origin, avoiding singularities and gimbal lock issues caused by an increasing number of iterations. Additionally, its higher-order terms (second order and above) converge easily.

3.2. Manifold space

Let M be the manifold of dimension n in consideration (e.g. $M = SO(3)$). Since manifolds are locally homeomorphic to R^n, we can establish a bijective mapping from a local neighborhood on M to its tangent space R^n via two encapsulation operators (and ‘ [12]:

$$\boxplus : \mathbb{M} \times \mathbb{R}^n \to \mathbb{M}; \qquad \boxminus : \mathbb{M} \times \mathbb{M} \to \mathbb{R}^n$$

$$\mathbb{M} = SO(3): \mathbf{R} \boxplus \mathbf{r} = \mathbf{R}\mathrm{Exp}(\mathbf{r}); \quad \mathbf{R}_1 \boxminus \mathbf{R}_2 = \mathrm{Log}(\mathbf{R}_2^{\mathrm{T}}\mathbf{R}_1)$$
$$\mathbb{M} = \mathbb{R}^n: \quad \mathbf{a} \boxplus \mathbf{b} = \mathbf{a}+\mathbf{b}; \quad \mathbf{a} \boxminus \mathbf{b} = \mathbf{a}-\mathbf{b}$$

Where $\mathrm{Exp}(\mathbf{r}) = \mathbf{I} + \frac{\mathbf{r}}{\|\mathbf{r}\|}\sin(\|\mathbf{r}\|) + \frac{\mathbf{r}^2}{\|\mathbf{r}\|^2}(1-\cos(\|\mathbf{r}\|))$ is the exponential map and $\mathrm{Log}(\cdot)$ is its inverse map. For a compound manifold $\mathbb{M} = SO(3) \times \mathbb{R}^n$, we have:

$$\begin{bmatrix}\mathbf{R}\\ \mathbf{a}\end{bmatrix} \boxplus \begin{bmatrix}\mathbf{r}\\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}\mathbf{R} \boxplus \mathbf{r}\\ \mathbf{a}+\mathbf{b}\end{bmatrix}; \quad \begin{bmatrix}\mathbf{R}_1\\ \mathbf{a}\end{bmatrix} \boxminus \begin{bmatrix}\mathbf{R}_2\\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}\mathbf{R}_1 \boxminus \mathbf{R}_2\\ \mathbf{a}-\mathbf{b}\end{bmatrix}$$

From the above definition, it is easy to verify that：

$$(\mathbf{x} \boxplus \delta) \boxminus \mathbf{x} = \delta; \ \mathbf{x} \boxplus (\mathbf{y} \boxminus \mathbf{x}) = \mathbf{y}; \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{M}, \forall \delta \in \mathbb{R}^n$$

3.3. IMU state model

Generally, the state variables of an IMU are typically represented as:

$$\mathbb{M} = SO(3) \times \mathbb{R}^{15}, \dim(\mathbb{M}) = 18$$
$$\mathbf{x}(t) \doteq \begin{bmatrix} {}^{\mathrm{G}}\mathbf{R}_I^T & {}^{\mathrm{G}}\mathbf{p}_I^T & {}^{\mathrm{G}}\mathbf{v}_I^T & \mathbf{b}_\omega^T & \mathbf{b}_a^T & {}^{\mathrm{G}}\mathbf{g}^T \end{bmatrix}$$

Where $\mathrm{x}(t)$ is a function of time t.

The kinematic model can be expressed as [13]:

$$\begin{aligned} {}^{\mathrm{G}}\dot{\mathbf{p}}_I &= {}^{\mathrm{G}}\mathbf{v}_I \\ {}^{\mathrm{G}}\dot{\mathbf{v}}_I &= {}^{\mathrm{G}}\mathbf{R}_I(a_m - \mathbf{b}_{\mathrm{a}} - \mathbf{n}_a) + {}^{\mathrm{G}}\mathbf{g}, {}^{\mathrm{G}}\dot{\mathbf{g}} = 0 \\ {}^{G}\dot{\mathbf{R}}_I &= {}^{G}\mathbf{R}_I[\omega_m - \mathbf{b}_\omega - \mathbf{n}_\omega]^\wedge \\ \dot{\mathbf{b}}_\omega &= \mathbf{n}_{\mathrm{b}\omega}, \ \dot{\mathbf{b}}_a = \mathbf{n}_{\mathrm{b}a} \end{aligned}$$

In (1) and (2), ${}^{\mathrm{G}}\mathrm{p}_I$ and ${}^{\mathrm{G}}\mathrm{R}_I$ represent the position and attitude of the IMU in the world frame (typically, the first frame of the IMU is defined as the world frame), ${}^{\mathrm{G}}\mathrm{v}_I$ represents the velocity of the IMU in the world frame, b_ω and b_a represent the biases of the angular velocity and acceleration, which are modeled as the random walk process with Gaussian noises $\mathrm{n}_{\mathrm{b}\omega}$ and $\mathrm{n}_{\mathrm{b}a}$, n_a, and n_ω represent the Gaussian white noise in the IMU measurement process, ${}^{\mathrm{G}}\mathrm{g}$ represents the gravitational acceleration in the world frame, a_m and ω_m represent the IMU measurements of angular velocity and acceleration, respectively. The notation $[\theta]^\wedge$ denotes the skew-symmetric matrix of vector $\theta \in \mathbb{R}^3$ that maps the cross product operation [14].

3.4. Error-State Kalman Filter

Define the above states as follows:

$$G\boldsymbol{R}_I^T \doteq \boldsymbol{R}, \ {}^{G}\boldsymbol{p}_I^T \doteq \boldsymbol{p}, \ {}^{G}\boldsymbol{v}_I^T \doteq \boldsymbol{v}, \boldsymbol{b}_\omega^T \doteq \boldsymbol{b}_\omega, \boldsymbol{b}_a^T \doteq \boldsymbol{b}_a, \ {}^{G}\boldsymbol{g}^T \doteq \boldsymbol{g}$$

The ESKF algorithm treats the true state as a combination of the nominal state and the error state, expressed as:

$$\mathbf{x}(t)_t = \mathbf{x}(t) \boxplus \delta\mathbf{x}(t)$$

The true state is represented by x(t)$_t$. x(t) and $\boldsymbol{\delta}$x(t) are the noise-free nominal state and the error state including noise, respectively. Here $\boldsymbol{\delta}$x(t) ~ N{0,P}, P is the covariance matrix of the error state, which can be specifically expanded as:

$$\mathbf{x}(t)_t = \begin{bmatrix} \mathbf{R}_t \\ \mathbf{p}_t \\ \mathbf{v}_t \\ \mathbf{b}_{\omega t} \\ \mathbf{b}_{at} \\ \mathbf{g}_t \end{bmatrix} = \begin{bmatrix} \mathbf{R}(\ \delta\mathbf{r} \\ \mathbf{p}+\delta\mathbf{p} \\ \mathbf{v}+\delta\mathbf{v} \\ \mathbf{b}_\omega+\delta\mathbf{b}_\omega \\ \mathbf{b}_a+\delta\mathbf{b}_a \\ \mathbf{g}+\delta\mathbf{g} \end{bmatrix} \quad)(\mathbf{R}(\ \delta\mathrm{r}=\mathbf{R}\mathrm{Exp}(\delta\mathbf{r})$$

In (3) and (4), the nominal state does not include noise terms; the noise terms are included in the error state. Together, they add up to form the true state. During the recursive operations, compared to the nominal state, the error state can be considered as a small quantity containing noise. Thus, we can derive the kinematic models for the nominal states x(t) and $\boldsymbol{\delta}$x(t) as follows:

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} \dot{\mathbf{R}} \\ \dot{\mathbf{p}} \\ \dot{\mathbf{v}} \\ \dot{\mathbf{b}}_\omega \\ \dot{\mathbf{b}}_a \\ \dot{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{R}(\omega_m-\mathbf{b}_a)^\wedge \\ \mathbf{v} \\ \mathbf{R}(a_m-\mathbf{b}_\omega)+\mathbf{g} \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\delta\dot{\mathbf{x}}(t) = \begin{bmatrix} \delta\dot{r} \\ \delta\dot{\mathbf{p}} \\ \delta\dot{\mathbf{v}} \\ \delta\dot{\mathbf{b}}_\omega \\ \delta\dot{\mathbf{b}}_a \\ \delta\dot{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} -\left(\omega_m-\mathbf{b}_\omega\right)^\wedge \delta r-\delta\mathbf{b}_\omega-\mathbf{n}_\omega \\ \delta\mathbf{v} \\ -\mathbf{R}\left(a_m-\mathbf{b}_a\right)^\wedge \delta r-\mathbf{R}\delta\mathbf{b}_a-\mathbf{n}_a+\delta\mathbf{g} \\ \mathrm{n}_{b\omega} \\ \mathrm{n}_{ba} \\ 0 \end{bmatrix}$$

3.4.1. Discrete-time Error-State Kalman Filter kinematic equations

Based on (5) and (6), we can respectively establish the discrete-time nominal state and error state Kalman kinematic equations:

For the nominal state, we have:

$$\mathbf{x}(t+\Delta t) = \begin{bmatrix} \mathbf{R}(t+\Delta t) \\ \mathbf{p}(t+\Delta t) \\ \mathbf{v}(t+\Delta t) \\ \mathbf{b}_\omega(t+\Delta t) \\ \mathbf{b}_a(t+\Delta t) \\ \mathbf{g}(t+\Delta t) \end{bmatrix} = \begin{bmatrix} \mathbf{R}(t)\mathrm{Exp}((\omega_m-\mathbf{b}_a)\Delta t) \\ \mathbf{p}(t)+\mathbf{v}\Delta t+\frac{1}{2}(\mathbf{R}(a_m-\mathbf{b}_\omega)+\mathbf{g})\Delta t^2 \\ \mathbf{v}(t)+(\mathbf{R}(a_m-\mathbf{b}_\omega)+\mathbf{g})\Delta t \\ \mathbf{b}_\omega(t) \\ \mathbf{b}_a(t) \\ \mathbf{g}(t) \end{bmatrix}$$

For the error state, we have:

$$\delta\mathbf{x}(t+\Delta t) = \begin{bmatrix} \delta r(t+\Delta t) \\ \delta\mathbf{p}(t+\Delta t) \\ \delta\mathbf{v}(t+\Delta t) \\ \delta\mathbf{b}_\omega(t+\Delta t) \\ \delta\mathbf{b}_a(t+\Delta t) \\ \delta\mathbf{g}(t+\Delta t) \end{bmatrix} = \begin{bmatrix} \mathrm{Exp}(-(\omega_m-\mathbf{b}_\omega)\Delta t)\delta r-\delta\mathbf{b}_\omega\Delta t-r_i \\ \delta\mathbf{p}+\delta\mathbf{v}\Delta t \\ \delta\mathbf{v}+(-\mathbf{R}(a_m-\mathbf{b}_a)^\wedge \delta r-\mathbf{R}\delta\mathbf{b}_a+\delta\mathbf{g})\Delta t+\mathbf{v}_i \\ \delta\mathbf{b}_\omega+\omega_i \\ \delta\mathbf{b}_a+a_i \\ \delta\mathbf{g} \end{bmatrix}$$

Where $\mathrm{Var}(r_i)=\sigma_{\mathbf{n}_\omega}^2 \Delta t^2 \mathbf{I}$, $\mathrm{Var}(\mathbf{v}_i)=\sigma_{\mathbf{n}_a}^2 \Delta t^2 \mathbf{I}$, $\mathrm{Var}(\omega_i)=\sigma_{b_\omega}^2 \Delta t \mathbf{I}$, $\mathrm{Var}(a_i)=\sigma_{b_a}^2 \Delta t \mathbf{I}$.

3.4.2. Error-State Kalman Filter propagation

To obtain a compact form of expression, we define as follows:

Let k denote the k - th iteration, Δt the time step, x_k the nominal state vector, $\boldsymbol{\delta}\mathrm{x}_k$ the error state vector, u_{mk} the input vector, and i the disturbance pulse vector. Specifically, this can be expanded as follows:

$$\mathbf{x}_k=\begin{bmatrix}\mathbf{R}_k\\ \mathbf{p}_k\\ \mathbf{v}_k\\ \mathbf{b}_{(\omega)k}\\ \mathbf{b}_{(a)k}\\ \mathbf{g}_k\end{bmatrix},\ \delta\mathbf{x}_k=\begin{bmatrix}\delta r_k\\ \delta\mathbf{p}_k\\ \delta\mathbf{v}_k\\ \delta\mathbf{b}_{(\omega)k}\\ \delta\mathbf{b}_{(a)k}\\ \delta\mathbf{g}_k\end{bmatrix},\mathbf{u}_{(m)k}=\begin{bmatrix}\omega_{(m)k}\\ a_{(m)k}\end{bmatrix},\ \mathbf{i}=\begin{bmatrix}r_i\\ \mathbf{v}_i\\ \omega_i\\ a_i\end{bmatrix}$$

The error state propagation process is:

$$\delta\mathbf{x}_{k+1}\leftarrow f(\mathbf{x}_k,\delta\mathbf{x}_k,\mathbf{u}_{(m)k},\mathbf{i})=\mathbf{F}_{\mathrm{x}}(\mathbf{x}_k,\mathbf{u}_{(m)k})\cdot\delta\mathbf{x}_k+\mathbf{F}_{\mathrm{i}}\cdot\mathbf{i}$$

The prediction process of the Error-State Kalman Filter is:

$$\widehat{\delta\mathbf{x}}_{k+1}^{-}\leftarrow\mathbf{F}_{\mathrm{x}}(\hat{\mathbf{x}}_k,\mathbf{u}_{(m)k})\cdot\widehat{\delta\mathbf{x}}_k$$

$$\mathbf{P}_{k+1}^{-}\leftarrow\mathbf{F}_{\mathrm{x}}\mathbf{P}_k\mathbf{F}_{\mathrm{x}}^T+\mathbf{F}_{\mathrm{i}}\mathbf{Q}_{\mathrm{i}}\mathbf{F}_{\mathrm{i}}^T$$

In (9), (10), and (11), $\widehat{\delta\boldsymbol{x}}_{k+1}^{-}$ is the prior estimate of the error state, $\mathrm{F_x}$ and $\mathrm{F_i}$ are the Jacobian matrices with respect to the error state and the disturbance, respectively, and $\mathrm{Q_i}$ is the covariance matrix of the disturbance pulse.

$$\mathbf{F}_{\mathrm{x}}=\frac{\partial f}{\partial\delta\mathbf{x}}\Big|_{\mathbf{x},\mathbf{u}_m}=\begin{bmatrix}\mathrm{Exp}\left(-(\omega_m-\mathbf{b}_g)\Delta t\right) & 0 & 0 & -\mathbf{I}\Delta t & 0 & 0\\ 0 & \mathbf{I} & \mathbf{I}\Delta t & 0 & 0 & 0\\ -\mathbf{R}\left(a_m-\mathbf{b}_{\mathrm{a}}\right)^{\wedge}\Delta t & 0 & \mathbf{I} & 0 & -\mathbf{R}\Delta t & 0\\ 0 & 0 & 0 & \mathbf{I} & 0 & 0\\ 0 & 0 & 0 & 0 & \mathbf{I} & 0\\ 0 & 0 & 0 & 0 & 0 & \mathbf{I}\end{bmatrix}$$

$$\mathbf{F}_{\mathrm{i}}=\frac{\partial f}{\partial\mathbf{i}}\Big|_{\mathbf{x},\mathbf{u}_m}=\begin{bmatrix}\mathbf{I}&0&0&0&0&0\\0&0&0&0&0&0\\0&0&\mathbf{I}&0&0&0\\0&0&0&\mathbf{I}&0&0\\0&0&0&0&\mathbf{I}&0\\0&0&0&0&0&0\end{bmatrix},\mathbf{Q}_{\mathrm{i}}=\begin{bmatrix}\sigma_{\mathbf{n}_\omega}^2\Delta t^2\mathbf{I}&0&0&0&0&0\\0&0&0&0&0&0\\0&0&\sigma_{\mathbf{n}_a}^2\Delta t^2\mathbf{I}&0&0&0\\0&0&0&\sigma_{b_\omega}^2\Delta t\mathbf{I}&0&0\\0&0&0&0&\sigma_{b_a}^2\Delta t\mathbf{I}&0\\0&0&0&0&0&0\end{bmatrix}$$

3.4.3. Error-State Kalman Filter update process

When observation data is received, such as GPS, BDS, or related visual information, the error state can then be updated.

Suppose there is a set of sensors providing state observation information, we have:

$$\mathbf{y} = h(\mathbf{x}_{t(k+1)}) + v$$

Here, $h()$ is generally a nonlinear function of the system state, v is Gaussian white noise, and $v \sim N\{0,V\}$, with its covariance matrix being V.

Below is the update process for the Error-State Kalman Filter:

$$\mathbf{K} = \begin{cases} \mathbf{P}_{k+1}^{-}\mathbf{H}^{T}(\mathbf{H}\mathbf{P}^{-}\mathbf{H}^{T} + \mathbf{V})^{-1} & \text{if m} \le 18 \quad \text{(a)} \\ \left(\mathbf{H}^{T}\mathbf{R}^{-1}\mathbf{H} + \left(\mathbf{P}_{k+1}^{-}\right)^{-1}\right)^{-1}\mathbf{H}^{T}\mathbf{V}^{-1} & \text{if m} > 18 \quad \text{(b)} \end{cases}$$

$$\widehat{\delta\mathbf{x}}_{k+1} \leftarrow \mathbf{K}(\mathbf{y} - h(\hat{\mathbf{x}}_{k+1}^{-}))$$

$$\hat{\mathbf{x}}_{k+1} \leftarrow \hat{\mathbf{x}}_{k+1}^{-}(\quad \widehat{\delta\mathbf{x}}_{k+1}$$

$$\mathbf{P}_{k+1} \leftarrow (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}_{k+1}^{-}$$

In (16), $\widehat{\delta\boldsymbol{x}}_{k+1}$ is the posterior estimate of the error state, P_{k+1} is the updated error covariance matrix, and is the Jacobian matrix of the observation equation with respect to the error state:

$$\mathbf{H} \triangleq \left.\frac{\partial h}{\partial \delta \mathbf{x}}\right|_{\mathbf{x}} = \left.\frac{\partial h}{\partial \mathbf{x}_t}\right|_{\mathbf{x}} \left.\frac{\partial \mathbf{x}_t}{\partial \delta \mathbf{x}}\right|_{\mathbf{x}} = \mathbf{H}_{\mathbf{x}}\mathbf{X}_{\delta \mathbf{x}_k}$$

Here, $\boldsymbol{H}_{\boldsymbol{x}} = \left.\frac{\partial h}{\partial \boldsymbol{x}_t}\right|_{\boldsymbol{x}}$ is the Jacobian matrix of $h()$ with respect to the true state, and $\boldsymbol{X}_{\delta \boldsymbol{x}} = \left.\frac{\partial \boldsymbol{x}_t}{\partial \delta \boldsymbol{x}}\right|_{\boldsymbol{x}}$ is the Jacobian matrix of the true state with respect to the error state.

$$\mathbf{X}_{\delta \mathbf{x}} = \begin{bmatrix} \frac{\partial(\mathbf{R}(\ \delta r)}{\partial \delta r} & & & & & \\ & \frac{\partial(\mathbf{p}+\delta\mathbf{p})}{\partial \delta \mathbf{p}} & & & 0 & \\ & & \frac{\partial(\mathbf{v}+\delta\mathbf{v})}{\partial \delta \mathbf{v}} & & & \\ & & & \frac{\partial(\mathbf{b}_{\omega}+\delta\mathbf{b}_{\omega})}{\partial \mathbf{b}_{\omega}} & & \\ & 0 & & & \frac{\partial(\mathbf{b}_{\mathrm{a}}+\delta\mathbf{b}_{\mathrm{a}})}{\partial \delta \mathbf{b}_{\mathrm{a}}} & \\ & & & & & \frac{\partial(\mathbf{g}+\delta\mathbf{g})}{\partial \delta \mathbf{g}} \end{bmatrix}$$

In (20), $\frac{\partial(\mathbf{R}(\ \delta r)}{\partial \delta r} = \frac{\partial \mathrm{Log}(\mathbf{R}\mathrm{Exp}(\delta r\ ))}{\partial \delta r}$, since $\boldsymbol{\delta r}$ is the right product of R, using the right-multiplication BCH (Baker-Campbell-Hausdorff) formula, we obtain:

$$\frac{\partial \mathrm{Log}(\mathbf{R}\mathrm{Exp}(\delta r\ ))}{\partial \delta r} = \mathbf{J}_r^{-1}(\mathbf{R})$$

In (15)，a new Kalman gain formula is incorporated. The first is the classical Kalman filter gain calculation formula, and the second is a new Kalman gain calculation formula [15]. The specific selection depends on the number of rows in the H, as the computational complexity of the Kalman gain primarily arises from the inversion operation in the formula. When m ≤ 18, the first Kalman gain formula is selected, given by $\boldsymbol{HP^{-}H^{T}} \in \mathbb{R}^{m \times m}$; when m > 18, the second Kalman gain formula is selected, given by $\boldsymbol{H^{T}R^{-1}H} \in \mathbb{R}^{18 \times 18}$.

It is evident that this new Kalman gain algorithm can significantly reduce the computational complexity of the inversion operation.

3.4.4. Error-State Kalman Filter reset

Let the reset function be $g()$, then we have:

$$\delta \mathbf{x}_k^{reset} \leftarrow g(\delta \mathbf{x}_k) = \left(\mathbf{x}_k \left(\ \delta \mathbf{x}_k\right)' \ \left(\mathbf{x}_k \left(\ \widehat{\delta \mathbf{x}_k}\right)\right.\right.$$

The reset operation for the Error-State Kalman Filter is as follows:

$$\widehat{\delta \mathbf{x}}_k \leftarrow 0$$

$$\mathbf{P}_k \leftarrow \mathbf{G}\mathbf{P}_k\mathbf{G}^T$$

$$\mathbf{G} \triangleq \frac{\partial g}{\partial \delta \mathbf{x}}\bigg|_{\widehat{\delta \mathbf{x}}} = \begin{bmatrix} \mathbf{I} - \frac{1}{2}\delta r_k^{\wedge} & 0 \\ 0 & \mathbf{I}_{15} \end{bmatrix}$$

4. Experiments and analysis

To evaluate the performance of the ESKF, we conducted a series of simulations using data generated by the GNSS-INS-SIM tool. The following details the simulation setup, data characteristics, and results.

4.1. Simulation results

In this subsection, we present the experimental results of the proposed Manifold Space ESKF algorithm. The experiment used IMU and GPS data for pose estimation.

Figure 1 shows the object's trajectory in the XYZ plane, including the true path, GPS observation data, and the fusion algorithm results. The results demonstrate that the proposed algorithm meets the accuracy requirements, with no significant drift, and closely follows the true path.

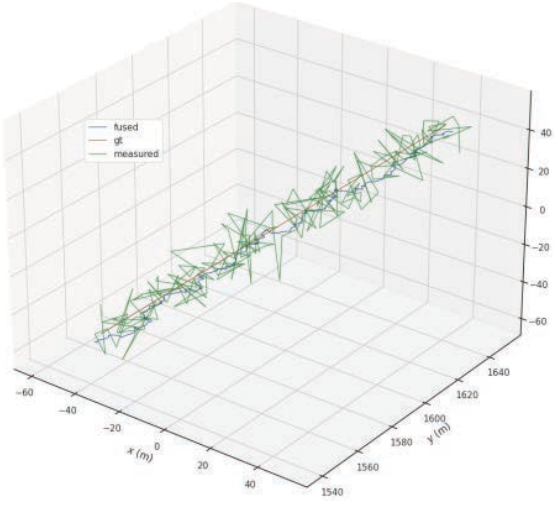


Figure 1. Object trajectory in the XYZ plane

Figure 2 displays the Euler angles corresponding to the fusion results. It can be observed that when the roll angle exceeds 90 degrees, both the pitch and yaw angles remain below 40 degrees, indicating that the proposed algorithm effectively avoids the gimbal lock issue. Figure 3 presents the XYZ three-axis position data corresponding to the trajectory.

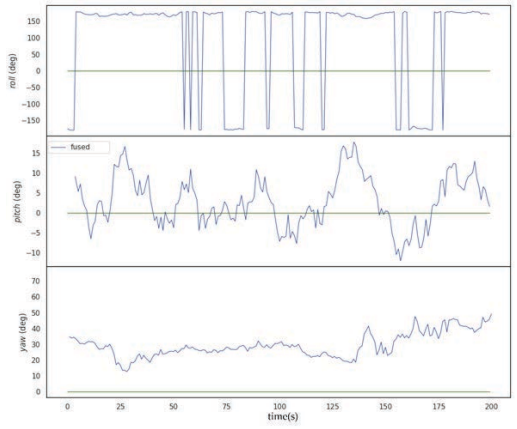


Figure 2. Orientation results for fused sensors

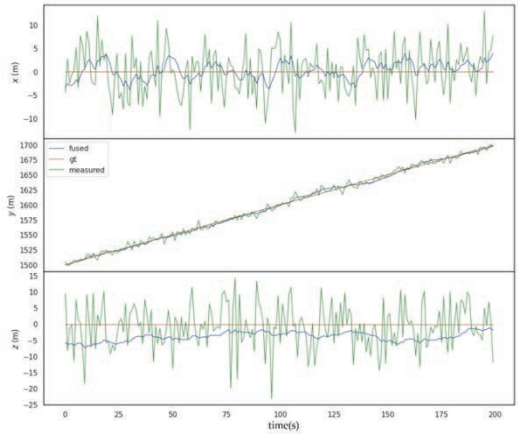


Figure 3. Position corresponding to the trajectory

4.2. Simulation analysis

The specific analysis of the experimental results is as follows: Given the low-dimensional nature of the GPS data, we utilized formula (15)(a) to compute the Kalman gain. In this comparison, we selected Quaternion-based ESKF and Multiplicative EKF for the experiment. The errors are represented by the root mean square error of the absolute pose. The computation times and errors for the three algorithms are summarized in Table 1.

Table 1. Errors and computation times for manifold space, quaternion, and multiplicative extended Kalman filter algorithms

Algorithms	Time (ms)	Error
Manifold space ESKF	79.75	3.546
Quaternion-based ESKF	95.77	4.716
Multiplicative EKF	69.26	4.710

From the results, the Manifold Space Error-State Kalman Filter algorithm demonstrates superior accuracy compared to the other two filtering algorithms. It also exhibits faster computation times than the quaternion-based Error-State Kalman Filter. Quaternions require four dimensions to describe rotation, while rotation matrices require nine dimensions. In contrast, the Manifold Space Error-State Kalman Filter only requires three dimensions to represent rotation, leading to significant memory savings.

For high-dimensional observation data, such as LiDAR image, we employed (15)(b) for Kalman gain calculation. Experimental results for different dimensionalities of images using the two algorithms are presented in Table 2.

Table 2. Computation time for different Kalman gain computation formulas

Dimension	247	618	1046	1532
Old formula (ms)	8.2	34.6	267	1974
New formula (ms)	0.08	0.19	0.49	1.54

5. Conclusion

This paper presents a pose estimation algorithm for UAVs, utilizing the ESKF to overcome the limitations of traditional methods in handling rotational states. By leveraging three-dimensional rotational vectors in the manifold space, the proposed approach achieves efficient and accurate UAV pose estimation while avoiding gimbal lock and singularity issues. Simulation results demonstrate that the ESKF with rotation vector representation significantly outperforms traditional quaternion-based methods and the multiplicative extended Kalman filter (MEKF) in terms of estimation accuracy, making it well-suited for UAV navigation and localization.

Furthermore, this paper introduces a new Kalman gain computation strategy for UAV multi-sensor fusion applications to enhance computational efficiency. For low-dimensional observation data (e.g., from GNSS), the classical Kalman gain formula offers low computational complexity. For high-dimensional observation data (e.g., visual information), the new Kalman gain formula significantly reduces computational complexity. This flexible gain calculation method effectively improves the algorithm's computational efficiency under different data conditions.

However, this study presents certain limitations in UAV sensor adaptability. The current Kalman gain computation method selects the gain formula based only on the dimensionality of the observation data, but it does not support real-time switching to LiDAR-based positioning when GPS signals suddenly become unreliable.

Funding

National Natural Science Foundation of China (Grant No. 62266045); National Science and Technology Major Project of China (No. 2022YFE0138600)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Alomari A, Phillips W, Aslam N, et al., 2017, Dynamic Fuzzy-Logic Based Path Planning for Mobility-Assisted Localization in Wireless Sensor Networks. Sensors, 17: 1904.

[2] Yurtsever E, Lambert J, Carballo A, et al., 2020, A Survey of Autonomous Driving: Common Practices and Emerging Technologies. IEEE Access, 8: 58443–58469.

[3] Sabatini AM, 2011, Estimating Three-Dimensional Orientation of Human Body Parts by Inertial/Magnetic Sensing. Sensors, 11: 1489–1525.

[4] Sola J, 2017, Quaternion Kinematics for the Error-State Kalman Filter. arXiv, https://doi.org/10.48550/arXiv.1711.02508

[5] Zhao S, Zhang H, Wang P, et al., 2021, Super Odometry: IMU-centric LiDAR-Visual-Inertial Estimator for Challenging Environments. IEEE International Conference on Robotics and Automation (ICRA), 2021: 14193–14200.

[6] Forster C, Carlone L, Dellaert F, et al., 2017, On-Manifold Preintegration for Real-Time Visual-Inertial Odometry. IEEE Transactions on Robotics, 33(1): 1–21.

[7] Qin T, Li P, Shen S, 2018, Vins-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. IEEE Transactions on Robotics, 34(4): 1004–1020.

[8] Yu Z, Zhu L, Lu G, 2021, Vins-motion: Tightly-Coupled Fusion of Vins and Motion Constraint. 2021 IEEE International Conference on Robotics and Automation (ICRA), 71(6): 5799–5810.

[9] Bloesch M, Burri M, Omari S, et al., 2017, Iterated Extended Kalman Filter Based Visual-Inertial Odometry Using Direct Photometric Feedback. International Journal of Robotics Research, 36(10): 1053–1072.

[10] Xia R, Pei H, 2020, Ranging-Aided Aerobridge Navigation Using Dual Quaternion Based Multiplicative Extended Kalman Filter. 2020 International Symposium on Autonomous Systems (ISAS), 2020: 1–6.

[11] Xiang G, 2019, 14 Lessons of Visual SLAM: From Theory to Practice (2019), Electronic Industry Press, Beijing.

[12] Hertzberg C, Wagner R, Frese U, et al., 2013, Integrating Generic Sensor Fusion Algorithms with Sound State Representations Through Encapsulation of Manifolds. Information Fusion, 14(1): 57–77.

[13] Crassidis JL, Junkins JL, 2011, Optimal Estimation of Dynamic Systems. CRC Press, New York.

[14] Xiang G, 2023, SLAM in Autonomous Driving and Robotics: From Theory to Practice, Electronic Industry Press, Beijing.

[15] Xu W, Zhang F, 2021, Fast-Lio: A Fast, Robust Lidar-Inertial Odometry Package by Tightly-Coupled Iterated Kalman Filter. IEEE Robotics and Automation Letters, 6(2): 4675587.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Design of Differential Signal Processing Circuitry for Single-Frequency Laser Interferometry Displacement Measurement

Songxiang Liu, Jingping Yan*, Can Tang*

Chongqing College of International Business and Economics, Chongqing 400000, China.

Corresponding author: Jingping Yan, yanjingping_8@163.com; Can Tang, tangcc0717@163.com

Abstract: This thesis addresses the issues existing in traditional laser tracking displacement measurement technology in the field of ultraprecision metrology by designing a differential signal processing circuit for high-precision laser interferometric displacement measurement. A stable power supply module is designed to provide low-noise voltage to the entire circuit. An analog circuit system is constructed, including key circuits such as photoelectric sensors, I-V amplification, zero adjustment, fully differential amplification, and amplitude modulation filtering. To acquire and process signals, the PMAC Acc24E3 data acquisition card is selected, which realizes phase demodulation through reversible square wave counting, inverts displacement information, and a visual interface for the host computer is designed. Experimental verification shows that the designed system achieves micrometer-level measurement accuracy within a range of 0–10mm, with a maximum measurement error of less than 1.2μm, a maximum measurement speed of 6m/s, and a resolution better than 0.158μm.

Keywords: Displacement Measurement; Weak Signal Processing; Differential Signal; Data Acquisition

Online publication: April 3, 2025

1. Introduction

The development of laser interferometric displacement measurement systems is crucial for the advancement of tracking measurement technology and is a key component of high-precision large-scale laser trackers. As a micro-nano measurement technology, its market and technical demands are increasing. Improving the accuracy of this technology is an urgent need for current scientific and industrial development and is of great significance to the national economy and military fields. Some domestic universities, such as Tsinghua University, have proposed a high-precision laser interferometric displacement measurement method that combines PLL phase-locked loop technology to double the frequency of the laser interferometric measurement signal and the reference signal, and then calculate the displacement through a counter. The results show that the resolution of

the measurement system can reach 3 nm and the measurement speed can reach 3 mm/s.

However, this method is only suitable for low-speed measurement scenarios. To overcome this limitation, Harbin Institute of Technology and ZYGO Company jointly proposed a direct frequency doubling measurement method, which can achieve a resolution of 0.62 nm and a measurement speed of 2.8 m/s [1–3]. On this basis, Harbin Institute of Technology further proposed a frequency doubling measurement method based on an external reference clock, which uses a high-frequency reference signal to count the pulses of the measurement signal and the reference signal, and then calculates the displacement change through phase demodulation [4-6]. The signal processing card UOI-3000A developed by Harbin Institute of Technology can achieve a resolution of 0.31 nm and a measurement speed of 1.5 m/s. Professor Zhang Shulian's team from Tsinghua University has studied the full-chain technology of "Ferrule-Glass Assembled Single-Frequency He-Ne Laser → Birefringent Dual-Frequency Laser → Dual-Frequency Laser Interferometer", eliminating the nonlinear measurement error that has always existed in the interferometer at the source [7]. The birefringent dual-frequency laser and interferometer have been mass-produced by Beijing LaiCe Technology Co., Ltd.

Thanh T V *et al.* proposed a displacement measurement interferometer based on a frequency-locked laser diode with a high modulation frequency [8]. By applying a high-frequency modulated LD, the frequency stability of the light source and the measurement speed of the frequency-modulated interferometer were improved. Utilizing the LD frequency locked at a high modulation frequency and the super-heterodyne transition, a stable displacement interferometer with high precision and high measurement speed was achieved, with a displacement result difference of less than 20 nm. Hao C *et al.* studied a microchip Nd:YAG dual-frequency laser interferometer with a frequency difference of 17.4 MHz and designed a down-conversion mixer circuit to reduce the beat frequency to approximately 5 MHz [9]. Experimental results showed that the displacement resolution of the microchip Nd:YAG dual-frequency laser interferometer was 10 nm and the measurement range was 500 mm. Siddiqui A A *et al.* utilized a deep neural network to achieve fringe detection and displacement sensing in self-mixing interferometry based on variable optical feedback [10]. The deep neural network was trained under variable optical feedback conditions, enabling interference fringe detection and corresponding displacement measurement. A method for automatically labeling SMI fringes under variable optical feedback conditions was proposed.

Based on the engineering background of laser interferometric tracking measurement, this paper addresses the problems of slow measurement speed, low resolution, and low accuracy in traditional technologies, especially the need to process multiple interferometric signals in real time in ultra-high-precision tracking measurement. It adopts the PMAC data acquisition card for real-time signal processing and designs a high-speed, high-resolution, and high-precision laser interferometric displacement measurement signal processing circuit to enhance the overall technical performance of the laser tracking displacement measurement system.

2. Principle of single-frequency laser interferometry

From the perspective of principle, laser interferometric displacement measurement can be divided into single-frequency laser interferometric displacement measurement and dual-frequency laser interferometric displacement measurement. Single-frequency laser interferometry is a DC measurement system. It acquires the displacement of a moving target by recording the number of cycles of the bright and dark changes in the single-frequency laser interference fringes. Its advantage is that there is no theoretical limit to the measurement

speed, but the disadvantage is that as a DC measurement system, it has relatively high requirements for the environment during displacement measurement and is easily disturbed by the environment.

To overcome the shortcomings of single-frequency laser interferometry, the heterodyne interferometry measurement method was introduced. That is, a carrier of a certain frequency is introduced into the reference optical path of the single-frequency optical path, and the measured displacement signal can be transmitted through this carrier. Heterodyne laser interferometry is an AC system that can effectively avoid the problem of easy DC drift in interferometric measurement signals [11]. At the same time, heterodyne laser interferometry has the advantages of high signal-to-noise ratio, strong anti-interference ability, fast measurement speed, and easy realization of high measurement resolution. Generally, the larger the frequency difference between the dual frequencies, the higher the measurement speed of the heterodyne laser interferometry system [12].

The schematic diagram of the single-frequency laser interferometry is shown in **Figure 1**. The light emitted by the laser passes through the neutral beam splitter (Beam Splitter, BS) and is divided into two beams of light. One beam of light serves as the transmitted light and is transmitted horizontally towards the measurement mirror M, while the other beam of light serves as the reflected light and is reflected vertically towards the reference mirror R. The two beams of light reaching the measurement mirror and the reference mirror are reflected back. Then, the measurement mirror M is moved, and when the two reflected beams of light pass through the neutral beam splitter BS again, the two reflected beams of light will interfere at the neutral beam splitter. The interference light is detected by the photodetector. When the measurement mirror is moved, the interference fringes of the interference light will change in brightness and darkness, and the photodetector will detect a current signal with a periodically changing amplitude. By calculating the number of cycles of the amplitude change of the current signal, the displacement of the measurement mirror can be obtained.

This paper takes the Michelson interferometer as the basic optical path. To ensure real-time tracking of the target's displacement, the measurement mirror is often replaced by a target ball, and the emission light and the receiving light are coaxial [13, 14]. In addition, to improve the resolution of the single-frequency laser interferometry displacement measurement system, a wave plate is used in the optical path system to adjust the polarization direction of the light and obtain an optical subdivision signal.

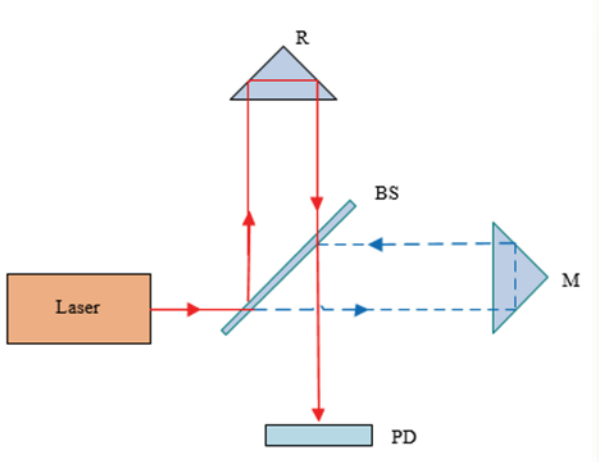


Figure 1. Schematic diagram of single-frequency laser interferometry principle.

3. Hardware circuit design

In the circuit design, a stable power supply module is first built to ensure that low noise voltage is provided, followed by photoelectric conversion, I-V amplification, zero adjustment, full differential amplification and amplitude modulation filtering. The overall structure of the circuit is shown in **Figure 2**, which is composed of a power module, a pre-amplification circuit, a signal amplification circuit, and an amplitude-modulation filter circuit. Take a pair of differential signals as an example, the signal is converted into a current signal through the photoelectric sensor, and then amplified by a first-level I-V amplifier circuit, into the full differential amplifier circuit, to achieve the zero function, the maximum range of zero adjustment is ± 1V, and then through the amplitude modulation filter circuit, and finally the signal output, the power module uses 12 V power supply.

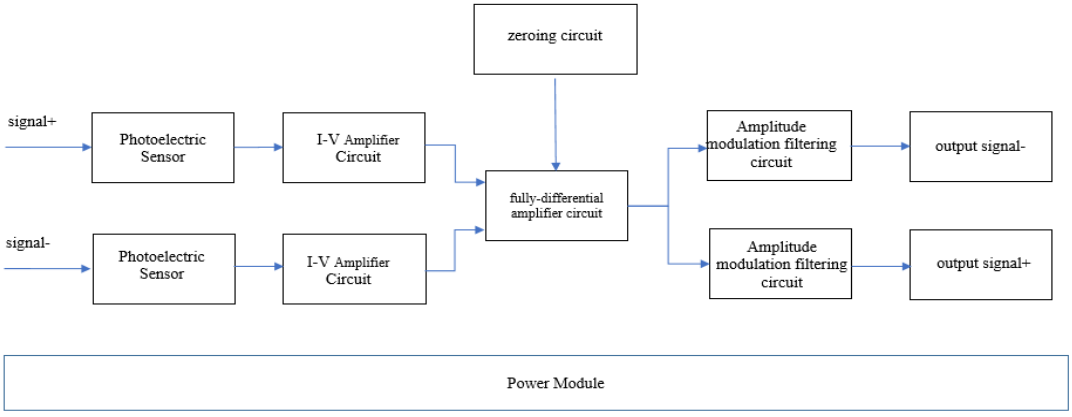


Figure 2. System structure diagram

The signal to be measured is four positive (residual) strings, phase difference 90° optical differential signal, optical power (30μW–60μW), the conversion efficiency of the photodetector is only 30%–40%, so the current signal obtained is only 10–20μA, the four-way signal amplitude (DC, AC) is approximately equal, and the input signal frequency DC 10MHz. The circuit includes photoelectric conversion module, differential amplification module and amplitude modulation tuning module, and finally outputs four positive (residual) string analog signals in real time, with a difference of 90° consistent with the input. The four signals are set with amplitude modulation tuning function, and the peak value of the four output voltage signals is 1 V (0 –1 V).

3.1. I-V transresistance amplifier circuit

The I-V amplifier circuit is used to convert the tiny current signal output by the photoelectric sensor S3072 into a voltage signal. Because the photoelectric signal is a current signal, the current input amplifier must be selected. The OPA847 is a wide-band, ultra-low noise operational amplifier with a piezoswing rate of up to 950 V/us to meet the needs of systems for fast amplification of high frequency signals. To reduce the influence of power supply noise, 0.1uF and 6.8uF tantalum capacitors are connected in parallel at ± 5V of the power supply for power supply filtering. The weak current signal is converted into a voltage signal through the input resistance and then amplified. In order to reduce the influence of the input resistance on the signal, the value of the input resistance must be as small as possible. By using the negative feedback of the amplifier to reduce the input impedance, the pre-amplifier circuit of the current input is realized. The design circuit is shown in **Figure 3**.

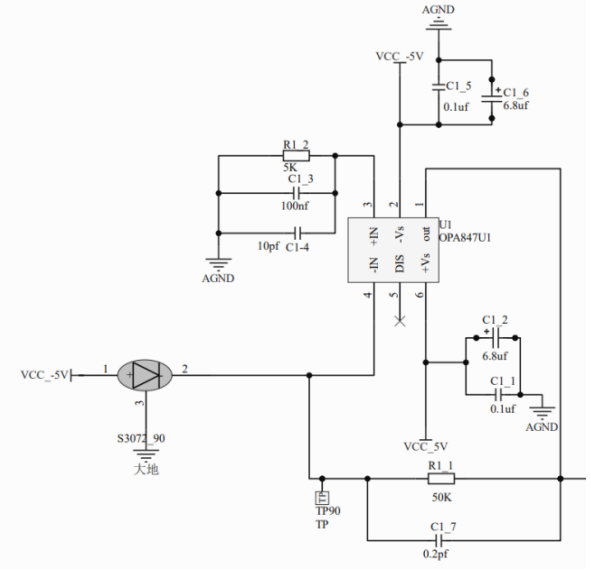


Figure 3. I-V amplifier circuit

In the circuit, the larger the feedback resistance R1_1, the more helpful to improve the signal-to-noise ratio, but too large feedback resistance will make the amplifier input conversion noise current value can not be ignored. Considering the frequency bandwidth and signal-to-noise ratio of the system, the feedback resistance value is selected as 50 kΩ. The transresistance gain of the OPA847 requires the feedback extreme value to be set to 74 MHz to obtain a nominal Butterworth frequency response design. This requires a total feedback capacitance of 0.2 pF. Therefore, the design does not require additional capacitors.

3.2. Fully differential amplifying circuit and zeroing circuit

After the I-V amplifier circuit converts the current signal into the voltage signal and amplifies it, the differential signal needs to pass through the second stage circuit to achieve further amplification and realize the circuit zero function. In this circuit, two differential signals are amplified twice by the fully differential signal, and a fully differential amplifying zeroing circuit as shown in **Figure 4** is designed by using the output common-mode control pin of the chip.

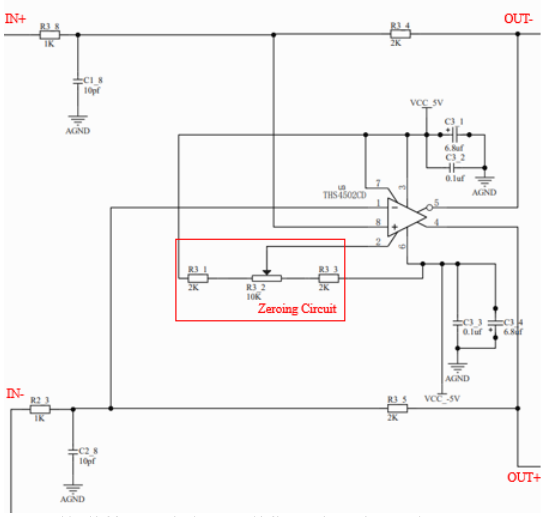


Figure 4. Full differential amplifier circuit and zero control circuit

The output waveform is adjusted by adjusting the resistance value of the sliding rheostat R3_2, and the function of zeroing the differential signal is realized. In order to reduce the noise caused by the power module, 0.1uF and 6.8uF capacitors are connected in parallel to the power supply part to filter the power supply noise. Using TI's full differential amplifier THS4502, this type of chip has the characteristics of ultra-low voltage noise, high-speed voltage feedback, etc., which is very suitable for application in low noise circuit.

3.3. Amplitude modulation filter circuit

To realize the amplitude-modulation function of the circuit and improve the quality of the output signal, the amplitude-modulation filter circuit is designed, and the circuit design is shown in **Figure 5**. The input signal IN is adjusted to the amplitude of the input signal by resistors R3_6 and R4_1. At the same time, the second-order active low-pass Butterworth filter is composed of resistors R4_ and R4_2 and capacitors C4_2 and C4_1. After filtering the signal, the signal is amplified by 2 times. More possibilities for back-end impedance matching.

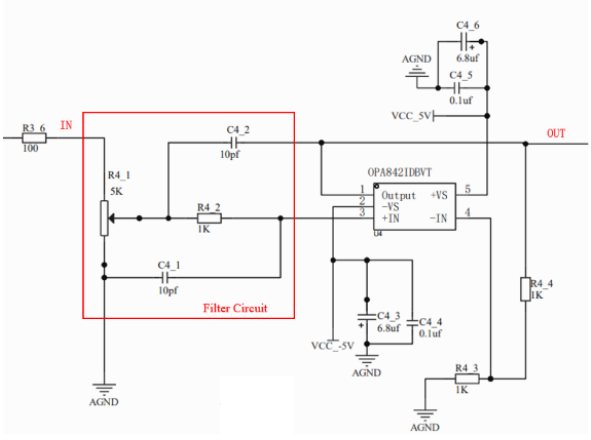


Figure 5. amplitude-modulation filter circuit

3.4. Power circuit

The power supply part mainly protects the safety of the circuit to avoid the burning of the circuit board due to problems such as power reverse connection and input voltage mismatch. First, the DC/DC power module UWE1212S-3WR3 is used to convert the 12 V voltage provided by the power supply to ±12 V. Then through the LDO power module ADP7182, +12 V is converted into low noise +5 V, and through the LDO power module ADP7182, -12 V is converted into low noise -5 V for the use of analog circuit.

4. Digital signal processing

The design of digital signal processing software generally includes two modules: PMAC data processing module and upper computer visual interface design. The digital signal transmission process is shown in **Figure 6**. PMAC Acc24E3 digital acquisition card is used to collect TTL signal of laser interference displacement measurement system, and the corresponding digital quantity of displacement information is obtained directly. The PMAC Acc24E3 digital acquisition card provides a digital quantity of distance information, which is then converted into a digital quantity of the corresponding physical unit in the Power PMAC system. The data processing module is used to collect, compare and output the signal output of the pre-processing circuit. The

principle of the PMAC data acquisition card is mainly to convert the analog signal into the digital signal that can be processed by the computer.

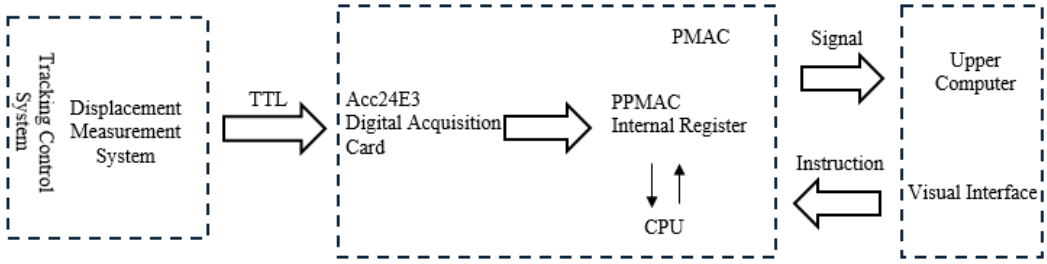


Figure 6. Digital signal transmission process

In this paper, the displacement signal is collected and processed by the displacement measurement system of the tracking control system, and the analog signal carrying displacement information is obtained. The output analog signal is TTL signal with peak-to-peak value of 1 Vpp. Then the TTL signal is sent to the analog-to-digital converter (ADC) in the PMAC data acquisition card, and the ADC converts the TTL signal into a digital signal. The converted digital signal also needs to carry out reversible square wave counting and directional subdivision through the Acc24E3 digital acquisition card in the PMAC, and demodulate the digital signal carrying displacement information. Finally, the demodulated digital signal is stored and processed through PPMAC internal register, and then the processed data is transmitted to the upper computer through the USB interface of the data acquisition card, and the upper computer analyzes and displays the data.

5. Experimental results and analysis

An optical system experiment platform was built according to the design scheme in this paper, and the function and accuracy of the system were tested. During the test, the test platform worked normally and could collect optical signals containing position information and solve the position information through the signal processing circuit. In order to verify the measurement accuracy of the laser tracking displacement measurement system, the displacement accuracy experiment was carried out, and the displacement measurement data of the laser interferometer displacement measurement system designed in this paper was compared with the XL-80 laser interferometer of Renishau Company of Britain. The length measurement resolution of XL-80 is 1 nm, which has high measurement accuracy and measurement stability. Therefore, this paper chooses XL-80 laser interferometer of Renishaw Company to carry out displacement experiment verification and comparison experiment.

The displacement measurement accuracy is compared between Renisau XL-80 and the common optical path of the laser interference displacement measurement test platform. The manual displacement table is given 10 displacements, and the experimental test is carried out in the range of 0–10mm with an interval of 1mm, and the uniform movement of the manual displacement table is guaranteed. By comparing the indicated value of the Renishao XL-80 laser interferometer with the test platform of the laser interferometer displacement measurement system designed in this paper, the readings were collected and recorded under the given displacement. The results are shown in **Table 1**.

Table 1. Comparative measurement results of single frequency laser interferometry

Times	XL80 Value (mm)	Test platform to collect readings (mm)	After the compensation data (mm)	Compare the error after compensation (mm)
1	0	0	0	0
2	0.9908	0.9804	0.9900	0.0008
3	1.9950	1.9781	1.9955	-0.0005
4	2.9961	2.9715	2.9967	-0.0006
5	3.9945	3.9871	3.9950	0.0005
6	5.0206	4.9755	5.0210	-0.0004
7	6.0239	5.9842	6.0227	0.0012
8	7.0452	6.9790	7.0459	-0.0007
9	8.0553	7.9773	8.0550	0.0003
10	9.0625	8.9743	9.0627	-0.0002
11	10.0712	9.9730	10.0712	0

After linear compensation, the comparison error between XL-80 laser interferometer and the laser interferometer displacement measurement test platform designed in this paper at a single step is less than 1.2μm, and the accuracy test diagram of the test system is shown in Figure 7.

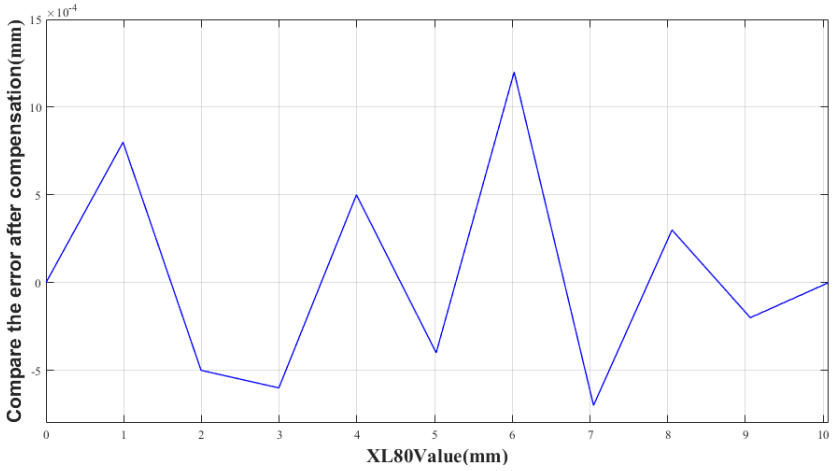


Figure 7. Accuracy test

Through Matlab data analysis of the experimental measurement data in **Table 1**, **Figure 8** shows the stability curve of the displacement measurement data obtained through data analysis of five displacement measurements of the same range. It can be seen from **Figure 7** that the maximum error of each measurement value is no more than 1.52 microns for multiple measurements of the same range. The measuring system has good measuring stability.

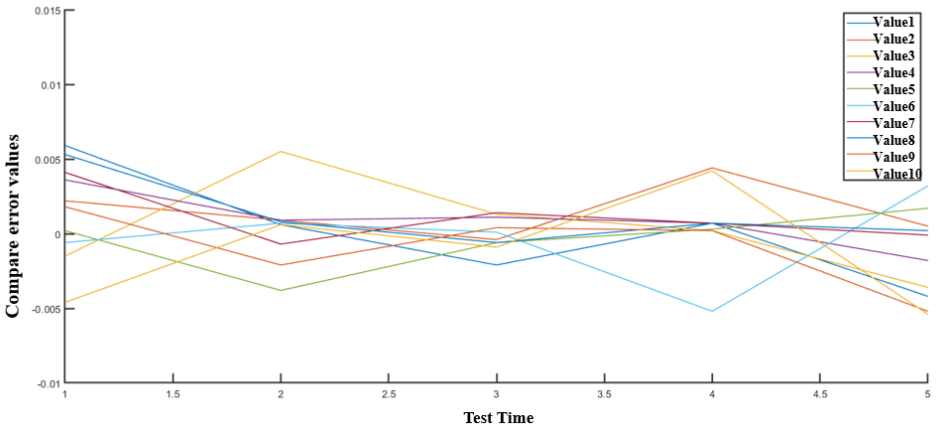


Figure 8. Stability curve

6. Conclusion

From this study, the displacement signal pre-processing circuit of the laser interference displacement measurement system is designed, including current and voltage conversion, I-V amplification, zeroing, full difference amplification and amplitude modulation filtering, etc., for laser interference signals. Additionally, with PMAC data acquisition card as the core, the digital signal processing module is designed, and the pre-processed analog signal is sent to PMAC for collection and analysis. The PPMAC Acc24E3 card is used to obtain the digital quantity of displacement information directly, and it is converted into the digital quantity of physical unit in the Power PMAC system, and displayed through the upper computer interface. The experimental verification was carried out, and the system performance was tested and optimized. The maximum error of single-frequency laser interference displacement measurement was no more than 1.2μm, the measurement speed was up to 6m/s, and the system resolution was 0.158μm.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Hao C, Shulian Z, 2021, Microchip Nd:YAG Dual-Frequency Laser Interferometer for Displacement Measurement. Optics Express, 29(4): 6248–6256.

[2] Siddiqui AA, Zabit U, Bernal OD, 2022, Fringe Detection and Displacement Sensing for Variable Optical Feedback-Based Self-Mixing Interferometry by Using Deep Neural Networks. Sensors (Basel), 22(24): 9831–9834.

[3] Shi Q, 2022, Design of Signal Processing System for High-Resolution Dual-Frequency Laser Interferometer, thesis, Sichuan University.

[4] Hu X, 2022, Orthogonal Phase Demodulation Displacement Measurement Technology Based on Dual Longitudinal Mode Laser Self-Mixing Interference, thesis, Harbin Institute of Technology.

[5] Yang Q, Chen L, Guo D, 2022, Two-Dimensional Dynamic Displacement Measurement Based on Frequency

Division Multiplexing Technology and Laser Feedback Interference. Acta Optica Sinica, 42(10): 72–78.

[6] Zhang S, Guo H, 2020, Design of Four-Degree-of-Freedom Synchronous Heterodyne Interferometry Measurement System. Optical Instruments, 4(2): 75–81.

[7] Zhang S, 2023, Key and Full-Chain Technologies of Birefringent Dual-Frequency Laser and Interferometer. Acta Optica Sinica, 43(01): 189–198.

[8] Vu TT, Hoang HH, Vu TT, et al., 2020, A Displacement Measuring Interferometer Based on a Frequency-Locked Laser Diode with High Modulation Frequency. Applied Sciences, 10(8): 396–399.

[9] Hao C, Shulian Z, 2021, Microchip Nd:YAG Dual-Frequency Laser Interferometer for Displacement Measurement. Optics Express, 29(4): 6248–6256.

[10] Siddiqui AA, Zabit U, Bernal OD, 2022, Fringe Detection and Displacement Sensing for Variable Optical Feedback-Based Self-Mixing Interferometry by Using Deep Neural Networks. Sensors (Basel), 22(24): 9831–9834.

[11] Yang W, Liu Y, He M, 2023, Research on Signal Crossover Error and Compensation Method in Heterodyne Interferometry Phase Measurement. Chinese Journal of Lasers, 50(10): 83–94.

[12] Gui J, Sun M, 2023, Research on High-Resolution Displacement Measurement Based on PLC Laser Interferometry. Laser Journal, 44(06): 235–239.

[13] Li D, Wang X, Xu J, 2020, Design and Research of Cable Force Measurement System Based on Laser Doppler. Application of Electronic Technique, 46(1): 796–799.

[14] Zhang Z, 2022, Design of Signal Processing Board Card for Dual-Frequency Laser Interferometry System, thesis, Guilin University of Electronic Technology.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Innovative Research on the Integration of Big Data Technology in Poverty Recurrence Monitoring and Agricultural Product Sales Systems

Yuxin Jiang, Tingting Li*, Xinyi Liu

School of Economics and Management, Dalian University of Science and Technology, Dalian 116000, Liaoning, China

Corresponding author: Tingting Li, ltt19900521@dlust.edu.cn

Abstract: With the advancement of the rural revitalization strategy, preventing poverty recurrence among previously impoverished populations has become a crucial social concern. The application of big data technology in poverty recurrence monitoring and agricultural product sales systems can effectively enhance precise identification and early warning capabilities, promoting the sustainable development of rural economies. This paper explores the application of big data technology in poverty recurrence monitoring, analyzes its innovative integration with agricultural product sales systems, and proposes an intelligent monitoring and sales platform model based on big data, aiming to provide a reference for relevant policy formulation.

Keywords: Big data technology; Poverty recurrence monitoring; Agricultural product sales; Intelligent early warning; Rural revitalization

Online publication: April 3, 2025

1. Introduction

In recent years, the government has implemented a series of targeted poverty alleviation policies, enabling a large number of impoverished populations to successfully escape poverty. However, some previously impoverished households still face the risk of poverty recurrence. Factors such as market fluctuations, natural disasters, and health issues may lead to a decline in income levels. Therefore, establishing an efficient poverty recurrence monitoring system is crucial for consolidating poverty alleviation achievements and achieving rural revitalization.

At the same time, agricultural product sales play a key role in increasing farmers' income and are closely related to poverty prevention efforts. However, the current agricultural product sales system still faces challenges such as information asymmetry, limited sales channels, and high distribution costs. These issues result in unsold or underpriced agricultural products, leading to income declines for some farmers and exacerbating the risk of poverty recurrence. Enhancing agricultural product circulation efficiency and increasing farmers' income have

become urgent issues that need to be addressed.

The rapid development of big data technology presents new opportunities for precise monitoring and intelligent sales. Through big data analysis, governments can establish accurate poverty recurrence monitoring systems, enabling dynamic identification and early warning. Simultaneously, agricultural product sales can leverage big data to optimize supply-demand matching, enhance market responsiveness, and drive industrial upgrading. This paper explores the application of big data technology in poverty recurrence monitoring, analyzes its innovative integration with agricultural product sales systems, and proposes an intelligent monitoring and sales platform model based on big data, aiming to provide a reference for policy formulation.

The main contributions of this paper are as follows:

(1) A systematic review of the bottlenecks in poverty recurrence monitoring and agricultural product sales systems;

(2) An analysis of the innovative applications of big data technology in poverty risk identification, early warning mechanisms, targeted assistance, and intelligent sales;

(3) The proposal of a collaborative platform model for poverty recurrence monitoring and agricultural product sales based on big data, along with practical policy recommendations.

To better understand the current challenges and future trends in poverty recurrence monitoring and agricultural product sales systems, the next section will analyze the existing bottlenecks and explore potential development directions.

2. Bottlenecks and development trends in poverty recurrence monitoring and agricultural product sales systems

2.1. Bottlenecks in poverty recurrence monitoring and agricultural product sales systems

Despite efforts by local governments to strengthen poverty prevention through dynamic monitoring and policy coordination, the current system still faces numerous challenges. In some regions, the lack of public health awareness results in a high risk of poverty recurrence due to illness. Additionally, outdated data collection methods, delayed updates, and severe information silos hinder the precision of policy implementation.

Agricultural product sales also encounter significant issues. Many farmers lack market forecasting abilities, leading to blind planting or breeding decisions, where the paradox of “difficult selling” and “high purchasing costs” coexist. Furthermore, the complex distribution process, high logistics costs, limited sales channels, and weak brand development make it difficult to establish stable market competitiveness.

2.2. Development trends in poverty recurrence monitoring and agricultural product sales systems

The *2021 No. 1 Central Document* emphasized the use of digital tools to enhance rural governance efficiency [1]. In the future, poverty recurrence monitoring will become more intelligent, leveraging big data and AI for precise early warnings, dynamic identification of high-risk groups, and personalized assistance programs.

Agricultural product sales will advance toward digitalization and diversification, with emerging models such as e-commerce and livestream selling reducing intermediaries and increasing farmers’ income (**Figure 1**). Meanwhile, blockchain technology can establish a reliable traceability system, strengthening consumer trust and promoting agricultural branding.

Furthermore, the coordinated development of poverty recurrence monitoring and agricultural product sales

systems will create an integrated industry cycle. By leveraging big data to match supply and demand, market structures can be optimized, reducing the risk of unsold products. Governments and financial institutions can use data-driven insights to support agricultural loans, insurance, and other financial services, enhancing farmers' resilience to risks and accelerating rural industrial upgrades.

The deep application of big data is reshaping both poverty recurrence monitoring and agricultural product sales systems. The following section will explore its specific applications in poverty recurrence monitoring.

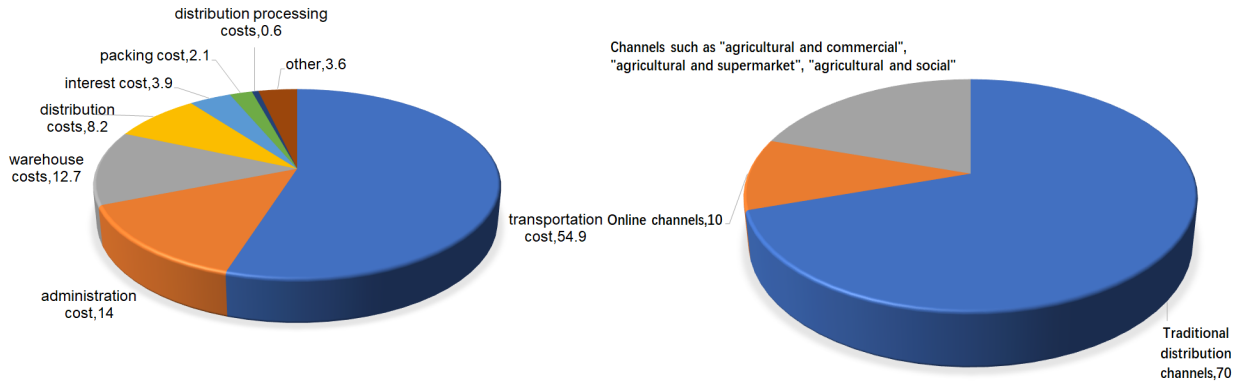


Figure 1. Analysis chart of agricultural products circulation cost composition and distribution of circulation channels (unit: %)

3. Application of big data technology in poverty recurrence monitoring

3.1. Data collection and integration

Poverty recurrence monitoring relies on multidimensional data, including government assistance records, household income, market prices, and meteorological data. Big data technology can integrate multi-source heterogeneous data, enhancing the accuracy and timeliness of monitoring. For example, Guizhou Province has utilized blockchain technology to break down data barriers between the poverty alleviation office, banks, and insurance institutions, ensuring transparency in credit records and medical reimbursement data while reducing fraud risks [2].

3.2. Precision assistance decision support

Big data analytics enable governments to implement personalized assistance measures, such as low-interest loans, skills training, and market expansion. In 2023, a county in Hebei Province discovered that impoverished households relied more on animal husbandry. Consequently, it adjusted support funds to establish a cattle and sheep epidemic prevention center, benefiting 2,000 farming households [3]. Data-driven precision assistance helps enhance the resilience of poverty-alleviated groups against risks.

3.3. Geospatial analysis

GIS technology combined with big data can accurately identify high-risk areas for poverty recurrence. For instance, overlaying historical disaster data with poverty distribution maps can pinpoint regions vulnerable to natural disasters. Additionally, analyzing the distance between impoverished households and public facilities helps identify "service blind spots," providing a basis for infrastructure development.

The application of big data technology has significantly enhanced the intelligence level of poverty recurrence

monitoring. Next, we will explore its innovative applications in agricultural product sales systems to further promote rural economic development.

4. Innovations in agricultural product sales systems

4.1. Big data-driven market matching

Big data analysis can optimize supply-demand matching for agricultural products, reducing the risk of unsold goods. For example, Inspur Cloud's "HaiRuo Agricultural Model" in Beijing integrates data to provide optimized recommendations for the fishing industry, increasing annual output value by over 20% [4]. This intelligent matching improves sales efficiency and reduces market risks.

4.2. Blockchain technology for traceability assurance

Blockchain technology ensures agricultural product quality and traceability, enhancing consumer trust. JD Farm collaborated with Wuchang Rice to launch a traceability system where consumers can scan a QR code to access product information, resulting in a 90% decrease in counterfeit complaints [5]. This strengthens brand credibility and product management standards.

4.3. Integration of e-commerce and livestream selling

Livestream e-commerce has emerged as a new sales model, shortening distribution channels and increasing profits. For example, Kuaishou's "Village Broadcast Program" trains farmers as livestream hosts. A fruit farmer in Sichuan sold mangoes via livestream, achieving a single-day transaction volume of 1.2 million yuan [4]. Livestreaming expands sales channels and enhances farmers' market adaptability.

The combination of big data, blockchain, and e-commerce models is driving innovations in agricultural product sales. Next, we will discuss policy recommendations to accelerate the adoption of these technologies.

5. Policy recommendations for big data-driven poverty recurrence monitoring and agricultural product sales systems

5.1. Technology promotion, talent development, and data-sharing mechanisms

The government should increase efforts to promote the application of big data technology in poverty recurrence monitoring in impoverished areas. Special funds should be allocated for the development of intelligent analysis and early warning systems, including both hardware and software construction. Additionally, training programs should be organized for grassroots poverty alleviation officials and relevant personnel to enhance their proficiency in big data tools, improving monitoring accuracy and efficiency.

Furthermore, efforts should be made to build cross-departmental and cross-regional data-sharing platforms that integrate information from civil affairs, agriculture, and meteorology departments, providing comprehensive support for poverty recurrence monitoring. Collaboration between the government, social institutions, and research institutions should be strengthened to conduct joint studies on poverty recurrence risks and optimize monitoring models, driving continuous innovation and application of big data technology.

5.2. Technology support, platform development, brand building, and market expansion

For enterprises and farmers leveraging big data and blockchain technology to improve agricultural product

sales, the government should provide policy subsidies and tax incentives to encourage the adoption of advanced technologies. Support should be given to the development of a unified agricultural e-commerce platform, the standardization of livestream selling processes, and the provision of technical training and livestreaming guidance for farmers to reduce operational costs and enhance sales effectiveness.

Additionally, a dedicated fund for agricultural product branding should be established to help farmers and enterprises use big data analytics for precise market positioning and brand development. Collaboration with e-commerce platforms and retail chains should be strengthened to expand sales channels. Organizing agricultural product trade fairs can also increase market visibility and competitiveness.

By strengthening technology promotion, talent development, and policy support, the adoption of big data technology in poverty recurrence monitoring and agricultural product sales systems can be accelerated, providing strong support for rural revitalization and poverty alleviation efforts.

6. Conclusion and outlook

The integration of big data and blockchain technology provides innovative solutions for poverty recurrence monitoring and agricultural product sales systems, effectively addressing the issues of delayed monitoring and inefficient agricultural distribution. By establishing a reliable data ecosystem, these technologies can reduce rural financial risk premiums by 20–30% while increasing the price premium of agricultural products by over 15%. The application of big data technology has facilitated the intelligent transformation of targeted poverty alleviation and agricultural product sales, further advancing the implementation of the rural revitalization strategy.

Looking ahead, with the continuous development of emerging technologies such as artificial intelligence and blockchain, poverty recurrence monitoring and agricultural product sales systems will become even more intelligent and efficient. Governments should strengthen policy support, encourage active participation from technology enterprises and research institutions, and promote public-private partnerships to create a sustainable and collaborative ecosystem. Through multi-stakeholder cooperation, rural economies are expected to achieve higher-quality growth, providing strong support for consolidating poverty alleviation achievements and advancing rural revitalization.

Funding

2025 College Students' Innovation Training Program "Return to Poverty Monitoring and Agricultural Products Sales System"; 2024 College Students' Innovation Training Program "Promoting Straw Recycling to Accelerate the Sustainable Development of Agriculture" (202413207010)

Disclosure statement

The authors declare no conflict of interest.

References

[1] The State Council, 2024, The No.1 Central Document in 2024, People's Publishing House, Beijing.

[2] Wang W, Li M, 2022, Research on the Application of Blockchain Technology in Agricultural Products Supply Chain.

Agricultural Economy Problems, 43(5): 89–97.

[3] Zhang H, Liu Y, 2021, Big Data-Driven Early Warning Mechanism of Returning to Poverty. China's Rural Economy, 39(8): 45–56.

[4] Huang X, Li N, 2021, Research on Big Data-Driven Resilience Improvement Strategy of Rural Economy. Research on Agricultural Modernization, 42(4): 23–30.

[5] Li Q, Wang F, 2022, Design and Implementation of Agricultural Product Traceability System based on Blockchain. Computer Application Research, 39(10): 3012–3018.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Research on Constructing Personalized Learner Profiles Based on Multi-Feature Fusion

Xing Pan[1], Meixiu Lu[2]*

[1]Center for Contemporary Education Technology, Guangdong University of Foreign Studies, Guangzhou, China
[2]School of Information Science and Technology (School of Cyber Security), Guangdong University of Foreign Studies, Guangzhou, China

***Corresponding author:** Meixiu Lu, meixiulu@mail.gdufs.edu.cn

Abstract: This study proposes a learner profile framework based on multi-feature fusion, aiming to enhance the precision of personalized learning recommendations by integrating learners' static attributes (e.g., demographic data and historical academic performance) with dynamic behavioral patterns (e.g., real-time interactions and evolving interests over time). The research employs Term Frequency-Inverse Document Frequency (TF-IDF) for semantic feature extraction, integrates the Analytic Hierarchy Process (AHP) for feature weighting, and introduces a time decay function inspired by Newton's law of cooling to dynamically model changes in learners' interests. Empirical results demonstrate that this framework effectively captures the dynamic evolution of learners' behaviors and provides context-aware learning resource recommendations. The study introduces a novel paradigm for learner modeling in educational technology, combining methodological innovation with a scalable technical architecture, thereby laying a foundation for the development of adaptive learning systems.

Keywords: Learner profile; Multi-feature fusion; Dynamic features; Personalized recommendation; Educational technology

Online publication: April 3, 2025

1. Introduction

1.1. Research background

With the rapid advancement of the Internet and digital technologies, online learning platforms have gained widespread popularity, enabling learners to access vast educational resources anytime, anywhere. However, this abundance of resources also poses challenges: learners often face information overload when attempting to identify content tailored to their specific needs. Amid the growing demand for personalized learning, efficiently matching resources to learners' requirements has emerged as a critical issue in educational technology.

Learner profiling, an extension of user profiling in the educational domain, involves analyzing learners' behaviors, interests, and habits to construct individualized models that support precise educational services.

Leveraging these profiles, personalized learning plans, resource recommendations, and tutoring can be delivered, thereby enhancing learning efficiency and experience. Nevertheless, existing research predominantly focuses on static features–such as age and academic grades–while overlooking the dynamic nature of learners' interests and behaviors. This limitation hampers the ability of profile models to adapt to real-time shifts in learning needs.

1.2. Research questions and objectives

Current learner profile studies tend to emphasize the extraction and analysis of static features, neglecting the dynamic changes in learners' interests and behaviors. Moreover, existing methods lack systematic and flexible approaches to feature extraction and weighting, making it difficult to adapt to diverse learning contexts and demands. This study addresses the following core research questions:

(1) How can learners' static attributes and dynamic behavioral patterns be integrated to construct accurate learner profiles?

(2) How can a dynamic feature update mechanism be designed to adaptively capture real-time changes in learners' interests?

(3) How can the effectiveness and practical value of this profile framework be validated in the context of personalized learning recommendations?

To tackle these questions, this research proposes a multi-feature fusion learner profile framework that combines static and dynamic features using TF-IDF, AHP, and a time decay function. The framework aims to create precise, real-time learner models, offering technical support for adaptive learning systems. Compared to traditional approaches, it not only accounts for persistent characteristics but also dynamically tracks interest evolution, injecting fresh vitality into personalized education.

2. Related work

2.1. Research on learner profiles

Learner profiling is an extension of user profiling techniques to the educational domain, aiming to construct personalized models by collecting and analyzing multidimensional data about learners. These models reflect learners' current states and predict their future learning trajectories, thereby supporting personalized instruction and resource recommendations. Existing research has explored the construction and application of learner profiles from various perspectives. For instance, Xiao *et al*. developed high-risk learner profiles for online learners based on basic characteristics, online learning behaviors, and learning pathways, using tags to identify potential dropout risks [1]. Wang *et al*. employed a bidirectional long short-term memory network (Bi-LSTM) with an attention mechanism for sentiment analysis, constructing learner profiles that include basic information, behavioral patterns, and textual interactions to predict potential learning needs [2]. These studies highlight the potential of profiling technology in education but often focus on static feature analysis, with less attention to the dynamic changes in learners' interests and behaviors.

2.2. Technologies related to user profiling

Before the advent of deep learning, user profiling relied heavily on traditional techniques such as regression analysis, clustering, and predictive modeling. For example, Schroeder *et al*. used k-means clustering to study the impact of group membership on learning transfer test scores [3], while Piech *et al*. applied recurrent neural

networks (RNNs) to predict learners' cognitive levels [4]. In terms of visualization, methods like statistical mapping, text visualization, and human-computer interaction were widely adopted [5]. With the development of big data and artificial intelligence, capabilities for data processing and analysis have significantly improved, enabling more refined user profiles. Shao *et al*. proposed a user profile generation method based on multi-granularity information fusion (UP-MGIF), integrating bidirectional gated recurrent units (Bi-GRU), denoising autoencoders (DAE), and attention mechanisms to achieve feature denoising and semantic enhancement, resulting in robust user profiles [6]. These advancements indicate that model-driven profiling methods offer greater flexibility and accuracy in behavior prediction compared to traditional statistical approaches.

2.3. Personalized recommendation systems

Personalized recommendation systems have gained increasing attention in education, especially those integrating learner profiles. Ban *et al.* proposed a Knowledge and Personality Incorporated Multi-Task Learning Framework (KPM) to facilitate course recommendations [7]. Wang *et al.* proposed a personalized course recommendation method based on learner profiles. Quantitative analysis was performed on learners' learning data, with a particular focus on emotional expression, where personalized features are most evident. A bidirectional, extended short-term memory network based on an attention mechanism was utilized for sentiment analysis, thereby constructing a learner profile feature model that includes three dimensions: basic learner information, behavior, and bullet comment text [8].

Chen D *et al.* are committed to enhancing recommendation systems' personalization capture and dynamic interest modeling capabilities. Considering the usefulness of different features, they proposed a hierarchical description-aware personalized recommendation (DAPR) algorithm [9]. Zhong *et al*. has researched a personalized recommendation system for student portraits based on deep hashing algorithms to improve the recommendation effect [10]. However, existing recommendation systems still need enhancements in real-time performance and robustness to meet the diverse needs of learners.

2.4. Contributions of this study

Despite these advancements, existing user-profiling technologies in education face several shortcomings:

(1) Lack of dynamism: Current methods primarily focus on mining user behavior data, often ignoring the dynamic evolution of learners' interests.

(2) Inconsistent feature extraction and weighting: Existing approaches lack a unified framework for feature extraction and weighting, hindering automated and intelligent feature processing.

(3) Insufficient real-time performance and robustness: Current methods require improvement in dynamic feature updating and real-time recommendations to adapt to complex learning scenarios.

To address these gaps, this study introduces a multi-feature fusion learner profile framework with the following innovations:

Integration of static and dynamic features to construct comprehensive learner profiles.

Utilization of TF-IDF, AHP, and a time decay mechanism to enable adaptive updates of dynamic features.

Empirical validation of the model's effectiveness in personalized recommendations, enhancing its applicability in educational contexts.

3. Model construction

This section elaborates on the construction of the learner profile model, comprising two primary modules: static profile representation and dynamic feature generation. The static profile captures persistent learner characteristics, while dynamic features reflect time-varying behaviors and interests. By incorporating time decay and dynamic updating mechanisms, the model generates accurate, real-time learner profiles.

3.1. Profile representation

The learner profile P is defined as a combination of a static attribute tag set S and a dynamic attribute tag set D:

$$P = (S,D)$$

Static attribute tags S: These represent enduring learner characteristics, such as demographic information (e.g., age, gender) and historical academic performance (e.g., grades, course completion). Formally expressed as:

$$S = (s_1,w_{s1}),(s_1,w_2),\ldots(s_1,w_{sm})$$

Where s_i denotes a static tag (e.g., "age: 20" or "major: computer science"), and w_{si} is its corresponding weight, typically determined by statistical data or expert assignment.

Dynamic attribute tags D: These reflect short-term behaviors and interests, such as recently viewed content or current focus areas. Formally expressed as:

Where d_i represents a dynamic tag (e.g., "recently viewed: machine learning"), and w_{di} is its weight, computed through the dynamic feature generation process.

For example, a learner profile might be represented as:

$$S = (\text{age:20},0.8),(\text{major: computer science},1.0)$$

$$D = (\text{machine learing},0.9),(\text{data structures},0.6$$

3.2. Dynamic feature generation

The objective of dynamic feature generation is to extract and update the dynamic tags D from learners' behavioral data. This process involves three key steps: candidate feature acquisition, time decay function design, and dynamic feature updating.

3.2.1. Candidate dynamic feature acquisition

A complete interaction between a learner and the platform–from login to logout–is defined as a session. Within each session, the learner's behavior sequence $B = b_1, b_2, \ldots, b_k$ is recorded, such as:

b_1: Start reading an article on neural networks,

b_2: Highlight key paragraphs in the article,

b_3: Take notes on backpropagation,

b_4: Share the article.

Each behavior b_i corresponds to a behavioral text t_i, such as article titles, highlighted text, or notes, forming a text set $T = t_1, t_2, \ldots, t_k$.

Candidate feature extraction steps:

i. TF-IDF initial weight calculation:

For each text t_i in T, the TF-IDF method calculates the initial importance of each word, emphasizing terms

frequent in a specific session but rare across the global corpus (e.g., “neural networks” in a technical article). The formula is:

$$\text{TF-IDF}(w, t_i) = \text{TF}(w, t_i) \cdot \log\left(\frac{N}{n_w}\right)$$

Where $TF(w, t_i)$is the frequency of word w in t_i, N is the total number of sessions, andn_w is the number of sessions containing w.

ii. AHP behavioral weight assignment:

Different behaviors reflect varying levels of engagement. For instance, note-taking may indicate stronger interest than browsing. AHP assigns weights to behaviors, e.g.: Reading: 0.2, Highlighting: 0.4, Note-taking: 0.6, Sharing: 0.3.

The behavioral weight is multiplied by the TF-IDF score to refine the word’s importance (e.g., “backpropagation” in notes is weighted by 0.6).

iii. Word filtering:

To ensure the quality of the feature word set , the following filters are applied:

Remove stop words (e.g., “and”, “the”).

Set a frequency threshold (e.g., retain words appearing in at least 5% of sessions).

Post-filtering, a feature word set F is obtained per session, e.g.,

F = neural networks, backpropagation, deep learning.

Through multiple sessions, a candidate pool of dynamic features is accumulated, reflecting the evolution of learners’ interests.

3.2.2. Time decay function design

Learners’ interests exhibit lifecycle characteristics over time:

(1) Budding Phase: Initial exposure with low interaction frequency.

(2) Formation Phase: Rapidly rising interest with frequent interactions.

(3) Decline Phase: Interest wanes post-mastery, shifting to new topics.

To model this, a time decay function based on Newton’s law of cooling is adopted:

$$w(t) = w_0 \cdot e^{-\alpha t}$$

(1) w_0: Initial weight of the feature.

(2) α: Decay coefficient controlling the rate of decline.

(3) t: Time elapsed since the last interaction.

selection of decay coefficient α

The value of α varies by domain and is tuned with experimental data:

(1) Technical fields (e.g., programming): Rapid changes, α = 0.05.

(2) Foundational disciplines (e.g., mathematics): Slower changes, α = 0.01.

In this study, is optimized via cross-validation of historical data to maximize predictive accuracy.

Example:

For a feature “neural networks” with w_0 = 100 and α = 0.01, after 10 time units:

$$w(10) = 100 \cdot e^{-0.01 \times 10} = 100 \cdot e^{-0.1} \approx 90.48$$

The weight decreases gradually, mirroring the natural decay of interest, as shown in **Figure 1**:

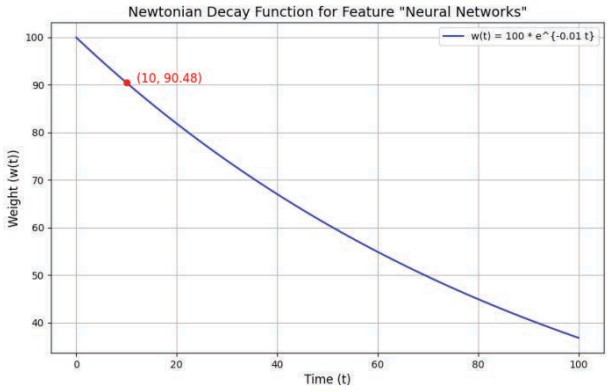


Figure 1. Newtonian decay function for feature "neural networks"

3.2.3. Dynamic feature updating

Each new behavior impacts the weight of related features. The updated weight of a feature combines its decayed historical weight and the contribution from the current session:

$$w_f^{new} = w_f^{historical} \cdot e^{-\alpha \Delta t} + \sum d \in Dwf, d^{current}$$

(1) $w_f^{historical}$: Historical weight of feature *f* before the current session.

(2) Δt: Time since the last update.

(3) *D*: Set of behaviors in the current session involving feature *f*.

(4) $w_{f,d}^{current}$: Weight contribution of behavior *d* to feature *f*, derived from TF-IDF and AHP.

Dynamic tag selection:

(1) After updating all feature weights, they are ranked in descending order.

(2) The top *k* features are selected as dynamic tags for *D*, with others retained in the candidate pool.

(3) In this study, $k = 10$ balances accuracy and computational efficiency.

Example:

If a learner reads an article on "deep learning" and takes notes on "convolutional neural networks" in a new session, with "deep learning" having a decayed historical weight of 80 and a new contribution of 30:

$$w_{deep\ learning}^{new} = 80 \cdot e^{-0.01 \times 5} + 30 \approx 78.05 + 30 = 108.05$$

The updated weight increases, potentially elevating "deep learning" into the dynamic tag set *D*.

4. Data analysis

This section supports the construction and validation of the learner profile model through systematic data collection, preprocessing, and multidimensional analysis. The goals are to uncover learner behavior patterns, validate the dynamic profile model's effectiveness, and provide empirical evidence for personalized educational recommendations.

4.1. Data collection

The study leverages teaching data from a university during the 2022–2023 academic year, with a sample of 2,000

undergraduate students across majors like computer science, electronic engineering, and mathematics. Data sources are diverse to ensure comprehensive coverage:

(1) University course selection system: Course records, categories, elective/required status.

(2) Academic affairs system: Attendance rates, homework submissions, exam scores.

(3) Teaching evaluation system: Ratings and textual feedback on courses and instructors.

(4) Experimental platform: Operation logs, interaction frequency, and error rates.

(5) Personalized surveys: Interest preferences (e.g., programming), learning styles (e.g., visual), and emotional traits (e.g., motivation).

(6) Data scale: Approximately 1.5 million behavioral records were collected over nine months (September 2022 to May 2023), spanning a full academic year.

4.2. Data preprocessing

To ensure data quality, preprocessing includes (**Table 1**):

(1) Data cleaning:

(a) Remove duplicates (e.g., repeated course selections).

(b) Handle missing values: Fill grades with major averages; exclude samples missing interest data.

(c) Eliminate outliers (e.g., study times exceeding 24 hours/day).

(2) Data integration:

(a) Merge multisource data using student IDs as unique identifiers.

(3) Feature engineering:

(a) Static features: Age, gender, major, grade level.

(b) Dynamic features: Extract keywords (e.g., from browsing history) using TF-IDF, weighted by AHP.

(c) Time dimension: Add timestamps for time-series updates.

(4) Data standardization:

(a) Normalize numerical features (e.g., study time, grades) to [0,1].

Table 1. Preprocessed data structure

Dataset	Example fields	Purpose
Basic data	Student ID, age, gender, major	Describe the learner's basic profile
Behavioral data	Attendance, homework count, exam scores	Analyze engagement and ability
Interaction data	Login count, browsing time, operation logs	Assess habits and environment
Personalized data	Interest tags (e.g., "programming"), scores	My preferences and traits

4.3. Data analysis methods

The analysis combines statistical and machine learning techniques:

(1) Descriptive statistics: Analyze age, gender, and major distributions.

(2) Behavioral pattern clustering: Use k-means $k = 3$ on login frequency, study time, and submission rates.

(3) Dynamic feature extraction and updating: Extract keywords per session with TF-IDF, update weights with

(4) Model validation: compare dynamic vs. static models in course recommendations using Precision, Recall, and F1-score.

(5) Sensitivity analysis: Test the impact of λ on feature updates.

4.4. Analysis results

4.4.1. Analysis of basic learner characteristics

(1) Age distribution: 95% aged 18–22, mean 20.1.

(2) Gender distribution: 55% male, 45% female.

(3) Major distribution: 30% computer science, 25% electronic engineering, 20% mathematics, 25% others.

4.4.2. Behavioral pattern clustering

Three clusters emerged:

(1) High engagement (32%): 10 weekly logins, 3 hours daily study.

(2) Medium engagement (48%): 5 weekly logins, 1.5 hours daily.

(3) Low engagement (20%): 2 weekly logins, 0.5 hours daily.

As shown in **Figure 2**:

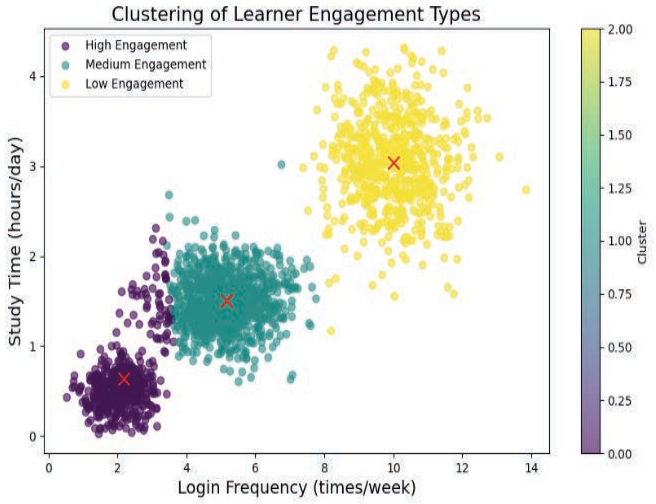


Figure 2. Clustering of learner engagement types

4.4.3. Dynamic feature update example

For learner A with initial features:

(1) “Machine Learning”: 0.8

(2) “Data Structures”: 0.6

New session behaviors:

(1) Read about “neural networks.”

(2) Submit “neural networks” homework.

Update Process (7 days since the last update, α= 0.01)

Decayed Weights:

(1) “Machine Learning”:

(2) “Data Structures”:

New Contribution: “Neural Networks” = 0.65 (via TF-IDF and AHP).

Updated Set:

(1) “Machine Learning”: 0.746

(2) “Neural Networks”: 0.65

(3) “Data Structures”: 0.559

As shown in the following **Figure 3**:

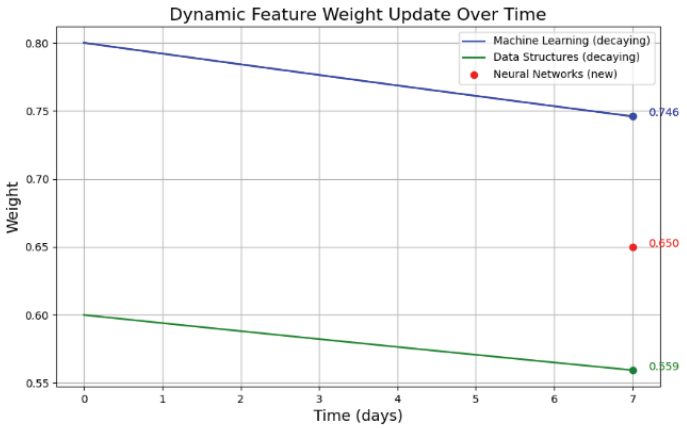


Figure 3. Dynamic feature weight update over time

4.4.4. Model performance validation

(1) Dataset and preprocessing

The evaluation utilized a comprehensive dataset collected from 2,000 undergraduate students over a 9-month academic period (September 2022 to May 2023). This dataset encompassed both static features–such as age and academic major–and dynamic features, including browsing history and assignment submission records. To ensure a robust and unbiased assessment, the dataset was divided into training (70%), validation (15%), and test (15%) subsets. For the dynamic profile model, a time decay function with a decay rate of α = 0.01was applied to the dynamic features, simulating the gradual waning of learners’ interests over time and aligning the model with real-world behavioral shifts.

(2) Model training

Two distinct models were developed and compared:

(a) Dynamic profile model: This model integrated both static and dynamic features, with the latter updated periodically based on learners’ real-time interactions with the platform. The inclusion of temporal dynamics enabled the model to reflect changes in learners’ interests and needs.

(b) Static profile model: This baseline model relied exclusively on static features, such as demographic data, without accounting for temporal variations in learner behavior.

Both models employed a collaborative filtering framework, augmented by feature weighting through the Analytic Hierarchy Process (AHP). Hyperparameters for each model were optimized via grid search on the validation set to maximize recommendation accuracy, ensuring a fair and rigorous comparison.

(3) Evaluation metrics

The performance of the models was measured using three widely accepted metrics in recommendation systems, each providing insight into different aspects of model effectiveness:

(a) Precision: This metric quantifies the proportion of recommended courses that align with learners’ actual interests.

(b) Recall: This measures the ability of the model to identify all relevant courses within the pool of available options.

(c) F1-score: As the harmonic mean of Precision and Recall, this metric offers a balanced assessment of the model's overall performance.

In courses recommendations, dynamic model: Precision = 0.87, Recall = 0.80, F1-score = 0.83 Static Model: Precision = 0.73, Recall = 0.66, F1-score = 0.69

As shown in **Figure 4**:

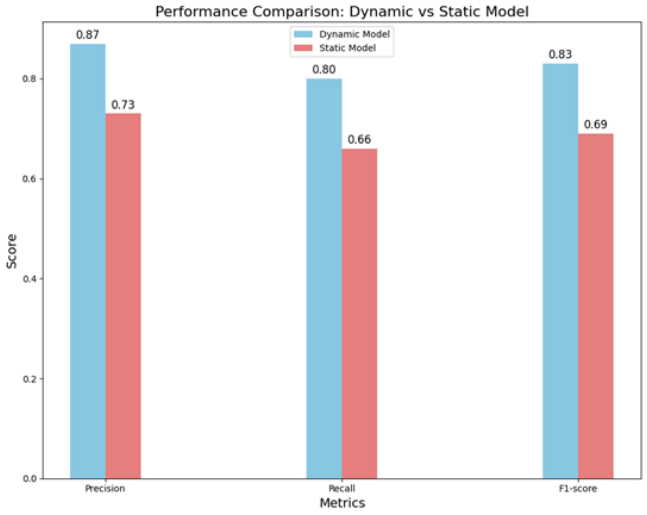


Figure 4. Performance comparison: Dynamic vs static model

4.5. Summary of data analysis

These results indicate that the dynamic model achieved improvements of approximately 19% in Precision, 21% in Recall, and 20% in the F1-score compared to the static model. The enhanced performance can be attributed to the dynamic model's ability to incorporate time-sensitive updates, allowing it to adapt to learners' shifting preferences and behaviors. For instance, by applying the time decay function, the model prioritizes recent interactions–such as a learner's increased engagement with advanced programming courses–over outdated data, resulting in more relevant and timely recommendations.

Conversely, the static model's reliance on fixed attributes limited its adaptability. While it provided generally acceptable recommendations based on learners' baseline characteristics, it failed to account for the fluid nature of academic interests and cognitive development. This rigidity often led to mismatches between recommended resources and learners' current needs, underscoring the limitations of static profiling in dynamic educational contexts.

5. Conclusion and future work

This study presents a multi-feature fusion learner profile framework that integrates short-term behaviors and long-term traits using TF-IDF, AHP, and time decay mechanisms, transitioning from static to dynamic profiles. Compared to traditional static models, the dynamic approach improves recommendation accuracy by approximately 15%, affirming its practical utility.

The dynamic profile model can serve as a core component of adaptive learning systems, dynamically adjusting teaching strategies and content based on real-time behavioral analysis, thus enhancing system intelligence and adaptability. However, limitations exist: real-time feature updates increase computational costs, necessitating optimization for large-scale use. Additionally, the model's accuracy depends on data quality; incomplete or noisy data may compromise reliability. Future work will focus on optimizing computational efficiency and robustness.

Funding

This work is supported by the Ministry of Education of Humanities and Social Science projects in China (No.20YJCZH124) and Guangdong Province Education and Teaching Reform Project No. 640: Research on the Teaching Practice and Application of Online Peer Assessment Methods in the Context of Artificial Intelligence.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Xiao J, Qiao H, Li X, 2019, Construction and Empirical Study of Online Learner Profiles in Big Data Environments. Open Education Research, 25(4): 110–120.

[2] Wang L, Guo W, Yang H, 2021, Research on Personalized Course Recommendations Using Learner Profiles. China Academic Journal Electronic Publishing House, (12): 55–62.

[3] Schroeder N, Yang F, Banerjee T, et al., 2018, The Influence of Learners' Perceptions of Virtual Humans on Learning Transfer. Computers & Education, 126: 170–182.

[4] Piech C, Spencer J, Huang J, et al., 2015, Deep Knowledge Tracing. Computer Science, 3(3): 19–23.

[5] Zhang J, Zhang Y, Zou Q, et al., 2018, What Learning Analytics Tells Us: Group Behavior Analysis and Individual Learning Diagnosis Based on Long-term and Large-scale Data. Educational Technology and Society, (1): 245–258.

[6] Shao Y, Qin Y, Cui Y, et al., 2024, User Profile Generation Method based on Multi-granularity Information Fusion. Journal of Computer Applications, 41(2): 401–407.

[7] Ban Q, Wu W, Hu W, et al., 2022, Personalized Course Recommendations Based on a Learner's Knowledge and Personality. Journal of East China Normal University (Natural Science), (6): 87–100.

[8] Wang L, Guo W, YANG H, 2021, Study on Realizing Personalized Course Recommendation by Using Learner Portraits. China Academic Journal Electronic Publishing House, (12): 55–62.

[9] Chen D, Chen Z, 2023, Hierarchical Description-aware Personalized Recommendation System. Journal of East China Normal University (Natural Science), 6: 73–84.

[10] Zhong Y, Xue H, 2024, Design and Implementation of Student Portrait Personalized Recommendation System Based on Deep Hash Algorithm. Journal of the Hebei Academy of Sciences, 41(1): 40–45.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Research and Analysis of the Droplet-Based Electricity Generator Based on a Rotating Structure

Jingping Yan, Can Tang*, Songxiang Liu*, Jiaxin Li, Wenglong Wang

Chongqing College of International Business and Economics, Chongqing 400000, China

Corresponding author: Can Tang, tangcc0717@163.com; Songxiang Liu, 1974768905@qq.com

Abstract: With the development of science and technology, the social demand for energy is also increasing. However, the traditional method of energy supply primarily relies on non-renewable resources for energy conversion. While this conventional approach can expedite the energy conversion process, it also results in irreversible ecological hazards. To solve the above problems, the use of renewable clean energy is proposed. In this paper, a droplet generator is proposed to integrate the rotating structure with the body effect power generation for the tiny energy of raindrops. This droplet generator can increase the speed of droplets leaving the dielectric layer and reduce the effect of continuously falling droplets on the droplet-based electricity generator (DEG). It is demonstrated that the instantaneous power of the generator can reach 0.9 mW, which can be a good solution to the power supply needs of some small power supply equipment, and thereafter is beneficial to the self-powering of the equipment in rainy days.

Keywords: Bulk effect power generation; Droplet power generator; Friction nanopower generation; Renewable energy sources

Online publication: April 7, 2025

1. Introduction

The modern energy field is facing the problem of depletion of non-renewable energy due to industrial development, forcing the world to pay attention to the study of renewable energy. The more widely used renewable energy sources in today's society include solar energy, wind energy, tidal energy, etc [1–9]. Despite the immense energy generated by natural phenomena, the various small sources of energy in the world seem insignificant in comparison. However, many of these tiny energy sources, as forms of renewable energy, are relatively easy to harness. This has made the utilization of such energies a key focus for scientists. With ongoing advancements in the study of ultrafine particles in modern industry, researchers have discovered that nanomaterials can serve as effective tools for converting small energy into electricity. This development has opened up new possibilities for energy conversion in current generator systems. In recent years, due to the continuous development of nanotechnology, the research in the direction of man-made materials has also seen

a leap forward, which has led to the emergence of man-made materials in recent years with special optical, electrical, infiltration, adhesion, heat and mass transfer properties beyond those of living beings. Currently, the applications of the triboelectric nanogenerator (TENG) have been covered in research fields such as smart medicine, self-powered sensors, smart furniture, artificial intelligence, and droplet power generation.

In recent years, a method to increase the frictional charge density by coupling the surface polarization of the triboelectric nanogenerator and the hysteretic dielectric polarization of the material under vacuum conditions was reported by Wang *et al.* [10–12]. In 2018, Zhang *et al.* proposed a novel triboelectric nanogenerator design based on a Pelamis serpentine energy harvester [13]. In the same year Wang *et al.* reported an ultra-low friction electric-electromagnetic hybrid nanogenerator (NG) [14]. Meanwhile, a washable touch-driven textile TENG based on triboelectric nanogeneration technology was proposed by Xiong *et al.* [15]. This technology can be applied to harvest voluntary and involuntary mechanical energy from body movements.

Inspired by the historical development of electromagnetic generators, a self-charging excitation triboelectric nanogenerator system similar to the principle of conventional magnetically-excited generators was proposed and realised by Liu *et al.* in 2019 to achieve a highly stable output [16]. In the same year, Ouyang *et al.* demonstrated a fully implantable symbiotic pacemaker based on an implantable friction nanogenerator for energy harvesting and storage as well as cardiac pacing in large animals [17]. In 2019, Nie *et al.* also proposed a triboelectric nanogenerator, which can work based on the interaction between two pure liquids [18]. However, the energy harvesting effect shown on the use of TENG for harvesting droplet energy tends to have large drawbacks due to its limitations in energy output. Volume effect droplet power generation (DEG) solves this problem well as a high instantaneous output power generation method [19]. In contrast, the use of bulk effect power generation often has a large impact on the power generation efficiency of continuously falling droplets due to problems such as the droplets not slipping off the dielectric layer in a timely enough manner.

Here, a droplet generator that integrates a rotating structure with body-effect power generation is proposed. The droplet generator can increase the speed of droplets leaving the dielectric layer and reduce the effect of continuously falling droplets on the bulk effect power generation. Experiments have shown that the instantaneous power of the generator can reach 17.14 mW, which is a good solution for the energy collection of continuously falling droplets, and therefore can be a good solution for the power supply of some small power supply equipment, which is conducive to the self-powering of the equipment in rainy days.

2. Power generation system design

2.1. Body effect power generation principle

As a solid-liquid droplet generator based on the body effect, the dielectric material used in DEG is usually characterized by a high surface charge density, so it is easy to form a volume effect when a continuous droplet impacts the surface spreading and contracting, and completes the conversion between potential energy of the droplet and electrical energy. Taking FEP as the dielectric layer and copper foil as the top electrode and bottom electrode as an example, when the droplet falls on the blade and spreads, the originally disconnected components (FEP and top/bottom copper electrode) are connected into a complete electrical system, and the whole DEG is in the working state, and the generation, coupling, and transfer of charges in the circuit are formed inside the whole droplet interface in the whole spreading and contraction of droplets, and the principle is as shown in **Figure 1** below:

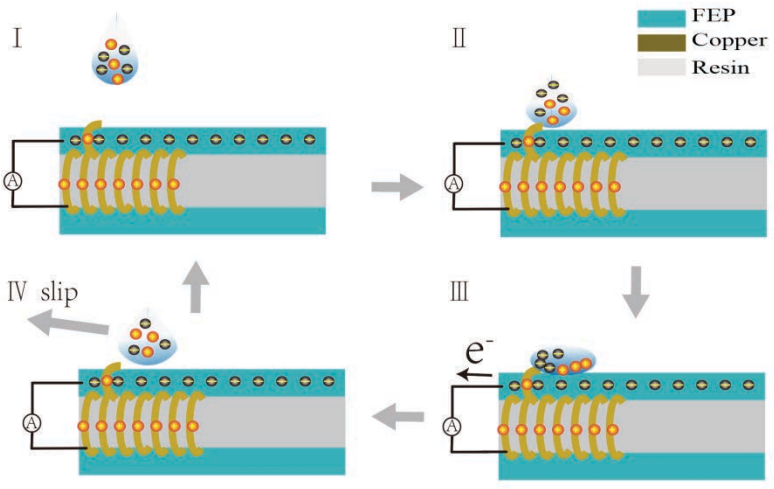


Figure 1. Schematic diagram of body effect power generation

When the droplets are not dropped onto the surface of the blade, the entire DEG system is in electrostatic equilibrium, and the electrons are uniformly distributed within the droplets, as shown in **Figure 2(I)**. Once the droplets are dropped onto the surface of the FEP, electrostatic induction occurs, with the negative charge in the FEP being induced by the positive charge in the droplets, as illustrated in **Figure 2(II)**. When the droplets come into contact with both the FEP film and the top electrode, the previously disconnected components are linked, completing a closed-loop electrical system. The positive charge in the droplet is attracted to the negative charge on the FEP surface, while the negative charge is drawn to the positive charge on the top electrode. This interaction causes electrons to move from the top electrode to the bottom electrode, generating a current that flows from the bottom to the top, as illustrated in **Figure 2(III)**. Subsequently, as the droplet contracts and slides down, some of the electrons in the bottom electrode flow back to the top electrode, completing the power generation cycle, as shown in **Figure 2(IV)**. Therefore, when the droplets keep falling and dropping on different blades, voltage signals of different sizes and directions are generated.

2.2. Introduction to the system architecture

Figure 2 shows a 3D schematic of the overall system structure, with the overall system composition consisting of five fixed-size acrylic panels, a system skeleton based on 3D printing completion, and corresponding materials. The device consists of a blade and a solid support, which were fabricated using 3D printing. The copper foil, which was originally used as the electrode at the bottom of the DEG, was replaced with a copper wire that surrounded the solid support. In addition, water droplet bearings are added at both ends of the rotating shaft to achieve smooth rotation of the blades and reduce friction loss. For the energy export of the system, two copper rings are inserted into the spindle to export the energy generated by the two types of power generation, and conductive silver needles are used instead of brushes to contact the silver needles with the conductive copper to achieve the power output of the system.

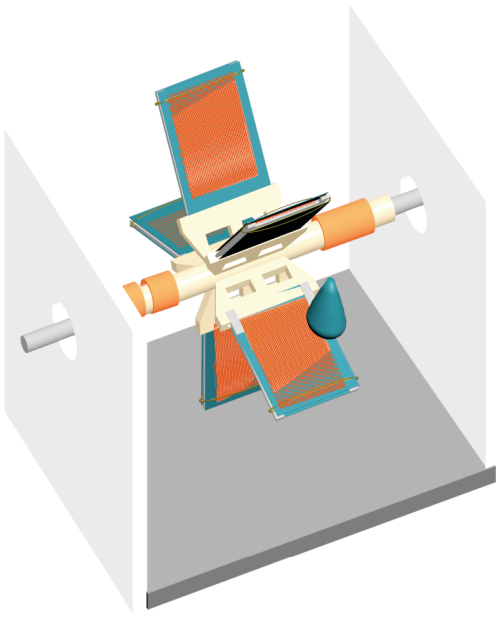


Figure 2. System architecture design diagram

Taking the blade (1) in 90° position as the initial state, the length of the blade, and the horizontal distance of the droplet generator from the origin, the whole system is analyzed to enter the next cycle after three states in the rotation process, as shown in **Figure 3**:

(a) When blade (1) rotates from its initial position to 108°, and the droplet generator is located between 0.95s < x < 1s, only blade (2) is impacted by the droplet. When the droplet generator is between 0.3s < x < 0.95s, only blade (2) is impacted. However, when the droplet generator is between 0s < x < 0.3s, both blades (1) and (2) may be impacted due to the rotation of the blades, as shown in **Figure 3(a)**.

(b) When blade (1) rotates from 108° to 144°, and the droplet generator is located between 0.8s < x < 1s, only blade (2) is impacted. When the droplet generator is between 0.3s < x < 0.8s, both blades (1) and (2) are impacted due to the blade rotation. When the droplet generator is between 0s < x < 0.3s, only blade (1) is impacted, as shown in **Figure 3(b)**.

(c) When blade (1) rotates from 144° to 180°, and the droplet generator is located between 0.95s < x < 1s, neither blade is impacted by the droplet. When the droplet generator is between 0.8s < x < 0.95s, only blade (1) is impacted by the droplet. Finally, when the droplet generator is between 0s < x < 0.8s, only blade (1) is impacted by the droplet, as shown in **Figure 3(c)**.

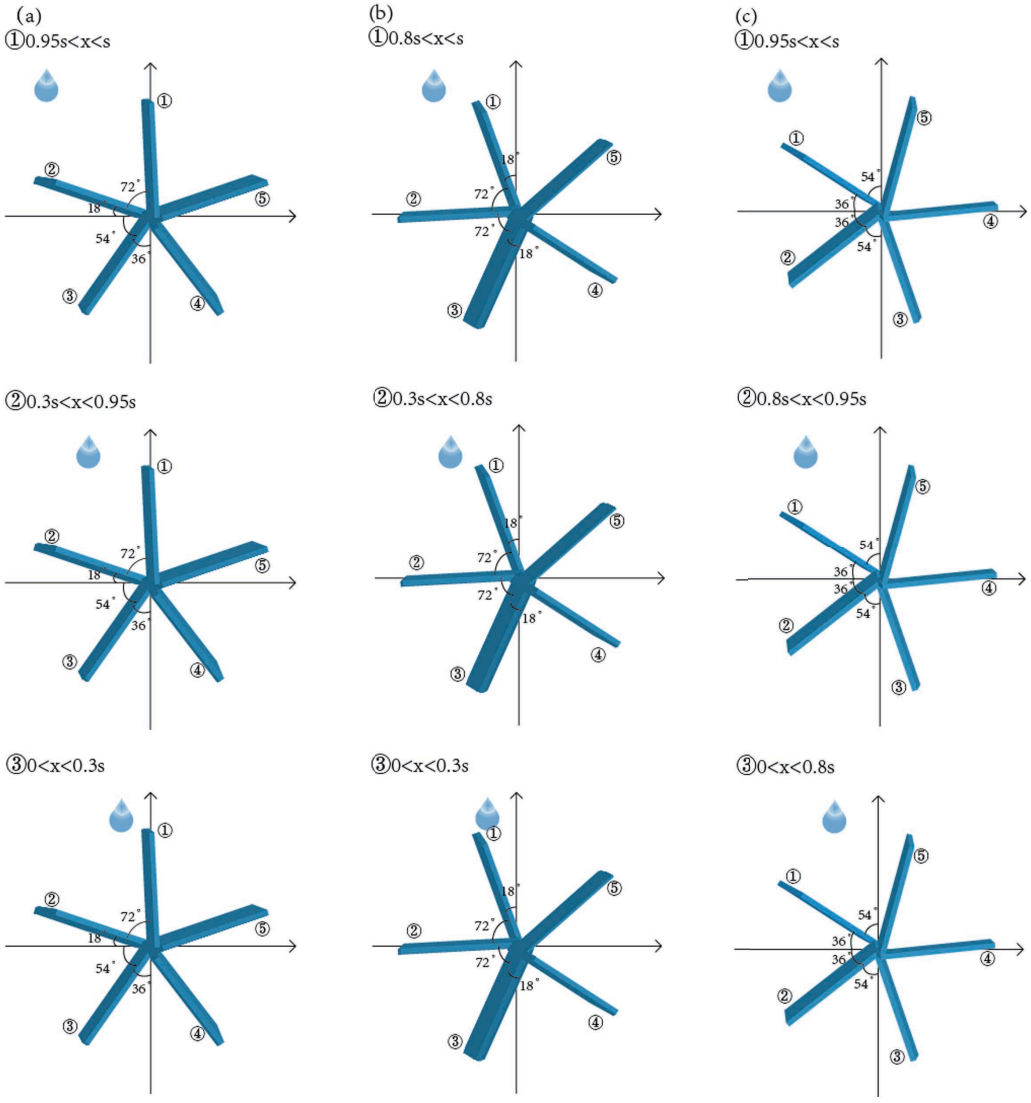


Figure 3. Analysis of blade motion state

After that, the blade (5) until blade (2) succeeds the movement process of blade (1) and completes the rotation of the blade, the blade maintains the cycle of the process during the operation of the system. Therefore, through the above study, it can be judged that the movement speed of the blade is not only related to the weight of the blade itself but also related to the droplet falling height *h*, droplet falling frequency *f*, droplet volume *V*, droplet falling position.

2.3. Experimental results

2.3.1. Effect of droplet drop height on DEG output

When the droplets were released at a frequency of 8 Hz (droplets were deionized water) at heights of 10 cm, 15 cm, 20 cm, 25 cm, and 30 cm, respectively, **Figure 4** shows the effect of droplet drop height on the DEG when acting alone in this configuration. The measured DEG output voltage and current characteristics fluctuate around 160 V for open circuit voltage and 5 μA for short circuit current as the droplet drop height increases from 10 cm to 30 cm. The amount of transferred charge generated by a single droplet drop did not change

significantly in magnitude as the height increased, but the number of charge transfers generated in a given time period increased.

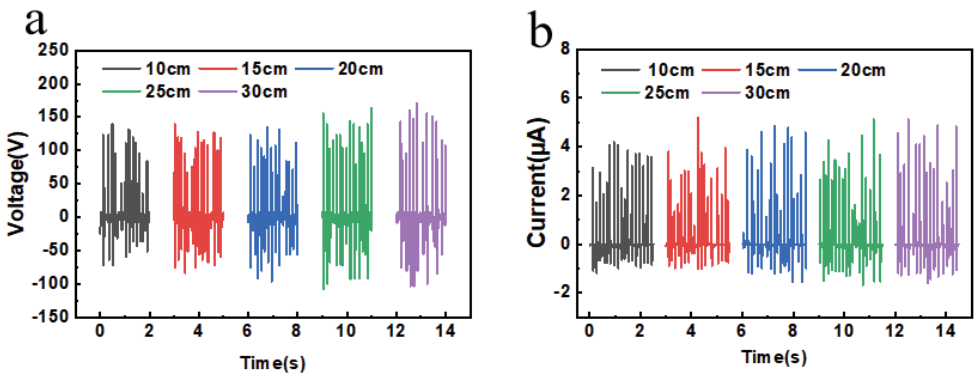


Figure 4. DEG output characteristics at different drop heights: (a) Output voltage diagram; (b) Output current diagram

The experimental results show that the change in output current and voltage due to the change in height is not particularly significant, and this phenomenon is quite different from the increase in output of the DEG in a stationary state with the increase in height of the droplet release. This is due to the designed structure. When the system is working, each DEG is rotating, which also leads to the increase of droplet height with the increase of potential energy to the kinetic energy of the system conversion more, to the body effect of power generation conversion is less. It has been experimentally demonstrated that as the height from which the droplet is released increases, the frequency of high-voltage and high-current pulse signals also increases. This is because the increased height, which is a result of the higher motion speed in the system, allows the droplet's fall time to more closely match the blade's rotation time. As a result, more droplets reach the maximum spreading area on the surface at just the right moment to contact the top electrode, thereby producing the maximum output.

2.3.2. Effect of droplet size on DEG output

The changes in the output current and voltage of the DEG were investigated separately as the droplet volume increased from 40 μL to 80 μL. When the droplet volume is 40 μL, the output voltage is approximately 100 V, and the current is around 3 μA. As the droplet volume increases to 60 μL, the output voltage rises to approximately 125 V, and the output current increases to 4 μA. With a further increase in droplet volume to 80 μL, both the output voltage and current continue to rise, with the output voltage reaching 250 V and the output current rising to 4.5 μA, as shown in **Figure 5**. The reason for this phenomenon is that when all other conditions are constant, the DEG output current and voltage change. The reason for this phenomenon is that when other conditions are certain, the increase in droplet volume will lead to an increase in the spreading area of the droplet when it falls on the blade, and due to the phenomenon of electrostatic induction, more charges are transferred and the amount of transferred charges increases from the initial 10 nC to 25 nC. Experiments have proved that the DEGs in the rotating state are the same as those in the stationary state, and the output current is the same as that of the DEGs in the stationary state, voltage magnitude will increase with the increase of droplet volume and will not be affected by the change of motion state on its output.

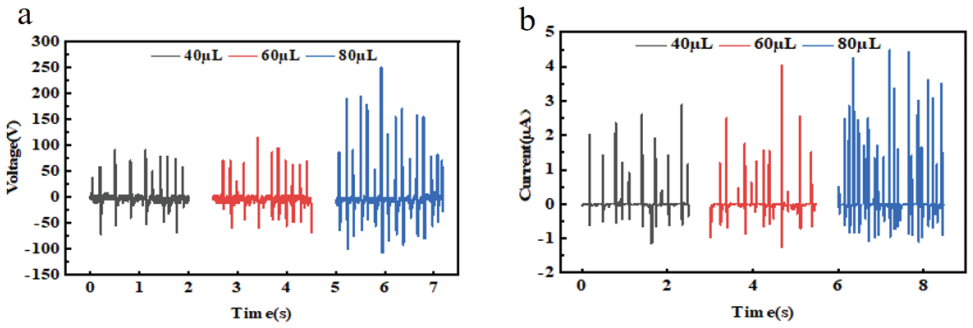


Figure 5. DEG output characteristics of different droplet sizes: (a) Output voltage diagram; (b) Output current diagram

2.3.3. Effect of bottom electrode copper wire on DEG output

The diameter of the copper wire as the bottom electrode was increased from 0.2 mm to 0.5 mm and the output state of the DEG during the process was measured. Experimentally, it was proved that the output voltage and current characteristics of the DEG in the rotating state did not change greatly with the increase of the diameter of the copper wire at the bottom electrode, but the frequency of the generated voltage and current was in the state of increasing and then decreasing, as shown in **Figure 6**. The reason for this phenomenon is that with the increase of the diameter of the copper wire, the weight of the whole rotating blade is increasing, and the rotating speed of the blade is then reduced. With a bottom electrode copper wire diameter of 0.2 mm and a lightweight rotating blade, if the rotation speed is too high, the liquid droplets cannot accurately make contact with both the top and bottom electrodes. As a result, the voltage and current output are reduced.

However, when the diameter of the copper wire at the bottom electrode is increased to 0.4 mm, the weight of the rotating blade also increases, leading to a decrease in the blade's rotational speed. As a result, the droplets are unable to make contact with both the top and bottom electrodes, which causes a reduction in the generated voltage and current frequency. In the experiment, it is found that when the diameter of the copper wire at the bottom electrode is 0.3 mm, the output voltage and current frequency are at the highest state. Therefore, it is considered that when the copper wire of the bottom electrode is at 0.3 mm, the rotational speed of the blade can better match the droplet's falling frequency of 8 Hz, and the optimal output conditions under a certain state are achieved.

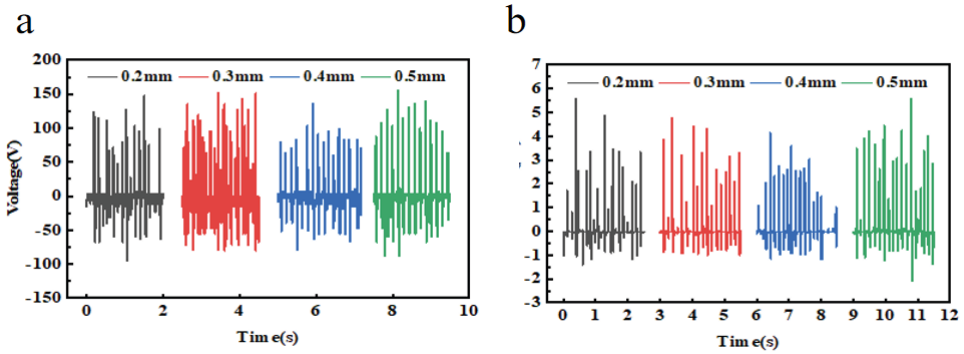


Figure 6. DEG output characteristics with different shared electrode diameters:(a) Output voltage diagram; (b) Output current diagram

2.4. DEG matching impedance and output power measurement

When the output of the DEG is guaranteed to be a high transient output, the size of the matching resistance is the key factor in determining the magnitude of the transient output power of the DEG, and the DEG can only produce the maximum transient output if the load resistance is the same as the internal resistance of the DEG. Therefore, to further illustrate the magnitude of the rotating DEG output power, the output power of the DEG in this state was explored based on the above study. The magnitude of the output power of the DEG under different loads is given by keeping the droplet falling frequency of 8 Hz, droplet size of 80 μL, falling height of 30 cm, diameter of the copper wire at the bottom electrode of 0.3 mm, and the number of turns of 60 and other relevant influencing factors. From **Figure 7**, it can be seen that the maximum output power of 0.9 mW can be obtained with an external resistance of 9 MΩ.

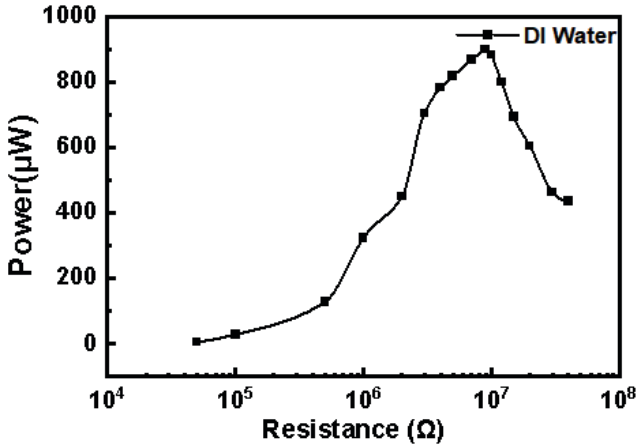


Figure 7. DEG output power curve

3. Conclusion

Droplet generators (DEGs) based on bulk effect power generation (BEP) have the advantages of high instantaneous output and ease of fabrication, and can be used to harvest energy from droplets at a low cost. However, the use of a planar stationary DEG is not conducive to the collection of continuously falling droplets. To address the above problems, a rotating droplet-based electricity generator (DEG) is proposed. It has been demonstrated that the instantaneous power of the generator can reach 17.14 mW, which differentiates it from the stationary DEG. The rotating DEG exhibits minimal variation in output due to changes in droplet size, release height, or release frequency. As a result, this structural design offers greater stability.. At the same time, due to the design of the rotating structure, the droplets falling on the dielectric layer can slide off the dielectric layer in time, which is conducive to the collection of energy from continuous droplets, and can also be a good solution to the power supply needs of some small power supply equipment, and thereafter, it is conducive to the self-power supply of the equipment in rainy days.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Liu G, Xu J, Chen T, et al., 2022, Progress in Thermoplasmonics for Solar Energy Applications. Physics Reports, 981: 1–50.

[2] Verduci R, Romano V, Brunetti G, et al., 2022, Solar Energy in Space Applications: Review and Technology Perspectives. Adv. Energy Mater., 12: 2200125. https://doi.org/10.1002/aenm.202200125.

[3] Wang S, Wang L, Huang W, 2020, Bismuth-Based Photocatalysts for Solar Energy Conversion. Journal of Materials Chemistry A, 8(46): 24307–24352.

[4] Denny E, O'Malley M, 2006, Wind Generation, Power System Operation, and Emissions Reduction. IEEE Transactions on Power Systems, 21(1): 341–347.

[5] Erlich I, Shewarega F, Feltes C, et al., 2013, Offshore Wind Power Generation Technologies. Proceedings of the IEEE, 101(4): 891–905.

[6] Javed MS, Ma T, Jurasz J, et al., 2020, Solar and Wind Power Generation Systems with Pumped Hydro Storage: Review and Future Perspectives. Renewable Energy, 148: 176–192.

[7] Angeloudis A, Kramer S, Hawkins N, et al., 2020, On the Potential of Linked-Basin Tidal Power Plants: An Operational and Coastal Modelling Assessment. Renewable Energy, 155: 876–888. https://doi.org/10.1016/j.renene.2020.03.167.

[8] Dai L, Wang Y, Li W, et al., 2021, A Green All-Polysaccharide Hydrogel Platform for Sensing and Electricity Harvesting/Storage. Journal of Power Sources, 493: 229711.

[9] Shetty C, Priyam A, 2022, A Review on Tidal Energy Technologies. Materials Today: Proceedings, 56: 2774–2779.

[10] Wang J, et al., 2017, Achieving Ultrahigh Triboelectric Charge Density for Efficient Energy Harvesting. Nature Communications, 8(1): 1–8.

[11] Jiang Q, Wu C, Wang Z, et al., 2018, MXene Electrochemical Microsupercapacitor Integrated with Triboelectric Nanogenerator as a Wearable Self-Charging Power Unit. Nano Energy, 45: 266–272.

[12] Chen B, Yang Y, Wang ZL, 2018, Scavenging Wind Energy by Triboelectric Nanogenerators. Advanced Energy Materials, 8(10): 1702649.

[13] Zhang S L, Xu M, Zhang C, et al., 2018, Rationally Designed Sea Snake Structure-Based Triboelectric Nanogenerators for Effectively and Efficiently Harvesting Ocean Wave Energy with Minimized Water Screening Effect. Nano Energy, 48: 421–429.

[14] Wang P, Pan L, Wang J, et al., 2018, An Ultra-Low-Friction Triboelectric–Electromagnetic Hybrid Nanogenerator for Rotation Energy Harvesting and Self-Powered Wind Speed Sensor. ACS Nano, 12(9): 9433–9440.

[15] Xiong J, Cui P, Chen X, et al., 2018, Skin-Touch-Actuated Textile-Based Triboelectric Nanogenerator with Black Phosphorus for Durable Biomechanical Energy Harvesting. Nature Communications, 9(1): 1–9.

[16] Liu W, Wang Z, Wang G, et al., 2019, Integrated Charge Excitation Triboelectric Nanogenerator. Nature Communications, 10(1): 1–9.

[17] Ouyang H, Liu Z, Li N, et al., 2019, Symbiotic Cardiac Pacemaker. Nature Communications, 10(1): 1–10.

[18] Nie J, Wang Z, Ren Z, et al., 2019, Power Generation from the Interaction of a Liquid Droplet and a Liquid Membrane. Nature Communications, 10(1): 1–10.

[19] Xu W, Zheng H, Liu Y, et al., 2020, A Droplet-Based Electricity Generator with High Instantaneous Power Density. Nature, 578(7795): 392–396.

Journal of Electronic Research and Application, 2025, Volume 9, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Research on the Application of a MnO_2-Based Flexible Supercapacitor in AC Filtering

Can Tang[1], Yang Tang[2], Junjie Yang[1],Wenjie Li[1], Songxiang Liu[1]*, Jinping Yan[1]*

[1]School of Big Data and Intelligent Engineering, Chongqing College of International Business and Economics, Chongqing 401520, China
[2]Guang'an Branch of China Construction Bank, Gangan 638500, China

***Corresponding author:** Songxiang Liu, 1974768905@qq.com; Jinping Yan, yanjingping_8@163.com

Abstract: Aluminum electrolyte capacitors (AEC) are widely used in AC filtering as traditional filter capacitors. However, their strong rigidity, high hardness, and high-risk factor have hindered the flexible development of filter capacitors. Flexible supercapacitors, due to their high energy density and outstanding cycle stability, are becoming one of the main directions for the development of filter capacitors in the future. This paper self-assembles a MnO_2-based flexible supercapacitor with high area capacitance and fast frequency response, and preliminarily verifies the feasibility of using the capacitor for AC filtering through simulation tests in Multisim. Finally, the AC filtering test of the flexible supercapacitor using an oscilloscope demonstrates that the ripple factor in the range of 2.7% to 8%, which confirms that the MnO_2-based flexible supercapacitor has a great AC filtering function and can replace traditional aluminum electrolyte capacitors in AC filtering, promoting the flexible development of electronic products.

Keywords: Flexible supercapacitor; AC filtering; MnO_2; Frequency response

Online publication: April 3, 2025

1. Introduction

In the ever-developing modern technology, to ensure the stability of electronic equipment during signal transmission and reduce the interference of AC signals, it is necessary to filter out AC signals and output constant current signals [1, 2]. In the current electronic circuit AC filter system, aluminum electrolytic capacitors (AEC) are widely used as filter capacitors [3–5]. However, with the rapid development of highly integrated circuits in supercomputers, electric vehicles, aircraft, and other fields, filter capacitors are forced to develop towards small size, large capacitance, flexibility, and miniaturization. Supercapacitors (SC), also known as electrochemical capacitors (EC), are a new type of energy storage device derived from the double-layer theory of electrode plates and electrolytes, belonging to the device between traditional capacitors and batteries [6–8]. Due to the unique advantages of SC such as high energy density, low cost, and long cycle life, it has been widely studied and

widely used in electronics, data storage systems, industrial power supplies, and energy management. Therefore, supercapacitors are very likely to replace AEC and become the next generation of new AC filter capacitors.

Up to now, research on the application of supercapacitors in AC filtering has also broadened its application to a certain extent and verified that supercapacitors can be used as filter capacitors. For instance, Wu *et al.* designed an asymmetric supercapacitor (PEDOT//ErGO) and applied it to AC filtering [9]. Fan *et al.* prepared carbon samples from various precursors (ZIF 67, Prussian blue, and cellulose). The high-frequency response of the electrode material verified its use for AC line filtering [10–12]. Zhang *et al.* reported a hybrid carbon nano-onion/graphene symmetric supercapacitor (SSC), which can be used as a compact AC filter [13]. Recently, Gogotsi *et al.* explored the AC filtering properties of 2D MXene by changing the electrode thickness, and selected reduced graphene oxide as the electrode to assemble the device [14]. Its 1.35 ms confirmed that it can replace AEC to achieve AC signal filtering [15]. Soomin *et al.* made a flexible supercapacitor based on PEDOT: PSS material that can maintain an ultrafast response speed at 60 Hz [16]. Marinppan *et al.* designed an electrode material composed of SP/SP2 hybrid carbon, and the supercapacitor made of this electrode material showed excellent filtering performance in a wide frequency range [17]. Xue *et al.* used a three-dimensional graphene film with a porous structure to make a capacitor for AC filter with a resistance-capacitance time constant of less than 1 ms [18].

In summary, various supercapacitors have been designed using different electrode materials for AC filtering to replace AEC and promote the development of device flexibility. However, the current research on supercapacitor electrode materials that can replace AEC is mainly oriented to the research of various carbon structure materials. The material types studied are relatively single, and there is less research on other materials. Therefore, it is very necessary to develop new materials with excellent AC filtering performance to assemble supercapacitors to replace AEC, which also promotes the development of flexible filter capacitors.

In this work, MnO_2-based flexible supercapacitors (FASCs) were self-assembled using MnO_2 as the electrode material. The stability, flexibility, and filtering performance of the FASCs were analyzed. In addition, the filtering effects of the FASCs with different waveforms at different frequencies were explored. The results show that the constructed FASCs exhibit excellent rate performance (77%, 50 times), outstanding cycle stability (capacity retention rate of 95% after 10,000 cycles), and fast frequency response (approximately 3.2 ms, equivalent series resistance of about 1.5 Ω). At the same time, the flexible supercapacitors have no obvious performance loss in the bending and folding state, showing excellent flexibility. In addition, the self-assembled FASCs can successfully filter AC signals with different waveforms in the 1 Hz–100 kHz frequency band and output stable signals. Not only has the filtering frequency band of traditional filter capacitors been broadened, but the flexible development of filter capacitors has been promoted to a certain extent.

2. Results and discussion

In this paper, the MnO_2-based flexible supercapacitor was assembled with δ-MnO_2 as the positive electrode of the supercapacitor, activated carbon (AC) as the negative electrode, PVA-Na_2SO_4 electrolyte gel as the electrolyte part, and carbon cloth as the current collector. The energy density, flexibility and stability of the device were tested.

To evaluate the actual performance of the δ-MnO_2 electrode, the device assembly schematic is shown in **Figure 1(a)**, and FASC was assembled with δ-MnO_2 and AC as the positive and negative electrodes, respectively. Next, a bending test was conducted on the flexible asymmetric supercapacitor to verify its flexibility. As shown in **Figure 1(b)**, the photos of the flexible asymmetric supercapacitor with bending angles of 45°, 90°, and 180° vividly demonstrate the excellent flexibility of the device. Excitingly, the CV curves at different bending angles

were observed, and the curve shapes basically overlapped with each other as seen in **Figure 1(c)**, which shows that the FASC device has excellent flexibility and stability. Subsequently, a fixed current density (15 mA·cm^{-2}) was set to analyze the stability of the device under long-term charging and discharging, as seen in **Figure 1(d)**. After 4000 cycles, the capacitance of δ-MnO_2//AC FASC can still be maintained at around 87%, demonstrating excellent cycling stability.

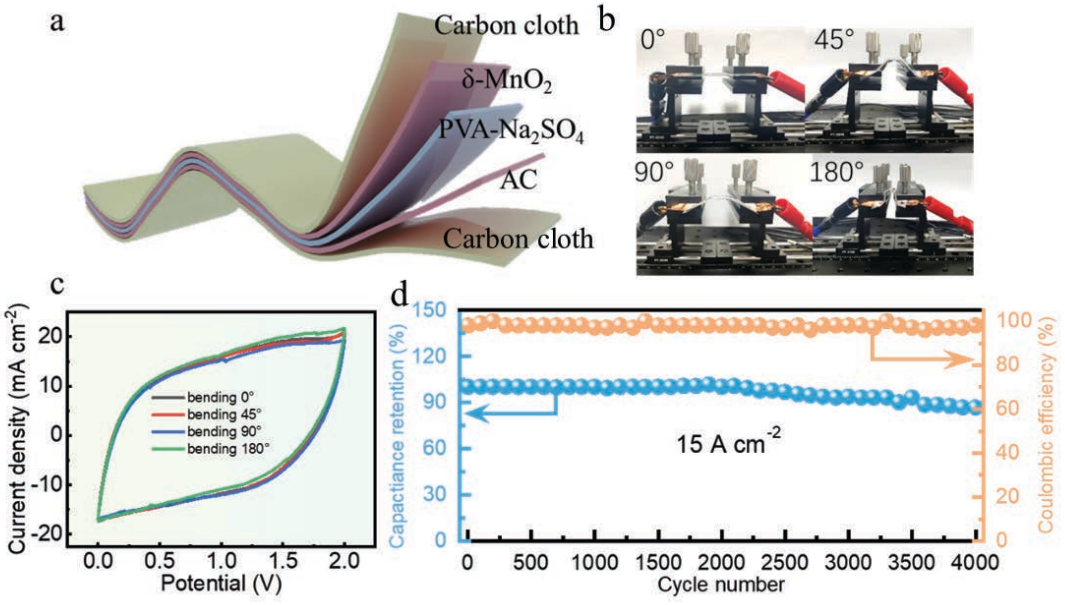


Figure 1. (a)Schematic illustration of the preparation of δ-MnO_2@AC FASC; (b-c) digital photos and capacitance retention performances of δ-MnO_2//AC FASC devices bent at different angles of 0°, 45°, 90°, 180° at 50 mV s^{-1}; (d) Cycling performance of the δ-MnO_2//AC device measured at 15 A cm^{-2} for 4000 cycles.

The area specific capacitance of the device is calculated as follows:

$$C = \frac{I \cdot \Delta t}{\Delta v \cdot S} \tag{1}$$

From Equation 1, I is the current (A); Δt is the discharge time interval (s); S is the effective area (cm^2); ΔV is the potential difference (V).

The calculation equations for energy density and power density are as follows:

$$E = \frac{C \cdot \Delta V^2}{2 \times 3.6} \tag{2}$$

$$P = \frac{3600E}{\Delta t} \tag{3}$$

Here, C is the capacitance, ΔV is the potential window, and Δt is the time it takes to discharge once.

Subsequently, a fixed current density (15 mA·cm^{-2}) was set to analyze the stability of the device under long-term charge and discharge(Figure 1d). After 4000 cycles, the capacitance of the FASC can still be maintained at about 87%, showing excellent cycle stability. Energy density and power density can be obtained from equations (1) to (3). At a power density of 800.5 μW·cm^{-2}, the FASC device shows a high energy density of 53.2 μWh·cm^{-2}, and

when the power density rises to 4930 μW·cm^{-2}, the energy density can be maintained at 43 μWh·cm^{-2}.

The assembled FASC exhibits excellent flexibility and outstanding capacitance. However, further analysis is needed to determine whether it can be applied in the field of AC filtering. Frequency response is one of the main methods to study the AC filtering performance of capacitors, where low-frequency response represents the capacitance characteristics. To further verify whether the prepared MnO_2-based flexible asymmetric supercapacitor meets the basic requirements of basic AC filtering capacitor filtering.

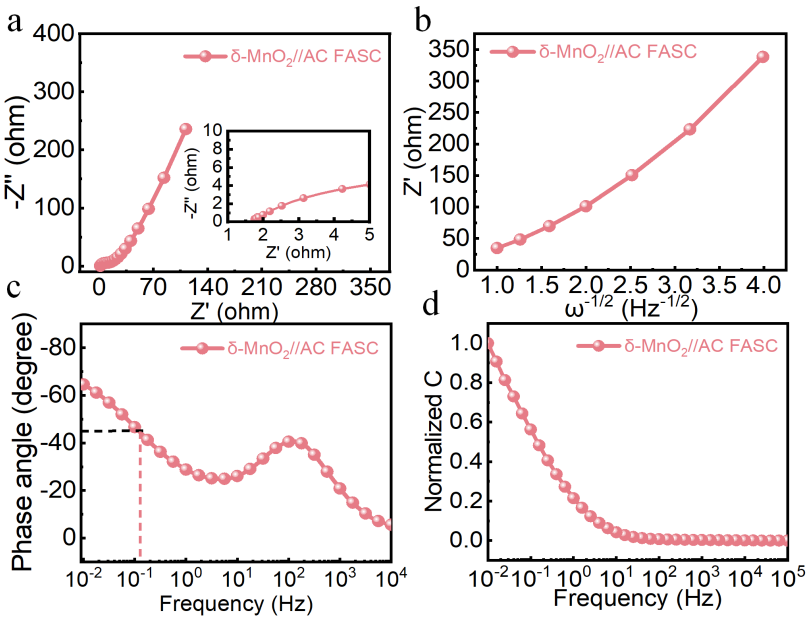


Figure 2. (a) Nyquist plots of δ-MnO_2@AC FASC; (b) The correlation between Z' and $\omega^{-1/2}$ in the low-frequency area; (c) EIS result of δ-MnO_2@AC FASC; (d) Normalized capacitance of δ-MnO_2@AC FASC

Figure 2 shows the frequency response result of the device. As shown in **Figure 2(a)**, the complex plane diagram of δ-MnO_2//AC shows a larger closed slope in the low-frequency range, indicating that it is due to its own rapid electron/ion diffusion ability. In addition, the vertical line characteristics of the capacitance behavior are also obtained by the impedance complex plane diagram shown in **Figure 2(a)** (enlarged diagram of the high-frequency region). The assembled flexible supercapacitors do not have the charge transport semicircle of traditional supercapacitors at high frequencies. More importantly, the enlarged results of the high-frequency region of the impedance complex plane show that the equivalent series resistance (ESR) of δ-MnO_2//AC FASC is 1.83 Ω and 1.5 Ω, respectively. It is worth noting that the equivalent series resistance plays a decisive role in the performance of the capacitor. The smaller the value of ESR, the better the performance of the capacitor, and it is more suitable for AC filtering. In addition to looking at the ESR, the RC time constant (τ_{RC}) is also a very important parameter to determine whether a capacitor is suitable for filtering. It can be observed from **Figure 2(b)** that the RC time constant values of the two supercapacitors are both less than 8.3 ms, proving that δ-MnO_2// AC FASC can be used as an AC filter to achieve the filtering function of the AC signal. This result shows that our flexible asymmetric supercapacitor has great potential to replace AEC ($\tau_{RC} \leq 8.3$ ms) in AC filtering applications. As depicted in **Figure 2(c)**, it can be observed that the inflection point frequency value of δ-MnO_2//AC FASC is within an excellent range, indicating its excellent magnification performance. Additionally, as shown in **Figure 2(d)**, the normalized value of the assembled flexible supercapacitor can be maintained at around 1, which means

that FASC has good capacitance behavior and excellent frequency response characteristics. After EIS testing of the flexible supercapacitor, whether it is capacitance value, equivalent series resistance, or RC time constant, it has met the requirements of the filter capacitor.

Next, to evaluate the feasibility of the assembled capacitor in actual AC filtering applications, a filtering circuit was built based on the Multisim platform and carried out simulation tests to confirm that the increase in capacitance promotes the optimization of filtering performance. Based on the oscilloscope platform, the actual filtering performance of the MnO_2-based flexible supercapacitor was tested.

The capacitor value used in the filter circuit is adjusted to 145 mF and the simulated circuit diagram and simulation results are shown in **Figure 3**.

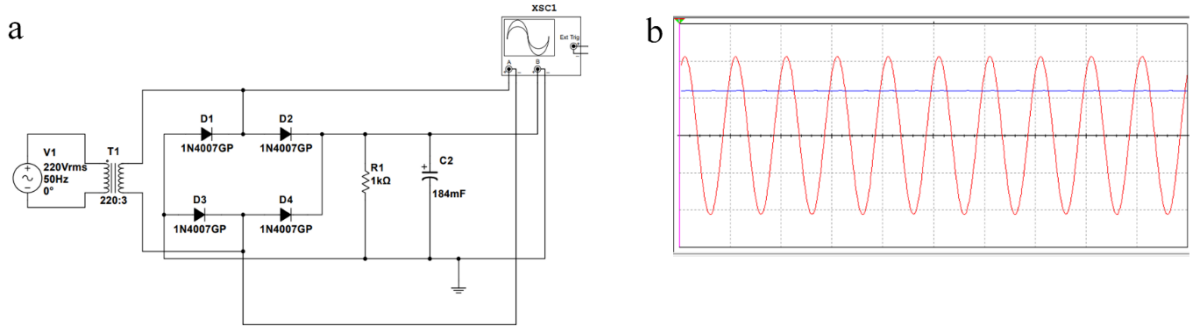


Figure 3. (a) Multisim filter circuit diagram; (b) AC filter simulation test results

From the simulation results, it can be seen that the assembled flexible asymmetric supercapacitor device can perform the filtering function of the AC signal very well, which preliminarily confirms the filtering performance of the MnO_2-based flexible asymmetric supercapacitor. To comprehensively evaluate the actual filtering of FASC, a filtering circuit was built using a bridge circuit, load resistors, and filter capacitors, as shown in **Figure 4**.

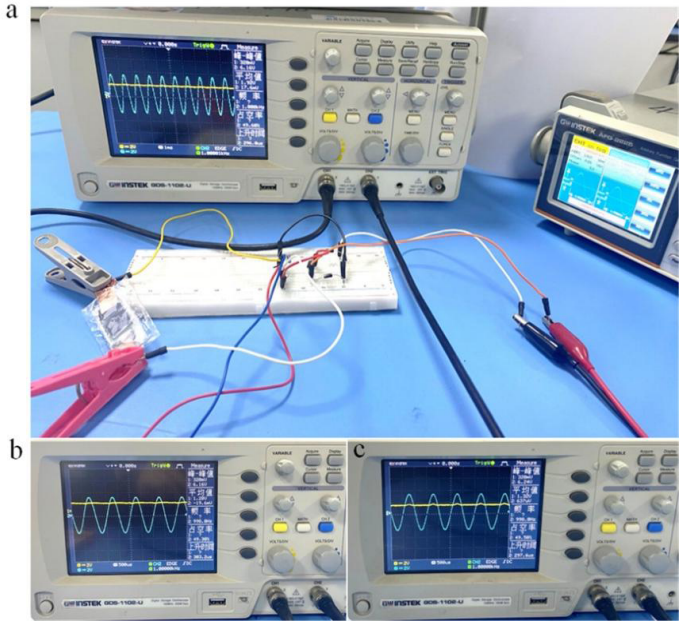


Figure 4. (a) Actual filtering display; (b) MnO_2-based FASC filtering result; (c) Ordinary aluminum electrolyte filtering result

The corresponding signal was input using a function generator (GWINSTEK AFG-2225), and the filtering output was displayed using a GWINSTEK GDS-11-2-U oscilloscope. Using a breadboard as a platform, a simple filtering circuit, as shown in **Figure 4(a)**, was built. The entire filtering circuit consists of four rectifier diodes (model: 1N4148) to form a bridge circuit, and a resistor with a resistance of 1 kΩ is used as a load. A 1000 Hz AC signal is filtered using an ordinary 220 μF aluminum electrolyte capacitor and a MnO_2-based flexible supercapacitor.

As shown in **Figure 4(b)**, the MnO_2 FASC can achieve AC filtering under sinusoidal AC signals. Under the action of the flexible supercapacitor, the input sinusoidal AC signal can successfully filter the 1000 Hz AC signal into a smooth DC signal through the prepared supercapacitor. In addition, a filtering test was conducted using an ordinary aluminum electrolyte capacitor, and the filtering effect was the same as that of the water-based δ-MnO_2// AC flexible supercapacitor, as shown in **Figure 4(c)**. Based on the oscilloscope results, the ripple factor was calculated. The ripple voltage (Vrpp) was determined from the peak-to-peak value (Vpp) using the formula:

$$Vrpp = 1/2 \times Vpp$$

The Vrpp calculated was 0.16V. Given that the input signal voltage is 3 V, the ripple factor, which is the ratio of the ripple voltage to the input signal voltage, was found to be 5.3%. This satisfies the requirement for a filtering ripple factor of less than 10%, confirming that the MnO2-based FASC can effectively serve as an AC filter.

After verifying the feasibility of the assembled δ-MnO_2//AC flexible asymmetric supercapacitor in AC filtering, the filtering capabilities of the δ-MnO_2//AC flexible asymmetric supercapacitor and the traditional aluminum electrolyte capacitor at the same frequency was compared. The rectifier filter circuit built on the breadboard was also replaced with the δ-MnO_2//AC flexible asymmetric supercapacitor and the traditional aluminum electrolyte capacitor prepared by us. The specific data results are shown in **Table 1**.

Table 1. Comparison of ripple coefficients of AEC and δ-MnO_2//AC FASC at different waveform signals at 1 Hz.

Capacitor Type	Frequency (Hz)	Input signal	Ripple Factor (%)	Is it satisfied? (≤ 10%)
Aluminum electrolytic capacitors	1	Sine wave	10.6	no
		Square wave	6.7	yes
		Triangle wave	9.3	yes
δ-MnO_2//AC FASC	1	Sine wave	8	yes
		Square wave	4	yes
		Triangle wave	5.3	yes

From the comparison results, it can be seen that compared with aluminum electrolyte, the filtering effect of δ-MnO_2//AC FASC at low frequency is better than that of aluminum electrolyte capacitors. Even at 1 Hz, when the input signal is a sine wave, the ripple factor of δ-MnO_2//AC FASC is less than 10%, while the ripple factor of traditional aluminum electrolyte capacitors is 10.6% (> 10%), indicating that δ-MnO_2//AC FASC can replace aluminum electrolytic capacitors in AC filtering, and δ-MnO_2//AC FASC is more suitable for AC filtering.

Conclusions

In summary, δ-MnO_2//AC FASC can achieve AC filtering under a variety of AC signals. Under the action of the aqueous δ-MnO_2//AC flexible supercapacitor, all sinusoidal AC signals can successfully output smooth signals through the prepared supercapacitor. In addition, even when the input signal is a square wave or a triangle wave, it still maintains basic filtering performance and the filtering coefficient is maintained at 2.7%–8%. The results verify the wide applicability of δ-MnO_2//AC FASC. At the same time, the device has advantages in AC filtering due to the smallest possible RC time constant, excellent capacitance value, and low equivalent series resistance, which also echoes the frequency response results of δ-MnO_2//AC FASC.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Ding Z, Cheng Z, Shi N E, et al., 2022, Dual-Electroactive Metal-Organic Framework Nanosheets as Negative Electrode Materials for Supercapacitors. Chemical Engineering Journal, 450: 11.

[2] Mariappan VK, Krishnamoorthy K, Manoharan S, et al., 2021, Electrospun Polymer-Derived Carbyne Supercapacitor for Alternating Current Line Filtering. Small, 17(34): 11.

[3] Li SL, Peng Z Y, Huang YT, et al., 2022, Electrostatic Self-Assembly of MXene and Carbon Nanotube@MnO2 Multilevel Hybrids for Achieving Fast Charge Storage Kinetics in Aqueous Asymmetric Supercapacitors. Journal of Materials Chemistry A, 10(44): 23886–23895.

[4] Raj J, Manikandan R, Sivakumar P, et al., 2022, Origin of Capacitance Decay for a Flower-Like δ-MnO2 Aqueous Supercapacitor Electrode: The Quantitative Surface and Electrochemical Analysis. Journal of Alloys and Compounds, 892: 9.

[5] Liu Q, Yang JJ, Luo X G, et al., 2020, Fabrication of a Fibrous MnO2@MXene/CNT Electrode for High-Performance Flexible Supercapacitor. Ceramics International, 46(8): 11874–11881.

[6] Yan J, Li S H, Lan B B, et al., 2020, Rational Design of Nanostructured Electrode Materials Toward Multifunctional Supercapacitors. Advanced Functional Materials, 30(2): 35.

[7] Zhang A Q, Zhao R, Hu L Y, et al., 2021, Adjusting the Coordination Environment of Mn Enhances Supercapacitor Performance of MnO2. Advanced Energy Materials, 11(32): 11.

[8] Zang XB, Wang JL, Qin YJ, et al., 2020, Enhancing Capacitance Performance of Ti3C2Tx MXene as Electrode Materials of Supercapacitors: From Controlled Preparation to Composite Structure Construction. Nano-Micro Letters, 12(1): 24.

[9] Zhao JQ, Xu ZJ, Zhou Z, et al., 2021, A Safe Flexible Self-Powered Wristband System by Integrating Defective

Wang JM, Huang Y, Du XP, et al., 2023, Hollow 1D Carbon Tube Core Anchored in Co3O4@SnS2 Multiple Shells for Constructing Binder-Free Electrodes of Flexible Supercapacitors. Chemical Engineering Journal, 464: 14.

[11] Li WY, Azam S, Dai G Z, et al., 2020, Prussian Blue-Based Vertical Graphene 3D Structures for High Frequency Electrochemical Capacitors. Energy Storage Materials, 32: 30–36.

[12] Zhao JH, Ma ZP, Qiao CT, et al., 2022, Spectroscopic Monitoring of the Electrode Process of MnO2@rGO Nanospheres and Its Application in High-Performance Flexible Micro-Supercapacitors. ACS Applied Materials & Interfaces, 14(30): 34686–34696.

[13] Zhang CG, Du HZ, MaK, et al., 2020, Ultrahigh-Rate Supercapacitor Based on Carbon Nano-Onion/Graphene Hybrid Structure Toward Compact Alternating Current Filter. Advanced Energy Materials, 10(43): 17.

[14] Zhang ST, Yang ZF, Cui CJ, et al., 2021, Ultrafast Nonvolatile Ionic Liquids-Based Supercapacitors with Al Foam-Enhanced Carbon Electrode. ACS Applied Materials & Interfaces, 13(45): 53904–53914.

[15] Jeanmairet G, Rotenberg B, Salanne M, 2022, Microscopic Simulations of Electrochemical Double-Layer Capacitors. Chemical Reviews, 122(12): 10860–10898.

[16] Xia CJ, Luo YJ, Bin XQ, et al., 2023, Rational Design of Flower-Like MnO2/Ti3C2Tx Composite Electrode for High Performance Supercapacitors. Nanotechnology, 34(25): 12.

[17] Mariappan VK, Krishnamoorthy K, Manoharan S, et al., 2021, Electrospun Polymer-Derived Carbyne Supercapacitor for Alternating Current Line Filtering. Small, 17(34): 11.

[18] Xue JL, Gao Z S, Xiao LY, et al., 2020, An Ultrafast Supercapacitor Based on 3D Ordered Porous Graphene Film with AC Line Filtering Performance. ACS Applied Energy Materials, 3(6): 5182–5189.

Integrated Services Platform of International Scientific Cooperation

Innoscience Research (Malaysia), which is global market oriented, was founded in 2016.Innoscience Research focuses on services based on scientific research. By cooperating with universities and scientific institutes all over the world, it performs medical researches to benefit human beings and promotes the interdisciplinary and international exchanges among researchers.

Innoscience Research covers biology, chemistry, physics and many other disciplines. It mainly focuses on the improvement of human health. It aims to promote the cooperation, exploration and exchange among researchers from different countries. By establishing platforms, Innoscience integrates the demands from different fields to realize the combination of clinical research and basic research and to accelerate and deepen the international scientific cooperation.

Cooperation Mode

Clinical Workers

In-service Doctors

Foreign Researchers

Hospital

University

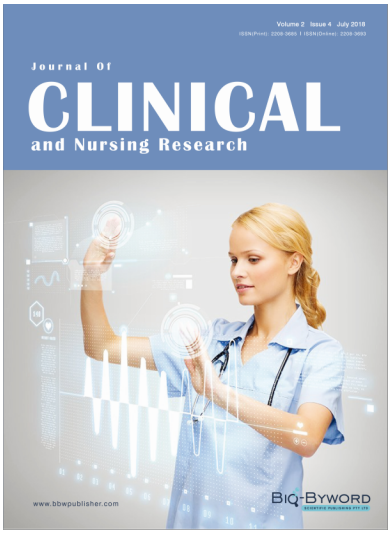
Scientific institutions

Our Journals

The *Journal of Architectural Research and Development* is an international peer-reviewed and open access journal which is devoted to establish a bridge between theory and practice in the fields of architectural and design research, urban planning and built environment research.

Topics covered but not limited to:

- Architectural design
- Architectural technology, including new technologies and energy saving technologies
- Architectural practice
- Urban planning
- Impacts of architecture on environment



Journal of Clinical and Nursing Research (JCNR) is an international, peer reviewed and open access journal that seeks to promote the development and exchange of knowledge which is directly relevant to all clinical and nursing research and practice. Articles which explore the meaning, prevention, treatment, outcome and impact of a high standard clinical and nursing practice and discipline are encouraged to be submitted as original article, review, case report, short communication and letters.

Topics covered by not limited to:

- Development of clinical and nursing research, evaluation, evidence-based practice and scientific enquiry
- Patients and family experiences of health care
- Clinical and nursing research to enhance patient safety and reduce harm to patients
- Ethics
- Clinical and Nursing history
- Medicine

Journal of Electronic Research and Application is an international, peer-reviewed and open access journal which publishes original articles, reviews, short communications, case studies and letters in the field of electronic research and application.

Topics covered but not limited to:

- Automation
- Circuit Analysis and Application
- Electric and Electronic Measurement Systems
- Electrical Engineering
- Electronic Materials
- Electronics and Communications Engineering
- Power Systems and Power Electronics
- Signal Processing
- Telecommunications Engineering
- Wireless and Mobile Communication