Shuihua Wang
Zheng Zhang
Yuan Xu (Eds.)

415

IoT and Big Data Technologies for Health Care

Second EAI International Conference, IoTCare 2021
Virtual Event, October 18–19, 2021
Proceedings, Part II

Part 2

EAI

Springer

Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 415

More information about this series at

Shuihua Wang · Zheng Zhang · Yuan Xu (Eds.)

IoT and Big Data Technologies for Health Care

Second EAI International Conference, IoTCare 2021
Virtual Event, October 18–19, 2021
Proceedings, Part II

Springer

Editors
Shuihua Wang 🔟
University of Leicester
Leicester, UK

Zheng Zhang
Harbin Institute of Technology
Shenzhen, China

Yuan Xu 🔟
University of Jinan
Jinan, China

Preface

We are delighted to introduce the proceedings of the second edition of the European Alliance for Innovation (EAI) International Conference on IoT and Big Data Technologies for HealthCare (IoTCare 2021). This conference brought together researchers, developers, and practitioners around the world who are leveraging and developing technology for Internet of Things (IoT) and big data applications in healthcare. The theme of IoTCare 2021 was the convergence of IoT and big data technologies for e-health, e-care, lifestyle, aging populations, smart personal living applications, etc.

The technical program of IoTCare 2021 consisted of 80 full papers in the oral presentation sessions at the main conference track on integrating healthcare with IoT. Aside from the high-quality technical paper presentations, the technical program also featured two keynote speeches and two technical workshops. The two keynote speeches were given by Manu Malek from Stevens Institute of Technology, USA, and Ivan Tyukin, Turing Artificial Intelligence Fellow, from the University of Leicester, UK. The two workshops organized were AI-based Internet of Medical Things (IoMT) and Information Fusion for the Devices of Internet of Things (InfusIoT). IoMT aimed to address the combination of IoT and AI to enable more personalized, preventative, and collaborative forms of IoT care. InfusIoT aimed to address the issue of information fusion for IoT devices to maintain high accuracy, adaptiveness, robustness, timeliness, reliability, and intelligence.

Coordination with the steering chairs, Imrich Chlamtac and Liangxiu Han was essential for the success of the conference. We sincerely appreciate their constant support and guidance. It was also a great pleasure to work with such an excellent organizing committee team for their hard work in organizing and supporting the conference. In particular, we are grateful to the Technical Program Committee, who have completed the peer-review process for the technical papers and helped to put together a high-quality technical program. We are also grateful to Conference Manager Natasha Onofrei for her support and all the authors who submitted their papers to the IoTCare 2021 conference and workshops.

We strongly believe that the IoTCare conference provides a good forum for all researchers, developers, and practitioners to discuss all science and technology aspects that are relevant to IoT in healthcare. We also expect that the future IoTCare conferences will be as successful and stimulating as this year's, as indicated by the contributions presented in this volume.

November 2021   Shuihua Wang
                                                                                                              Yu-Dong Zhang

Organization

Steering Committee

Imrich Chlamtac	University of Trento, Italy
Liangxiu Han	Manchester Metropolitan University, UK

Organizing Committee

General Chair

Shui-Hua Wang	University of Leicester, UK

General Co-chair

Shuai Liu	Hunan Normal University, China

Technical Program Committee Chair

Yu-Dong Zhang	University of Leicester, UK

Technical Program Committee Co-chairs

Jin Sun	Yangzhou University, China
Ruidan Su	Shanghai Advanced Research Institute, Chinese Academy of Sciences, China
Zheng Zhang	Harbin Institute of Technology, China
Juan Manuel Górriz Sáez	Universidad de Granada, Spain
Muhammad Attique Khan	HITEC University, Pakistan

Sponsorship and Exhibit Chair

Zhanhan Tu	University of Leicester, UK

Local Chair

Xiang Yu	University of Leicester, UK

Workshops Chairs

Yuan Xu	University of Jinan, China
Zedong Zheng	De Montfort University, UK

Publications Chairs

Chong Zeng	University of Leicester, UK
Xujing Yao	University of Leicester, UK

Web Chair

Siyuan Lu	University of Leicester, UK

Technical Program Committee

Chun Guang	Henan Vocational College of Industry and Information Technology, China
Chunli Guo	Inner Mongolia University, China
Cui Jianfeng	Xiamen University of Technology, China
Dan Zhang	Xinyang Vocational and Technical College, China
Dongye Liu	Inner Mongolia University, China
Feng Cheng	Xizang Minzu University, China
Gaocheng Liu	Inner Mongolia University, China
Hao Xu	Xinyang Vocational and Technical College, China
Huadong Wang	Inner Mongolia University, China
Jamal Hussain Shah	COMSATS University Islamabad, Pakistan
Jiangyi Zhang	Jiangnan University, China
Jianwei Zhang	Zhengzhou University of Light Industry, China
Junaid Ali Khan	HITEC University, Pakistan
Kashif Javed	National University of Science and Technology, Pakistan
Keming Mao	Northeastern University, China
Li Heng	Henan Finance University, China
M. Hassaballah	South Valley University, Egypt
Ma Lei	Beijing Polytechnic, China
Mingcheng Peng	Jiangmen Vocational and Polytechnic College, China
Mudassar Raza	COMSATS University Islamabad, Pakistan
Muhammad Attique Khan	COMSATS University Islamabad, Pakistan
Muhammad Sharif	COMSATS University Islamabad, Pakistan
Muhammad Younus Javed	HITEC University, Pakistan
Na Ta	Inner Mongolia University, China
Rajinikanth Venkatesan	St Joseph's College of Engineering, India
Rameez Naqvi	COMSATS University Islamabad, Pakistan
Reham Mostafa	Mansoura University, Egypt
Robertas Damasevicius	Vytautas Magnus University, Lithuania
Shuai Liu	Hunan Normal University, China
Shuai Yang	Changchun University of Technology, China

Shuihua Wang   University of Leicester, UK
Siyuan Lu   University of Leicester, UK
Sui Dan   Cal Poly Pomona, USA
Tallha Akram   COMSATS University Islamabad, Pakistan
Tenghui He   Hunan Normal University, China
Tian Hong   Baotou Iron and Steel Vocational Technical
                              College West Gate, China
Tong Xuanyue   Nanyang Institute of Technology, China
Vishnu Varthanan   Kalasalingam Academy of Research and
                              Education, India
Weiling Bai   Inner Mongolia University, China
Wenda Xie   Jiangmen Polytechnic, China
Xiang Yu   University of Leicester, UK
Xiaogang Zhu   Nanchang University, China
Xinchun Zhou   Baoji University of Arts and Sciences, China
Xinyu Liu   Hunan Normal University, China
Xuechao Zhang   Hulunbuir Vocational Technical College, China
Xujing Yao   University of Leicester, UK
Yanning Zhang   Beijing Polytechnic, China
Yi Chen   Nanjing Normal University, China
Yuling Jin   Chizhou Vocational and Technical College, China
Zhengchao Dong   Columbia University, USA

Contents – Part II

Information Fusion for the Devices of Internet of Things

AI-Based Internet of Medical Things

Contents – Part I

Integrating Healthcare with IoT

Data Transmission Reliability Detection of Hybrid Information System Based on Smart Contract

Huan-yu Wang[1][✉] and Xiao-gang Ma[2]

[1] College of Earth Environment and Science, Southwest Jiao Tong University,
Chengdu 611756, China
`why1924984571@163.com`
[2] Shandong Vocational College of Science and Technology, Weifang 261053, China

Abstract. There is a large amount of data in the information system, but the calculation process of the data transmission reliability detection method of the previous manual detection system is complex, and it is difficult to ensure the accuracy of the detection. Therefore, a data transmission reliability detection method of hybrid information system based on Intelligent contract is proposed. On the basis of designing the detection rules and design rule reuse rate compiler of the front-end input module, the system data stream used for detection is set. The machine learning algorithm is used to test the reliability of data transmission in the hybrid information system. The experimental results show that the hit rate of data transmission reliability detection based on smart contract is 96.9%, and the false alarm rate is 15.3%. Compared with mythirl tool and manual detection method, it improves the accuracy of data transmission reliability detection.

Keyword: Smart contract · Information system · Data transmission · Reliability monitoring

1 Introduction

Smart contract is the networked representation of traditional contract or protocol, which needs to be stored on the blockchain platform; At the same time, according to the contract code logic, the user input is received and the appropriate output is obtained. Traditional contracts need mutual trust agreements signed by two or more parties, so smart contracts should also meet the trust of the whole network blockchain user nodes [1]. Smart contract can have a secure and stable storage medium and operating environment, precisely because the blockchain of smart contract has the characteristics of decentralization, information transparency and difficult to tamper, which greatly expands the scope of application. However, the key feature of smart contract is that its execution does not rely on any credit endorsement, and the execution of various terms in the contract does not need to rely on an authoritative third party.

The traditional data transmission reliability detection method of hybrid information system is difficult to effectively use the contract, which leads to the poor reliability of the detection results. Therefore, this paper proposes a data transmission reliability detection method of hybrid information system based on Intelligent contract, which transforms the fault analysis problem into a logical analysis problem, rather than a physical analysis problem [2]. For the logical behavior of data transmission reliability, it is difficult to analyze its failure mode. Usually, several defects can be mapped into a fault model, or a many to one mapping. The data transmission reliability detection of hybrid information system based on smart contract is to analyze from the aspects of architecture security analysis, source code vulnerability analysis, binary vulnerability analysis and operating system vulnerability analysis, so as to find the vulnerability of data transmission reliability. This can not only effectively detect the code defects of the contract, but also ensure the high security of the hybrid information system based on smart contract, so it has a wide application prospect [3].

2   Data Transmission Reliability Detection Method of Hybrid Information System Based on Smart Contract

With the rapid development of integrated circuit technology, the traditional bus or point-to-point interconnection structure has become increasingly serious in scalability, communication efficiency, power consumption and other aspects, and data transmission has become a prominent performance bottleneck. Therefore, the data transmission of the hybrid information system describes the failure effect through the changes derived from the signals in the circuit or system [4]. For the reliability detection of data transmission in the hybrid information system of smart contract, the fault model is used in the circuit detection instead of the failure mode.

2.1   Front End Input Module

The front end of hybrid information system based on smart contract is mainly responsible for analyzing and inputting the types of smart contract, and selecting the detection rules of different data transmission according to the development language or bytecode type [5]. The front-end input module of the hybrid information system is mainly to receive the smart contract provided by the user, analyze the specific types of the contract content, and select the appropriate data transmission detection rules. At present, data transmission reliability detection of hybrid information system based on smart contract supports Ethereum, EOS, ont and other blockchain platforms. The types and suffixes of smart contracts corresponding to these platforms are shown in Table 1.

Because Ethereum bytecode is a series of bytes array generated directly, it is not saved in specific file, so it can not be judged directly by suffix [6]. Therefore, the user can select the correct data transmission reliability detection rules by specifying the contract type parameters. After selecting the type of intelligent contract, the front-end input module of hybrid information system needs to judge the current version according to the declaration information in the intelligent contract, so as to ensure that the subsequent modules can be correctly executed.

Table 1. Smart contract information supported by front end input module

Blockchain platform	Smart contract	Source code	Bytecode
Ethernet	Solidity	.sol	/
	Vyper	.vy	/
EOS	C++	.cpp	.wasm
ONT	Python	.py	.avm

The use of rigorous mathematical reasoning and logical proof is the key to test the attribute security and function correctness of intelligent contract. The hybrid information system based on Intelligent contract can completely cover the behavior during code execution, thus eliminating all possible input defects in traditional testing methods. At the same time, it ensures the absolute correctness of intelligent contract in a certain range, and provides some assistance for formal verification of intelligent contract through the semantic representation of virtual machine.

In the front-end design of hybrid information system of intelligent contract, CORBA is a specified document, which is mainly used to describe the transmission data. To some extent, CORBA can provide a good technical standard for hybrid information system distributed in abnormal environment. CORBA specifies that documents can interact effectively with data in different programming languages or operating systems. In addition, CORBA provides that simultaneous interpreting of different transmission data can be considered as an object in the front end design of mixed information system, whether it runs on the client side or server side.

Each CORBA specification document will publish an interface within the hybrid information system, which lists its methods and the data types it supports. At the same time, CORBA is an image of server application in hybrid information system, and it is not related to its language program. Java, as well as c+ language, are the main programming languages in the system, and each server needs to follow the interface definition principles.

Based on the specification of CORBA, DCOM can effectively solve the limitation of the front end of hybrid information system. DCOM can make the data transfer location transparent, and in general, the way that the customer connects the component and invokes the component is the same.

In addition, in DCOM hybrid information system, its location is obviously independent, which can effectively simplify the detection process in the process of data transmission. The location independence flow of DCOM in the system is shown in Fig. 1.

In the design of the front-end module of hybrid information system, DCOM detects the reliability of data transmission by keeping an index count for each component. But in the process of data transmission, hybrid information system based on Intelligent contract can not only transmit to one customer, but also can be shared by multiple customers.
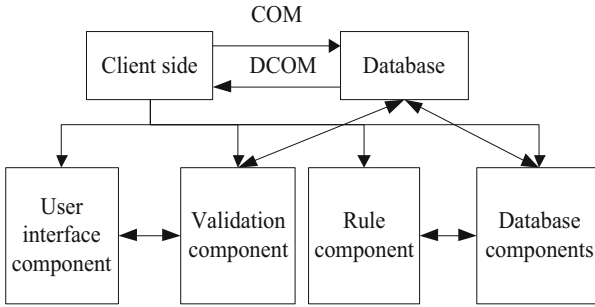
Fig. 1. Location independence

Due to the independence of DCOM in the system, it is necessary to transfer the distributed components of the front end of the system to other servers. Meanwhile, DCOM, which has no state or need not share with other components, can run on different servers and even generate multiple copies of data transmission detection in hybrid information system. When DCOM is used in the front end of the system, it is very easy to change the process of data transmission, but the same data transmission channel does not need to change the transmission path or recompile it.

In the front-end design of the hybrid information system of intelligent contract, the whole system shares a database and adopts the distributed application server. During data transmission, the application server is set up, and each system program server needs to be configured with two network connections. One connects to the database server through its internal network, while the other connects to the LAN to serve the client. The network structure of hybrid information system is shown in Fig. 2.
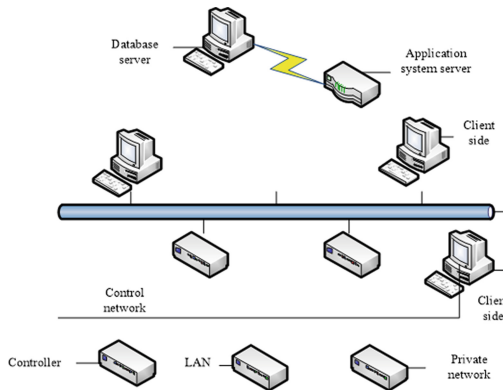
Fig. 2. Network structure of hybrid information system

In the front-end design of the hybrid information system of smart contract, a three-tier distributed architecture is adopted. The C/S part is developed by Delphi 7, while the B/S part is developed by Delphi 7 ASP.NET development. In the whole mixed information system, there are both C/S programs and B/S mixed mode in some interfaces, that is,

browser is embedded to access B/S pages on the server side. To a certain extent, the system structure based on CS/BS mixed mode can greatly improve the superiority of DCOM through the operation effect of HTTP mode.

2.2　Compiler Module

The data transmission of hybrid information system based on smart contract first needs to collect a large number of verified smart contract transmission data, and label them manually and by tool scanning. Before compiling the transmission data in the hybrid information system, we need to abstract the source code, bytecode and operation code, and pave the way for the subsequent feature extraction.

In the compiler module of hybrid information system based on smart contract, it mainly includes: lexical analysis, semantic analysis, intermediate code generation, code optimization, code generation and other steps [7]. Among them, lexical analysis and syntax analysis reorganize the input smart contract into an abstract syntax tree, and the code generation part outputs the bytecode of the contract, which can be used for the analysis of data transmission reliability detection.

Through the one vs. rest strategy based on the smart contract hybrid information system to transmit data for multi label classification. At the same time, in order to keep the balance of the number of samples between training sets, we need to balance the vulnerability of dependence on data transmission order, the vulnerability of not checking return value in the process of data transmission, the vulnerability of dependence on data transmission timestamp, and the vulnerability of code reentry in the process of data transmission by smote and smotetomek. The flow chart is shown in Fig. 3.
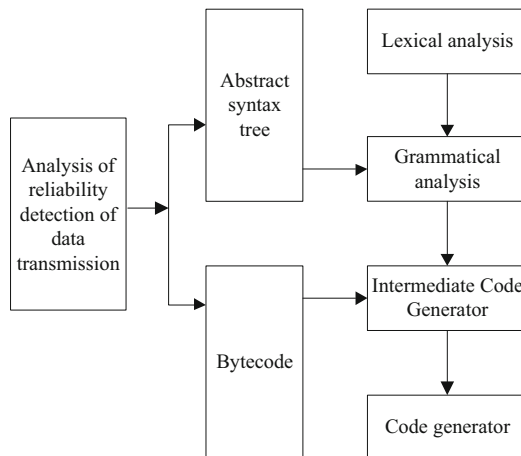
Fig. 3. Compiler module flow of data transmission reliability monitoring system

In the hybrid information system based on smart contract, the compiler module is designed mainly around the development language of smart contract. In the data transmission detection and analysis of the system, the smart contract source code of

the front-end input module can be divided into four types: C++, python, solidness and Vyper [8]. By using llvm and other compiler framework systems, different development languages are converted into the same set of intermediate code to generate the same bytecode. At this time, the data transmission reliability analysis module only needs to design rules for different types of smart contracts at the level of abstract syntax tree, and can share a set of rules at the level of intermediate code and bytecode to improve the reuse rate of rules.

However, most of the hybrid information systems based on smart contracts have official compilers, and the data transmission reliability test results of different contracts are often different. Using a common compilation framework will inevitably lead to problems such as difficult development and long-term maintenance. In addition, the data transmission detection of the hybrid information system also needs to consider the case that the input is the on chain bytecode of the smart contract [9]. Therefore, in the hybrid information system based on smart contract, the design scheme of compiler module for data transmission reliability detection and analysis is to integrate the official compiler module of each smart contract. At the same time, the implementation interface is provided for the new compiler to facilitate the subsequent expansion. The structure of the compiler module in the data transmission detection system is shown in Fig. 4.
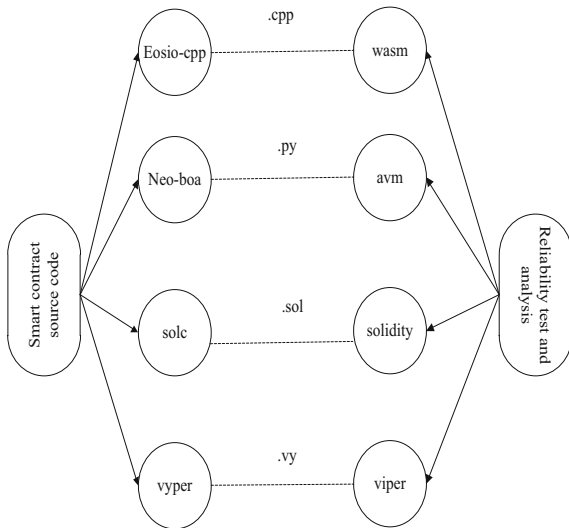
Fig. 4. Compiler module structure of data transmission reliability testing system

According to the hybrid information system of smart contract, the official compiler tools of smart contract are eosin CPP, Neo boa, Solc and Vyper. Eosin CPP is a tool chain of eosin. CDT, which can compile EOS smart contract into wasm bytecode and generate corresponding ABI file. Neo boa compiler module supports Python language subset. The purpose is to compile Python smart contract on ont platform into. AVM format and run it on Neo virtual machine. Both Solc and Vyper are compiler tools for

Ethereum smart contract. The former is for solidness smart contract, and the latter is Viper smart contract. Both of them provide the output interface of abstract syntax tree.

2.3 Data Parameter Setting of Mixed Information System

As a program analysis technology, the hybrid information system based on smart contract provides powerful and accurate formal verification. In the mixed information system, the execution data symbol is a function that uses the symbolic value instead of the input data to transform the program input into the symbolic input. In addition, in the hybrid information system, it can not only run directly on the virtual machine bytecode, but also provide the function of compiling solid source code.

In this paper, we analyze and deal with some specific vulnerabilities of smart contract, such as exception of error handling, transaction order dependence, timestamp dependence, and reentry. In view of the arrival execution path and the output of reliability parameters in the hybrid information system, different smart contracts are used to detect vulnerabilities.

The data flow analysis is a technology used to collect the information of computer programs at different points, which is widely used in the process of program compilation and optimization. Because some characteristics or properties of program data flow are closely related to the reliability of data transmission [10–13]. Therefore, data flow analysis can also be used as an important data transmission reliability detection and analysis technology.

In the data transmission reliability detection and analysis of hybrid information system based on smart contract, data flow analysis technology can be directly applied to the analysis of system software, and analyze the loopholes or defects of various data transmission reliability detection. In addition, data flow analysis is also a supporting technology of data transmission reliability detection vulnerability analysis, which can provide important data support for other vulnerability analysis methods.

In the vulnerability analysis of data transmission reliability detection, data flow analysis can be divided into flow insensitive analysis, flow sensitive analysis and path sensitive analysis according to the analysis accuracy of program path. One form of data flow analysis is to establish a data flow equation for a node in the control flow graph, and then solve it repeatedly through iterative calculation until it reaches a fixed point. The function of data flow analysis is relatively powerful, and it is generally combined with control flow information to solve the active variables with simple operation.

When setting the reliability detection parameters of data transmission based on Intelligent contract, the basic idea is an automatic vulnerability detection technology based on defect injection. It mainly uses the input of a large amount of abnormal data to the target program and monitors the running state of the target program, and uses whether there is any abnormality in the program as a sign to determine whether there is a potential vulnerability. Among them, the abnormal data detected is also called semi effective data. Without changing the parameter type, it will directly lead to the crash of hybrid information system or error triggered related loopholes in the process of data transmission.

In practical application, the hybrid information system based on Intelligent contract is deployed on the public chain. Its deployment and execution contract need to be verified by the generation of new area blocks. Therefore, on the basis of building private chain,

the contract to be tested is deployed to the private chain for testing. Under the premise of deploying public chain, we need to have private chain account when the mixed information system is set up initially, so as to reduce the waiting for data transmission detection on the public chain. Therefore, the specific flow of data transmission reliability detection of intelligent contract is shown in Fig. 5:
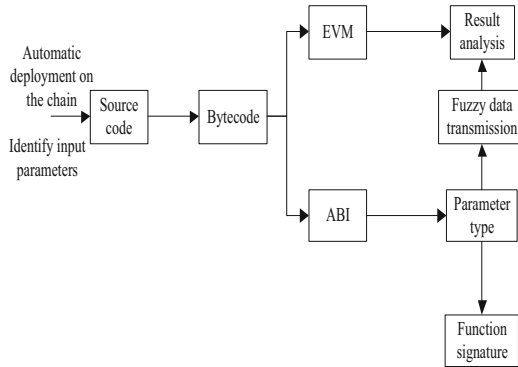
Fig. 5. Data transmission reliability detection process based on smart contract

To sum up, the deployment and execution of data transmission on the blockchain are realized through shell programming. Therefore, an automatic deployment tool based on smart contract is designed, including compiler call, smart contract deployment with geth, and contract code conversion. In the process of analyzing the data transmission reliability of hybrid information system, it is necessary to extract the data type of each ABI function parameter and the signature of the function used in each ABI function. For the fixed length input parameter type in the hybrid information system, it can realize the effective input generated randomly in the system. For the input parameter type with non fixed length, a positive number should be randomly generated as the length, and then selected randomly from the input field.

The data transmission reliability test of hybrid information system based on smart contract is carried out through Etherscan website etherscan.io/contractsVerified. You can view a contract list page on the web page. The list of data transmission reliability detection and analysis based on smart contract in hybrid information system shows the contract address, contract name, contract compiler version, encrypted data transmission capacity owned by the contract, and contract verification time and other important information of each smart contract. Click each contract link in the list to get to the details page of the corresponding smart contract.

In order to obtain a large number of required source code data of smart contract, it is necessary to use web crawler technology to automatically capture a large number of program scripts of Internet information according to preset rules. Web crawler starts from one or more seed URLs and gradually extends to the whole web page of the target website. Generally, crawling strategies can be divided into depth first strategy and breadth first strategy. This time, the depth first strategy is adopted for crawling, and the specific process is shown in Fig. 6.
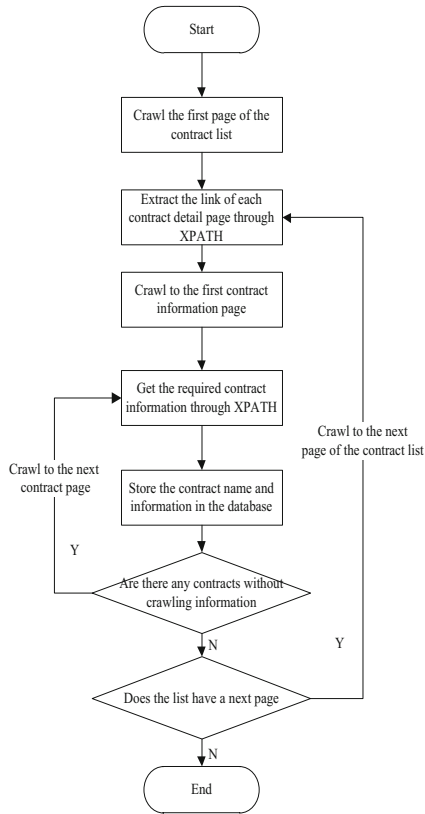
Fig. 6. Smart contract crawling process

XPath is a language used to determine the location of elements in XML or HTML markup language. Web crawlers usually use XPath to find the target elements in HTML. In addition, in order to speed up the crawling process, we need to optimize the crawling process with multiple threads.

In order to make different parallel threads cooperate with each other better, and prevent the crawling efficiency from decreasing due to the repeated data crawling, it is necessary to add crawling status fields in the database, with 0, 1 and 2 representing the three states of not crawling, crawling and crawling respectively. At the same time, the data transmission reliability detection of hybrid information system based on smart contract needs to lock the token table every time the crawling state is changed, so as to prevent multiple threads from crawling the same token repeatedly and affect the crawling rate.

2.4 Application of Machine Learning Algorithm to Realize Detection

The reliability detection of data transmission in hybrid information system based on Intelligent contract needs to compile the source code into bytecode through the Solc

compiler after climbing to the source code of the intelligent contract solidity. However, solidity language is a young language, and its version updates are very frequent. In the crawled intelligent contract data, different versions of solidity source code will be successfully compiled into hexadecimal byte code. At the same time, according to the corresponding relationship between byte code and operation code specified in intelligent contract, the corresponding relationship is mapped into data dictionary, and then byte code is analyzed into the operation code data stream according to data dictionary. According to statistics, after data preprocessing, the average length of each contract operating code is 4364.

At the same time, in the process of data transmission reliability detection of intelligent contract hybrid information system, it will appear on the basis of existing legal test input, and can generate reliable data transmission channel through random mutation strategy, boundary value variation strategy and phase variation strategy. However, it is blind to generate completely random data transmission in hybrid information system. To some extent, not only the efficiency of reliability detection is greatly reduced, but also the coverage of transmission data code can not be effectively improved.

Based on the analysis of the stain in the intelligent contract, the input data is screened and classified, and the genetic algorithm provides guidance for the reliability detection of the transmission data in the system. By simulating the optimal solution algorithm of data transmission reliability in hybrid information system, the recombining, mutation and selection of the algorithm are introduced into the solution calculation of the optimal solution.

While retaining the optimal individual locally, the better individuals are obtained through cross, mutation and selection, and the optimal detection results of data transmission reliability are obtained through iteration. The specific data transmission reliability detection flow is shown in Fig. 7:
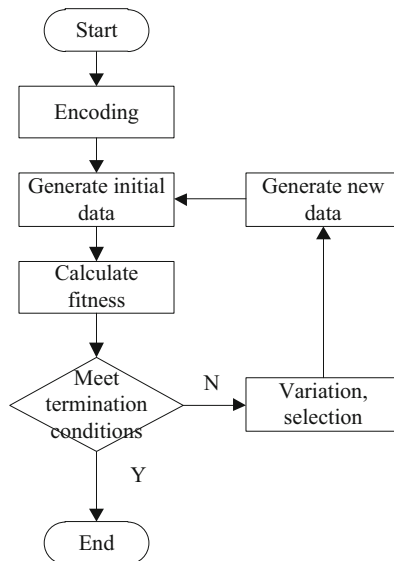
Fig. 7. Detection algorithm flow

If we use n-gram algorithm to extract features directly for all opcodes, because there are many kinds of opcodes and a large number of different operands, it will lead to too many features and cause dimension disaster. Through the analysis of the characteristics of solidness opcodes, it is found that many opcodes have similar functions, and some opcodes have similar vulnerability effects on data transmission reliability detection. The abstract rules of opcode are shown in Table 2.

Table 2. Opcode abstraction rules

Abstract opcode	Original opcode
ARITHMETIC_OP	ADD, MUL, SUB, DIV, SDIV, SMOD, MOD, EXP
CONSTANT1	BLOCKHASH, TIMESTAMP, NUMBER, DIFFICULTY, GASLIMIT
CONSTANT2	ADDRESS, ORIGIN, CALLER
COMPARISON	LT, GT, SLT, SGT
LOGIC_OP	AND, OR, XOR, NOT
DUP	DUP1-DUP16
SWAP	SWAP1-SWAP16
PUSH	PUSH5-PUSH32
LOG	LOG1-LOG4

In the data transmission reliability detection of hybrid information system based on smart contract, each PUSH instruction is followed by an operand. The number of operands varies from 1 byte to 32 bytes. These operands have little connection with the data transmission reliability detection and can be deleted.In addition, BLOCKHASH, TIMESTAMP, NUMBER, DIFFICULTY, GASLIMIT and COINBASE in class block information represent the information of the block, so they are classified into one category. After opcode abstraction, more than 140 opcodes are abstracted into more than 50, which greatly reduces the difficulty of feature dimension and machine learning model training in the subsequent feature extraction process.

By constructing the characteristic matrix FS, each smart contract has a corresponding characteristic vector in the matrix. The specific expression is as follows:

$$FS = \begin{bmatrix} f_{1,1}\ f_{1,2}\ \cdots f_{1,1619} \\ f_{2,1}\ f_{2,2}\ \cdots f_{2,1619} \\ \cdots \\ f_{i,1}\ f_{i,2}\ \cdots f_{i,1619} \\ \cdots \end{bmatrix} \tag{1}$$

where $FS = \begin{bmatrix} f_{i,1}\ f_{i,2}\ \cdots f_{i,1619} \end{bmatrix}$ is the eigenvector of the i th contract. The eigenvalue of the eigenvector of a contract is $f_{i,j}$, which is calculated according to the ratio of the

number of occurrences of each bigram in the contract to the total number of bigrams in the contract. Furthermore, it can be expressed by formula 2:

$$f_{i,j} = \frac{n_{j,ci}}{n_{all,ci}} \tag{2}$$

where $f_{i,j}$ represents the characteristic frequency of the j bigram in the i contract, and is a decimal number between 0 and 1. In addition, $n_{j,ci}$ represents the number of j bigrams in the i contract; $n_{all,ci}$ represents the total number of all bigrams in the i contract. If the j bigram feature is not present in the i contract, the corresponding $f_{i,j}$ value is 0.

By analyzing the virtual machine operation principle of opcode rules, the abstract rules of opcode are proposed, so that the number of bigram features extracted from the abstracted opcode data stream based on n-gram algorithm is greatly reduced. In the process of data transmission reliability detection of hybrid information system based on smart contract, it is necessary to explore all executable paths in the contract. In addition, the loop body in the smart contract sometimes needs a lot of iterations and consumes a lot of detection time. Therefore, the proposed data transmission detection method based on machine learning algorithm is more suitable for fast and automatic detection of a large number of smart contract vulnerabilities.

3   Experiment

3.1   Experimental Preparation

The test of data transmission reliability of hybrid information system based on Intelligent contract. Because the development language of hybrid information system is python, the test computer needs to install Python 3.6 and configure the environment. Developers can get the reliability test and analysis results of intelligent contract data transmission by executing Python script.

The development of hybrid information system based on smart contract is based on B/S architecture, the server provides services, the core code of the system is placed on the server, the client uses the browser to access, and there is no need to install a special client program. In addition, the front page of the experimental system is built with open source DWZ framework, the background program is written with java code, and based on s2sh framework, the data storage uses MySQL database, and its server program runs in Linux environment. The specific experimental running environment is shown in Table 3.

In the experiment, when users choose to upload the source code of the smart contract solidness, they need the system background to compile the source code to generate bytecode, which is completed by the solidness compilation module. In the background of the system, the solid compiler module integrates the solcj extended class library to compile the solid source code. The solcj extended class library is the Java implementation of the solid compiler Solc.

Table 3. System operation environment

Project	Model or version
Server model	ThinkServer RD640
CPU	Xecn E5-2660
Memory	320 GB
Operating system	Ubuntu 18.04
MYSQL	5.7.26
Solc	V0.4.25
JDK	1.8.0_161-b12

3.2 Experimental Process

In the experimental process of data transmission reliability detection of smart contract hybrid information system, the data transmission reliability of smart contract uploaded by users can be detected. The system requires the user to upload the source code or bytecode of the smart contract in the user interface. If the user chooses to upload the source code of the smart contract, the system background will automatically compile the source code into bytecode.

In addition, the bytecode is preprocessed in the hybrid information system, including bytecode analysis, opcode abstraction and other steps, and then the n-gram algorithm is used to extract the features of the data to obtain the feature vector of the target smart contract. Then the system loads the machine learning model trained by the existing data set in the background to detect and analyze the data transmission reliability of the smart contract, The test results can be saved in the database. The specific process is shown in Fig. 8.

In the information output module, according to the results of data transmission reliability analysis and line number mapping library, the vulnerability information of data transmission reliability detection is packaged again and output according to a certain format. Among them, in the hybrid information system, the output text format of vulnerability information of data transmission reliability detection is: contract, description, title, type, etc., while line represents the line number of data transmission reliability detection code. In addition, in order to facilitate ordinary users to use the security detection system, the front-end web page is designed to display and use the function of data transmission reliability detection and analysis of hybrid information system based on Intelligent contract, so as to improve the interactivity.

The data transmission reliability detection module is the core module of the hybrid information system. In order to get the best model for reliability detection, xgboost multi label classification model is used to detect the data transmission reliability of the target smart contract code; Then the feature vector data generated by the data preprocessing module is input into the multi label classification model loaded by the data transmission reliability detection module for reliability detection, and the prediction results are output.
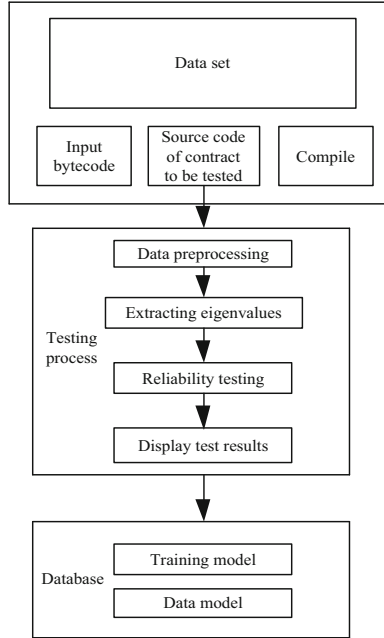
Fig. 8. Smart contract vulnerability detection process

3.3 Experimental Result

Data transmission reliability detection of hybrid information system based on smart contract refers to the number of data transmission reliability vulnerabilities that can be detected by the analysis tool of smart contract hybrid information system, and then reflects the perfection of vulnerability Library of hybrid information system. In order to detect the accuracy of data transmission reliability detection of hybrid information system based on smart contract, the two dimensions of hit rate and false alarm rate are tested and analyzed by calculating the deviation degree between the actual output and the expected result. In the case of detecting the same number of data transmission reliability test cases, the output results of the traditional monitoring method and the monitoring system based on smart contract are compared. The specific experimental comparison results are shown in Table 4.

According to the calculation results, the hit rate of data transmission reliability detection based on smart contract is 96.9%; The hit rate of mythirl tool is 36.3%; The hit rate of manual detection is only 26.7%. In addition, the false alarm rate of data transmission reliability detection of hybrid information system based on smart contract is 15.3%, which is far lower than that of mythirl tool and manual detection. Therefore, the experimental results show that the detection ability of data transmission reliability detection of hybrid information system based on smart contract is strong, and the accuracy of data transmission reliability detection is improved.

Table 4. Comparison of detection accuracy of various methods

/	Detection system based on smart contract	Mythirl	Manual detection
Number of test cases	184	184	184
Total number of questions	655	655	655
Number of hits	635	240	175
Number of false positives	115	226	167
Hit rate	96.9%	36.6%	26.7%
False positive rate	15.3%	48.5%	48.8%

4   Conclusion

The data transmission reliability test of a mixed information system based on smart contracts is a tool that provides symbolic abstraction methods through various security attributes in smart contracts, which can effectively detect the reliability of data transmission. This tool symbolically studies the dependency graph between the contract bytecodes, abstracts accurate semantic paradigms from it, and matches the three semantically prescribed patterns.

Compared with other methods of detecting the reliability of data transmission in the system, it greatly reduces the amount of manual tasks, and explores all contract paths by using a custom compliance model. To a certain extent, the false negatives caused by symbolic execution tools are effectively filtered, thereby improving the reliability of all possible path execution.

References

1. Huang, Y., Li, T., Luo, C., Fujita, H., Horng, S.-J., Wang, Bin.: Dynamic maintenance of rough approximations in multi-source hybrid information systems. Information Sciences 530 (2020)
2. Li, Z., Zhang, P., Xie, N., Zhang, G., Wen, C.-F.: A novel three-way decision method in a hybrid information system with images and its application in medical diagnosis. Engineering Applications of Artificial Intelligence 92 (2020)
3. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mob. Netw. Appl. **24**(1), 5–17 (2018). https://doi.org/10.1007/s11036-018-1134-8
4. Zeng, J., Li, Z., Zhang, P., Wang P.: Information structures and uncertainty measures in a hybrid information system: gaussian kernel method. Int. J. Fuzzy Sys. **22**(1), 212–231 (2020)
5. Xiong, W., Xiong, L.: Data resource protection based on smart contract. Computers & Security 98 (2020)
6. Yu, G.: Information structures and uncertainty measures in a hybrid information system with images. Soft Computing **23**(24), 12961–12979 (2019)
7. Liu, S., Fu, W., He, L., Zhou, J., Ma, M.: Distribution of primary additional errors in fractal encoding method. Multimedia Tools and Applications **76**(4), 5787–5802 (2014). https://doi.org/10.1007/s11042-014-2408-1

8. Technology - Technology Research: Studies from international business machines corporation yield new data on technology research (Automatic smart contract generation using controlled natural language and template). Comp. Netw. Comm. (2019)

9. Wang, P., Zhang, P., Li, Z.: A three-way decision method based on gaussian kernel in a hybrid information system with images: an application in medical diagnosis. Appli. Soft Compu. J. **77**, 734–749 (2019)

10. Hun, K.Y.: A study on smart contract for personal information protection. J. Digi. Converg. **17**(3), 215–220 (2019)

11. Xuan, S., et al.: An incentive mechanism for data sharing based on blockchain with smart contracts. Comp. Elec. Eng. **83**, 106587 (2020)

12. ZhongAn Information Technology Service Co. Ltd.: Patent Application Titled "Smart Contract Upgrade Method And System Based On Alliance Chain" Published Online (USPTO 20190278767). Computer Technology Journal (2019)

13. Liu, S., Pan, Z., Cheng, X.: A novel fast fractal image compression method based on distance clustering in high dimensional sphere surface. Fractals **25**(4), 1740004 (2017)

Research on Network Security Authentication Method Based on Data Mining Technology

Xiao-gang Ma[1][(✉)] and Huan-yu Wang[2]

[1] Shandong Vocational College of Science and Technology, Weifang 261053, China
`maxiaogang698@aliyun.com`
[2] College of Earth Environment and Science, Southwest Jiao Tong University,
Chengdu 611756, China

Abstract. In order to solve the problems of low authentication accuracy, long authentication time and poor authentication security in traditional network security authentication methods, this paper uses data mining technology to design a new network security authentication method. First, analyze the types of attacks on the network by illegal nodes on the network and the principles of authentication, and then mine the data to be authenticated through the binary network. In order to reduce the mining error, the acquired data is punished and integrated. In this process, in order to ensure the effective iteration of the data, the neural network algorithm in the machine learning algorithm is introduced for in-depth mining. The experimental results show that the authentication accuracy of this method can reach up to 98%, and the authentication time is always less than 2 s. The above results show that: after adopting this method, the network security performance can be improved.

Keywords: Network security · Data mining · Machine learning algorithm · Safety certification

1 Introduction

With the rapid development of information technology, the Internet has become an inseparable part of people's daily life, and has been widely used in various fields such as business, communications, entertainment, education, and personal daily affairs [1]. Therefore, a large amount of valuable information needs to be disseminated through the network. However, this also provides a way for various illegal intrusions against computer information systems [2].

At present, the applications of wireless public networks in people's lives are becoming more and more popular and diversified. The large amount of data and complex structure of the wireless public network means that it has characteristics such as complexity and dynamics. In this way, internal and external security risks have brought great pressure and challenges to network security work [3, 4]. And with the rapid development and extensive coverage of the Internet in recent years, the forms of cyber attacks have

become more diverse, and security threats have become deeper and deeper. The issue of network security needs urgent attention.

Reference [5] proposed a security authentication method for wireless sensor networks. In this method, the biometric key of the user is obtained by fuzzy extraction technology, and then the mutual authentication and session key negotiation between the legitimate user and the sensor node are completed with the help of the hash algorithm, which reduces the calculation cost. Finally, heuristic security analysis, BAN logic and ROM model are used to verify the security and effectiveness of the proposed scheme. Reference [6] proposed a computer network security routing method based on genetic algorithm. This method quantifies link security based on authentication access control and encryption mechanism, and then constructs a multi-objective secure routing model based on QoS parameters. According to the allocation of the common buffer pool and the minimum reserved bandwidth, the multi-objective secure routing model is selected to optimize the target. The objective function minimization problem of computer network secure routing is transformed into a maximization problem. Finally, the operator is selected for crossover and mutation, and the individual with the best fitness value is determined by genetic algorithm to realize the optimization of computer network secure routing. Reference [7] put forward the research method of encryption authentication and network security for Ethernet. The security of Ethernet is the research object of this method. On the basis of establishing the Ethernet model, AES-128 encryption algorithm and HMAC-SHA1 security authentication method and technology are introduced into Ethernet, so as to effectively prevent external intrusion, data jump and other risks, and further improve the security performance of Ethernet and bus network.

Although the above methods have achieved certain results, there are still problems such as low certification accuracy and long certification time. For this reason, this paper proposes a network security authentication method based on data mining technology. The design idea of the new method is as follows:

(1) Based on the analysis of the types of network attacks by illegal nodes and the principles of authentication, the data to be authenticated are mined through the binary network.
(2) In order to reduce the error of mining, the obtained data are punished and integrated. In this process, in order to ensure the effective iteration of the data, the neural network algorithm of machine learning algorithm is introduced to carry out deep mining.

2 Method Research

2.1 Data Mining Type Analysis

The process of data mining can be roughly divided into five stages: problem definition, data collection, data preprocessing, data mining implementation, and the interpretation and evaluation of mining results [8]. The principle of data mining is shown in Fig. 1.

Step 1: Problem definition. At this stage, a clear pre-achieved goal and user needs need to be determined.
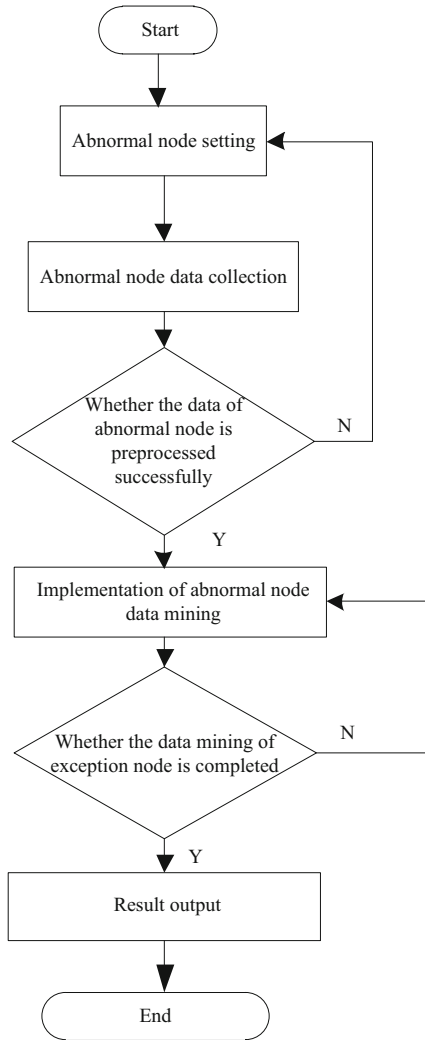
Fig. 1. Schematic diagram of data mining

Step 2: Data collection. After determining the target data objects involved in the data mining task, extract the data sets related to the mining task from the relevant data sources.
Step 3: Data preprocessing. Data preprocessing includes operations such as noise point elimination, data formatting, and data reduction.
Step 4: Data mining implementation. According to the task of data mining and existing methods (classification, clustering, outlier mining, etc.), data mining algorithms are selected, and the results are obtained by mining the obtained data.
Step 5: Interpretation and evaluation of results. At this stage, it is necessary to evaluate and analyze the results obtained by data mining in order to discover meaningful knowledge patterns.

Before adopting data mining technology, it is also necessary to have a detailed understanding of the data that needs to be mined. The types of risk data included in the network security certification are many and complex. Failure to effectively classify them will result in failure of the certification security or affect its safe operation. Therefore, this article first effectively divides the data types that need security authentication on the network, and implements effective factor design methods for them. The types of data that need to be mined for network security certification include the following:

(1) The continuous or offline value of the real number field is the most common target data form in the data mining method:
(2) Discrete integer field values are also quite common. In addition to expressing the meaning of their own attributes, discrete integer field values are also used to escape character data and enumerated data to support mining work.

2.2 Data Mining Algorithm Analysis

In network security certification, data mining technology is an important technology to improve its security. There are many mining algorithms in this technology. In order to achieve network security certification, this article analyzes the data mining algorithm in detail.

The binary network mining algorithm is an algorithm that can reduce the algorithm space. It determines the data relationship in the network security authentication through the network topology relationship [9], which can be expressed as:

$$A_i = \frac{1}{n} \sum_{i=1}^{n} \frac{a_i b \delta}{n(b_i)} \tag{1}$$

$$f(x) = \sum_{j=1}^{m} A_i f'(x) \tag{2}$$

Due to some errors in the data relationship obtained above, this paper punishes the trust relationship between data in network authentication through the high-order relationship, the process is as follows:

$$f(x) = \left(Ab^2 + Bb^3\right)f(a_i) \tag{3}$$

$$f'(x) = \sum_{i=1}^{n} \left(Ab^2 + Bb^3\right)f(a_i) \tag{4}$$

On the basis of binary network algorithm mining, it is believed that there is a greater degree of association and similarity between network security authentication data. Therefore, it is necessary to process the similarity of the network security authentication data, namely:

$$A_i = \frac{1}{k} \sum_{i=1}^{m} \frac{a_i \delta_i}{k(a_i \delta_i)} \tag{5}$$

In the process of network security authentication, due to the constant changes in the network environment, the number of data in the network that needs to be authenticated and the degree of data security need to be updated constantly, which increases the complexity of the authentication process, resulting in poor results and efficiency of authentication. For this reason, machine learning algorithm is introduced in the authentication process of network authentication data.

Machine learning algorithms are more familiar with the intelligent algorithm of neural network. By constructing the network level, the network security authentication data is reasonably trained, and the weight information is transmitted through the key cells in the network, namely:

$$C(x) = \sum_{x=1}^{n} p_n\left(v_n^t X\right) \tag{6}$$

After the above-mentioned security authentication data weight is transmitted, the information in the hidden layer is effectively converted, and the processing method of the intermediate layer can be obtained:

$$u_i = \frac{1}{1 + \exp\left(\sum_{i=1}^{n} w_i + \theta_j\right)} \tag{7}$$

After the hidden information is revealed, the non-secure data that needs to be authenticated can be obtained after valid authentication, namely:

$$p(a, b) = \frac{1}{1 + l_i uy} \tag{8}$$

Logical regression and support vector machines both train the parameters of the model through feature samples. When the parameters approached the optimal training solution, the training was stopped, and the results were outputted according to the model of unknown sample vectors. Typical logistic regression training methods include Newton's method, conjugate gradient method, quasi Newton's method and so on. At present, the Hessian matrix is used to solve the logistic regression model through the quasi-Newton method, which overcomes the disadvantage that the parameter gradient descent direction of the traditional Newton method cannot converge along the direction of the optimal solution. At the same time, the traditional Newton method is transformed into the least square iterative optimization process, which greatly improves the convergence speed of network security authentication [10].

In the research of data mining technology, the mining of non - secure data is the key. Therefore, the fully authenticated data is first mined through a binary network. In order to reduce the error of mining, we punish the acquired data, and automatically introduce the neural network algorithm in machine learning algorithm. Through deep mining, we complete the deep mining of data in the process of network security authentication.

2.3 Data Analysis and Network Security Certification

Network security certification is actually a process of accumulating normal data node data and non-secure data and ordering the certification process. The schematic diagram of the network security authentication node is shown in Fig. 2.
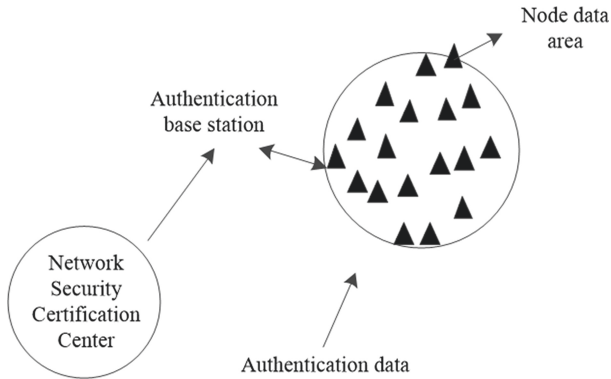
Fig. 2. Schematic diagram of network security authentication node

(1) Identity password stealing attack: The illegal node obtains the identity password of the legal node to access the system resources authorized to the legal node.
(2) Traffic analysis attack: The illegal node analyzes the information exchanged between the communicating parties, trying to judge or recover the original information.
(3) Retransmission attack: The illegal node intercepts the information sent by the sender, and then retransmits it to the receiver.
(4) Modification or forgery attack: The illegal node intercepts the information sent by the sender, replaces or modifies the information and then transmits it to the receiver, or pretends to be a legitimate node to send information [11].
(5) Denial of service attacks: prevent legitimate nodes from legally managing and using system resources.
(6) External attacks and internal attacks: External attacks refer to attacks initiated by external nodes when WSN communicates with external nodes. This external node is a newly added node.

In response to the above-mentioned attacks on illegal nodes, network security services should provide the following authentication functions:

(1) Reliability: The source of the information is credible, that is, the recipient of the information can confirm that the information obtained is not sent by a counterfeit.
(2) Integrity: The integrity of the information is required to be guaranteed during the transmission process, that is, the information recipient can confirm that the acquired information has not been modified, delayed or replaced during the transmission process [12].
(3) Non-repudiation: It is required that the sender of the information cannot deny the information it sends, and similarly, the receiver of the information cannot deny the received information.
(4) Access control: Illegal nodes cannot access the system resources of the network, and legal users can only access the resources authorized and designated by the system.

3   Experiment Analysis

In order to verify the effectiveness of the network security authentication method under the above-designed data mining technology, the method in this paper is compared with the method in Reference [5] and the method in Reference [6]. Through multiple iterations, the verification and analysis of the effectiveness of the method in this paper are realized.

3.1   Lab Environment

In the design of the experimental environment, since network authentication is an abstract process, the commonly used MATLAB platform is used for simulation analysis. The experimental operating system is WINDOWS 10.

3.2   Experimental Parameter Design

In order to meet the requirements of the experiment, the parameters of the experiment are designed, and the specific parameters are shown in Table 1.

Table 1.   Experimental parameter design

Parameter	Numerical value
Network security node/a	1000
Number of secure nodes	600
Number of non-secure nodes/piece	400
Authentication interval/s	0.5 s
Certification error/%	<2
Number of certification tests/time	100

According to the design of the above-mentioned experimental environment and parameters, the accuracy of the authentication node data, the authentication time-consuming and the post-authentication network security performance of the method in this paper, the method in Reference [5] and the method in Reference [6] are compared.

3.3   Analysis of Results

Analysis of the Accuracy of Node Authentication in Different Methods
The network security certification is mainly to authenticate the data of non-secure nodes with threats. Therefore, in the experiment, the method in this paper, the method in Reference [5] and the method in Reference [6] are compared to authenticate the sample security node data. In the experiment, the results are obtained multiple times to ensure the accuracy of the obtained results. The results obtained are shown in Fig. 3. Shown.
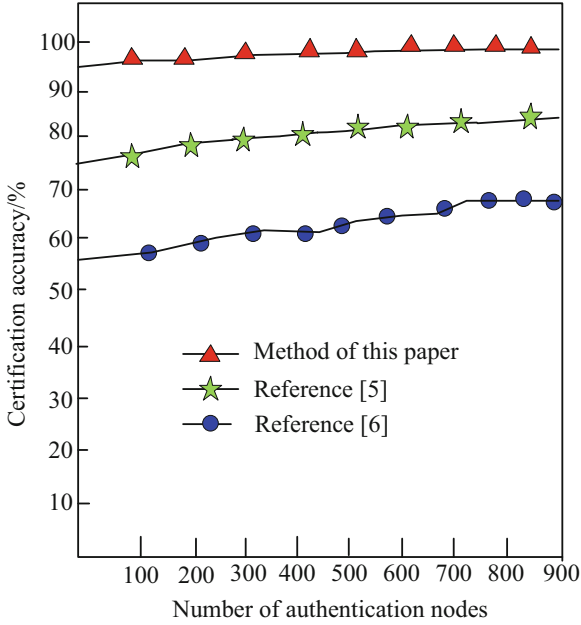
Fig. 3. Comparison results of node authentication accuracy of different methods

Analyzing the data in Fig. 3, it can be seen that with the continuous change of the number of authentication network nodes. The method in this paper, the method in Reference [5] and the method in Reference [6] have certain differences in the accuracy of the sample network node authentication. Among them, when the number of authenticated data nodes is 300, the accuracy of the method in this paper for authentication of sample network nodes is about 95%, and the accuracy of the method in Reference [5] for authentication of sample network nodes is about 87%, Reference [6] The accuracy of the method for authenticating sample network nodes is about 56%; when the number of authenticated data nodes is 600, the accuracy of the method in this paper for authenticating sample network nodes is about 96%. The method in Reference [5] is effective for authenticating sample network nodes. The accuracy is about 80%, and the accuracy of the Reference [6] method for the authentication of the sample network node is about 62%; when the number of authentication data nodes is 900, the accuracy of the method for the authentication of the sample network node is about 97%. The Reference [5] method has an accuracy of about 89% for the sample network node authentication, and the method in this paper has an accuracy of about 65% for the sample network node authentication. It can be seen that the authentication accuracy of the three methods shows an upward trend, but In comparison, the accuracy of the method in this article is higher than the method in Reference [5] and the method in Reference [6].

This is because the method in this paper determines the node attributes of authentication before authentication, and uses data mining technology to dig it deeply, which improves the authentication accuracy of the proposed method and has certain advantages.

Time-Consuming Analysis of Node Authentication in Different Methods
In network security certification, the time-consuming certification can show the performance of the network. If the certification time is shorter, the network performance is better. For this reason, comparing the time consumption of the method in this paper, the method in Reference [5] and the method in Reference [6] to authenticate the sample secure node data, the results obtained are shown in Fig. 4.
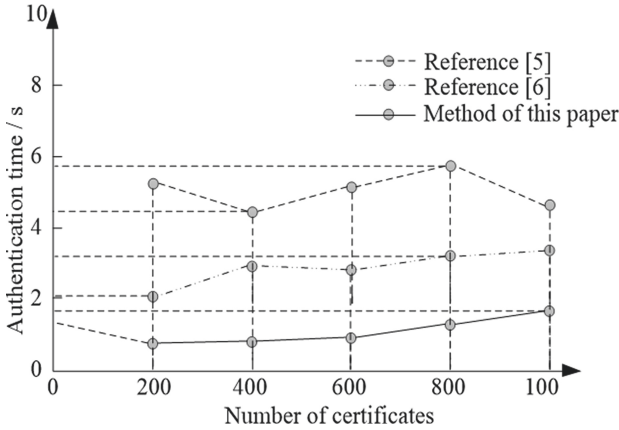
Fig. 4. Time-consuming analysis of node authentication by different methods

Analyzing the data in Fig. 4, it can be seen that under the same experimental environment, there are certain differences in the time-consuming authentication of the sample network nodes using the method of this paper, the method of Reference [5] and the method of Reference [6]. It can be seen from the curve in Fig. 4 that the verification time of the method in this paper is always less than 2 s, but the safety authentication time of the other two methods is lower than that of the method in this paper. This is because the method in this paper uses artificial intelligence algorithms to mine hidden data during data mining, which reduces the complexity of mining and improves the effectiveness of the method in this paper.

Security Analysis of Different Method Nodes after Authentication
The purpose of network security certification is to improve the security performance of the network. Therefore, the experiment analyzes the network security performance of the sample network nodes after the method of this paper, the method of reference [5] and the method of reference [6]. Among them, the higher the degree of security, the more secure the network performance. The results obtained are shown in Table 2.

Table 2. Comparison of security after authentication of different methods of nodes (%)

Number of iterations/time	Method of this paper	Method of Reference [5]	Method of Reference [6]
10	95	89	87
20	95	87	88
30	94	87	89
40	94	85	87
50	95	85	87
60	95	89	87
70	92	87	85
80	95	89	85
90	94	86	87
100	92	85	84

Analyzing the experimental data in Table 2, it can be seen that with the increase of the number of experimental certifications, the method of this paper, the method of Reference [5] and the method of Reference [6] have changed the security performance of the sample network node after authentication. Although the three methods have certain advantages in the security performance of the network security certification, the security performance of the method in this paper is better than that of the Reference [5] method and the Reference [6] method. This is due to the method in this paper. Continuously authenticate the mined data, strengthen the authentication details of non-secure data, and then enhance the effectiveness of the method in this paper.

4   Conclusion

Authentication of network security is the key to protect users' rights and interests. Therefore, this article uses data mining technology to effectively authenticate network security. Firstly, by analyzing the types of node data in the network and its characteristic attributes, in-depth mining of the unsafe nodes that exist through data mining technology, the research of network security authentication is realized. Compared with traditional methods, this method has the following advantages:

(1)   The accuracy of using this method for network security node authentication is about 98%, which can effectively authenticate network security nodes;
(2)   The time-consuming for network security node authentication using this method is always less than 2 s, which can quickly and effectively authenticate network security nodes;
(3)   The network security performance is improved after the method of this paper is used to authenticate the network security node, and the network security node can be effectively authenticated.

Although at the present stage, the method in this paper has achieved certain success and has certain help for network security authentication, there is still room for improvement. Therefore, in the future research center will carry out more in-depth research, and constantly enhance the strength of network security certification.

References

1. Hao, T.: The identity authentication of Wi-Fi system based on network security. Annals of telecommunications - annales des télécommunications **41**(10), 1–8 (2020)
2. Muqtadir, S.: Network security based on chaotic maps based authentication in vanets. J. Appli. Sci. Computa. **31**(5), 44–57 (2019)
3. Gao, S., et al.: A lightweight fingerprint-based device authentication architecture for wireless industrial automation networks. In: 2019 1st International Conference on Industrial Artificial Intelligence (IAI). IEEE (2019)
4. Guiga, L., Roscoe, A.: Neural network security: hiding CNN parameters with guided grad-CAM. In: 6th International Conference on Information Systems Security and Privacy (2020)
5. Alexander, R.: Using the latin square design model in the prioritzation of network security threats: a quantitative study. J. Inf. Secur. **11**(2), 92–102 (2020)
6. Haider, N., Baig, M.Z., Imran, M.: Artificial intelligence and machine learning in 5G network security: opportunities, advantages, and future research trends. Research Gate **23**(9), 303–319 (2020)
7. Zhao, M., et al.: CPTmilODIM method for bipolar fuzzy multi˚ttribute group decision making and its application to network security service provider selection. Int. J. Intell. Syst. **17**(4), 18–26 (2021)
8. Ahmad, M.B., Agarwal, P.: Viability of adaptive network security exercising tradeoff between performance and security. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science) **13**(5), 893–900 (2020)
9. Shuai, L., Liu, G.C., Zhou, H.Y.: A robust parallel object tracking method for illumination variations. Mob. Netw. Appl. **4**(1), 5–17 (2019)
10. Zhao, D., Song, H., Li, H.: Fuzzy integrated rough set theory situation feature extraction of network security. J. Intell. Fuzzy Sys. (1), 1–12 (2021)
11. Liu, S., Liu, G.C., Zhou, H.Y.: A robust parallel object tracking method for illumination variations. Mob. Netw. Appl. **24**(1), 5–17 (2019)
12. Deng, C.H., et al.: Research on attribute security verification method of space network based on complex network theory. Electr. Desi. Eng. **28**(3), 79–83 (2020)

High Reliability Design of Student Status Information Acquisition System in Ideological and Political Classroom Under Multi-target Tracking

Gui-xiu Xie[1(✉)] and Tao Lei[2]

[1] Guangdong Polytechnic College, Zhaoqing 526100, China
xx210428@163.com
[2] Physical and Electronic Engineering School, South China Normal University, Guangzhou 510006, China

Abstract. Aiming at the low reliability of the traditional ideological and political classroom student status information acquisition system, a high-reliability design of the ideological and political classroom student status information acquisition system under multi-target tracking is proposed. Under multi-objective tracking, the hardware design of the system is completed by designing the student status information storage module in the ideological and political classroom, designing the student status information analysis processor in the ideological and political classroom, and designing the student status information query module in the ideological and political classroom. By extracting the ontology of student status information in ideological and political classrooms and designing a program for obtaining student status information in ideological and political classrooms, the software design of the system has been completed and the acquisition of student status information in ideological and political classrooms has been achieved. The experimental results show that the ideological and political classroom student status information acquisition system under multi-target tracking has higher reliability in terms of accuracy, usefulness and recall.

Keywords: Multi-target tracking · Ideological and political class · Student status · Access to information · High reliability

1 Introduction

In recent years, with the development of education in colleges and universities, it is urgent to introduce advanced management technology of student status information, improve the level of education, improve the efficiency of obtaining student status information, so as to consolidate the comprehensive competitiveness of university education. In the daily operation of colleges and universities, the acquisition of student status information is a crucial task, and it is also a direct reflection of the level of Ideological and

political classroom management, which determines the quality and social image of Ideological and political classroom education in Colleges and universities. Therefore, how to actively introduce appropriate information technology and build a university student status information management platform is a key issue related to the development of university education [2].

In the United States, Europe and other developed countries and regions, many universities with technical strength have participated in the development of student status information collection system of Ideological and political classroom, especially in some developed countries in North America and Japan. Software development agencies cooperate to develop a practical system suitable for college students' management needs [3]. Literature [4] predicts students' performance through online learning behavior analysis. In terms of the development mode of the student status information collection system in the ideological and political classroom, most of the systems have developed from the original student status information collection system in the ideological and political classroom based on C/S mode to the current student status information collection system in the ideological and political classroom based on B/S mode, It realizes the integrated storage and distribution of data information, and the operability and concept of the system are gradually improving, which can greatly improve the efficiency and accuracy of the acquisition of student status information in the ideological and political classroom. It can realize data interconnection and information sharing with other systems. Literature [5] uses decision tree analysis to predict students' performance in higher education institutions. At present, in order to continuously improve the level and efficiency of student management, many foreign universities have joined some ERP design concepts to develop more systematic products. It can be seen that in many years of technology development and application practice, many foreign countries have accumulated rich experience in the development of the student status information collection system in the ideological and political classroom, and are still in the process of continuous improvement. It should be pointed out that the management mode of many foreign universities is different from that of China. Therefore, although the design concept and method can be used for reference in our country, it should be completely transformed into a system adapted to the students' environment in our country, rather than directly used [6].

The application of information technology in the acquisition of student status information in the ideological and political classroom of colleges and universities began in the late 1980s. Some qualified universities gradually apply the computer management mode to the acquisition of student status information in the ideological and political classroom [7, 8]. Many colleges and universities have begun to realize the efficiency and convenience brought by the management system based on information technology. Therefore, they began to deploy a more powerful and better performance student status information collection system [9] combined with their own internal network. Chinese universities can also be divided into different types. Different types of colleges and universities have their own characteristics and needs of obtaining student status information, so there is no general system to meet the needs for obtaining student status information of all colleges and universities [10–12].

Therefore, this paper proposes a high reliability design of the student status information acquisition system in the ideological and political classroom under multi-target

tracking. By using multi-target tracking algorithm, the state information of students in Ideological and political class is obtained, and its reliability is analyzed. Experiments show that the system can further improve the level of information management in colleges and universities, improve the status of teaching management. This is of great significance for strengthening the management of college students and providing a more stable and safe campus environment.

2 The Hardware Design of the Student Status Information Acquisition System in Ideological and Political Classroom

Under multi-target tracking, cloud computing is introduced to design the hardware of the student status information acquisition system. Cloud computing is a brand-new network service method that improves the traditional desktop-centric task processing method and transforms it into a network-core task processing, the use of Internet technology to achieve rapid information transmission [13]. The design purpose of the cloud computing-based ideological and political classroom student status information acquisition system is to realize the voluntary expansion of the system, unified management and efficient
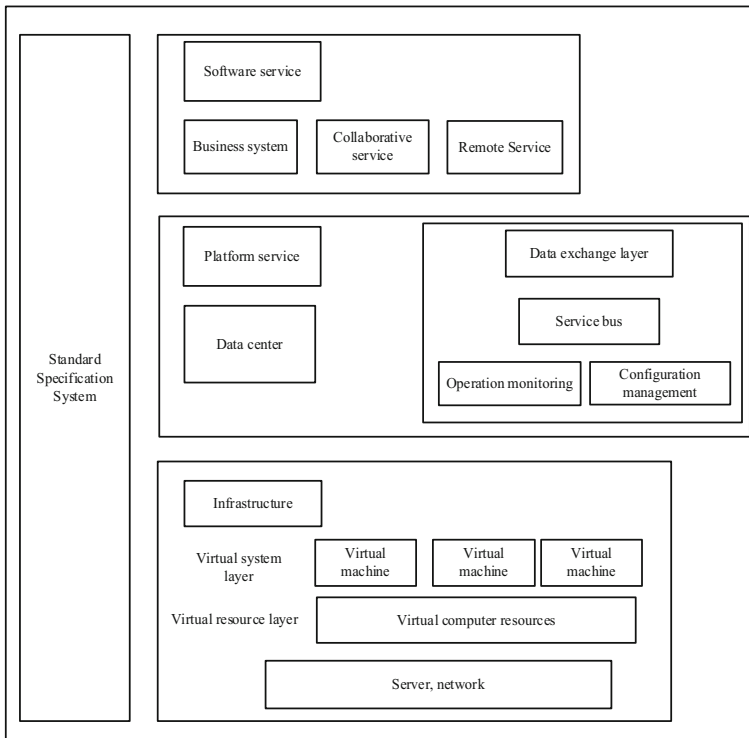
Fig. 1. Architecture diagram of student status information acquisition system in ideological and political classroom

query of large amounts of data [14]. The system is designed according to big data goals and business requirements, and the specific architecture is shown in Fig. 1.

It can be seen from Fig. 1: According to system requirements, the student status information acquisition system for ideological and political classrooms must be designed in a cloud computing environment. Business management expansion and horizontal expansion of system storage capacity can be carried out. Therefore, the ideological and political classroom student status information acquisition system designed with the concept of cloud computing has the advantage of large capacity, which can accommodate a large amount of ideological and political classroom student status information [15].

According to the overall system architecture, the system hardware structure is designed, as shown in Fig. 2.
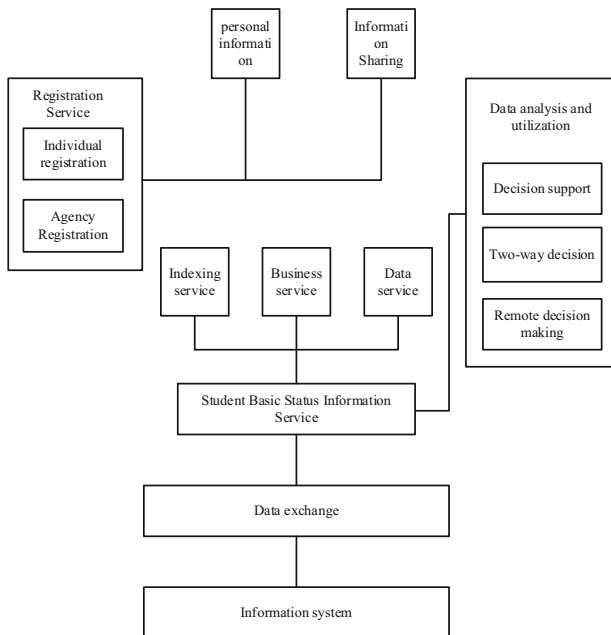
Fig. 2. The hardware structure diagram of the student status information acquisition system in the ideological and political classroom

It can be seen from Fig. 2 that the overall architecture of the student status information acquisition system in the ideological and political classroom is composed of the institutional layer and the platform layer. Among them: the institutional layer includes the information system, which is the main source of student status information in the ideological and political classroom. The platform layer is a data exchange platform with functions such as basic information services, registration services, information sharing services, and data analysis.

2.1 Ideological and Political Classroom Student Status Information Storage Module Design

Traditional ideological and political classroom student status information storage adopts a centralized serial storage method. The state information of students in ideological and political classrooms stored in this way is limited, and it is difficult to store the state information of students in ideological and political classrooms [16]. The multi-target database is the sum of the dynamic information related to the application on the physical storage medium of the system, and usually exists on the medium in a fixed structure.

For the storage of student status information in ideological and political classrooms under multi-target tracking, the cloud computing environment should be the center, and different devices should be connected to jointly complete the query of student status information in ideological and political classrooms. This process requires the establishment of a data center to store massive amounts of information in the server. This provides convenience for the analysis and processing of student status information in ideological and political classrooms. The structure design of student status information memory in ideological and political classrooms is shown in Fig. 3.
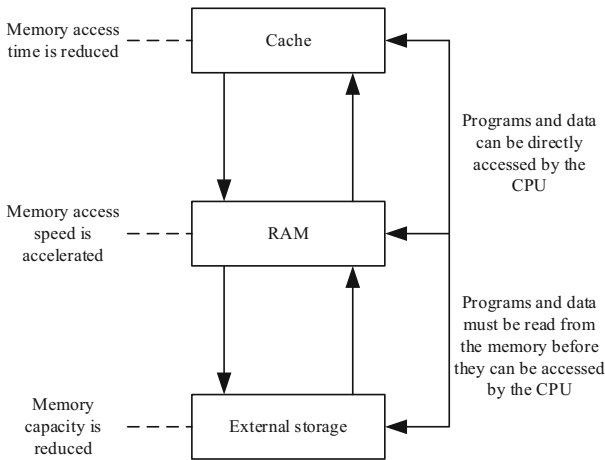
Fig. 3. Ideological and political classroom student status information memory structure diagram

It can be seen from Fig. 3: If there is a large amount of ideological and political classroom student status information, the information storage time will be greatly increased. Then it is necessary to increase the storage speed of student status information in ideological and political classrooms to ensure that the storage of dynamic information can save time. At this time, the stored state information of students in the ideological and political classroom can be directly accessed by the CPU. If the memory speeds up the storage of student status information in ideological and political classrooms, the internal fast conversion function needs to be used to convert the information into a memory form. At this time, the stored state information of students in the ideological and political classroom can be accessed by the CPU. If the mass information storage memory becomes smaller,

the information storage speed needs to be increased. The corresponding program is used to convert the data into the form of memory, and the data is accessed by the CPU through the cache [17].

The memory circuit diagram is shown as in Fig. 4.
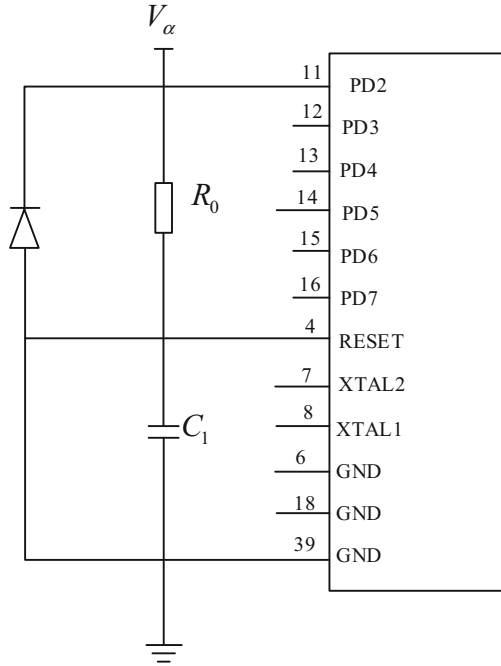
Fig. 4. Ideological and political classroom student status information memory circuit diagram

2.2 Ideological and Political Classroom Student Status Information Analysis Processor Design

This paper uses the ARM23MPCPR chip as the processor core. This chip is a new generation of V6 system processor chips based on the traditional ARM12. Different applications require different requirements. The processor can provide up to 6 combinations, and its performance can reach more than three times that of traditional processors. The processor structure is shown in Fig. 5.

It can be seen from Fig. 5 that the given processor is a relatively standard multi-core processor. The processor is composed of 4 independent cores of the traditional ARM12, which can improve performance without changing the frequency, except for this advantage. In addition, the other various application indicators of this processor are very good, suitable for most of the electronic system requirements on the market now, while still operating normally, there is still room to provide multiple multimedia functions [18]. In various application environments, this processor can show high performance and
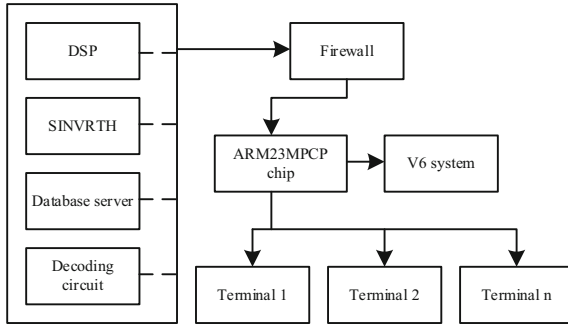
Fig. 5. The structure of the processor for analyzing the status of students in the ideological and political classroom

absolute advantages. The ARM23MPCPR chip can provide a multi-channel structure and support a hybrid system. This design greatly improves the flexibility, adaptability, and throughput of the processor. And computing power to meet various needs.

In terms of application, most manufacturers use traditional DSP and decoding circuits to implement encoding and decoding, and mainly use a single-core processor to complete the operation in the process. Although this method can decode efficiently, it has a large energy consumption problem. The architecture of this processor is a single-core architecture, which is similar to a single-core processor, but it is far superior to a single-core processor in terms of performance, including energy consumption and resource storage [19]. This processor incorporates the advantages of the Corntp-65 processor, uses the latest ARM architecture and V8 system, and integrates 4 72-bit storage units and processing units. It has very smooth acceleration performance and great flexibility.

The processor circuit diagram is shown as in Fig. 6.
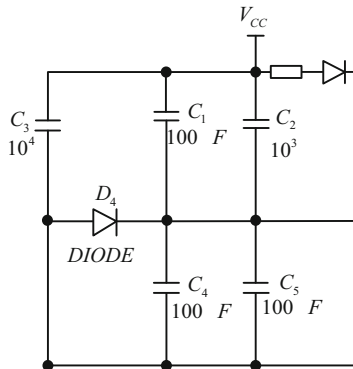
Fig. 6. The circuit diagram of the processor for analyzing the status of students in ideological and political classroom

2.3 Ideological and Political Classroom Student Status Information Query Module Design

According to the above-mentioned information analysis and processing results, the ideological and political classroom student status information query module is designed, as shown in Fig. 7.
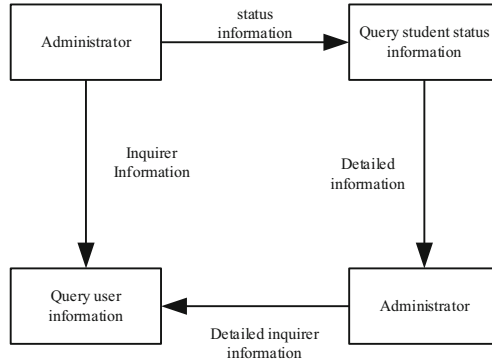
Fig. 7. Query module of student status information in ideological and political classroom

It can be seen from Fig. 7 that when the administrator wants to query the status information of the students in the ideological and political classroom, he needs to know the administrator's identity information first. In this process, the administrator's identity is verified, and it is determined whether it is a reader or an administrator. If it is an identity visitor, the identity information of the visitor needs to be known to facilitate the inquiry of the status information of the students in the ideological and political classroom.

The memory structure of student status information in the ideological and political classroom designed with multi-target tracking can save time. The internal fast conversion function is used to convert the student status information in the ideological and political classroom into a memory form, which is directly accessed by the CPU. The analysis and processing module analyzes certain historical records existing on the system and restores dormant files. According to the analysis and processing results of the student status information in the ideological and political classroom, the student status information query module in the ideological and political classroom is designed to ensure the safety of the system design.

3 Software Design of Student Status Information Acquisition System in Ideological and Political Classroom

3.1 Extraction of Student Status Information Ontology in Ideological and Political Classroom

In the process of designing the student status information acquisition system in the ideological and political classroom under multi-target tracking, the multi-target tracking

data is analyzed and processed first. The TF-IDF index of student status information in the ideological and political classroom is calculated based on the multi-target tracking data. The distribution model of student status information in the ideological and political classroom is established. On this basis, the ideological and political classroom student status information vector is calculated to extract the student status information ontology in the ideological and political classroom. The specific process is as follows:

Assume that K represents the subject collection of student status information in the ideological and political classroom searched by the user. M represents multi-target tracking data collection. N represents the number of ideological and political classroom student status information used in the multi-target tracking data. α, β represents a priori parameter. First, analyze and process the multi-target tracking data, filter the status information of students in ideological and political classrooms under multi-target tracking [20], and use Eq. (1) to calculate the TF-IDF index of student status information in ideological and political classrooms under multi-target tracking.

$$\mathrm{TF}_{i,j} - \mathrm{IDF}_i = \frac{f_{i,j}}{\sum\limits_{k=1}^{j} f_{k,j}} \times \log_2 \frac{N}{n_i} \tag{1}$$

where: TF represents the frequency of student status information in ideological and political classrooms under multi-target tracking data. IDF represents the inverse document frequency of student status information in ideological and political classrooms under multi-target tracking data. n_i represents the state information of students in the ideological and political classroom under the multi-target tracking data i the behavior set that has occurred. j represents a certain status information record under the multi-target tracking data. $f_{i,j}$ represents the number of times that the student status information i in the ideological and political classroom under the multi-target tracking data appears in j. $f_{k,j}$ represents the number of times that the topic k searched under the multi-target tracking data appears in j. From the Dirichlet distribution with the prior parameter α, extract the distribution model of student status information in the ideological and political classroom corresponding to each document M under the multi-target tracking data, and express it by formula (2):

$$\theta_m \sim \mathrm{Dir}(\alpha), \quad m \in [1, M] \tag{2}$$

In formula (2): θ_m represents the distribution model of student state information in the ideological and political classroom under the multi-target tracking data; $\mathrm{Dir}(\alpha)$ represents the Dirichlet distribution with the prior parameter α under the multi-target tracking data. Assuming that n_m represents the number of information that appears in the user's m state, use Eq. (3) to calculate the probability of generating ideological and political classroom student state information in the user's entire state library:

$$p(z|\alpha) = \prod_{m=1}^{M} p(z_m|\alpha) = \prod_{m=1}^{M} \frac{\Delta(n_m + \alpha)}{\Delta(\alpha)} \tag{3}$$

In formula (3): Δ represents the Laplacian operator; $p(z_m|\alpha)$ represents the probability of generating state information for students in the ideological and political classroom

in the m state of the user. The distribution model corresponding to the student state information K of each ideological and political classroom under the multi-target tracking data is extracted from the Dirichlet distribution with the prior parameter β, and the calculation is carried out using Eq. (4):

$$\phi_k \sim \text{Dir}(\beta), \quad k \in [1, K] \tag{4}$$

In formula (4):ϕ_k represents the distribution model of student state information in ideological and political classrooms under multi-target tracking data. $\text{Dir}(\beta)$ represents the Dirichlet distribution with the prior parameter β under the multi-target tracking data. Use formula (5) to calculate the probability of generating state information for all ideological and political classroom students under multi-target tracking data:

$$p(w|z, \beta) = p(w'|z'|, \beta) = \prod_{k=1}^{K} p(w_{(k)}|z_{(k)}, \beta) \tag{5}$$

In formula (5): $p(w'|z', \beta)$ represents the generation probability of student status information in ideological and political classrooms with a priori parameter of β under multi-target tracking data, and $p(w_{(k)}|z_{(k)}, \beta)$ represents the generation probability of student status information k in ideological and political classrooms under multi-target tracking data analysis. (3) Formula (5) is analyzed to obtain the joint distribution of the state information of all ideological and political classroom students under the multi-target tracking data, which can be expressed by formula (6):

$$p(w, z|\alpha, \beta) = p(w|z, \beta)p(z|\alpha) \tag{6}$$

Assuming that $Z_{m,n}$ represents the hidden content of student state information in the ideological and political classroom under the multi-target tracking data, combined with the distribution model of the student state information in the ideological and political classroom under the multi-target tracking data, the n ideological and political classroom corresponding to the multi-target tracking data m is calculated student status information is calculated using the following formula:

$$Z_{m,n} \sim \text{Mult}(\theta), \quad m \in [1, M], n \in [1, N_m] \tag{7}$$

$$W_{m,n} \sim \text{Mult}(Z_{m,n}) \tag{8}$$

In formula (8): $W_{m,n}$ represents the amount of student status information acquired in the ideological and political classroom under the multi-target tracking data. $\text{Mult}(Z_{m,n})$ represents the spatial threshold of the Multinomial conjugate distribution of the student state information distribution model in the ideological and political classroom under the user behavior data. $\text{Mult}(Z_{m,n})$ represents the spatial threshold of the Multinomial conjugate distribution with an implicit content of $Z_{m,n}$. Combining the Gibbs sampling method to solve the hidden content Z of the student state information in the ideological and political classroom under the multi-target tracking data, extract the student state information samples in the ideological and political classroom, and calculate the weight under the multi-target tracking data.

3.2   Program Design for Acquiring Status Information of Students in Ideological and Political Classroom

After constructing the hardware system used in this paper, the software environment is set up. The basic operating system of the software processor is the Linux system. Using u-boot configuration, the software system can perceive the status information of students in the ideological and political classroom in real time. The interface shows the collected status information characteristics of the students in the ideological and political classroom.

The set software flow chart is shown in Fig. 8.
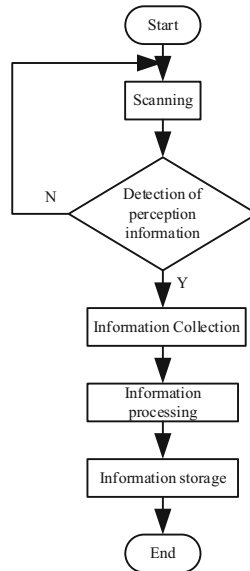
Fig. 8. Ideological and political classroom student status information acquisition program

It can be seen from Fig. 8 that the system software process designed in this paper is divided into three steps.

Step 1: Collect student status information in ideological and political classrooms. A variety of sensors are used to collect student status information in ideological and political classrooms, including learning information, behavior information, and action information. The information retrieved in the human-computer interaction interface is interacted with each other to achieve the purpose of auxiliary processing. Corresponding decision rules should be set during the collection process, and learning information, behavior information, and action information should be collected in a periodic manner. Multiple sensors are used to analyze whether the collected information meets the acquisition conditions. If the acquisition conditions are met, the acquisition service must be started, and if the acquisition conditions are not met, information collection cannot be performed. The collection period of behavior information is 5 s, and the collector has its

own buffer. The number of acquisition frames is diverse, and the data can be processed and stored.

Step 2: The collected information is processed. The collected ideological and political classroom student status information is processed. The effect of student status information in the ideological and political classroom has a direct impact on the effect of the system. When students are in different positions, their status is also different. Therefore, it can be judged whether the student is in the best condition by seeing the student's behavior.

Step 3: The results of student status information in the ideological and political classroom are stored. After the collected information is obtained, the collected information is stored in the memory. Useful information is selected for display. The storage system and the alarm system should be closely connected. Once the student's learning status is perceived to be poor, an alarm should be sounded immediately to remind students to take effective measures to improve their learning status.

In summary, a detailed design of the student status information acquisition system in the ideological and political classroom is carried out.

4   Experimental Comparative Analysis

4.1   The Purpose of the Experiment

In order to test the actual effect of the ideological and political classroom student status information acquisition system under this multi-target tracking, the method in this paper is compared with the traditional ideological and political classroom student status information acquisition system and the experimental results are analyzed.

4.2   Experimental Parameter Settings

The experimental parameters are set as shown in Table 1.

Table 1. Experimental parameters

Item	Parameter
Configuration software	U-boot
Operating Voltage/V	350
Operating current/A	200
System structure	Application layer, service layer, sensor layer
System functions	Data collection, data processing, information storage
Operating system	Linux 系统
Operating frequency/Hz	500
CPU memory/GB	12
Collection methods	Real-time collection

4.3   Results and Analysis

The traditional ideological and political classroom student status information acquisition system and the ideological and political classroom student status information acquisition system under multi-target tracking are used to carry out the design experiments of the student status information acquisition system in the ideological and political classroom. The two different acquisition systems are used to compare the accuracy of student status information acquisition in ideological and political classrooms, and the average absolute error MAE (%) is used as the accuracy comparison result to measure the student status information acquisition systems in different ideological and political classrooms, and the calculation is carried out using Eq. (9):

$$MAE = \frac{\sum\limits_{i=1}^{G} |p_i - q_i|}{G} \tag{9}$$

In formula (9): p_i represents the system's predicted scoring value for the obtained state information of the i ideological and political classroom student. q_i represents the actual scoring value of the obtained student status information of the i ideological and political classroom by the user. G represents the number of data in the ideological and political classroom student status information acquisition data set under multi-target tracking. The average absolute error of the acquisition of student status information in two different systems of ideological and political classrooms is compared, and the results are shown in Fig. 9.
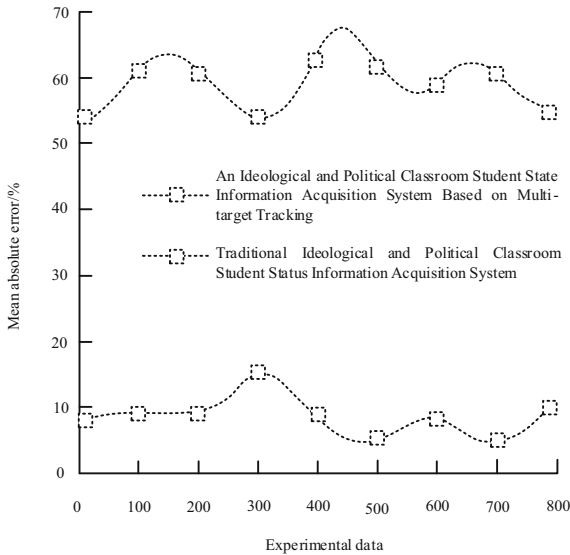
Fig. 9. Comparison results of average absolute error

Analysis of Fig. 9 shows that the average absolute error of using the ideological and political classroom student status information acquisition system under multi-target tracking to obtain the ideological and political classroom student status information is lower than that of the traditional ideological and political classroom student status information acquisition system. This is mainly because in the ideological and political classroom student status information acquisition system design process using the ideological and political classroom student status information acquisition system under multi-target tracking, the student status information in the ideological and political classroom is first analyzed and processed. The TF-IDF index of student status information in the ideological and political classroom under multi-target tracking is calculated. The Dirichlet distribution of the prior parameters is combined to establish the state information distribution model of ideological and political classroom students under multi-target tracking. Under multi-target tracking, the weight output of the student status information acquisition index in the ideological and political classroom is obtained. It makes the ideological and political classroom student status information acquisition system under multi-target tracking have a higher design accuracy for the ideological and political classroom student status information acquisition system under multi-target tracking.

The traditional ideological and political classroom student status information acquisition system and the ideological and political classroom student status information acquisition system under multi-target tracking are used to carry out the design experiments of the student status information acquisition system in the ideological and political classroom. The usefulness (%) and recall rate (%) of the two different systems for acquiring student status information in ideological and political classrooms were compared. Recall rate refers to the percentage of the total when users believe that the information obtained is useless. Usefulness refers to the degree of active availability of information acquisition. The comparison result is used to measure the comprehensive effectiveness of the information acquisition system design under the multi-target tracking of different systems. The comparison results are shown in Table 2 and Table 3.

Table 2. Comparison of usefulness of information acquisition systems under different systems (%)

Number of experiments/time	An ideological and political classroom student state information acquisition system based on multi-target tracking	Traditional ideological and political classroom student status information acquisition system
5	93.7	67.8
10	95.4	67.1
15	95.1	66.7
20	92.8	68.5
25	93.5	65.6
30	94.9	65.9

Table 3. Comparison of recall rates of information acquisition systems under different systems (%)

Number of experiments/time	An ideological and political classroom student state information acquisition system based on multi-target tracking	Traditional ideological and political classroom student status information acquisition system
5	7.1	22.6
10	6.8	20.7
15	7.0	19.6
20	6.6	19.9
25	6.9	21.5
30	7.3	20.9

Through the analysis of Table 2 and Table 3, it can be seen that the comprehensive performance of using the ideological and political classroom student status information acquisition system under multi-target tracking to obtain the student status information in the ideological and political classroom under multi-target tracking is higher than that of the traditional ideological and political classroom student status. Information acquisition system. This is mainly because in the process of acquiring information by using the ideological and political classroom student status information acquisition system under multi-target tracking, Internet users are abstracted as structured data information ontology. The nth state corresponding to the mth behavior of the user is calculated. The information ontology under the random behavior data of Internet users is extracted. On this basis, the user obtains the weight output of the information acquisition index by calculating the correlation between the query and the result record status information. The comprehensive performance of the student status information acquisition system in the ideological and political classroom under multi-target tracking is better.

5 Conclusion

This paper proposes a high-reliability design for the acquisition system of student status in ideological and political classrooms under multi-target tracking. Under multi-target tracking, the hardware and software of the student status information acquisition system in the ideological and political classroom are designed to achieve the acquisition of student status information in the ideological and political classroom. The results show that the information acquisition system designed in this paper has relatively high reliability, but in practical applications, the scalability of the system needs to be further studied.

References

1. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mob. Netw. Appl. **24**(1), 5–17 (2018). https://doi.org/10.1007/s11036-018-1134-8

2. Xu, R., et al.: Towards emotion-sensitive learning cognitive state analysis of big data in education: deep learning-based facial expression analysis using ordinal information. Computing **102**(10), 765–780 (2019)
3. Farhan, M., et al.: IoT-based students interaction framework using attention-scoring assessment in eLearning. Futur. Gener. Comput. Syst. **79**, 909–919 (2018)
4. Zhang, W., et al.: Student performance prediction via online learning behavior analytics. In: 2017 International Symposium on Educational Technology (ISET), pp. 153–157. IEEE 2017
5. Hamoud, A., Hashim, A.S., Awadh, W.A.: Predicting student performance in higher education institutions using decision tree analysis. Int. J. Intera. Multi. Artifi. Intellige. **5**, 26–31 (2018)
6. Liu, S., Pan, Z., Cheng, X.: A novel fast fractal image compression method based on distance clustering in high dimensional sphere surface. Fractals **25**(4), 1740004 (2017)
7. Kuanglu, Y., et al.: Fast information acquisition using spectra subtraction for Brillouin distributed fiber sensors. Optics express **27**(7), 9696 (2019)
8. Sun, S., et al.: A distributed incremental information acquisition model for large-scale text data. Clust. Comput. **22**(1), 2383–2394 (2019)
9. Liu, S., et al.: Parallel Fractal Compression Method for Big Video Data. Complexity 2016976 (2018)
10. Sang, H., Zhou, Y., Zhou, G.: Interacting multiple models integrated tracking splitting target tracking method. Command Control & Simulation **42**(2), 38–45 (2020)
11. Yang, H., et al.: Object tracking algorithm based on discriminative correlation filter and depth information. J. Chinese Comp. Sys. **41**(4), 736–740 (2020)
12. Ma, J., et al.: Target tracking method based on fusion of human structural features and kernelized correlation filter algorithm. J. Comp. Appl. (S01), 56–60 (2020)
13. Ming, Z., et al.: A performance based method for information acquisition in engineering design under multi-parameter uncertainty. Information Sciences (2020)
14. Li, H., et al.: Design and optimization of a hybrid sensor network for traffic information acquisition. IEEE Sensors Journal **20**(4), 2132–2144 (2020)
15. Chen, X., Li, J., Sui, Y.: A new stitching method for dark-field surface defects inspection based on simplified target-tracking and path correction. Sensors **20**(2), 448 (2020)
16. Lia, J., Wang, Z., Xu, M.: DeepMTT: a deep learning maneuvering target-tracking algorithm based on bidirectional LSTM network. Information Fusion **53**, 289–304 (2020)
17. Xu, X., et al.: Research on key technologies of smart campus teaching platform based on 5G network. IEEE Access **7**, 20664–20675 (2019)
18. Sharma, P., et al.: Decentralized gaussian filters for cooperative self-localization and multi-target tracking. IEEE Trans. Signal Processing, 67 (2019)
19. Wei, X., Jing, L.: Data transmission and real-time information acquisition based on fine-grained algorithm and wireless network. Peer-to-Peer Netw. Appl. **13**(2), 1–11 (2019)
20. Ullah, I., et al.: Multisensor-based target-tracking algorithm with out-of-sequence-measurements in cluttered environments. Sensors 18(11) (2018)

Weak Vibration Signal Extraction Method of Mechatronics Equipment Based on Stochastic Resonance

Dong-bao Ma[1]([✉]), Xue-mei Li[1], Ming-fei Qu[1], and Xiao-zheng Wan[2]

[1] College of Mechatronic Engineering, Beijing Polytechnic, Beijing 100176, China
`dongbaoma1109@126.com`
[2] Institute of Oceanographic Instrumentation, Qilu University of Technology (Shandong Academy of Sciences), Qingdao 266061, China

Abstract. Existing weak vibration signal extraction methods are limited by the small frequency parameters of the signal, and it is difficult to determine the critical amplitude of the signal, resulting in the delay of weak vibration signal extraction. Therefore, this paper designs a weak vibration signal extraction method for mechatronics equipment based on stochastic resonance. Based on the particle dynamics equation of Brownian motion, a bistable stochastic resonance model is constructed. By using the model, the parameters of potential well function are adjusted to achieve the optimal matching of signal, noise and nonlinear function to achieve stochastic resonance. According to the relationship between the output signal-to-noise ratio and the parameters of the potential well function, the output signal-to-noise ratio of the weak vibration is calculated, and the appropriate noise intensity is determined to enhance the weak periodic signal. The single oscillator is extended to an array equation group composed of four oscillators with different initial phase of driving signal, and the weak vibration signal with any initial phase is extracted. The experimental results show that: compared with the traditional method, the time of weak vibration signal extraction of this method is less, which shows that the weak vibration signal extraction of this method is more real-time.

Keywords: Stochastic resonance · Mechatronics · Weak vibration signal · Signal extraction

1 Introduction

With the rapid development of science and technology in our country, mechatronics equipment has been applied to all walks of life of the national economy, which strongly reflects the manufacturing level and engineering innovation ability of our country. Mechanical and electrical integration equipment components due to vibration fluctuations and other reasons, the failure will lead to the catastrophic collapse of the system, thus reducing the reliability and availability of the equipment, in turn increasing the downtime, resulting in a lot of financial losses, and each failure poses a threat to the

safety of workers [1]. Therefore, the health monitoring of equipment components is the key task to ensure the reliability of industrial process. However, many fault features are often difficult to extract in the background of strong noise, or in order to prolong the life cycle of the equipment, it is necessary to extract the early weak vibration signal, so as to diagnose the early fault of the equipment or the weak vibration signal in the background of strong noise, to minimize the impact of mechanical and electrical integration equipment fault.

Weak vibration signal extraction is an important branch of signal processing technology. It uses physics, electronics signal processing theory and other methods to study the characteristic frequency of the target signal, the variation law and characteristics of noise, so as to extract the characteristics of weak vibration signal submerged by strong noise. Due to the small amplitude of weak vibration signal and the strong noise produced by equipment operation. In addition, when extracting the characteristics of weak vibration signal, due to the influence of the noise of measuring instruments and sensors, the amplitude of the target signal to be extracted is often weaker, so that the weak vibration signal cannot be extracted [2]. The key point of weak vibration signal extraction is to extract weak characteristic frequency under strong noise background or early weak fault of mechanical equipment, so as to prevent mechanical fault.

Weak vibration signal extraction mainly depends on the quality of the signal. The signal-to-noise ratio of weak fault features in strong noise background is very low, and the signal quality is very poor. At present, the weak vibration signal extraction method is to extract the weak vibration signal from the angle of noise suppression, but at the same time, the characteristics of the weak signal are also suppressed, so that the signal-to-noise ratio is lower. When the frequency of the target signal is very close to that of the noise signal, the useful signal is damaged while processing the noise signal. At this time, it is difficult to extract the weak vibration signal. Due to the limitations of weak signal extraction technology, people continue to explore the weak vibration signal extraction and detection technology, so as to extract the weak vibration features in the strong noise background more efficiently and accurately. Most of the traditional weak vibration signal extraction methods suppress the strong noise signal, while stochastic resonance has the opposite advantages compared with the traditional noise reduction methods Using noise energy to transform weak vibration signal can enhance weak vibration characteristics and weaken part of the noise, so it can play a good role in the extraction of vibration signal in strong noise.

Based on the above research background, this paper designs a method of weak vibration signal extraction for mechatronics equipment based on stochastic resonance, which has important practical value for fault diagnosis under strong noise background. The overall work of the new method is as follows:

(1) Based on the particle dynamics equation of Brownian motion, a bistable stochastic resonance model is built, and the parameters of the potential well function are adjusted by the model, so that the optimal matching of signal, noise and nonlinear function is achieved, so as to realize the stochastic resonance.

(2) According to the relationship between the output signal-to-noise ratio and the potential well function parameters, the output signal-to-noise ratio of weak vibration is

calculated, so as to determine the appropriate noise intensity and realize the noise enhancement of weak periodic signal.

(3) A single oscillator is extended into an array system of equations composed of four oscillators with different initial phases of the driving signals, from which weak vibration signals with arbitrary initial phases are extracted.

2 Method Design

2.1 Construction of Bistable Stochastic Resonance Model

The extraction of weak vibration signal of mechatronics equipment relies on the matching of noise, periodic signal and nonlinear function to achieve resonance. In the nonlinear function, when the periodic driving and noise signal match, the weak periodic driving and random noise interference will produce a synergistic effect. This effect not only does not make the increased noise make the system output signal disordered, but also greatly improves the signal-to-noise ratio of the output response, so as to achieve the purpose of detecting weak vibration signal [3]. In order to match the relationship among the three, the bistable stochastic resonance model is needed firstly, which lays the foundation for the research of weak vibration signal feature extraction in this paper.

The bistable stochastic resonance model is based on the particle dynamics equation of Brownian motion. When a particle moves irregularly in a medium, it will collide with the medium molecules. At the microscopic level, Brownian particles collide with medium molecules randomly. The bistable stochastic resonance model can be expressed as:

$$\frac{ds}{dt} = -J(s) + A\cos(\omega t + \phi) + \sigma(t) \tag{1}$$

In Eq. (1), s is the trajectory of the particle; t is the running time; A, ω, ϕ is the amplitude, angular frequency and phase of the periodic signal; $J(s)$ is the symmetric bistable potential well; $\sigma(t)$ is the zero mean additive white Gaussian noise, which meets the following conditions:

$$\langle \sigma(t)\sigma(0) \rangle = 2Q\varepsilon(t) \tag{2}$$

In Eq. (2), $\sigma(0)$ is the white noise at the beginning of the motion; Q is the noise intensity; $\varepsilon(t)$ is the standard deviation of the distribution. The calculation formula of $J(s)$ is as follows:

$$J(s) = -\frac{1}{2}\alpha s^2 + \frac{1}{4}\beta s^4 \tag{3}$$

In Eq. (3), α, β is the bistable well parameter. By adjusting the parameters, the bistable well with different shapes can be obtained, as shown in Fig. 1.

In the case of the default periodic force, the trajectory function will jump between the local steady state (-1 or 1). The jump caused by noise satisfies Kramers rate. In the case of periodic signal, the probability of particle jumping is determined by the equilibrium position [4]. The response amplitude of bistable potential well has a function
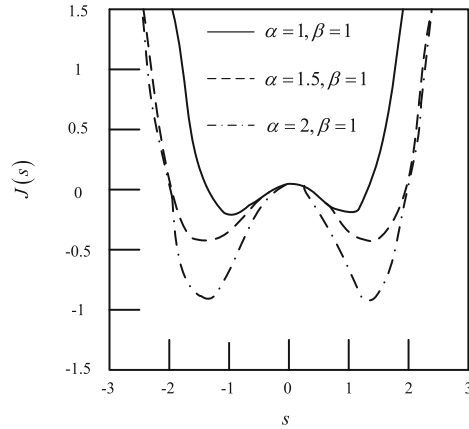
Fig. 1. Bistable well with different shapes.

relation with noise intensity. Under the periodic excitation of different amplitude, the output amplitude of stochastic resonance changes with the increase of noise intensity, not monotone, and decreases with the increase of noise intensity when the maximum value is reached. According to the derivation process, the bistable stochastic resonance model is constructed.

2.2 Adjusting Well Function Parameters Based on Stochastic Resonance

According to the analysis in the previous section, stochastic resonance is actually the optimal matching relationship among signal, noise and nonlinear function. However, in practical engineering applications, the situation of signal and noise is often unknown, and the optimal matching relationship between them and nonlinear function is not always. At this time, if we want to realize weak signal detection through the stochastic resonance of the oscillator, we need to change one, two or all of the characteristics of signal, noise or nonlinear function [5]. For the measured signal, the feasible method is to adjust the parameters of the function so that the signal, noise and nonlinear function can achieve the optimal matching and achieve stochastic resonance. The occurrence of stochastic resonance has an optimal limit on the amount of noise input into the nonlinear function. When different amounts of noise are input, the output response of the system varies greatly. However, when the noise is classified and different types of noise, such as white noise, color noise and additive noise, are combined with the periodic driving input function, the output response of the function is different [6].

In the process of signal response, not only the input noise of function affects the occurrence of stochastic resonance, but also the selection and adjustment of well parameters can not be ignored. When the noise intensity is 0, the whole bistable well will no longer keep balance, and the well will tilt back and forth at a certain frequency driven by periodic signal. As long as the signal amplitude is kept less than the critical value of the function, Brownian particles can only move in a certain well at the same frequency. However, after noise is input to the function, even if the signal amplitude is less than the

critical value of the function, particles can easily transition from one potential well to another [7]. At this time, the output response is switched between the two wells according to the modulation frequency of the signal. When the signal amplitude exceeds 0, the signal introduces a periodic change to the switching of the potential well, which effectively synchronizes the switching caused by noise, so that the small periodic component of the output result is enhanced. The cascade bistable system is formed by connecting several continuous bistable wells in series, and its structure is shown in Fig. 2.
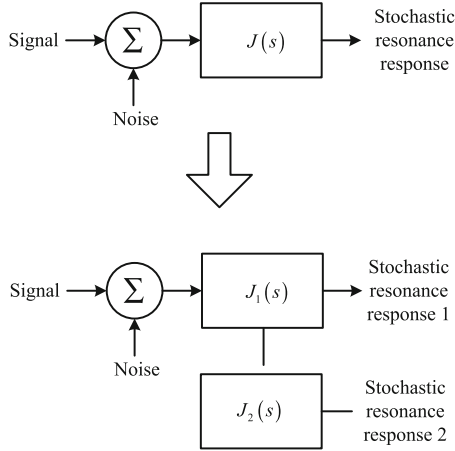
Fig. 2. Cascaded piecewise linear stochastic resonance potential well

The cascading method of continuous bistable well is used to transfer the high frequency energy to the low frequency region. The high frequency component is filtered out gradually, and the low frequency component is highlighted to achieve good noise reduction effect. When stochastic resonance occurs, the output response signal ratio of the continuous bistable well reaches the maximum. At this time, the output power spectrum consists of two parts, which can be expressed as:

$$P(f) = P_1(f) + P_2(f) \tag{4}$$

In Eq. (4), $P(f)$ represents the output power spectrum; $P_1(f)$ represents the power spectrum caused by the input periodic signal, which is equal to the frequency of the input signal; $P_2(f)$ represents the power spectrum caused by noise, which is in the form of Lorentz distribution. In a certain range, the larger the noise intensity is, the smoother the characteristic curve of Lorentz distribution is; the smaller the noise intensity is, the steeper the curve is. The output power of noise has the characteristics of Lorentz distribution [8], and most of the energy is concentrated in the low frequency band. Therefore, only with the help of the noise energy can Brownian particles cross the barrier and make a reciprocating transition between two potential wells in the continuous bistable system at the signal frequency, thus stochastic resonance occurs in the low frequency band. With the increase of the parameters of the potential well function, the value of the optimal signal-to-noise ratio decreases gradually, and the noise intensity

corresponding to the optimal signal-to-noise ratio also increases. That is to say, the smaller the parameters of the potential well function, the better the output response of the bistable potential well can be achieved with less noise energy. Based on the above adjustment process, stochastic resonance is used to determine the well function parameters to obtain the best signal-to-noise ratio.

2.3 Calculate the Signal-to-Noise Ratio of Weak Vibration Output

According to the relationship between the output signal-to-noise ratio and the parameters of the potential well function, the weak vibration output signal-to-noise ratio is further calculated. Stochastic resonance is a nonlinear phenomenon. When a certain amount of noise and weak periodic vibration signal are added to the input of nonlinear function, the output response end is not affected by the added noise, but can improve the signal-to-noise ratio of the output end. When the amount of noise in the input periodic signal is appropriate, there is an optimal signal-to-noise ratio at the output response end of the nonlinear function. The essence of the above process is that the energy of the noise is transferred to the energy of the weak vibration signal through the nonlinear function, so that the energy of the weak vibration signal is enhanced, which fully reflects the cooperative relationship among noise, periodic signal and nonlinear potential well function [9]. According to the above analysis, the amplitude of weak vibration periodic signal can be enhanced under noise excitation. But in practical engineering applications, more attention is paid to the proportion of signal and noise. The formula for calculating the power spectral density of stochastic resonance output is as follows:

$$R(\omega) = \int_{-\infty}^{+\infty} e^{-t} \langle \langle s(t + \Delta t)s(t) \rangle \rangle dt \tag{5}$$

In formula (5), $R(\omega)$ represents the power spectral density of the random resonance output; $s(t + \Delta t)s(t)$ represents the total average of the noise realization; and $\int_{-\infty}^{+\infty} e^{-t} \langle \langle \ \rangle \rangle dt$ represents the average of the initial phase of the input. When other frequency components are taken into account, the formula of SNR is obtained, namely:

$$snr = \pi \left(\frac{As}{Q} \right)^2 r_k \tag{6}$$

In Eq. (6), r_k is the relaxation rate of the transition. The SNR curve is very similar to the output amplitude curve, as shown in Fig. 3.

According to Fig. 3, the signal-to-noise ratio also increases with the increase of noise, reaches the maximum and then decreases. When the noise intensity is small, the oscillator can only vibrate in one potential well; when the noise intensity gradually increases, the oscillator can occasionally jump between potential wells; when the noise intensity is appropriate, the weak periodic signal can be enhanced with the help of noise, and the period of the output signal is more obvious; when the noise intensity is too large, the periodic component is polluted by noise. Therefore, only when the intensity of noise is appropriate, can the weak periodic signal be enhanced by noise.
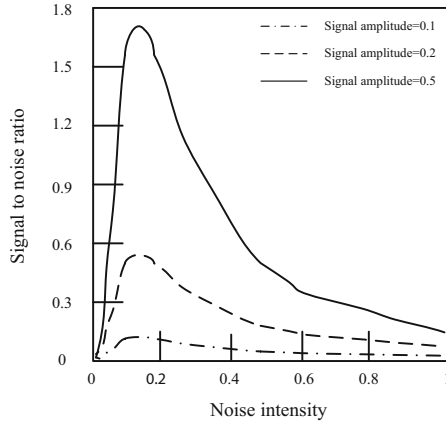
Fig. 3. Relationship between output SNR and noise intensity of bistable stochastic resonance

2.4 Extracting Weak Vibration Signal of Mechatronics Equipment

In addition to the characteristics of strong noise background, the signals collected in practical engineering usually contain other frequency components. For weak vibration fault, the acquisition signal is not only fault frequency, but also the components of frequency conversion, mode frequency and DC component exist simultaneously. When using the method of stochastic resonance to extract the characteristic frequency of signal, the above interference frequency includes frequency conversion, mode frequency, other characteristic frequency and DC component, which have adverse effects on the signal we want to extract. The practicability of stochastic resonance can not be extended. The energy of high frequency signal output at all levels can be transferred to the low frequency part in turn by using the cascade of bistable potential wells. At the same time, the DC component in the original signal has a great influence on the occurrence of stochastic resonance. The probability of system transition will be reduced when the DC component increases to a certain extent.

In order to minimize this effect, the input signal is preprocessed before the random resonance occurs. The main purpose is to eliminate the adverse effects of the peak spectrum and DC component on the stochastic resonance. When the potential well response reaches random resonance, increasing the noise intensity of input signal again will make the potential well unable to recognize the periodic component after the output response. The over resonance phenomenon occurs in the potential well output. For the input signal under the strong noise background, according to the relationship between the signal-to-noise ratio of the output signal and the amplitude and the noise intensity of the input signal, the bistable potential well is prone to over resonance. By adjusting the noise in the input signal, the output signal can reach the optimal random resonance [10].

Because of the existence of the detection window, only a single vibrator can detect the weak vibration signal of the initial phase within the detection window. It is impossible to detect the weak vibration signal whose initial phase is beyond the scope of the detection window. But the range of detection window in the range of [180, 180] varies with the change of the initial phase of the driving signal. Therefore, the single vibrator is extended to an array equation system composed of four different initial phase drivers to detect the weak vibration signals in any initial phase. In conclusion, as long as the amplitude of the characteristic signal to be measured is large enough, at least one potential well can be transformed from chaos state to large-scale periodic state in the array, so as to detect the weak vibration signal to be measured. In other words, as long as the amplitude of weak vibration signal is large enough, the array vibrator can extract any initial phase characteristic signal with the same frequency as the driving signal.

3   Experimental Study

In order to verify the effectiveness of the weak vibration signal extraction method of mechatronics equipment based on stochastic resonance, the following experiments are designed.

3.1   Experimental Preparation

The weak vibration test data of mechatronics equipment is provided by the Electrical Engineering Laboratory of West storage University of the United States. An acceleration sensor is placed on the bearing pedestal of the motor drive end and fan end. A 16 channel data recorder is used to collect the vibration signal of the rolling bearing of the motor. The sampling frequency of the signal includes 12 kHz and 48 kHz.

3.2   Analysis of Weak Vibration Signal Extraction Test Results

In order to test the application effect of the method of weak vibration signal extraction of mechanical and electrical integration equipment based on stochastic resonance, the signal extraction time of different methods is tested by comparing with the existing signal extraction methods. Three test points are arranged on the platform rotating axis in order to collect vibration signals of different components of the equipment, and the sampling frequency of the signals is 12 kHz. The test results are shown in Fig. 4.

According to the comparison results in Fig. 4, the weak vibration extraction time of different components of the equipment is obtained. The extraction time of weak vibration of each component obtained by this method changes in the range of 3–5 s, and the fluctuation amplitude is small. However, the time of the existing extraction methods fluctuates in the range of 6–8 s, which indicates that the extraction time is delayed, which is not conducive to the real-time extraction of weak vibration signals. The weak vibration signal extraction time of mechatronics equipment at different sampling frequencies is further counted, as shown in Table 1 and Table 2.

(a)Element 1

(b)Element 2

(c)Element 3
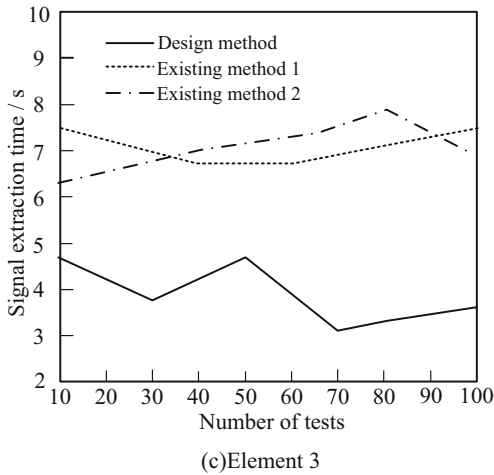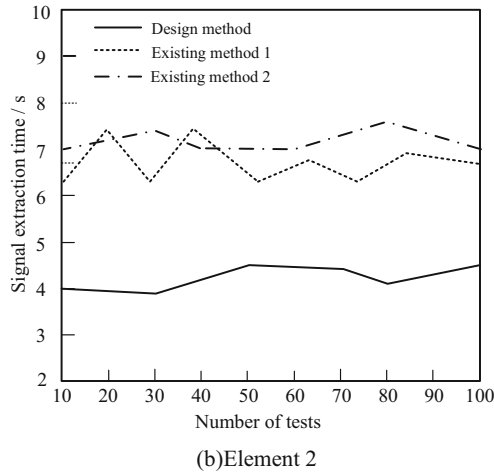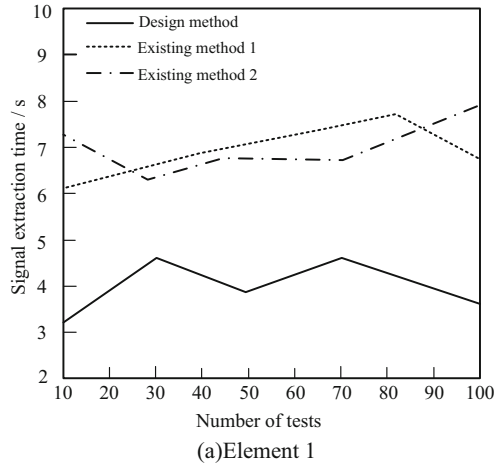
Fig. 4. Signal extraction time test of Mechatronics components.

Table 1. Test results of signal sampling frequency 12 kHz

Serial number	Extraction time of weak vibration signal/s		
	Method of this paper	Existing methods 1	Existing methods 2
1	8	18	19
2	9	16	18
3	8	18	19
4	10	17	22
5	11	18	21
6	8	21	23
7	9	18	22
8	9	19	23
9	8	21	22
10	7	19	19
11	9	18	20
12	9	18	21
13	11	19	21
14	9	18	20
15	10	19	19
16	11	21	18
17	9	16	19
18	8	17	20
19	9	16	21
20	11	16	22
Average value	9.15	18.15	20.45

Table 2. Test results of signal sampling frequency 48 kHz

Serial number	Extraction time of weak vibration signal/s		
	Method of this paper	Existing methods 1	Existing methods 2
1	6	15	16
2	5	14	16
3	6	15	17
4	5	13	15
5	4	15	16
6	6	13	16
7	5	12	15
8	6	13	14
9	4	14	13
10	7	15	15
11	6	13	15
12	5	14	16
13	6	12	16
14	4	13	14
15	6	14	13
16	5	15	13
17	6	14	14
18	7	12	15
19	6	13	16
20	5	13	15
Average value	5.5	13.6	15

According to the test results in Table 1 and Table 2 when the signal sampling frequency is 12 kHz, the weak vibration signal extraction time of this method is 9.15 s, which is 9 s and 11.3 s less than the existing methods; when the signal sampling frequency is 48 kHz, the weak vibration signal extraction time of this method is 5.5 s, which is 8.1 s and 9.5 s less than the existing methods. To sum up, the method in this paper effectively reduces the extraction time and makes the extraction process of weak vibration signals more real-time.

4   Concluding Remarks

In this study, a method of weak vibration signal extraction for mechatronics equipment was designed based on stochastic resonance. Based on the construction of bistable stochastic resonance model, the method used the model to adjust the potential well function parameters, so that the optimal matching of signal, noise and nonlinear function could be achieved, and the stochastic resonance could be achieved. Then, according to the relationship between the output signal-to-noise ratio and the parameters of the potential well function, the output signal-to-noise ratio of the weak vibration is calculated, and the appropriate noise intensity is determined to realize the noise enhancement of the weak periodic signal. Finally, the single oscillator is extended into a set of array equations composed of four oscillators with different initial phases of the driving signals, and the weak vibration signals with arbitrary initial phases are extracted. In this study, the experimental results proved that the method of weak vibration signal extraction is more real-time, shorten the extraction time.

Due to the complexity of actual working conditions and conditions, there is still room for further optimization of weak vibration signal extraction. When constructing the signal extraction model, membership function can be introduced to select the optimal parameter range to further improve the extraction effect of each classified signal.

References

1. Zhang, G., et al.: Weak vibration signal detection method for exponential stochastic resonance systems. J. Vibr. Shock **38**(9), 53–61 (2019)
2. Wang, W., et al.: Weak fault feature extraction of rolling bearing based on secondary clustering segmentation and Teager energy spectrum. J. Vib. Shock **39**(13), 246–253 (2020)
3. Zhang, H., Zhou, J.: Weak fault feature extraction of rolling bearing combined adaptive MCKD with CEEMDAN. J. Electron. Meas. Instrum. **33**(4), 79–86 (2019)
4. Wang, J., et al.: Bearing fault diagnosis method based on K-medoids clustering. Aerosp. Control Appl. **46**(4), 24–28 (2020)
5. Yang, X., Zhang, X.: Study on the analysis and extraction method of gear local fault vibration signal. J. Xi'an Univ. Technol. **35**(1), 28–33 (2019)
6. Zeng, T., et al.: A micro-oscillation signal extraction method for radar target in strong clutter environment. J. Signal Process. **35**(6), 935–943 (2019)
7. Chen, B., Liu, G.: Extraction for HHT features of pressure vibration signals of oil pulsating flow with different particle concentration. Opt. Precis. Eng. **27**(11), 2484–2497 (2019)
8. Liu, S., Pan, Z., Cheng, X.: A novel fast fractal image compression method based on distance clustering in high dimensional sphere surface. Fractals **25**(4), 1740004 (2017)
9. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mobile Netw. Appl. **24**(1), 5–17 (2018). https://doi.org/10.1007/s11036-018-1134-8
10. Liu, S.: Distribution of primary additional errors in fractal encoding method. Multimed. Tools Appl. **76**(4), 5787–5802 (2014). https://doi.org/10.1007/s11042-014-2408-1

Fast Integration System of English Online Learning Resources Based on Multi Sensor Network

Hai-yun Han[1](✉) and Bing-bing Han[2]

[1] School of Humanities and Social Sciences, Sanya Aviation and Tourism College, Sanya 572000, China
sdgsahd1232@aliyun.com
[2] Panjin Vocational and Technical College, Panjin 124000, China

Abstract. Common English online learning resources rapid integration system, resulting in the system in the traffic load gradually increases, there is a throughput reduction phenomenon, can not meet the needs of users, this paper proposes a multi-sensor network based English online learning resources rapid integration system. In terms of hardware, the browser/server architecture and client/server architecture are combined to configure the hardware architecture of the system; according to the way of multi-sensor network collecting English online learning resources, the multi-sensor network structure is determined, and the hardware design of the system is completed. In terms of software, based on OAIS and multi-sensor network, the overall structure of English online learning resource integration system is designed; according to the three steps of resource collection and submission, resource storage and management, resource release and service, the system function module is designed to complete the system software design. Design the test environment and system performance test interface, experimental results: after the traffic load increases, the designed system has no obvious change compared with the two groups of commonly used system throughput, low loss rate and false detection rate, and high integration efficiency.

Keywords: Multi sensor network · English · Online learning · Resource integration

1 Introduction

Network education is a promising new field of education, attracting a large number of social funds, enterprises and celebrities (including IT enterprises, social education organizations, investors, subject experts, it talents, etc.) to develop and construct network curriculum resources. The core of network education is network course, and the integration of network course resources plays a basic and overall role in the quality of network education [1]. However, the lack of integration of network course resources is very prominent, which greatly reduces the efficiency and quality of network teaching

and learning. Therefore, on the basis of network curriculum resources, in the process of communication and interaction between network educators and network educates through network environment, the links and elements of curriculum content, curriculum objectives, teaching strategies, design standards, curriculum management, curriculum sharing, network teaching means and network teaching evaluation are organically integrated, and the internal components of network curriculum resources are integrated and standardized in one [2]. In this paper, a new method of anchor control is proposed. And constantly from the network environment to obtain new knowledge and information, human and material, sharing and cooperation, so as to greatly improve the efficiency of online education and the quality of online learning innovation process.

Overseas research on the integration of digital resources is earlier, and has achieved more research results, which have been applied to practice. We have designed and developed systems such as encompass, chameleon iportal, iport, etc., and built the portal website for the integration of MetaLib and SFX resources [3]. Researchers and software companies in related fields in China have conducted a lot of research on resource integration, and also launched related application systems, such as USP, TRS IIP and other platforms, webbridge, CALIS, CSDL and other systems [4]. According to the theoretical background of the integration of English online learning resources, literature [5] discusses the current situation of College English Teaching in the information technology environment, and then from the perspective of humanism and constructivism, integrates College English teaching resources by means of information technology, enriches English learning context, and improves learners' Autonomous Learning ability. Literature [6] analyzes the existing online learning resources and the characteristics of Higher Vocational Students' English learning, integrates ubiquitous learning resources, designs ubiquitous resource platform, and promotes English learning. On the basis of previous studies, this paper introduces multi-sensor network and proposes to design a fast integration system of English online learning resources based on multi-sensor network.

Based on the above problems, a fast integration system of English online learning resources based on multi-sensor network is proposed. In the design of the system, the structure of the multi-sensor network is determined, the hardware structure of the system is designed, and the corresponding functions of the software module are designed to complete the design of the system and improve the rapid integration of English online learning resources.

2 Hardware Design of Fast Integration System of English Online Learning Resources Based on Multi Sensor Network

2.1 Design System Hardware Architecture

This time, the multi-sensor network technology is applied to the integration of English online learning resources. Considering the integration speed of online learning resources, the system has the function of Integrating English online learning resources. Based on the resource integration system designed by previous researchers, the hardware architecture of the rapid integration system of English online learning resources is designed, as shown in Fig. 1.
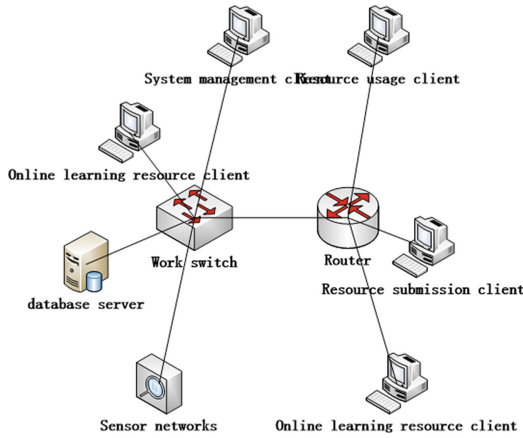
Fig. 1. Hardware Architecture of the System

The system adopts the combination of browser/server architecture and client/server architecture to configure the hardware architecture of the system. For the application of English online learning resources, as well as the audit management of English online learning resources, the client/server architecture can better improve the operation efficiency of resource management. For resource query, download and other functional applications, the browser can be used/the server architecture pattern can make users more convenient and flexible to use the required functions. Therefore, in the application layer of the system, the software component of English online learning resource submission and audit is deployed on the user's client, while the function component of resource query and download is deployed on the website server, and the client calls the system functions through the browser. Each functional component of the business logic layer is deployed in the application server of the system, which is called by the client software and website server in the form of service [7]. The sensor network is also deployed in the application server, and the components in the sensor network provide support for the business logic layer components. The database of the system is deployed in the database server, including relational database and file database, which is responsible for the storage of teaching resources.

2.2 Determine the Structure of Multi-sensor Network

In the rapid integration system of resources in multi-sensor network, each sensor node samples the English online learning resources according to its own sampling period, and then transmits the collected raw data to the estimator, which estimates according to the information transmitted by the sensor. Of course, a sensor node can not only transmit the status information of English online learning resources collected by itself, but also act as a routing node to help other sensor nodes transmit data. In this case, the sensor node does not do data processing, only for acquisition and transmission, the structure is shown in Fig. 2.
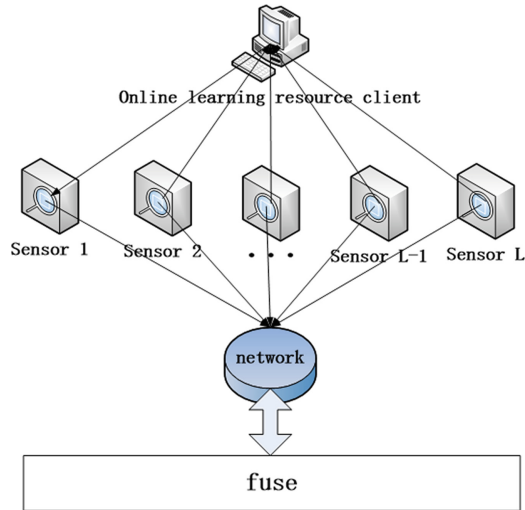
Fig. 2. Multi sensor network structure

In Fig. 2, the sensor is used to perceive the measurement unit of English online learning resources, to perceive the state of English online learning resources, and to transform the information into measurement signals and send them to the estimation center. Communication network is the communication medium between the components. The fusion center receives the data from the sensor, and then estimates these data. With the development of intelligent sensor technology, the current sensor equipment has a certain storage capacity and computing capacity. In order to reduce the computational pressure of the fusion estimation center, the sensor uses its own computing power to preprocess the collected raw data, so that the sensor can obtain the local optimal estimation, and then transmit the local optimal estimation to the estimation center. The information fusion estimation center performs the fusion estimation on the basis of the local optimal estimation, so as to obtain the global optimal estimation the optimal estimation of the model.

By using the estimation system of the above structure, users can monitor English online learning resources well. In practical application, it is often necessary not only to know the real status of English online learning resources, but also to respond to some unexpected status of English online learning resources. Communication and control originally belong to different research directions. Communication theory focuses on how to ensure the effective and reliable transmission of two endpoints, regardless of the purpose of information, and rarely consider the feedback of information [8]. The purpose of common control system is to achieve a certain performance target through the information. Sometimes the influence of communication limitation may be ignored when only considering control. However, with the introduction of wireless communication and the increasing scale of the system, researchers have to consider the combination of communication and control. Therefore, a control unit (controller) and an execution unit (actuator) are added to the system for acquiring state information. In this way, the whole

system can not only achieve real-time monitoring, but also achieve the effect of online decision-making and real-time response.

3 Software Design of Fast Integration System of English Online Learning Resources

Based on the system hardware framework and the determined multi-sensor network, this paper designs the system English online learning resources fast integration functional structure, as well as information acquisition, resource storage, resource management and other functional modules, so as to promote the system to have the function of fast integration of English online learning resources.

3.1 Design the Functional Structure of the System

According to the overall structure of the designed English online learning resource integration system, and based on the open archive information system model (OAIS) and multi-sensor network, this paper puts forward the functional structure of the computer network teaching resource integration system, as shown in Fig. 3.
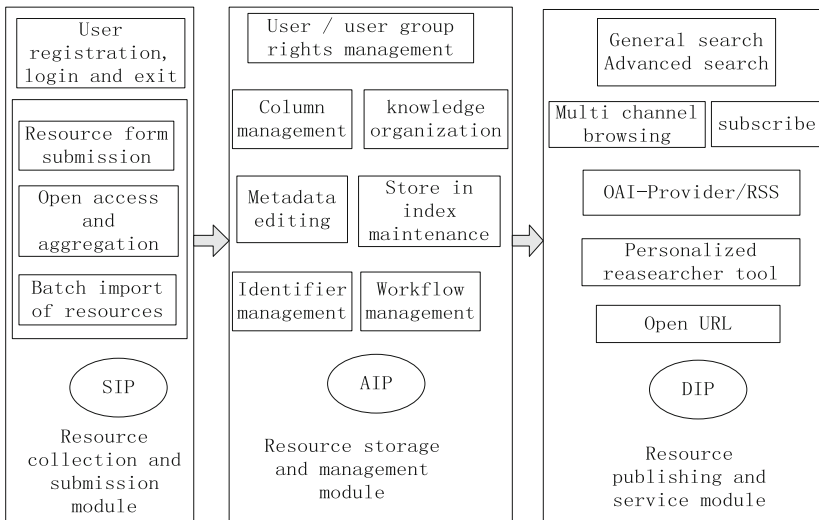
Fig. 3. System function structure

According to the functional structure diagram shown in Fig. 3, the workflow of English online learning resource integration system based on multi-sensor network is mainly divided into three steps: resource collection and submission; resource storage and management; resource release and service. The overall flow of the system is shown in Fig. 4.
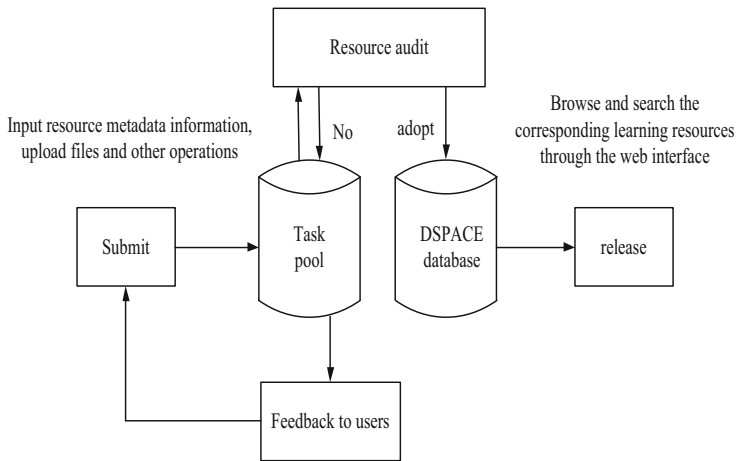
Fig. 4. Overall process of resource integration system

3.2   System Function Module Design

Resource Collection and Submission Module
In the collection and submission module of the designed resource integration system, in addition to the system administrator, the resource maintenance personnel need to register to fill in personal information, and add it to the corresponding departments and topics. After logging in to the system, they enter "my work area", and gradually complete the description, indexing and uploading of the resources to be submitted through the interface. At the same time, you can modify and import resource information in my workspace. The workflow is shown in Fig. 5.

According to the workflow of the acquisition and submission Module shown in Fig. 5, its main functions are as follows:

1. User registration, login or exit. This is the most basic function of the system. Ordinary users can browse, retrieve and open download resources anonymously, but for new users, if they want to upload teaching resources, e-mail notification or resource management functions, they must be registered and authorized by the system administrator before they can use them [9]. For the registered users, they must go through the verification of account and password before they can log in to the system and use these functions.

   (1) User registration function: unregistered users can register a user name, and provide a password and simple personal information, such as e-mail address, you can register successfully. It should be noted that the user name cannot be repeated. Users who have successfully registered can log in to the system according to the user name and password just registered. Therefore, the input data items of the registration function include e-mail user name and password.
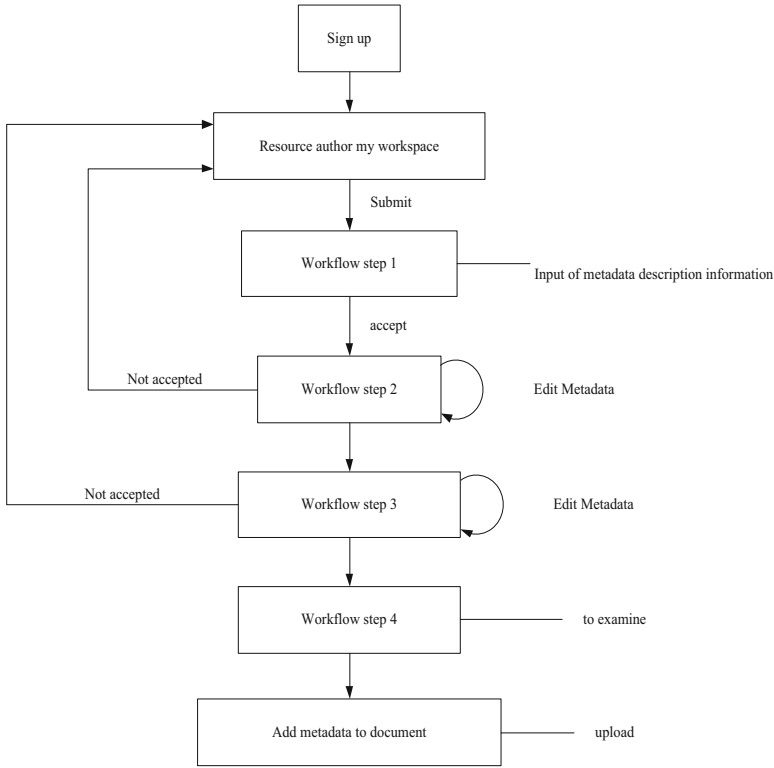
Fig. 5. Workflow of acquisition and submission module

(2) User login function: Only registered users can log in to the system. Login system is very simple, just need to provide login user name and password can login system. Therefore, the input data items of login function include e-mail user name and password.

(3) User exit function: after logging in to the system, the user has the corresponding operation authority of the system. On the basis of this system, more operation functions can be developed. The logged in users can log off their personal login information through the exit operation, so that the user no longer has the internal management function of the system. After the user leaves the system, other people will not be able to operate, which improves the security of the system.

2. Upload and submit resources. In the module of resource collection and submission, the multi-sensor network supports two ways of resource submission: one is based on Web page, which supports any user with submission authority to submit the required content to the system. Generally, through the traditional web interface, the data are uploaded to the system step by step, starting with the selection of the corresponding collection, and then filling in the corresponding metadata information and the selected uploaded documents by the submitting party.

The other is batch import, which supports batch import of resource files conforming to dSPACE SIP format into the system. Generally, the resources are uploaded in batch on the server side of the system through the program provided by the multi-sensor network. When the data is uploaded to the collection, each collection can set up a special auditor according to the needs. In order to ensure the quality of the resources, the system will audit the resource items uploaded by the submitting party. If the resources are denied in the process of audit, the system will notify the submitting party by e-mail. On the contrary, if it is approved, the resources will be formally included in the system database.

3. Open access and aggregation. By rediscovering, reorganizing and integrating the information in the scattered teaching resources around a specific resource topic, we can more quickly establish a digital resource integration system that is in line with the habit of teachers and students in our college to obtain and use thematic information, and at the same time, the content of the information resource database will be more widely found and used.

Resource Storage and Management Module

The archive and management module of resource integration system is to audit the format and metadata description of resources submitted by users, and modify the non-standard metadata description [10]. The resources that meet the requirements will be submitted to the resource library and published to the public. The resources that do not meet the requirements will be returned to the user. The user will see the management personnel's modification requirements for the resource description in his own workspace. At the same time, the module can also set and modify user permissions and workflow. Its main functions are as follows:

The original intention of this system is to solve the problem of English online learning resources storage. It can store any format of literature, such as documents, books, databases, computer programs, multimedia publications and audio and video. Therefore, two storage modes of "digital stream save" and "functional save" are designed for the system. The former is to save a digital document according to its original shape, which may not be read out after several years due to software upgrade, and can only be identified by using special software or coding by special experts; the latter requires that its usability be maintained all the time when the technical format and media medium change. Obviously, "functional preservation" is a more rational storage mode. In order to improve the storage function of the system, the file format is divided into the following three levels.

(1) Supported format: use the format migration technology to save this file functionally; (2) known format: this format file means that it is unable to complete the functional saving, but as a popular format, you can try to complete the format migration through the conversion tool provided by a third party to achieve functional saving; (3) unsupported format: according to the current technology Unable to complete functional save.

2. Resource management. The design of the system will integrate all kinds of teaching resources together, the main purpose is to facilitate management. When the qualified resources are submitted to the resource database, the related resources will be organized and managed according to the hierarchical classification of "learning stage – learning

project – resource entry", and the topic tree navigation will be provided to facilitate resource management.

3. System management and user rights management. The system administrator can manage the resources, users and user groups in the system. In order to perform certain operations on an object, users must have corresponding permissions. The design of the system uses the "default reject" policy. The operation of authorization system identification includes:

(1) Read: know that an object exists, and browse all the metadata related to it. (2) Write: modify metadata related to objects. But it cannot be deleted. (3) Add: adds objects to the thematic space. (4) Remove: removes objects from the thematic space. (5) Workflow: a workflow that can be added to a set, such as having the right to reject a specific submission to enter a set.

4   System Test

In order to verify the design of English online learning resources rapid integration system, two groups of common English online learning resources rapid integration system are selected. By means of system comparative test experiment, the English online learning resources of a school are taken as the experimental object of the system comparative test, and the design of English online learning resources is verified on the a1419 computer operating platform This paper introduces the method of pattern recognition. The throughput, loss rate, false detection rate and integration efficiency of three groups of English online learning resources rapid integration system are compared.

4.1   Experimental Preparation

During the experiment, we need to detect the system throughput, loss rate, false detection rate and integration efficiency, so we use tamosoft throughput test software to complete the detection. Tamosoft throughput test can simulate TCP and UDP data streams in real time, calculate important indicators, and generate test results in digital or graphical form. In order to ensure that the system can be successfully tested and compared on the a1419 computer operating platform, the selected experimental test environment is shown in Table 1.

According to the a1419 computer operating platform environment shown in Table 1, input the original values of three groups of systems to test the performance of the system. In order to ensure the accuracy and preciseness of the test results, except for the experimental objects, other experimental variables are the same.

4.2   Experimental Result

First Group of Experimental Results
In the experiment, according to the experimental results of the online learning system, six groups of parameters are selected.

Table 1. Test environment

Environment	Name	Specifications
Software environment	Operating system	Windows 2000
	Browser	IE 6.0
	Database	Posttgre SQL 8.3
	Open source software	DSpace 1.5.2
Hardware environment	CPU	Intel Pentium 4
	CPU main frequency	1.6 GHz
	RAM	80G
	Network card	100/1000 M self-adaption
	Hard disk	40TB SRS
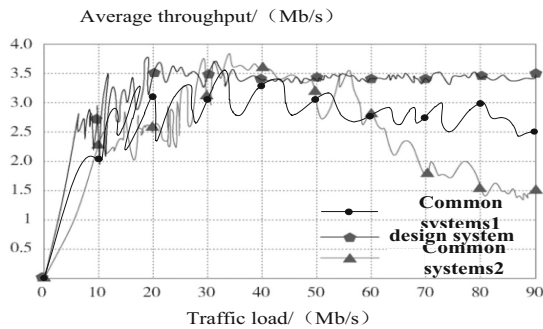
Fig. 6. Test results

It can be seen from Fig. 6 that when the traffic load is kept at 0–45 mb/s, the throughput of common system 2 is almost the same as that of the design system, which increases with the increase of network traffic and the increase of network utilization, but the traffic load increases to 45. After mb/s and higher, the network congestion caused by too many packets in the network, the network throughput curve of the original system began to change gradually, showing a downward trend. It is proved that when the network traffic load is too large, the original system can not keep normal operation and meet the user's resource demand; common system 1 differs from the design system, and the traffic load is different from the design system After increasing to 50 mb/s and higher, it shows a downward trend. It is proved that the original system can not keep normal operation and meet the resource demand of users when the network traffic load is too large. However, the throughput curve of the system is stable, and it will not cause network congestion due to the increase of traffic load and meet the needs of users.

The Second Group of Experimental Results

120000 digital resources are randomly selected from an English online learning platform for integration. The system in this paper and two common systems are tested. After the integration of the resources of the online learning platform, the loss rate and error detection rate of resources are compared. The performance of the three groups of systems is compared. See Table 2 for the comparison.

Table 2. Comparison of the results of system loss rate and error detection rate of three groups (%)

System	Number of resources	Loss rate	Noise factor
Design system	20000	0	0
	40000	0	0
	60000	0	0
	80000	0.14	0
	100000	0.21	0.13
	120000	0.33	0.12
Common systems 1	20000	0.38	0.31
	40000	0.79	0.42
	60000	1.21	0.56
	80000	1.56	0.95
	100000	1.89	1.47
	120000	2.43	2.38
Common systems 2	20000	0.27	0.21
	40000	0.46	0.29
	60000	0.78	0.37
	80000	0.95	0.46
	100000	1.23	0.53
	120000	1.45	0.79

It can be seen from Table 1 that after the design system is used to integrate the resources of the English online learning platform, the error detection rate and loss rate of the resources are far lower than those of the two groups of commonly used systems. Among them, the error rate of the proposed system integration resources is only about 0.14% when the number of resources is 80000, while the error rates of the other two methods are about 1.56% and 0.95% respectively. With the continuous increase of data, the error rates of the three methods have a rising trend, but the error rate of this method is lower. Therefore, the integration success rate of the design system is higher and the system performance is better.

The Third Group of Experimental Results

On the basis of the second group of experiments, the time of three groups of systems for the integration of English online learning platform resources was recorded, and the results were analyzed and compared to get the integration efficiency of different systems. The comparison is shown in Fig. 7.
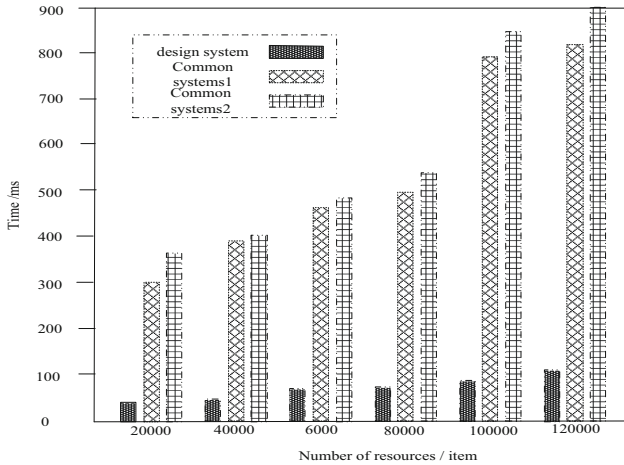
Fig. 7. Comparison of resource integration efficiency

It can be seen from Fig. 7 that the design system takes the shortest time to integrate the digital resources of university library, and with the growth of the number of resources, the overall rise of the time is relatively slow, and there is no phenomenon that the growth of the number of resources takes too long. However, the time of the other two groups of common systems grows faster when the number of resources is higher, indicating that the performance of the other two groups of common systems is unstable It's settled. Among them, the highest speed of Resource integration in this paper is about 100 ms, while the highest speed of other two systems is about 800 ms and 900 ms, respectively. In contrast, the speed of the proposed method is faster. Therefore, the design system has high integration efficiency and stable performance.

5   Conclusion

In conclusion, the design of a rapid integration system of English online learning resources based on multi-sensor network, which makes full use of the multi-sensor network, collects English online learning resources. The system test results show that the design has practical significance for the improvement of data throughput. The application of the design system to the integration of the resources of the current online learning platform can effectively improve the data processing ability of the integration of the resources of the online learning platform. In addition to the continuous improvement of dynamic resource configuration and access interface, the integration degree of

new integrated resources should be increased through further research on integration mode and integration technology, so that users can use all kinds of resources of public technology service platform more conveniently and effectively. In this system design, the corresponding sensor position is designed to improve the speed of resource collection.

References

1. Li, D., Liu, P.: Research on physical resource integration technology of training data center. Radio Eng. **51**(1), 74–77 (2021)
2. Han, Z.: University library digital resource integration system based on grid technology. Mod. Electron. Tech. **44**(3), 65–68 (2021)
3. Zhang, P., et al.: Exploration of resources integration of Beijing-Tianjin-Hebei city cluster based on urban innovation. Resourc. Dev. Market **37**(3), 327–332, 371 (2021)
4. Hong, Z.: Application of cloud computing in teaching resources integration of vocational colleges. Heilongjiang Sci. **12**(1), 74–75 (2021)
5. Liu, S., He, T., Dai, J.: A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese. Mobile Netw. Appl. **26**(5), 1891–1903 (2021). https://doi.org/10.1007/s11036-020-01725-x
6. Wang, Y.: Research on the necessity of integrating and sharing water conservancy information network and information resources. Wuxian Hulian Keji **17**(8), 108–109 (2020)
7. Yu, W.: On the integration of employment resources in colleges and universities from the perspective of precise supply. Jilin Normal Univ. J. (Hum. Social Sci Edn.) **48**(2), 94–99 (2020)
8. Li, K.: Design of teaching material resource integration system based on mobile technology. Microcomput. Appl. **36**(9), 67–69 (2020)
9. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mobile Netw. Appl. **24**(1), 5–17 (2018). https://doi.org/10.1007/s11036-018-1134-8
10. Yang, Y., Jao, J.: Resource integration optimization of elder care service platforms based on the convenient depth description of service modes. Chinese J. Manage. **17**(5), 725–733 (2020)

Quality Evaluation of Human Resource Management Information System Based on Intelligent Optimization Algorithm

Bo Sun[✉] and Hao-nan Chu

School of Labor Relations and Human Resource, China University of Labor Relations, Beijing 100048, China
fsdfdf55@163.com

Abstract. The conventional quality evaluation method of human resource management information system can not early warn the deterioration index with small weight, so a quality evaluation method of human resource management information system based on intelligent optimization algorithm is designed. Firstly, the experts' opinions are widely solicited, information is exchanged repeatedly, and the evaluation index system is established by determining the index type and the scale of the index set, and the weight of the index is assigned by using the quantitative index screening method; the McCall model is used as the quality evaluation model to accurately and quantitatively evaluate the software system index, and finally the ant colony algorithm is selected for the quality intelligent optimization evaluation So far, the research on the quality evaluation method of human resource information system based on intelligent optimization algorithm is completed. The simulation results show that the evaluation results of the method based on intelligent optimization algorithm are consistent with the data collection results, which can accurately reflect the real operation state of the system, and verify the effectiveness of the evaluation model.

Keywords: Intelligent optimization · Human resource · Information system · Quality evaluation

1 Introduction

In the background of the information age, the quality of the human resource information system determines the competitiveness of enterprises to a certain extent. Therefore, the evaluation of human resource information system is very important for enterprise information and enterprise development. However, due to the difficulty of quantitative and qualitative analysis, although many human resource information systems have been developed, the evaluation of human resource information system is a blank of human resource information system [1].

For example, in reference [2], AHP fuzzy comprehensive evaluation method is used in the design of quality evaluation system. The purpose of quality evaluation of human resource management information system is to evaluate the technical ability, work performance and utilization rate of the system, improve the management level and improve the economic benefits of the enterprise. System evaluation measures the current performance of the system and provides a basis for further improvement of the system in the future. The maintenance of the system is to ensure that the information system can continuously coordinate with the requirements of user environment, data processing operation, enterprise or other relevant departments.

In the existing quality evaluation method of human resource management information system, it is impossible to early-warning the deterioration index with less weight. Therefore, this paper studies the quality evaluation of human resource information system. For the first time, the paper uses the evaluation method of AHP, expert scoring and intelligent optimization algorithm to analyze the human resource information system comprehensively, quantitatively and qualitatively, and compares the evaluation results.

2   Quality Evaluation of Human Resource Management Information System Based on Intelligent Optimization Algorithm

2.1   Establishment of an Evaluation Indicator System

The design of the evaluation index system is to correctly evaluate the human resource information system, provide data support for enterprises and relevant departments, make the human resource information system more perfect and produce greater economic benefits. Designing a scientific, reasonable and systematic evaluation index system is the basis and premise of accurate and effective evaluation results. The evaluation index system should comprehensively reflect the objectives and requirements of the system to be evaluated, be as scientific and reasonable as possible, conform to the actual situation, and be basically acceptable to the relevant personnel and departments. Therefore, the establishment of the evaluation index system needs to be based on the comprehensive analysis of the system. Firstly, the draft of the index should be drawn up. After extensive solicitation of expert opinions, repeated exchange of information, statistical processing and comprehensive induction, the index type and the scale of the index set should be determined. On the one hand, the more types and quantities of indicators, the more comprehensive and accurate the evaluation conclusion will be. On the other hand, the larger the scale of the indicator set, the more likely the selected indicators are to be associated with each other, and the greater the workload and the more types of indicators, the higher the complexity of the indicator system and the greater the risk. In the selection of evaluation index set, each index should be independent of each other, and there is no mutual calculation relationship.

To determine the evaluation index system is a complex work with many related factors, which requires the participation and guidance of technical experts who have a good understanding of the evaluation objectives. Generally, the Delphi method is used, that is, by asking questions to experts, collecting expert opinions, summarizing and

synthesizing, and then feeding back to experts, and repeating the process for many times until the evaluation index system is determined [3, 4]. Finally, the evaluation index system of the system is determined. The index design process is carried out from four aspects of software characteristics, hardware characteristics, network characteristics and investment according to the method of layer by layer decomposition, thus forming the hierarchical structure of the index system, as shown in Table 1.

Through the evaluation of all these indicators, the goal of information system evaluation can be achieved. At the same time, after the completion of all the evaluation of these indicators, managers can accumulate a lot of information system development experience according to the evaluation results.

Table 1. Evaluation indicators architecture

First level indicators	Secondary indicators	Third level indicators
Software features X1	Functionality X11	Accuracy X111
		Security X112
		Applicability X113
	Reliability X12	Maturity X121
		Fault tolerance X122
		Recoverability X123
	Maintainability X13	Analyzability X131
		Transformability X132
		Testability X133
	Accessibility X14	Intelligibility X141
		Easy to learn X142
		Easy to operate X143
	Portability X15	Consistency X151
		Replaceability X152
		Easy to install X153
	Efficiency X16	Time characteristics X161
		Resource characteristics X162
Hardware features X2	Functional X21	Security X211
		Applicability X212
	Reliability X22	Stability X221
		Maturity X222
	Maintainability X23	Analyzability X231
		Extensibility X232
Network characteristics X3	Reliability X31	Stability X311
		Fault tolerance X312

Table 1. (*continued*)

First level indicators	Secondary indicators	Third level indicators
	Security X32	Physical security X321
		Technical safety X322
	Efficiency X33	Network bandwidth X331
		Communication equipment X332
Investment X4	Real income X41	Efficiency X411
		Accuracy X412
		Cost X413
	Hidden incomes X42	Brand image X421
		Process reengineering X422
	Budget performance X43	Timeliness X431
		Accuracy X432

After the determination of the indicator system, the quantitative indicator screening method can be used to assign the weight of the indicators, that is, the statistical method is used to screen out the calculation weight of each indicator, and the information of the original indicator set is not lost [5–7]. Firstly, the conditional generalized variance minimization method is selected, assuming that there are m indexes $X1, ..., Xm$ of n group observation data

$$X = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1m} \\ x_{21} & x_{22} & ... & x_{2m} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{nm} \end{bmatrix} \tag{1}$$

In matrix X of formula (1), each row is a group of observed data, and the covariance between the mean value and variance of x_i and x_i and x_j can be obtained:

$$\overline{X}_i = \frac{1}{n} \sum_{n=1}^{n} x_{ai}$$
$$S_{ii} = \frac{1}{n} \sum_{a=1}^{n} (x_{ai} - \overline{X}_i)^2 \tag{2}$$
$$S_{ij} = \frac{1}{n} \sum_{a=1}^{n} (x_{ai} - \overline{X}_i)(x_{ai} - \overline{X}_j)$$

Matrix $S = (S_{ij})_{p \times p}$ constitutes the variance and covariance matrix of these m indicators, and $|S|$ is its determinant, which can reflect the changes of indicators, also known as generalized variance [8]. When $m = 1$, $|S| = |S_{ii}|$ is the variance of variable X_i. There has been a proven theorem: the value range of generalized variance is $0 < |S| < 1$. If all the values between X_m are independent of each other X, then the value of generalized variance $|S|$ is the largest. If the data between X_i are linearly correlated, then the value of

generalized variance $|S|$ is 0. The value of $|S|$ reflects the importance of the indexes to be evaluated, so as to realize the weight between the indexes Assignment. After the selection of indicators is completed, the weight of each index needs to be analyzed. This paper uses analytic hierarchy process to calculate, constructs a multi-level and multi-structure model for each element of the problem, compares the elements of the same structure with the previous structure, establishes a discrimination matrix, judges the importance of the two factors according to the judgment scale, and calculates the relative weight of the elements to the level. When comparing the importance of n elements B_1, B_2, ..., B_n to the previous layer, it is necessary to establish the proportion of them in the previous layer. For any two elements B_i and B_j, the proportion of influence degree is represented by b_{IJ}. The scale table of importance constructed is shown in Table 2:

Table 2. Scale of importance

Scale value	Meaning	Scale value	Meaning
1	Equally important	2	The tradeoff between 1 and 3 judgment
3	More important	4	The tradeoff between 3 and 5 judgments
5	Important	6	The tradeoff between 5 and 7 judgments
7	Very important	8	The compromise between 7 and 9 judgments
8	Extremely important	Remarks	B_i is compared with B_j to get b_{IJ}; B_j is compared with B_i to get $\frac{1}{b_{IJ}}$

That is, two judgment matrices $B = (b_{IJ})_{n \times n}$ are obtained. Then, the relative weight between each element needs to be calculated. After the weight of the previous level index is assigned to the current index, the weight of each indicator is recorded as w_1, w_2, ..., w_n, and the relative weight of B_i and B_j is:

$$b_{IJ} = \frac{w_i}{w_j}(i, j = 1, 2, ..., n) \tag{3}$$

Due to the diversity and one sidedness of human factors, and many evaluation factors, it is necessary to use the fuzzy theory of artificial intelligence technology for specific calculation and discrimination. Fuzzy theory is to use fuzzy logic to describe the level of things in real life, and quantify it fuzzy, then get the degree of belonging, and use the degree of belonging to complete the evaluation of curriculum design [9, 10]. The judgment matrix can be expressed as:

$$B = \begin{bmatrix} \frac{w_1}{w_1} & \frac{w_1}{w_2} & \cdots & \frac{w_1}{w_n} \\ \frac{w_2}{w_1} & \frac{w_2}{w_2} & \cdots & \frac{w_2}{w_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{w_n}{w_1} & \frac{w_n}{w_2} & \cdots & \frac{w_n}{w_n} \end{bmatrix} \tag{4}$$

The weight vector can be expressed as:

$$W = (w_1, w_2, ..., w_n)^T \tag{5}$$

The judgment matrix can be expressed as:

$$B = W \cdot \left(\frac{1}{w_1}, \frac{1}{w_2}, ..., \frac{1}{w_n} \right) \tag{6}$$

Fuzzy matrix is an important tool in the research of fuzzy relationships, which is based on the idea of the weighted average method, and can accurately calculate the weight of each index.

2.2 Design Quality Evaluation Model

Information system quality evaluation model is the basis of system quality evaluation. With the continuous development of the systematization industry, people are more and more aware of the importance of system quality evaluation. As the basis of system quality evaluation, more and more experts at home and abroad began to invest in this aspect of research work from the late 1970s. According to the evaluation index system established in this paper, the quality evaluation model used in this paper is the McCall model, which can divide the system quality evaluation into three levels, namely factor, quality and criterion [11–13]. The model divides the evaluation of software quality into 11 quality factors in three major aspects, including software correctness, reliability, efficiency, integrity, availability, maintainability, testability, flexibility, portability, reusability and interoperability, as shown in Fig. 1:
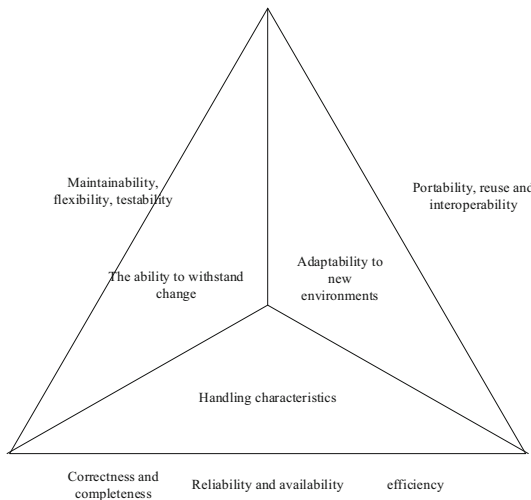
Fig. 1. McCall system quality assessment model

In addition, there are corresponding evaluation criteria for each quality factor. In the McCall model, 23 evaluation criteria are involved, which are extracted from software attributes. The 23 evaluation criteria are completeness, consistency, accuracy, fault tolerance, simplicity, modularity, generality, extensibility, instrumentality, self description, execution efficiency, storage efficiency, access control, access review, operability, training, communication ability, software system independence, machine independence, communication commonality, data commonality and simplicity. By evaluating the corresponding evaluation criteria of each quality factor, the evaluation of the software in the quality factor is obtained. According to the evaluation model in Fig. 1, the membership function of geometric elements is constructed. According to the characteristics of quantitative data, the triangular membership function is selected to calculate the average threshold of quantitative data, determine the range of comments set, and map to the function interval. According to the data characteristics of gesture recognition operation time, taking the minimum and maximum time of the task as the boundary points, the membership function of "very short time" can be obtained:

$$ux = \begin{cases} 1 & x < 1.0 \\ -2x + 3 & 10 \le x < 1.5 \\ 0 & 1.5 \le x \end{cases} \tag{7}$$

The membership function of "long time" can be expressed as:

$$ux = \begin{cases} 2x - 4 & 2.0 \le x < 2.5 \\ 1 & 2.5 \le x \\ 0 & else \end{cases} \tag{8}$$

Then, the membership function of "shorter time" can be obtained respectively:

$$ux = \begin{cases} 2x - 3 & 1.0 \le x < 1.5 \\ -2x + 4 & 1.5 \le x < 2.0 \\ 0 & else \end{cases} \tag{9}$$

The membership function of "long time" can be expressed as:

$$ux = \begin{cases} 2x - 3 & 1.5 \le x < 2.0 \\ -2x + 5 & 2.0 \le x < 2.5 \\ 0 & else \end{cases} \tag{10}$$

For the evaluation of these criteria, McCall model gives some measures. According to the above membership function, the membership function image of quality evaluation time is shown in Fig. 2
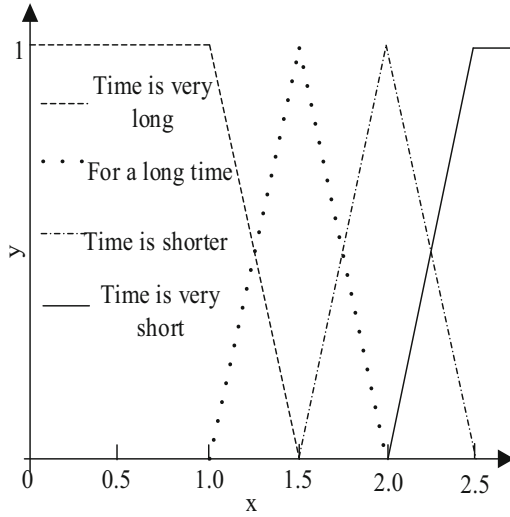
Fig. 2. Membership function diagram of gesture recognition operating time

The membership function is constructed according to the corresponding membership relationship of the input quantitative data, and the index weight of each fault in the system is calculated. Based on the membership function, the comprehensive fuzzy matrix of fault identification performance measurement is established. After the comprehensive evaluation is obtained, the comprehensive evaluation of user experience is output. The fuzzy multi-layer matrix evaluation is established through the above steps, and the comprehensive evaluation model is established through the above steps, It can objectively and accurately evaluate the quality of human resource management system The calculation method given in the model is comprehensive and accurate, and the software system analyzed by the model can be accurately evaluated quantitatively.

2.3 Introduction of Intelligent Optimization Algorithm

In this paper, the ant colony algorithm is selected to complete the quality evaluation of the system. In fact, the ant colony algorithm is to imitate the foraging behavior of ants to find the optimal path of quality evaluation. Based on the large number of individuals of ant colony, they communicate with each other through pheromones to transfer information for their peers. A large number of ants form a feedback system, which can find degradation indicators with any weight. Therefore, it has high efficiency and time complexity in the evaluation process, and has great value. It solves the problem that the slight deterioration index is easy to ignore. This paper studies the problem of system quality evaluation. The main goal is to check the deterioration index of human resource management information system in the process of operation, to create higher profits for enterprises.

Ant colony algorithm has a good search ability, but its initial information is rapidly lacking, and the convergence speed is relatively slow. Therefore, this paper combines ant colony algorithm with genetic algorithm, and proposes an improved hybrid ant colony algorithm for iterative solution. In ant colony algorithm, the number of ants needs to be set first. According to the characteristics of ant colony algorithm, when the number of ants is m and the global search ability and convergence speed are guaranteed, there is the following relationship between the number of ants and the scale n of scheduling problem:

$$m = \sqrt{n} \sim \frac{n}{2} \tag{11}$$

The scale n of scheduling problem mainly includes the number of system evaluation indexes. Ant algorithm is a kind of evolutionary algorithm based on population, which is applied in path optimization. At the beginning, all ants will choose different paths to finding food. In the search process, ants will communicate with each other by pheromone, and choose the better path to search for food for the second time. If it iterates repeatedly, an optimal path will be found. The path search process is shown in Fig. 3.
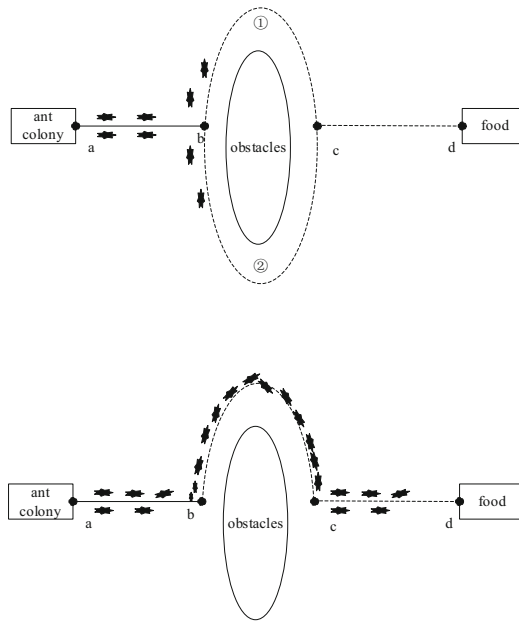
Fig. 3. Selection of ant foraging path

Combined with ant colony algorithm, the intelligent optimization algorithm flow of system quality evaluation in this paper is obtained, as shown in Fig. 4:
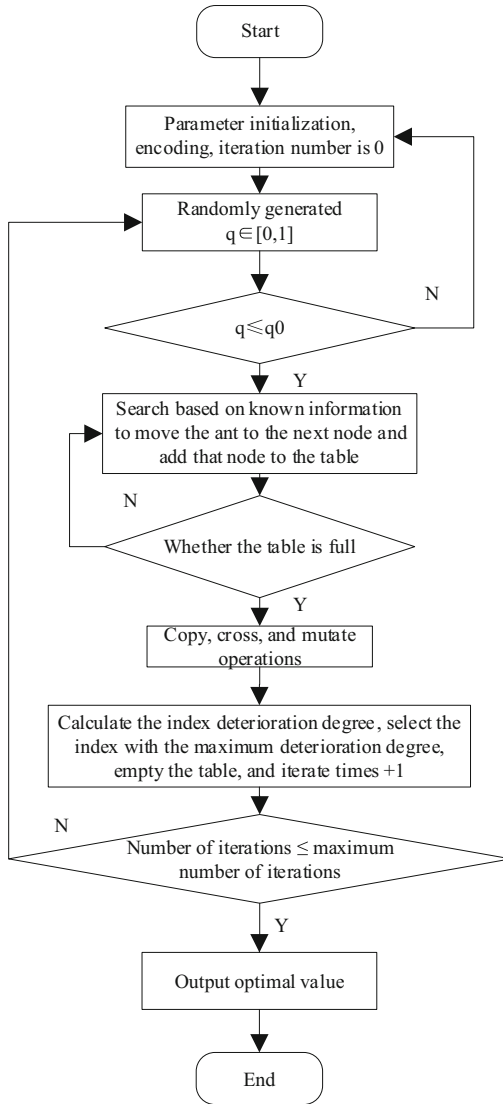
Fig. 4. Improved ant colony algorithm

The selection, crossover and variation of genetic algorithm are three genetic operations based on gene coding. The degradation index is coded by imitating chromosome coding, and a set of deterioration index codes composed of natural numbers are obtained. The core idea of selection is to copy, and the optimal solution in the reproduction inheritance parent continues to improve, so as to avoid the loss of high-quality solution, and the crossover can generate new individuals and increase diversity, prevent premature stagnation, mutation operation on the optimal individual, and save the optimal solution.

3 Method Testing

3.1 Design of Test Method

In order to verify the effectiveness of the quality evaluation method designed in this paper, the design method needs to be tested. A running human resource management information system is established through the simulation platform. A time node is set every 10 min to monitor the data of each evaluation index. At the same time, the traditional quality evaluation method and the designed evaluation method are used to evaluate the simulation system at the same time node. Finally, the evaluation results of the two systems are compared. The experimental platform is xsimstudio, and the topology is shown in Fig. 5:
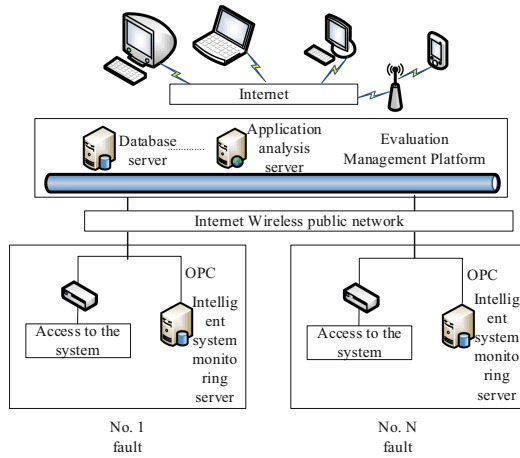
Fig. 5. Experimental topology

Based on the theory of AHP, the four experts of the project evaluation working group conducted a comprehensive study on the evaluation indexes of the system, and then compared the indexes of each level of the project to form a judgment matrix, and calculated the eigenvector and the maximum eigenvalue of the judgment matrix. Finally, the consistency detection of each judgment matrix was completed, and the results are shown in Table 3:

Table 3. Evaluation matrix under results

Assessment results	Software features	Hardware features	Network characteristics	Investment
Software features	1	5	7	3
Hardware features	1/5	1	3	1/3
Network characteristics	1/7	1/3	1	1/5
Investment	1/3	3	5	1

According to the eigenvector:

$$W = (W_1, W_2, ..., W_n)^T$$
$$= (0.5214, 0.1852, 0.0551, 0.3541)^T \qquad (12)$$

According to the maximum eigenvalue, the consistency test is passed. According to the above form, the judgment matrix of software characteristics, hardware characteristics, network characteristics, investment and other subordinate branches at all levels is obtained. The system with the above parameters runs normally on the platform, and each evaluation index is detected every 10 min, and the data is recorded to obtain the index data at time $T_1 - T_5$. The calculation results in Table 4 are obtained:

Table 4. Evaluation indicator monitoring data for each time point

Index	T_1	T_2	T_3	T_4	T_5
X111	0.54	0.73	0.96	0.21	0.19
X112	0.58	0.43	0.70	0.87	0.25
X113	0.95	0.30	0.50	0.25	0.77
X121	0.60	0.25	0.12	0.40	0.75
X122	0.87	0.73	0.59	0.23	0.96
X123	0.95	0.38	0.15	0.96	0.88
X131	0.28	0.78	0.94	0.11	0.55
X132	0.62	0.53	0.83	0.90	0.61
X133	0.03	0.65	0.22	0.49	0.54
X141	0.67	0.28	0.44	0.28	0.70
X142	0.23	0.14	0.91	0.82	0.29
X143	0.81	0.80	0.58	0.82	0.81
X151	0.70	0.70	0.43	0.83	0.28

(*continued*)

Table 4. (*continued*)

Index	T_1	T_2	T_3	T_4	T_5
X152	0.58	0.34	0.74	0.77	0.21
X153	0.23	0.55	0.76	0.39	0.15
X161	0.30	0.49	0.84	0.30	0.31
X162	0.33	0.73	0.93	0.44	0.15
X211	0.04	0.59	0.63	0.97	0.53
X212	0.35	0.23	0.09	0.32	0.16
X221	0.06	0.76	0.99	0.13	0.84
X222	0.95	0.82	0.54	0.45	0.76
X231	0.41	0.89	0.69	0.01	0.59
X232	0.77	0.59	0.87	0.14	0.07
X311	0.75	0.60	0.60	0.90	0.69
X312	0.01	0.80	0.48	0.11	0.90
X321	0.45	0.46	0.41	0.38	0.12
X322	0.23	0.86	0.48	0.45	0.04
X331	0.07	0.91	0.10	0.13	0.87
X332	0.27	0.87	0.73	0.99	0.54
X411	0.17	0.83	0.43	0.26	0.97
X412	0.08	0.40	0.22	0.46	0.66
X413	0.49	0.67	0.98	0.62	0.42
X421	0.27	0.96	0.21	0.99	0.12
X422	0.20	0.15	0.11	0.66	0.01
X431	0.91	0.83	0.01	0.21	0.41
X432	0.34	0.46	0.37	0.65	1.00

From the detection data of the above five time points, at $T_3 - T_4$ time, the change trend of X121 increased significantly, and the deterioration phenomenon appeared. At the subsequent time points, the deterioration of X131 and X211 also gradually appeared, indicating that the operation state of the system was abnormal. Therefore, the users of the system at T_4 time should be vigilant, check the problem, and serious problems appear at T_6 time.

3.2 Simulation Results and Analysis of Single Fault Case

Under the above experimental conditions, the traditional evaluation method and the evaluation method designed in this paper are used to evaluate the simulation index data of six time points, and the evaluation vector is analyzed. Firstly, it is set in the case of single fault. The evaluation results of the two methods are shown in Table 5:

Table 5. Comparison of single fault assessment results

Point of time	Evaluation vectors obtained by traditional methods	The evaluation vector obtained by this method
T_1	(0.37,0.63,0,0)	(0.37,0.63,0,0)
T_2	(0.16,0.84,0,0)	(0.17,0.80,0.03,0)
T_3	(0.24,0.76,0,0)	(0.25,0.75,0,0)
T_4	(0.18,0.71,0.1,0.01)	(0.24,0.69,0.06,0.01)
T_5	(0.14,0.57,0.27,0.02)	(0.18,0.53,027,0.02)
T_6	(0.25,0.26,0.46,0.0,3)	(0.23,0.50,024,0.03)

The format of the evaluation vector is (a_1, a_2, a_3, a_4) and $a_1 + a_2 + a_3 + a_4 = 1$. when a_4 is 0, the running state of the human resource management information system is normal. When $0 < a_4 \leq 0.02$, the running state of the human resource management information system is abnormal. When $a_4 > 0.02$, the running state of the human resource management information system is serious. From the evaluation results of the simulation experiment, although the traditional evaluation model and the evaluation model designed in this paper have some differences in the specific value of the vector, the result of the running state judgment is more in line with the actual situation, the evaluation result is "abnormal" at $T_4 - T_5$ time, and the evaluation result is serious quality problem at T_6 time. The experimental results show that the design is feasible in the case of a single fault The performance of the quality evaluation method of human resource management information system is very good compared with the traditional methods, which can accurately reflect the real running state of the system.

3.3 Simulation Results and Analysis of Double Fault Situation

Under the same experimental conditions, two failures occurred in the same time. In the case of double faults, the evaluation results of the two methods are shown in Table 6:

Table 6. Comparison of double fault assessment results

Point of time	Evaluation vectors obtained by traditional methods	The evaluation vector obtained by this method
T_1	(0.24,0.76,0,0)	(0.24,0.76,0,0)
T_2	(0.17,0.80,0.3,0)	(0.17,0.79,0.04,0)
T_3	(0.18,0.76,0.06,0)	(0.25,0.70,0.05,0)
T_4	(0.31,0.61,0.08,0.00)	(0.24,0.69,0.06,0.01)
T_5	(0.17,0.64,0.18,0.01)	(0.18,0.53,027,0.02)
T_6	(0.20,0.31,0.46,0.03)	(0.23,0.47,0.24,0.03)

From the evaluation results of simulation experiment, the traditional evaluation model is still evaluated as "qualified" at T_4 time, as "abnormal" at T_5 time, as "serious quality problem" at T_6 time; the evaluation model designed in this paper is evaluated as "abnormal" at $T_4 - T_5$ time, as "serious" at T_6 time, which is consistent with the results of data acquisition. Compared with the traditional evaluation method, there are some differences, but the final result is not much different, which shows that in the case of double failure, the performance of the designed method is better than the traditional method, and it can accurately reflect the real operation state of the system.

3.4   Simulation Results and Analysis of Multi-fault Situation

Under the same experimental conditions, the system is set to have multiple faults in the same time. In the case of multiple faults, the evaluation results of the two methods are shown in Table 7:

Table 7. Comparison of multi-fault assessment results

Point of time	Evaluation vectors obtained by traditional methods	The evaluation vector obtained by this method
T_1	(0.25,0.75,0,0)	(0.25,0.75,0,0)
T_2	(0.24,0.76,0,0)	(0.25,0.75,0,0)
T_3	(0.25,0.75,0,0)	(0.25,0.75,0,0)
T_4	(0.25,0.75,0,0)	(0.24,0.69,0.05,0.02)
T_5	(0.25,0.47,0.28,0)	(0.23,0.50,024,0.03)
T_6	(0.25,0.26,0.02)	(0.23,0.50,024,0.03)

From the evaluation results of simulation experiment, the traditional evaluation model is still evaluated as "qualified" at time $T_3 - T_4$, and it is only "abnormal" at time T_5. The evaluation model designed in this paper is evaluated as "abnormal" at time T_3, and serious at time T_4, which is consistent with the result of data collection. It shows that in the case of a single fault, the designed human resource management of detection is very important compared with the traditional methods, the information system quality evaluation method can accurately reflect the real running state of the system.

To sum up, the evaluation results of the two evaluation methods are consistent and relatively accurate in the case of single fault and double fault, but in the case of multiple faults, the shortcomings of the traditional evaluation methods are revealed, and the evaluation results are quite different from the actual situation, which shows that the quality evaluation method based on intelligent optimization is better than the traditional method, It has certain validity.

4 Conclusion

The quality of human resource management information system is related to the development technology of the system, the quality of the developers, the organization of the development, the control of the development process and the utilization rate of the development equipment. The above analysis is based on a single factor, but the various factors interact. It is very complex to accurately measure the cost, benefit and quality of the management information system, which should be considered comprehensively according to the characteristics of the enterprise itself. At the same time, the existing success and failure cases should be treated dialectically, and the cost-benefit prediction provided by IT department should be fully demonstrated. The most important thing is the accurate positioning and evaluation of the enterprise's own situation. Otherwise, the detailed cost-benefit analysis and its conclusion may lead to huge deviation after the operation of the system.

In addition, system evaluation itself has many inherent difficulties, namely, single evaluation is easy, comprehensive evaluation is difficult; quantitative evaluation is easy, qualitative evaluation is difficult; hard index (Science and technology, production, level) evaluation is easy, soft index (society, economy, organization) evaluation is difficult. Even if the evaluation index can be designed, the evaluation practice will be a difficult thing because of the inherent relationship between the index system and decision-making and the constraints of the environment. Therefore, further study on the theory of system evaluation, combined with the current situation of social and economic development to make the application of evaluation theory and method, still needs relevant scholars to continue to make efforts.

References

1. Li, W., Long, F., Sun, X.: Quality elements definition and quantitative evaluation of software quality for shipboard C3I system. Ship Electron. Eng. **39**(4), 114–116 (2019)
2. Zhang, W., et al.: Realtime power quality evaluation system of the electric propulsion ship based on AHP-fuzzy comprehensive evaluation method. Chin. J. Ship Res. **14**(6), 48–57 (2019)
3. Zhu, X., Wang, J.: Research on several key technologies in machine translation quality automatic evaluation system. Sci. Educ. Article Cultu. (12), 115–117+125 (2019)
4. Lin, T., et al.: Wind turbine performance evaluation based on support vector machine optimized by improved fruit fly optimization algorithm. Renew. Energy Resour. **37**(1), 132–137 (2019)
5. Ma, W., et al.: Flexibility evaluation and optimal dispatch model of distribution network considering soft open point. Power Syst. Technol. **43**(11), 3935–3943 (2019)
6. Fu, K., Yan, Y.: Research on university teaching management evaluation based on adjusting and optimizing BP neural network. Modern Electron. Technique **42**(17), 152–154 (2019)
7. Wang, H., et al.: Evaluation on state of health of storage battery for power grid based on artificial intelligence. Guangdong Electr. Power **32**(4), 79–84 (2019)
8. Hu, P., Shuai, B., Wu, Z.: Intelligent method for evaluating percolation robustness of hazardous materials transportation network. China Saf. Sci. J. **29**(2),57–62 (2019)
9. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mobile Netw. Appl. **24**(1), 5–17 (2018). https://doi.org/10.1007/s11036-018-1134-8

10. Wang, Y.-x., et al.: The design and implementation of human resource management information system basing on web service. Comput. Eng. Softw. **40**(2), 63–66 (2019)
11. Liu, S., Pan, Z., Cheng, X.: A novel fast fractal image compression method based on distance clustering in high dimensional sphere surface. Fractals **25**(4), 1740004 (2017)
12. Wang, H., et al.: Research on the application status and prospect of information system in human resource management. China Comput. Commun. **31**(23), 54–56 (2019)
13. Liu, S., He, T., Dai, J.: A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese. Mobile Netw. Appl. **26**(5), 1891–1903 (2021). https://doi.org/10.1007/s11036-020-01725-x

Human Resource Social Insurance Data Dynamic Update System Based on Wireless Communication

Hao-nan Chu and Bo Sun[(✉)]

School of Labor Relations and Human Resource, China University of Labor Relations,
Beijing 100048, China
`fsdfdf55@163.com`

Abstract. The current human resource social insurance data updating system cannot realize real-time data transmission, which leads to its low efficiency. Therefore, a new dynamic updating system of human resource social insurance data based on wireless communication is designed. Hardware protection and anti-jamming modules are added to the hardware. The wireless communication protocol was established and the social security information data of human resources were dynamically mined. The wireless transmission technology is used to realize the real-time transmission of social insurance data. The system test results show that the efficiency of the system designed in this study has been significantly improved.

Keywords: Wireless communication · Human resource · Social insurance · Insurance data · Dynamic update

1 Introduction

Social insurance of human resources refers to a social security system in which the state raises funds in various ways in accordance with the law, guarantees workers and provides economic compensation. Workers are the main driving force for the country to move forward. When they are unable to work normally for various reasons, they should be protected by the state for their basic living security. As the main reasons that lead to workers unable to work normally are too old, illness, work injury, childbearing or enterprise reasons leading to unemployment, etc., so social insurance mainly includes social pension, medical, work-related injury, childbirth, unemployment and other insurance [1]. The social insurance system is the core of the whole social security system. The social insurance system with Chinese characteristics can fully adjust the interests of both sides and promote the harmonious development of society. It plays a very important role in promoting the reform, development and stability, and effectively promotes the sustainable, healthy and stable development of economy. In recent years, computer technology and network technology are becoming more and more common, greatly liberating the productive forces, improving production efficiency, information construction

has become a new development point of social security system construction. The construction of social security informatization is a comprehensive problem. It integrates various technologies, integrates computer science, network engineering, software engineering and other disciplines and fields, and is also established outside the networking and information sharing system between various social security departments, units and social security departments. Through the centralized management and control platform, the comprehensive analysis of information resources can be effectively realized.

Through the development of human resources social insurance management platform, the efficient management of social security data is realized. However, social security data presents a real-time change mode, and user information and insurance status may change at any time. In order to ensure the timeliness and accuracy of the data in the human resources social insurance management platform, it is necessary to update the human resources social insurance data dynamically [2]. Data update is the process of replacing the old data item or record in the data file or database with the new data item or record, which is realized by deleting, modifying and re inserting.

At present, some research results have been obtained for insurance data updating. For example, in the literature [3], Yang put forward based on the security network coding technology of the medical insurance information system data update method, using piecewise linear coding technology for data processing, information fusion based on spectral feature extraction and the fuzzy control method for control system data update process, complete medical insurance information system data update. However, this data updating system has the problem of slow updating speed. Applying the traditional data updating system to the human resources and social insurance data management will lead to serious quality problems of social security data.

Therefore, wireless communication technology is introduced in this study. Wireless communication refers to the long-distance transmission and communication between multiple nodes without conductor or cable transmission. Radio, radio, etc. can be used for wireless communication. Through the use of wireless communication technology to achieve cross regional social security information sharing and transmission, to ensure that the new social insurance information is transmitted to the data storage environment at the first time, so as to improve the dynamic update efficiency of human resources social insurance data.

2 Hardware System Design of Social Insurance Data Dynamic Update

The design of human resources social insurance data dynamic update system adheres to the idea of platform and plug-in. The functions to be realized fully meet the existing needs of the social security system, and fully consider the scalability and compatibility problems brought by the possible future needs. Based on the above system design principles, carry out the design and development of human resources social insurance data dynamic update based on wireless communication technology, so that every step of system development has rigorous design basis and scientific design guidance, so as to improve the reliability and availability of the system. According to the structure, the social insurance data dynamic update system can be divided into three parts: hardware,

database and software. The basic structure of the wireless communication part in the hardware system is shown in Fig. 1.
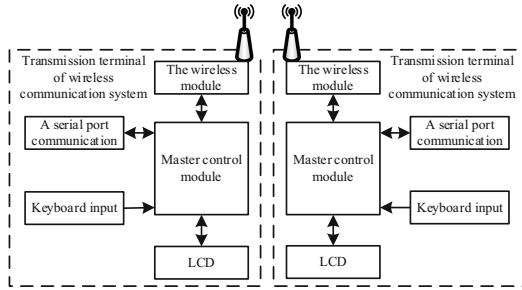
Fig. 1. Hardware structure of wireless communication module.

The micro control unit of the main control module in Fig. 1 is the core of the whole system. All logic operations and most control functions are completed by it. At the same time, it also provides many peripheral interfaces, including wireless module, LCD module, keyboard input module and serial communication module. Wireless communication module is the specific part of wireless communication between terminals. It can transmit and receive data wirelessly through a wireless transceiver antenna. The MCU of the sending terminal transmits the data to the wireless module of the sending terminal, which will send the data out wirelessly; when the wireless module of the receiving terminal receives the information, it will transmit it to the MCU of the receiving terminal for processing. Serial communication provides a way of communication between wireless communication terminal and PC. In practical application, the specific data of the upper PC can be written into the data storage space of the MCU chip of the wireless communication terminal through serial communication, and the data in the wireless communication terminal can also be returned to the upper PC machine for recording, storage and analysis. In addition, the keyboard input module provides the way of manual input for the whole system [4]. The current working state of wireless communication terminal can be changed by manual control, and the data transmission and reception can be controlled manually.

In the process of optimization design of social insurance data dynamic update hardware system, on the basis of traditional data update hardware system, it introduces the relevant equipment needed by wireless communication technology, optimizes some processing elements, adjusts the connection circuit between various hardware devices, and obtains the final hardware system optimization results.

2.1 Main Chip Selection of Hardware System

2.1.1 Wireless Transceiver Chip

The wireless data transmission module based on nRF2401 is adopted. NRF2401 can set 40bit address through software, which is suitable for point to multipoint data transmission; Built in CRC error correction hardware circuit and protocol, improve the reliability

of the system, and no longer need to use software to transmit data error control coding, simplify the software programming; the working temperature range is –40–+85 °C, suitable for use in industrial field; Data transmission rate is 1 Mbps, low power consumption, small size. NRF2401 has three wire serial interface, and all working parameters such as transmit power and channel can be programmed. The current consumption of nRF2401 is low. When the output power is –5 dBm, the current consumption of nRF2401 is 10.5 ma; In the receiving mode, the power current consumption is 18ma, and has a low power consumption mode. NRF2401 has 125 channels, which can meet the needs of multi frequency and frequency hopping. The data transmission rate is 1Mbps and the channel conversion time is less than 200 µs. The power supply voltage is 1.9–3.6 V, which meets the needs of low power consumption design. A special voltage stabilizing circuit is set in the chip, and all kinds of power supply, including DC/DC switching power supply, have good communication effect.

2.1.2 Digital-to-Analog Conversion Chip

MAX197 is a high-speed A/D conversion chip with 12 bit measurement accuracy introduced by Maxim company. It only needs a single power supply, and the conversion time is very short. It has 8 input channels. It also provides a standard parallel interface – 8-bit three state data I/O port, which can be directly connected with most single-chip computers. It is very convenient to use. MAX197 can complete a/D conversion function independently without external components. It can be divided into internal sampling mode and external sampling mode. The sampling mode is determined by D5 bit of control register. In the internal sampling control mode, the sampling interval is started by the write pulse, and the A/D conversion is started after the instantaneous sampling interval [5]. In the external sampling mode, two write pulses control the sampling and a/D conversion respectively. When the first write pulse appears, write acqmod to 1 and start the sampling interval. When the second write pulse appears, write control word acqmod is 0, MAX197 stops sampling and starts a/D conversion. The time interval between the two write pulses is a sampling time. When a conversion is finished, the corresponding int pin of MAX197 is set to low level to inform the processor that the conversion result can be read.

2.1.3 Single Chip Microcomputer Chip

AT89C52 microprocessor produced by ATMEL company is selected as MCU, which is compatible with 80C51 and 80C52. It has 8K bytes of on-line reprogrammable memory and can program the chip many times. It also has 256 bytes of ram, which is more suitable for processing large amounts of data. The main clock frequency of the chip reaches 24 MHz, which shortens the interrupt response time and improves the execution speed of the program. Because the program memory is designed in the MPU chip, it can save space and improve the reliability of the system. AT89C52 has 40 pins, two external bidirectional I/O ports, two external and middle fractures, three 16 bit programmable timing counters, two full duplex serial communication ports and two read-write port lines. AT89C52 can be programmed according to the conventional method or online programming [6]. The connection of some pins of single chip microcomputer is shown in Table 1.

Table 1. Description of pin function of single chip microcomputer.

Pin type	Pin name	Pin function
Power and clock pins	VCC	Power terminal
	GND	Ground terminal
	XTAL1	The input of oscillator inverting amplifier and internal clock generator
	XTAL2	The output of the oscillator inverting amplifier
Multiplex pin of control line or other power supply	RET	Reset input
	PSEN	Program memory external access control signal
Input and output pins	P0	8-bit open drain bidirectional I/O口
	P1	8-bit bidirectional converter with internal lift resistor I/O口
	P2	8-bit bidirectional converter with internal lift resistor I/O口
	P3	8-bit bidirectional converter with internal lift resistor I/O口

2.2 System Circuit Design

2.2.1 Power Circuit

Power supply module is mainly to provide appropriate and stable working voltage for the system, so it is a key factor for the stability of the whole system. In the social insurance data dynamic update system, the standard working voltage of MCU chip and LCD module is 5 V, while the standard working voltage of MCU chip is 3.3 V, so it is necessary to design the power supply module to provide the system with stable power supply of 5 V and 3.3 V. The modified power circuit is shown in Fig. 2.
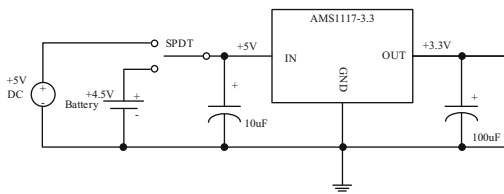
Fig. 2. Power supply circuit diagram.

2.2.2   Data Storage Circuit

The data storage device mainly solves two problems: On the one hand, it solves the contradiction between the large amount of data obtained by the acquisition system and the small data memory of the single chip microcomputer; on the other hand, it acts as the "buffer" between the data acquisition system and the wireless data transmission module. At24cu8 has ezprom with IZC bus interface. It has small volume, strong anti-interference ability and 8 kB data storage space. The chip is divided into 64 pages with 16b per page. Data reading and writing can be completed by controlling the clock line SCL and data line SDA [7]. The specific connection of the data storage circuit is shown in Fig. 3.
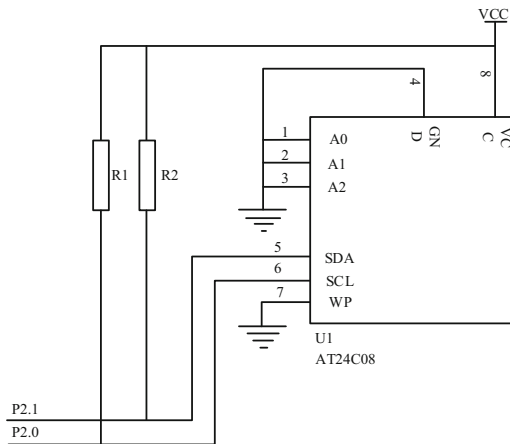
Fig. 3. Data storage circuit diagram.

2.2.3   Reset Circuit

In the system, the reset circuit mainly completes the power on reset of the system and the user's key reset function when the system is running. The reset circuit can be composed of simple circuits, or other relatively complex circuits with more perfect functions can be used. The common reset circuit can work normally in the experimental environment with low cost. The industrial control environment requires a high level. The special reset chip can be used to improve the anti-interference performance of the system and reduce the possibility of system misoperation. The system uses reset circuit.

2.2.4   Clock Circuit

The clock circuit is mainly used to generate the oscillation clock of the microprocessor chip, and each operation instruction is carried out in each working clock. Theresa's RL78 series chips have integrated a 32 MHz high-speed oscillator, which can be used as the working clock. In addition, two more clock circuits can be connected externally, namely the main clock circuit and the sub clock circuit. The CPU working clock can freely

select the internal high-speed oscillator or the main and sub-clock circuits as its clock oscillation provider. Special attention must be paid to the clock input of the RTC circuit to select the sub-clock circuit, so the sub-clock circuit design uses a 32.768 kHz passive crystal oscillator; and the main clock circuit, refer to the RL78 user manual, select an 8 MHz passive crystal oscillator.

2.2.5   Liquid Crystal Display Circuit

Choose the LCD 1602, which is an industrial character liquid crystal, which can display 16×2 or 32 characters at the same time. It is a dot matrix liquid crystal module specially used to display letters, numbers, symbols, etc. [8]. Its LCD screen is composed of several dot matrix character bits, such as 5×7 or 5×11. Each dot matrix character bit can display a character. There is a dot space between each bit, and there is a space between each line, which plays the role of character space and line space. Because of this, it cannot display graphics well. Select LCD 1602 as LCD screen.

2.3   Hardware Protection and Anti Interference Design

All input signals of single chip microcomputer in the system adopt photoelectric isolation measures. For example, til300, a high-precision linear optical coupler, is used in the voltage and current acquisition channel to isolate the digital signal from the analog signal. This measure effectively resists the electrical connection between the digital signal and the analog signal, and only keeps them in signal connection. At the same time, it also prevents the external clutter information from entering the data acquisition channel, which affects the normal parameter acquisition process. In addition, RC low-pass filter circuit is added to the data acquisition channel, which can also prevent the noise signal from transmitting in the channel and reduce the electromagnetic interference.

The size of the system circuit board should be appropriate, and the placement of high-frequency components should be given priority. The components should be arranged according to the direction of signal transmission as far as possible. For the components with large potential difference between each other, they should be placed relatively far away. For the power layout, they should be close to the outside of the circuit board as far as possible, and heat dissipation devices should be added when necessary. The double-layer board should avoid long straight parallel lines as far as possible, and the upper and lower layers should be arranged in different directions. When the power supply display is arranged, the width of the power supply line should be increased as far as possible if the situation permits.

3   Database Design of Social Insurance Data Dynamic Updating System

After fully understanding the human resource social insurance business, the business model of the system is further abstracted, and the complex application is decomposed into sub processes which are easy to understand, so that the system analyst can better communicate with users and ensure the correctness of the system logic. The logical

design of the system database adopts the dimension modeling technology, which is driven by the analysis topic and establishes the data model according to the analysis topic. The dimension model is composed of a fact table in the center and multiple dimension tables distributed around it. In each dimension table, a primary key must be set in advance. The primary key of the fact table is composed of multiple fields, which correspond to the foreign keys of the dimension table respectively, so as to conduct data Association [9]. Usually, this structure is called star model, and the logical design of the database is usually the design of the star model of sub topics, so as to determine the basic storage structure of each data table in the database. The personnel information database table is as follows (Tables 2 and 3).

Table 2. Personnel information database.

Field name	Interpretation	Data type
E_ID	Personnel number	Char(7)
E_Name	Name of personnel	ncarchar(50)
E_Sex	Gender	Char(1)
E_Tel	Contact number	ncarchar(50)
PDMECD	Medical certificate number	varchar(10)
PDIDCD	ID number	varchar(18)
PDEDST	Endowment insurance	varchar(1)
PDIDST	Work related injury insurance	varchar(1)
PDMEST	Medical insurance	varchar(1)
PDLBST	Unemployment insurance	varchar(1)
PDBRST	Maternity insurance	varchar(1)
PDSTCD	Status mark	varchar(4)

In addition, the storage of insurance information database table is as follows.

Table 3. Social insurance information.

Field name	Interpretation	Data type
E_D	Personnel number	nvarchar(7)
IN_Data	Insurance time	Char(11)
E_Name	Full name	nvarchar(50)
IN_Samt	Insurance amount	nvarchar(100)
IN_Level	Insurance level	nvarchar(10)

4   Software Function Design of Social Insurance Data Dynamic Update System

With the support of hardware equipment and database, the software logic program is designed for the purpose of realizing the dynamic update of social insurance data.

4.1   Set up Wireless Communication Transmission Protocol

In order to ensure the reliability and accuracy of wireless transmission, the transmission protocol of wireless communication network is designed from frame format, error detection and correction and link protocol. Due to the limitation of physical layer, in the actual transmission, the first and last boundary method with character filling is combined with other methods to improve the reliability. When a frame arrives, first check whether the frame header meets the requirements, and find out the length of frame one from the frame header, and determine the end of the frame by byte counting. Only when the frame qualifier appears at the end of the frame and the check sum is correct, the frame will be accepted as a valid frame. However, there is no frame format requirement for the frame to be sent, only the location and length of the address field are required. The frame format transmitted in the wireless communication network is shown in Fig. 4.
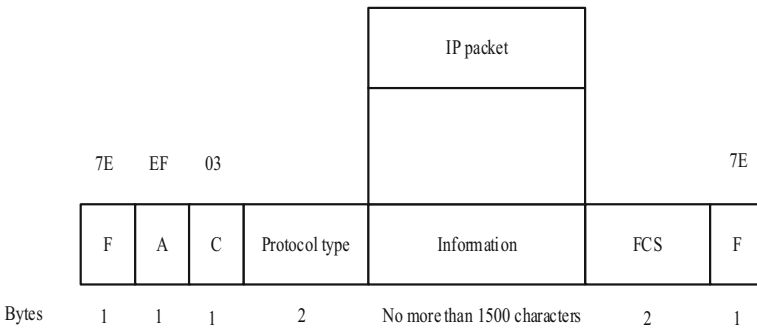
Fig. 4.   Schematic diagram of wireless communication transmission frame format.

From the aspect of error control, the commonly used method is to provide feedback information about the receiving situation to the data sender. Typically, the protocol requires the receiver to send back a special control frame as an affirmative or negative acknowledgement of the input. The timer and serial number are used to ensure that each frame can be transmitted to the destination correctly.

4.2   Dynamic Mining of Human Resource Social Insurance Information Data

The social insurance information data of human resources that need to be mined include the information of the insured, the payment information of the insured, the unit information of the insured and the payment information of the unit of the insured. The information data of human resource social insurance mainly comes from the information registration,

information change, information cancellation and other functional modules of human resource social insurance management system. After the station microcontroller is powered on and reset, it firstly completes the initialization of each port and module. In order to reduce the power consumption of the acquisition station, after the initialization, the wireless data transmission module is set to the receiving mode, and the microcontroller works in the power saving mode. When the data acquisition station needs to work, the master station sends a wake-up signal with the address of the substation to the substation. The substation compares the address code information contained in the received signal with the address code information of its own station. If it is the address of its own station, the microcontroller will be awakened to enter the normal working mode. Start the A/D conversion module to read and store the measured data. When the acquisition station receives the command from the master station to collect data, it takes the data out of the data buffer of the MCU, switches the receiving state of the module to the transmitting state, and transmits the data. After a period of delay, after the master station receives the data, the wireless module then restores the receiving state.

4.3 Wireless Transmission of Social Security Data

In the wireless communication network, the transmission process of collecting social security data can be divided into two parts: sending and receiving. In the transmission process, nRF2401 is set as the transmitting mode. The collected social security data is packaged into packets, and CRC is calculated and loaded. NRF2401 automatically loads the data head and transmits at the rate of 250 kb/s to judge whether the transmission is completed. This process is automatically completed in nRF2401. It only needs to delay 200 μm in the single chip microcomputer, and finally adjust the transmitter device to the standby mode [10–12]. Configure PTR4000 as receiving mode at the receiving end MCU first. When the data received by the PTR4000 is consistent with the local address and the CRC check is correct, DR1 outputs a high level, and the microcontroller reads the received data through DATA and CLK1. After the data is read, DR1 becomes low. In the process of social security data transmission, a transmission guarantee mechanism is set up, which adopts coloring and bleaching technology. Firstly, nodes v_i and v_j trace back to the source and destination nodes respectively along the original transmission channel, $S(v_0, v_i)$ is red, $S(v_j, v_{DAPi})$ is blue, and DAPs are defined as permanent blue and colorless. Since then, agent AG_Red starts from v_i, Ag_Blu starts from v_j and explores to establish the optimal additional channel according to their respective composite metrics. To ensure that the additional channel transmission has low delay and jitter, that is, optimality. In the process of agent travel, the composite metric consists of two parameters: the residual bandwidth rate and the receiving node queue free rate. The calculation formula of the queue occupancy rate of the receiving node is as follows:

$$\mu_k(t) = 1 - \left[\frac{q_k - \int_{t_{0-}}^{t_0} v_k^l(t)dt + \int_{t_0}^{t_{0+}} v_k^r(t)dt}{B_k} + \xi \right] \tag{1}$$

where μ_k is the receiving node, v_k is the spare rate of the queue, q_k is the current queue length of node v_k, $v_k^l(t)$ and $v_k^r(t)$ are the packet sending rate and receiving rate of node v_k communication buffer at time t, ξ is the queue buffer reserved for flow control message

transmission, and B_k is the communication buffer size of node v_k. When t_i, the agent is in node v_h. By calculating the participation bandwidth rate of the wireless link between v_h and adjacent nodes and the queue spare rate of v_h neighbor nodes, the neighbor node with the largest composite metric is selected as the next hop travel destination node. The local search target formula is as follows:

$$\begin{cases} \max\{M(e_{hk})|M(e_{hk}) = \alpha\mu_k(t) + \beta\gamma(e_{hk})\} \\ v_k.color = white \\ e_{hk} \in E(D(v_h, v_k) \le \min(\lambda_h, \lambda_k), v_h, v_k \in V) \end{cases} \tag{2}$$

When the agent determines the optimization target based on local search and travels to the next hop node v_k, the node is dyed with the same color as the agent. Then the node goes into the bleaching process, that is, as time goes on, v_k. The TTL decreases and the nodes are bleached until they are uncolored. The state transition Markov process of nodes in the execution process of dynamic additional channel guarantee mechanism.

4.4   Real Time Social Security Data Processing and Integration

The newly generated social security data may duplicate the data in the database, that is, there are different interpretations of the same object under different data sources. Therefore, we should merge the data according to certain principles or methods, and then load the data. At present, the method of clearing duplicate records is mainly based on the standard method of detecting two identical records. In general, the idea of "sorting & merging" is used in the detection of similar duplicate records, among which the near matching ordered algorithm is widely used. Before matching, do some preprocessing, such as attribute cleaning and sorting, take the method of reducing the matching field when matching records, and match according to the weight and priority value of the field. Then, some post-processing is performed on the detected duplicate records, such as conflict resolution. After the social security data is processed, it is integrated. In the process of data integration, the name, code and number of data items are unified according to the designed data conversion rules, so as to discard the duplicate data and form a unified data storage platform.

4.5   Realize the Dynamic Update of Social Insurance Data

The dynamic update of social insurance data adopts three steps: insert, delete and replace. It detects the new data in the social insurance management platform, inserts the new data into a more appropriate location without destroying the order of data storage ID, and realizes the addition of new social insurance data. The whole process of deleting data in batch is very similar to that of inserting data in batch. Its purpose is to delete user information, expired information and invalid information. When the social security information is to be deleted in batch, the system collects the social security information into the cache just like inserting and deleting the social security information in batch. It only adds a deletion mark on the social security information, then performs string conversion, and sorts the subject and object at the same time. When the system merges the social security information in the cache into the main database, the system deletes

the social security information in the main database that is the same as those marked for deletion from the main database. In fact, when the system wants to delete some social security information in the main database, it does not need to move the subsequent social security information to fill in the space left by the deleted social security information. In the same way, the change of social insurance data can be realized. Through the above three steps, the dynamic update of human resource social insurance data can be realized.

5   System Test

In order to test the update performance and operation performance of the human resources social insurance data dynamic update system based on wireless communication, the system test experiment is designed, and through the comparison with the traditional update system, the operation advantages of the design update system are reflected.

5.1   System Test Environment

The system is developed on TripleBit, an open source system and highly scalable RDF storage system, and the dynamic library used is boost. The development language of the system is C++, compiled with g++, and the development environment is Linux. The main and high-end servers and storage devices are used in the hardware of the system. IBM p595 minicomputer is used. The server is divided into three logical partitions. The data warehouse partition has 12 cpu and 48 gb memory. The storage adopts IBM 4800 storage device and RAID 5 array mode. The DS4800 storage device is connected with the host through 4 GB optical fiber card. The partition operating system uses AIX 5.3.0.0 to install the database DB2 v9.1.0.4, and the Essbase multidimensional database is installed in the partition.

5.2   System Test Sample Data Set

The system test mainly uses multiple data sets, which are the social security information management database of multiple municipal cities. The selected database uses a more standardized method to generate data sets of different sizes. Select a human resource social insurance data as the initial data, and separate other data sets for processing, which will act as the new data of the system. The configuration of the system test data set is shown in Fig. 5.

5.3   Describe the System Test Process

The designed dynamic update system of human resources social insurance data is imported into the experimental environment, and the coding program is converted into the program code that can be run directly by the computer, and the running interface of the system is obtained. In order to form the experimental comparison, in addition to the designed update system, the traditional social security data update system and the social security data update system based on data mining technology are also designed. The two comparison systems are coded and driven in the same way, and the corresponding update system operation interface is obtained.
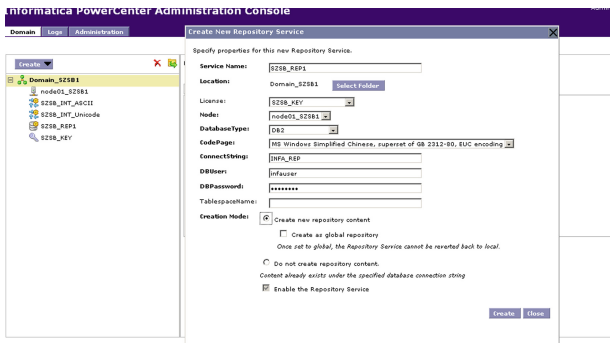
Fig. 5. Sample data set configuration diagram.

5.4 System Test Results

The test index set in the experiment is divided into two parts: update function and running performance. The update function is mainly whether the social security data can be completely updated, while the running performance is mainly to test the data update speed of different systems. The system test results are obtained under the test mode of taking the average value from multiple experiments, as shown in Table 4.

Table 4. System test results.

Experiment number	Initial data volume/GB	New social security data/GB	Traditional social insurance data update system		Social insurance data updating system based on Data Mining		The design of human resources social insurance data dynamic update system based on wireless communication	
			Amount of social security data after updating/GB	Single update interval/s	Amount of social security data after updating/GB	Single update interval/s	Amount of social security data after updating/GB	Single update interval/s
1	6.74	0.77	7.41	1.1	7.45	0.5	7.50	0.3
2	9.28	0.84	10.00	1.3	10.06	0.7	10.10	0.2
3	7.34	1.42	8.62	1.5	8.69	0.6	8.74	0.4
4	8.25	0.93	9.10	1.2	9.13	0.8	9.17	0.1
5	6.63	0.51	7.02	0.9	7.06	0.4	7.12	0.3
6	7.44	0.46	7.75	1.1	7.81	0.7	7.88	0.1
7	8.21	0.68	8.81	1.4	8.84	0.5	8.86	0.2
8	7.37	0.97	8.27	1.3	8.30	0.6	8.32	0.2

It can be seen from Table 4 that after the three systems complete the update function, the integrity of social security data is 99.2%, 99.5% and 99.8% respectively. From the perspective of update speed, the average update time of the three update systems is 1.225s, 0.600s and 0.225s. It can be seen that the design of human resources social

insurance data dynamic update system based on wireless communication has higher update efficiency.

The time delay of different methods was tested and the comparative results were shown in Table 5.

Table 5. Comparison of delay (Unit: ms)

Number of data transmission bits/Kbit	Traditional social insurance data update system	Social insurance data updating system based on Data Mining	The design of human resources social insurance data dynamic update system based on wireless communication
500	0.776	0.945	0.156
1000	1.102	1.265	0.267
1500	2.265	2.135	0.298
2000	3.882	3.376	0.329

It can be seen from Table 1 that the dynamic updating of human resources and social insurance data in this system has a relatively low delay, which effectively improves the real-time performance of data updating.

6   Concluding Remarks

This study uses data visualization technology to visualize historical social security data of different regions and industries on the map. Analyze the development trend of social security in different regions and different industries. The social security data is analyzed by using the drill-down and pull-down display methods of data visualization. Combined with wireless communication technology, real-time social security information collection and transmission. The system test results show that the update speed and update data quality of the designed system are optimized. It has higher application value in the actual human resources social insurance management system.

References

1. Jobski, K., Drks, M., Bantel, C., et al.: Regional differences in opioid prescribing in germany – results of an analysis of health insurance data of 57 million adult people. J. Pain Res. **13**, 2483–2492 (2020)
2. Tetzlaff, F, Epping, J, Sperlich, S., et al.: Widening income inequalities in life expectancy? Analysing time trends based on German health insurance data. J. Epidemiol. Commun. Health, jech-2019–212966 (2020)

3. Yang, R., Su, D., Liu, C., et al.: Medical insurance data updating method based on network coding technology. Inf. Technol. **44**(04), 107–110 (2020)
4. Gradeci, K., Labonnote, N., Sivertsen, E., et al.: The use of insurance data in the analysis of surface water flood events – a systematic review. J. Hydrol. **568**, 194–206 (2019)
5. Elmer, D., Endrei, D., Pónusz, R., et al.: PSY17 epidemiological disease burden of atopic dermatitis based on routinely collected health insurance claims data. Value Health **23**, S373–S374 (2020)
6. Liu, S., Bai, W., Zeng, N., et al.: A fast fractal based compression for MRI images. IEEE Access **7**, 62412–62420 (2019)
7. Liu, S., Li, Z., Zhang, Y., et al.: Introduction of key problems in long-distance learning and training. Mob. Netw. Appl. **24**(1), 1–4 (2019)
8. Pu, W., Chen, N., Zhong, Q.: SDCUP: software-defined-control based erasure-coded collaborative data update mechanism. IEEE Access **8**, 180646–180660 (2020)
9. Wang, L.S., Leung, Y.Y., Valladares, O., et al.: Genome center for Alzheimer's disease whole exome (WES) and whole genome sequencing (WGS) data update. Alzheimer's Dement **15**(7), P623–P624 (2019)
10. Himmelstein, D.U., Woolhandler, S., Fauke, C.: U.S. health care in the trump era: a data update. Int. J. Health Serv. **49**(3), 402–411 (2019)
11. Akay, M.A., Murat, G., Hayrünisa, K.E., et al.: Burns in children: data update with 145 patients. Kocaeli Med. J. **9**(3), 86–92 (2020)
12. Lai, C.C., Liu, C.M.: A mobility-aware approach for distributed data update on unstructured mobile P2P networks. J. Parallel Distrib. Comput. **123**, 168–179 (2019)

Design of Wireless Audio Real Time Transmission Model Based on Body Area Network Technology

Qing-li Niu[1(✉)] and Hong Xu[2]

[1] College of Information Engineering, Zhengzhou University of Science and Technology, Zhengzhou 450064, China
`jsj32989317@163.com`
[2] School of Software, Zhengzhou University, Zhengzhou 314500, China

Abstract. In order to achieve real-time compression of audio signals and establish a more stable wireless network transmission mode, this study designed a wireless audio real-time transmission model based on body area network technology. First, with the support of wireless body area network technology, a complete information aggregation processing system is built, and then the body area network environment is built by configuring the downloaded audio resources. Based on this, the digital signal processor is selected, and a complete data structure of multiple audio frames is established, so as to realize the real-time transmission of wireless audio. Experimental results show that compared with the traditional multi-channel audio transmission technology, the proposed model has higher effectiveness and stronger signal matching ability, which proves that it can compress audio signals in real time and establish a more stable wireless network transmission environment at the same time.

Keywords: Body area network technology · Wireless audio · Real-time transmission · Convergence processing layer · Resource download · Resource configuration · Signal processor · Protocol message

1 Introduction

Audio is a technical term used to describe sound-related devices in the audio range and their functions. All sounds that human beings can hear are called audio, which may include noise, etc. After the sound is recorded, whether it is voice, singing, musical instruments can be processed by digital music software, or it can be made into a CD. At this time, all the sounds have not changed, because CD is a type of audio file. Audio is just the sound stored in the computer. If there is a computer and the corresponding audio card, which is often called the sound card, we can record all the sounds, and the acoustic characteristics of the sound, such as the sound level, can be stored in the form of computer hard disk files. Audio signals are the carrier of regular frequency and amplitude changes of sound waves with voice, music and sound effects. According to

the characteristics of sound waves, audio information can be classified into regular audio and irregular sound. Regular audio can be divided into voice, music and sound effects. Regular audio is a continuous changing analog signal, which can be represented by a continuous curve, called sound wave. The three elements of sound are tone, intensity and timbre.

Body area network is a basic technology that can monitor and record human health signals for a long time. Its early application is mainly used to continuously monitor and record the health parameters of patients with chronic diseases and provide some way of automatic therapy control. In the future, body area network can also be widely used in consumer electronics, entertainment, sports, environmental intelligence, animal husbandry, ubiquitous computing, military or security fields. Body area network is a special wireless communication system, which takes the devices around the human body, such as watches, sensors and mobile phones, as well as the internal human body (implantable devices) [1].

Therefore, in order to realize the real-time compression of audio signals and establish a more stable wireless network transmission mode, this study designed a wireless audio real-time transmission model based on volume domain network technology. The design idea of this model is as follows:

(1) Build a complete information aggregation processing system by using the wireless body area network system architecture, and then complete the construction of the body area network environment by configuring the downloaded audio resources.
(2) The digital signal processor is selected and processed. According to the data structure of multiple audio frames, the smooth application of wireless audio real-time transmission model based on volume domain network technology is realized.

2 Building Body Area Network Environment

The body area network application environment is composed of system architecture, information aggregation processing layer, audio resource download and configuration system. The specific construction method is as follows.

2.1 Wireless Body Area Network System Architecture

Wireless body area network system architecture is divided into three levels, namely information collection layer, information aggregation processing layer and cloud information layer.

The audio information acquisition layer contains many sensor nodes, which have the function of collecting audio information, mainly for the audio wavelength, amplitude, frequency, etc. The collected information is then transmitted to the upper control node. In this study, a simple star topology is used as the system structure of the information acquisition layer in the volume domain network.

The audio information aggregation processing layer is similar to a personal server, and its requirements on energy consumption and size are different from the acquisition sensor of the audio information acquisition layer. In the star topology, the monitored

audio information is directly received from the front-end information acquisition nodes. In general, personal PDA devices with display and processing functions are selected as the aggregation nodes in the information aggregation processing layer. On the one hand, the wireless audio information collected by the front end can be stored and analyzed. On the other hand, the information can also be sent to the cloud data information layer through the wireless network [2].

The schematic diagram of the wireless body area network system architecture is shown in Fig. 1. The audio information cloud mainly includes a server that supports remote transmission, which is mainly used to receive the monitoring audio feature values forwarded by the sink node, process and analyze the received monitoring data, and store it in the background database or upload it to the network.
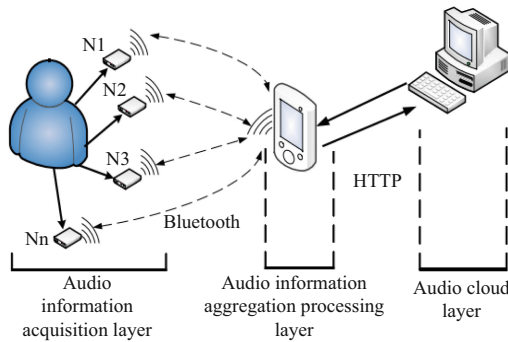
Fig. 1. Schematic diagram of wireless body area network system architecture

2.2 Information Aggregation Processing Layer

The audio information aggregation processing layer, as its name implies, refers to the aggregation node. In addition to receiving the monitoring information transmitted by the information collection layer, it also has the functions of UI display and audio data processing. For this reason, the body area network host is selected as the information aggregation processing layer. The convergence center is mainly considering that the positioning and remote communication functions of the body area network host can meet the characteristics of wireless audio instantaneity in emergency situations, and the body area network host has occupied with its complete openness and unlimited user equality. Most of the Internet application network [3]. For the communication connection with the Bluetooth module of the information collection layer, the body area network host has already supported Bluetooth development in the 2.0 version of the software development kit. In the communication with Bluetooth, eight categories are mainly used, and the location is under the Bluetooth package. The eight categories are shown in Table 1.

Table 1. Application capability description of information aggregation processing layer

Class name	Explain
BluetoothAdapter	Bluetooth adapter for local audio
BluetoothClass	Bluetooth
BluetoothClass.Device	Bluetooth transmission devices
BluetoothClass. Device.Major	Bluetooth transmission device management
BluetoothClass. Service	On the class of Bluetooth transmission service
Bluetooth Device	Bluetooth device for wireless remote audio
BluetoothServerSocket	Wireless audio monitoring Bluetooth connection class
BluetoothSocket	Bluetooth connection class, mainly the client

When communicating with Bluetooth module, we mainly use Bluetooth adapter class, Bluetooth device class and Bluetooth socket class. Bluetooth adapter class is used to obtain the local wireless audio adapter structure; Bluetooth device class is used to represent a remote Bluetooth device, that is, the Bluetooth module of the Bluetooth information collection layer; Bluetooth socket class is used to represent the output stream of the Bluetooth connection.

2.3   Download and Configuration of Audio Resources

In the wireless body area network system application development process, in addition to mastering the development technology of audio resource download, it is also necessary to understand the structure of audio configuration relationships.

The audio resource application program is generally constructed by four functional modules: active object, intention, service and content provider. But not all applications must use the above four functional modules, and can also be composed of one or more of the above. The only thing in common is that no matter which function module is used, it must be registered in advance in the configuration file of the program.

Activity
Activity is the most basic application component of the body area network system. Although each application is not necessarily composed of the above four components, as long as there is an interface display, the application must contain the activity application component. In a popular sense, each activity corresponds to an independent screen in the body area network system. It is an interface used to realize human-computer interaction, and is used to store display controls to realize various functions. Each activity can be regarded as a class. You can see that most applications are composed of many activities, in order to distinguish which one Activity is the current interface in the application program. The body area network system especially sets seven operation states for activity: create, start, pause, pause resume, stop, stop resume and destroy. These seven different operation states constitute the life cycle of each activity [4].

Intent

An application is often composed of multiple activities. In order to realize the jump between different activities, the body area network system uses the Intent component to complete the communication of different activities, such as starting a new activity, and can pass data to this new activity. The Activity o Intent can also be used to call applications provided by the body area network system itself, such as calls to dial-up programs and sending text messages. The data transferred in the Intent includes 7 types, namely operation, data, data type, operation category, additional information, components and signs.

Service

In the application, there are some programs running in the background, such as sending SMS or playing a music player in the background. These operation components without interface display are actually a service, and the main functions provide some necessary support for activity Service also has its own life cycle. Compared with activity, the basic life cycle of service only needs to use two operation methods: staxt sexvioe() and stop sexvioe().

Content Provider

The body area network system provides five storage methods for data, the most commonly used is to save in files and SQLite data, but when some applications developed need to call the data of existing applications, such as a program to obtain the phone book Since different applications are working in different processes, it is generally difficult to transfer data from one virtual machine to another virtual machine [5]. Therefore, in order to meet the data sharing between different programs, the body area network system specifically defines the content provider, a special data storage class, and provides a set of standard interfaces for obtaining and operating data. Content Provider also supports data sharing. Adding, querying, updating, and deleting operations facilitate the mutual access of data between different programs.

3   Wireless Audio Real Time Transmission Model

Based on body area network environment, according to the selection of digital signal processor, RTP/RTCP protocol message format definition, audio frame data structure connection processing flow, the design of wireless audio real-time transmission model based on body area network technology is completed.

3.1   Selection of Digital Signal Processor

In the design of wireless audio real-time transmission model based on body area network technology, a popular multimedia audio and video processing chip TMS320DM642 is selected for the digital signal microprocessor. It is a 32-bit fixed-point DSP chip of TMS320C6000 series which is suitable for processing digital audio and video stream launched by Texas TI company in 2003, instead of tms320c5000 series which is commonly used in embedded device system The design needs to process multi-channel audio

data, and the tms320c5000 series digital signal processor chip has at most three McBSP ports to receive digital audio data. Moreover, the McBSP port has other uses, which cannot be used to collect digital audio data completely [6]. TMS320DM642 chip adds a multi-channel audio mcasp port, which can accommodate up to eight audio data streams, just to meet the requirements of multi-channel audio processing, four digital audio input streams and four audio data output streams.

TMS320DM642 also has a 10/1OOM Ethernet interface that meets the IEEE802.3 standard, eliminating the need for an Ethernet controller, thereby simplifying circuit design and reducing circuit costs. Therefore, the core processor chip designed in this research selects TMS320DM642 as the audio processor. Facing the rapid development of multimedia networks, the requirements of multi-channel data transmission and the expansion of on-chip memory capacity, in 2003, Texas Instruments (TI) launched a new generation of digital multimedia processor TMS320DM642 to meet the requirements of the market., It is a 32-bit fixed-point DSP chip belonging to TI's C6000 series. TMS320DM642 retains the original core structure of C64x, using advanced Harvard structure memory [7]. Its main features are as follows:

1. The working frequency of the DSP chip is input by the external crystal oscillator, and the internal PLL frequency doubling setting can achieve the clock frequency of 720 MHz, the number of executable instructions per second is 5760mips, and 8 32-bit instructions are executed per cycle, which has high-speed digital processing speed.
2. There are six ALUs in DSP, and each Alu supports 32-bit arithmetic operation per clock cycle.
3. A multi-channel audio serial interface (mcasp), 8 serial data pins, can accept up to 8 channels of audio data input or output, complete support I2S, S/PDIF and aes-3 digital audio stream audio transmitter.
4. An Ethernet controller interface (EMAC) conforming to IEEE 802.3 standard is integrated, including MII and MDIO.
5. A 64 bit external memory interface (EMIF), which can be seamlessly connected with external nonvolatile memory such as flash, SDRAM and EPROM. The maximum addressable space of external memory is 1024MB.

3.2   Message Format of RTP/RTCP Protocol

RTCP sends control packets periodically. RTP needs RTCP to provide guarantee for its quality of service. RTCP and RTP are all transmitted by UDP. RTCP control packets contain short control information and are encapsulated in a UDP packet. RTCP includes five different types of data packets (Table 2).

RTP/RTCP protocol message format definition includes the following processing steps:

1. Version: indicates which RTP version is used. If version 2 is used, the value is set to 2, occupying the first two bits of the first byte.
2. Fill: the end of RTP packet can contain one or more fill bits, which are not included in the load and occupy the third bit.

Table 2. Five packet types of RTCP

Types of	Data pack	Purpose
200	SR	Audio sender report
201	RR	Audio receiver report
202	SDES	Audio origin description
203	BYE	End audio transmission
204	APP	Specific audio transmission applications

3. Extension: if the fixed header of RTP package is followed by an extended header, this bit will be assigned to occupy the fourth bit.
4. CSR counter: the number of CSR after RTP packet header, occupying the last 4 bits of the first byte.
5. Mark: its content is specified by a specific protocol and occupies the first bit of the second byte.
6. Load type: it is an important data bit in RTP packet. It identifies the data type of RTP packet and defines the load format, occupying the last seven bits of the second byte.
7. Sequence number: It is an important data bit in the RTP packet. The sequence number increases by 1 after sending an RTP packet. The receiver can use this to detect the number of lost packets and restore the data packets out of order, occupying two bytes.
8. Timestamp: It is also an important data in the RTP packet, indicating the sampling time of the first byte of the RTP packet data. The value of the timestamp must be continuously accumulated. The timestamp matches the serial number to correctly play the audio to achieve jitter removal and synchronization, occupying four bytes.
9. Synchronization source identifier: used to distinguish RTP data packets from different sources. A single RTP session can only have a unique SSRC value, which is also four bytes.
10. Contribution source list: used to indicate the RTP sources contributed in the data packet. The source SSRC identifier in the data packet is identified, and the receiving end can correctly identify the participant, including 0–15 items, each with four bytes [8].

3.3 Audio Frame Data Structure

IP audio means that digital audio signal is transmitted by IP data structure in Ethernet network, which provides a short delay, low cost and high precision solution for digital audio real-time transmission system. The goal of IP audio transmission is to deliver high-precision clock signals and professional uncompressed digital audio signals over the Ethernet network, and to carry out complex routing at the network layer. IP based network audio transmission process uses transport layer UDP protocol and RTP protocol, of course, also need to use RTCP protocol.

Ethernet (EtherNet) is a baseband local area network specification jointly developed by Xerox, Intel and DEC. It is the most widely used communication protocol standard for local area networks. In recent years, most of the network audio real-time transmission solutions that have emerged are based on Ethernet technology standards, which are currently the most popular embedded device technology used, and have replaced token ring, FDDI and ARCNET LAN standards to a large extent. The Ethernet-based TCP/IP protocol occupies a dominant position in the field of audio transmission.

Ethernet is the basis for determining the efficiency and performance of the audio network. The analog audio signal transmitted on the standard Ethernet cannot be easily converted into a data format. In the audio network transmission, the audio signal requires a relatively strong timeliness, and the data sent The delay of the packet will cause incoherence and loss of the audio signal. Ethernet is an asynchronous technology and does not have the concept of real-time. Transmission management is uncertain, which means that Ethernet cannot guarantee the timely delivery of data packets. Therefore, in order to transmit audio data in real time and stably, the network must have a certain deterministic time-sensitive transmission technology [9]. For example, EtherSound, CobraNet and Dante technologies can provide such technologies.

The audio frame data structure definition form is shown in Fig. 2.
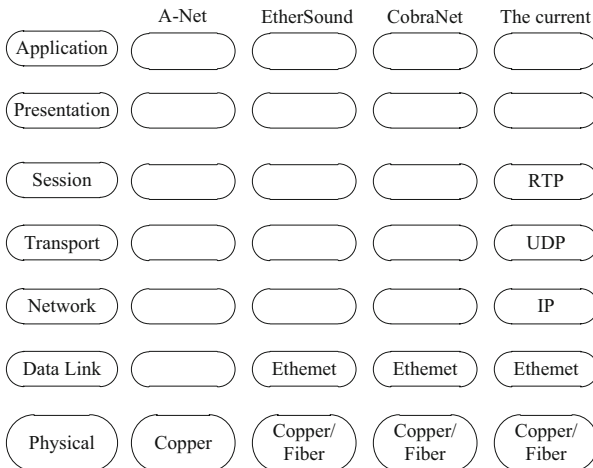
Fig. 2. Definition form of audio frame data structure

A complete Ethernet frame usually consists of synchronization bit PR, split bit SD, destination MAC address Da, source MAC address SA, type field type, MAC data field and frame check sequence. Among them, synchronization bit and split bit are also called preamble and frame start character. The content of the first seven bytes is 0xaaaaaa, and the content of the eighth byte is oxab. They play the role of synchronization at the receiving end, marking the formal start of transmission of an Ethernet frame. Followed by the six byte destination address DA and six byte transmission source address Sa of Ethernet frame transmission, the two byte type field indicates the protocol type used for network layer transmission [10, 11]. The next 461500 bytes are the data packets

transmitted, which are the key parts of frame transmission. The last four bytes are the frame validation fields of Ethernet frame, which verify whether the data bits transmitted before are correct or not.

CobraNet transmits PCM data of uncompressed digital audio stream without data encoding and decoding, so it is used for real-time audio transmission in radio and television industry. CobraNet data packet is located in data link layer, namely MAC layer, including clock data packet, audio data packet and reservation data packet. The structured of audio data packet is shown in Fig. 3.

Target audio MAC	Audio timestamp	Channel 1 format	
The audio source MAC	Bundle audio message number	Channel 2 format	......
Transport protocol type		Channel 3 format	CRC
CobraNET audio message		Channel 4 format	
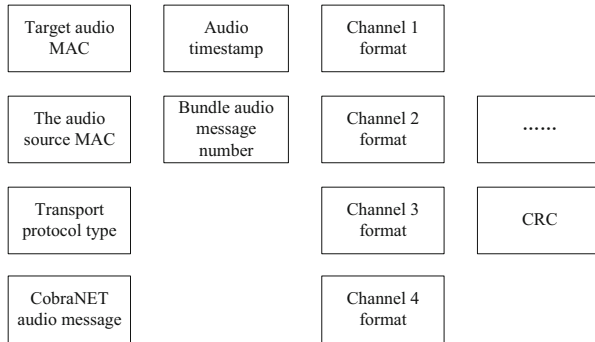
Fig. 3. Audio frame packet definition format

4 Model Application Ability Detection

In order to verify the practical application value of the wireless audio real-time transmission model based on volume area network technology designed above, the following comparative experiments are designed.

The experimental group and the control group were connected to the audio transmitter respectively. Among them, the experimental group was equipped with wireless audio real-time transmission model based on body area network technology, while the control group was equipped with multi-channel audio transmission technology. In the same experimental environment, the specific changes of the experimental indicators were analyzed.

It is known that the audio data transmission time can reflect the effectiveness of wireless audio information transmission. Generally, the shorter the consumption time, the higher the effectiveness of wireless audio information transmission, and vice versa. The specific experimental conditions are shown in Table 3.

Analysis of Table 3 shows that as the total amount of audio data increases, the transmission time of the experimental group maintains a trend of first increasing and then stable, and the global maximum can only reach 2.40 s. The transmission time of the control group has always maintained a rising trend, and the global maximum value reached 4.52 s, which is an increase of 2.12 s compared with the maximum value of the experimental group.

Table 3. Comparison of audio data transmission duration

Total audio data/(Gb)	Audio data transmission time/(s)	
	Experience group	Control group
1.0	2.31	4.18
2.0	2.33	4.21
3.0	2.35	4.24
4.0	2.36	4.27
5.0	2.38	4.29
6.0	2.39	4.33
7.0	2.40	4.35
8.0	2.40	4.38
9.0	2.40	4.52

The PSR index can describe the real-time matching ability of wireless audio signals. Generally, the larger the PSR index value, the stronger the real-time matching ability of wireless audio signals, and vice versa. Table 4 records the actual changes in the PSR indicators of the experimental group and the control group.

Table 4. Comparison of PSR indexes

Experiment time/(min)	PSR index/(%)	
	Experience group	Control group
5	57.38	41.32
10	76.19	42.08
15	58.24	42.79
20	78.87	43.16
25	56.73	44.87
30	79.62	45.20
35	56.25	46.13
40	80.36	46.65
45	56.41	47.41
50	79.50	48.54

According to the analysis of Table 4, with the extension of the experimental time, the PSR index of the experimental group kept the trend of rising and falling alternately, and the global maximum reached 80.36%. The global maximum of PSR in the control group

was only 48.54%, which decreased by 31.82% compared with that in the experimental group.

By comprehensive analysis of the above results, it can be seen that the transmission consumption time of audio data is effectively controlled after the application of the model in this paper, and the value of PSR index also increases significantly, which proves that the model in this paper can realize real-time compression of audio signals while establishing a stable wireless network transmission environment. The reason for the above results is that the model in this paper has an independent information aggregation processing layer system, which can establish a complete audio frame data structure, thus improving the PSR index and reducing the time consumption of transmission process.

5   Conclusion

Compared with the multiplexed audio transmission technology, the wireless audio real-time transmission model designed in this study has an independent information aggregation processing layer system supported by the volume domain network environment, which can realize the selection of digital signal processor and establish a complete audio frame data structure at the same time. From the practical point of view, after the application of the model in this paper, the audio data transmission time is shortened and the PSR index value increases, which proves that the model in this paper can meet the real-time compression requirements of audio signals more effectively and has a strong application and promotion significance.

In the following research, we will consider enhancing the anti-noise ability of the audio signal while using the model to efficiently transmit the audio data, so as to complete the further optimization of the model.

References

1. El-Zoghdy, S.F., El-Sayed, H.S., Faragallah, O.S.: Transmission of chaotic-based encrypted audio through OFDM. Wireless Pers. Commun. **113**(10), 1–21 (2020)
2. Kaur, A., Dutta, M.K.: High embedding capacity and robust audio watermarking for secure transmission using tamper detection. ETRI J. **40**(1), 133–145 (2018)
3. Bakeva, V., et al.: Gaussian channel transmission of images and audio files using cryptcoding. IET Commun. **13**(11), 1625–1632 (2019)
4. Telagam, N., et al.: USRP based digital audio broadcasting using OFDM in virtual and remote laboratory. Int. J. Onl. Eng. (iJOE) **15**(13), 77–85 (2019)
5. Anzinger, S., et al.: Acoustic transmission line based modelling of microscaled channels and enclosures. J. Acoust. Soc. Ame. **145**(2), 968–976 (2019)
6. Jain, A., et al.: A route selection approach for variable data transmission in wireless sensor networks. Clust. Comput. **23**(3), 1697–1709 (2020)
7. Xu, Y., et al.: An information integration and transmission model of multi-source data for product quality and safety. Inf. Syst. Front. **21**(1), 191–212 (2019)

8. Liu, S., et al.: A fast fractal based compression for MRI images. IEEE Access **7**, 62412–62420 (2019)
9. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)
10. Liu, S., et al.: Prediction of gene expression patterns with generalized linear regression model. Front. Genet. **10**, 120 (2019)
11. Zheng, Y.: Formal process virtual machine for smart contracts verification. Int. J. Perform. Eng. **14**(8), 1–9 (2018)

Time-Frequency Analysis of Vibration Signal Distribution of Rotating Machinery Based on Machine Learning and EMD Decomposition

Xiao-zheng Wan[1(✉)], Song Zhang[2], Ji-ming Zhang[1], Hui Chai[1], and Huan-yu Zhao[1]

[1] Institute of Oceanographic Instrumentation, Qilu University of Technology (Shandong Academy of Sciences), Qingdao 266061, China
wxz18661817997@163.com

[2] College of New Energy, China University of Petroleum, Qingdao 266580, China

Abstract. The conventional time-frequency analysis method of the vibration signal distribution of rotating machinery has a low correlation coefficient between the time-frequency features and the vibration signal, which leads to a low accuracy in identifying fault types of rotating machinery. For this reason, this paper proposes a time-frequency analysis method for the vibration signal distribution of rotating machinery based on machine learning and EMD decomposition. After collecting the vibration signal of the rotating machinery, it is decomposed by EMD, the collected signal is decomposed into a set of steady-state and linear data series, and the decomposed signal is filtered and denoised. Then set the time window to extract the local time domain feature components of the vibration signal, and then use the artificial neural network to identify the fault type of the rotating machinery. Experimental results show that this method improves the correlation coefficient between time-frequency characteristics and vibration signals, and improves the accuracy of identifying fault types in rotating machinery.

Keywords: Rotating machinery · EMD decomposition · Machine learning · Time-frequency characteristics

1 Introduction

Mechanical rotation will appear stationary or non-stationary characteristics of two cases, these two cases have a certain periodicity. But in the process of rotating machinery movement will appear vibration. Mechanical vibration contains the vibration of various frequency components from low to high frequency, and the vibration of rotating machinery is the same.

In order to better analyze the working efficiency and quality of machinery, it is necessary to study the vibration signal caused by machinery in work, and extract useful information from the vibration signal.

In the process of collecting mechanical vibration signal, the time domain signal should be analyzed emphatically. Time-frequency signals can reflect the time variation

characteristics and frequency variation signals simultaneously, which provides useful information for vibration signal analysis. Therefore, it is of great significance to study the time-frequency analysis method of vibration signal distribution of rotating machinery for analyzing the running state of equipment and distinguishing characteristic signals from non-characteristic signals.

Reference [1] uses the Fourier transform analysis method to analyze and describe the above signal frequency changes, but the method cannot reflect the mechanical vibration signal in real time. Reference [2] decomposes the signal into sine or cosine functions, showing the characteristics of frequency change with time, but this method cannot show the characteristics of frequency change with time.

In view of the problems existing in traditional methods, this research proposes a time-frequency analysis method for the vibration signal distribution of rotating machinery based on machine learning and EMD decomposition. EMD has good time-frequency aggregation, which can distinguish time resolution from frequency resolution, avoiding the occurrence of linear time-frequency constraints. Machine learning can convert one-dimensional time-domain signals into two-dimensional The time-frequency plane improves the information processing effect of non-stationary signals. The experimental results also prove that this study gives full play to the technical advantages of machine learning and EMD decomposition, and improves the extraction effect of time-domain characteristics of vibration signals, so that the average accuracy of the analysis results of time-frequency distribution of vibration signals of rotating machinery by this method reaches 98.1%.

2   Method Design

2.1   Collect Vibration Signals of Rotating Machinery

Design a data collector to collect vibration signals of rotating machinery. The host device, slave device, power amplifier, power dynamic real-time simulation device, and electronic transformer are selected to form a vibration signal collector. Among them, the power dynamic real-time simulation device performs real-time closed-loop simulation of rotating machinery. The selected simulation equipment is a portable RTDS real-time digital simulator, which has a power component library, and refers to the on-site conditions of the rotating machinery to set the power component parameters to correctly simulate various fault states that occur during the operation of the rotating machinery. The host device is respectively connected with the simulation device and multiple slave devices, and its communication module is configured with an optical fiber interface and an 8-bit modular interface, and transmits the simulation data of the simulation device to the slave device through the optical fiber connection. The slave device is also equipped with optical fiber interface and 8-bit modular interface, which is connected to the port of signal analysis and processing, and the analog data switch information sent by the host is input to the terminal, and the D/A conversion card is set to transmit through the internal D/A module. The number of analog quantities is 12, which converts the analog data into a standard 5V voltage analog signal, which is input to the power amplifier [3]. The optimized D/A module conversion circuit is shown in Fig. 1.
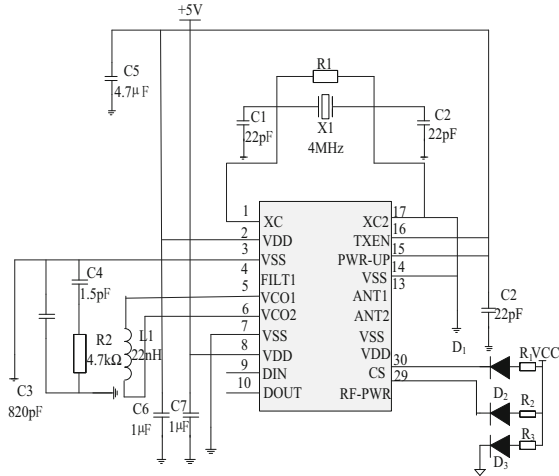
Fig. 1. D/A conversion circuit of rotating machinery vibration simulation data

The analog signal of the slave device is amplified by the power amplifier to meet the current and voltage of the signal collection, and the analog signal is received through the external synchronous clock interface to synchronously obtain the vibration signal of the rotating machine. Finally, the high-current generator is used as the current source, and the high-voltage generator generates the rated voltage. The rated current is passed through the optical fiber to obtain the current signal digital quantity, which is input to the electronic transformer calibrator to detect the output rotating machinery vibration signal and output in real time. The analog signal, including whether the waveform is synchronized, whether the polarity and transformation ratio are consistent, completes the verification of the vibration signal of the rotating machinery. For the communication module of the system host device, a communication network is built, and the line sampling signal is transmitted to the time-frequency analysis terminal. The switch information selects fiber optic Ethernet and GOOSE network, and configures the switch according to the interval. The components of the Ethernet switch are shown in Fig. 2.

The Ethernet switch adopts dual power supply, provides 1 Gbps rate optical fiber Ethernet interface, control recovery time < 5 ms, group ring any port, guarantees zero packet loss of the switch under strong electromagnetic interference, and transmits rotating machinery vibration signals [4]. At this point, the collection of vibration signals of rotating machinery is completed.

2.2 EMD Decomposes Vibration Signals of Rotating Machinery

Through EMD decomposition, the non-stationary and non-linear rotating machinery vibration signal is decomposed into a set of steady-state and linear data series, namely the eigenmode function. EMD obtains the intrinsic vibration mode through the characteristic time scale, and then decomposes the time series data from the intrinsic vibration mode, and decomposes a multi-component signal into multiple eigenmode functions. Each eigenmode function is transformed by Hilbert. The Hilbert spectrum is calculated, and
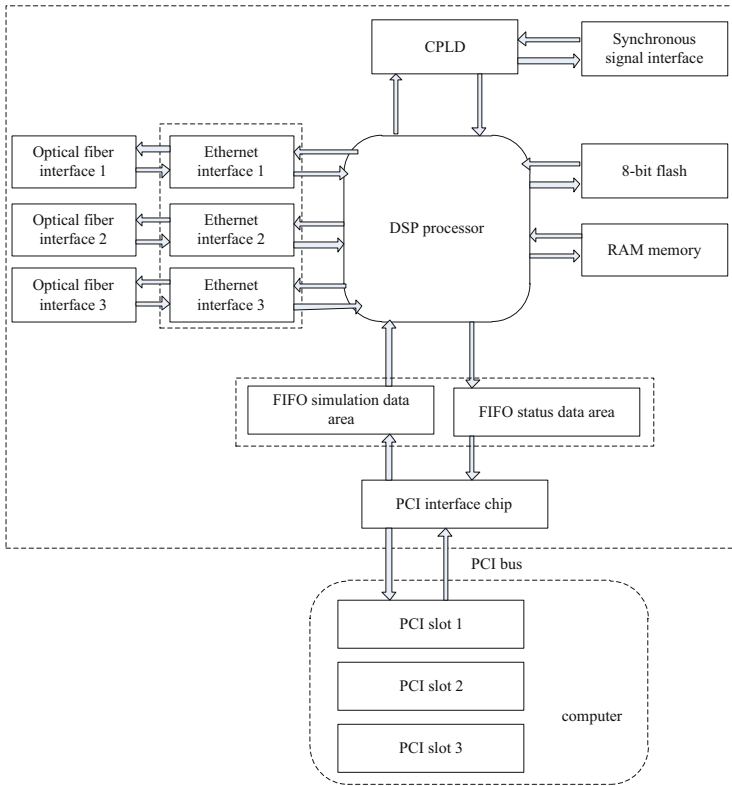
Fig. 2. Rotating machinery vibration signal communication switch

its structure is determined according to the scale characteristics of the original signal itself. In addition, the eigenmode function must meet the following two conditions: For a column of data, the number of extreme points and zero-crossing points must be equal or at most a little different; At any point, the envelope composed of local maximum points and the envelope composed of local minimum points have an average value of zero [5]. Suppose the time series data of the vibration signal of the rotating machinery is x(t), find all the maximum and minimum points of x(t), and use the cubic spline function to fit them to the upper and lower envelopes of the original data series. The average value of the upper and lower envelopes is the average envelope m_1. Subtract m_1 from the original data sequence to obtain a new data sequence h_1 with the low frequency removed, namely:

$$h_1 = x(t) - m_1 \tag{1}$$

Under normal circumstances, h_1 is not necessarily a stationary data series, so the above process needs to be repeated for it [6]. If the average envelope of h_1 is m_{11}, the data sequence h_{11} after removing the envelope representing the low-frequency components, namely:

$$h_{11} = h_1 - m_{11} \tag{2}$$

Repeat the above process to make the obtained average envelope tend to zero, so that the first eigenmode function component c_1 is obtained, which represents the highest frequency component in the signal data sequence. Subtract c_1 from $x(t)$ to obtain a new data sequence r_1 with the high frequency components removed, and perform EDM decomposition on r_1 to obtain the second eigenmode function component c_2, and repeat until the last data sequence r_n can no longer be decomposed. At this time, r_n represents the trend and mean value of the data sequence $x(t)$ [7]. The above process can be expressed as:

$$
\begin{aligned}
r_1 &= x(t) - c_1 \\
r_2 &= r_1 - c_2 \\
&\vdots \\
r_n &= r_{n-1} - c_n
\end{aligned}
\tag{3}
$$

Among them, r_{n-1} is $n-1$ the th vibration signal data sequence, and c_n is the n th eigenmode function component. From the above formula:

$$
x(t) = \sum_{i=1}^{n} c_i + r_n \quad i = (1, 2, \cdots, n)
\tag{4}
$$

The original data sequence can be expressed as the sum of the eigenmode function component and a residual term, and a steady-state linear data sequence set containing the vibration information of the rotating machine can be obtained. So far, the EMD decomposition of the vibration signal of the rotating machine is completed.

2.3   Filter Processing Vibration Signal of Rotating Machinery

The wavelet packet denoising method is used to filter the vibration signals of rotating machinery. From the energy point of view, the wavelet packet noise reduction algorithm is adopted, a scale function is introduced, and each scale signal is intercepted through the filter bank, and the wavelet system generated by the expansion of the scale function is obtained. The calculation formula of wavelet function ϕ is:

$$
\varphi = \left(a_j - 2c_j \right) D
\tag{5}
$$

Among them, D is the impulse response of the orthogonal mirror filter, a_j is the scale coefficient of the j-th layer of the scaling function, and c_j is the wavelet coefficient of the j-th layer. Select wavelet function ϕ, wavelet decomposes noise-containing rotating machinery vibration signal, assigns signal characteristics to wavelet decomposition coefficients of different scales, and compresses the signal to obtain wavelet packets. According to the order of energy, arrange the wavelet packets, select the first A wavelet packets with larger energy, and reconstruct the vibration signal of the rotating machinery [8]. To reduce the error between the reconstructed signal and the original signal, the calculation formula is:

$$
C = \min \left\{ \frac{1}{A} \sum_{\lambda=1}^{A} \left[d(\lambda) - g(\lambda) \right]^2 \right\}
\tag{6}
$$

Among them, C is the error minimization objective function, and $d(\lambda)$ and $g(\gamma)$ are the reconstructed signal and the original signal of the λ-th wavelet packet, respectively. On this basis, the wavelet packet basis which can form a set of orthogonal basis is extracted and decomposed. After decomposition, a sequence of information function is given, and then the bottom-to-top search method is adopted to find the optimal wavelet packet base that can express the signal characteristics. The specific search process is shown in Fig. 3.
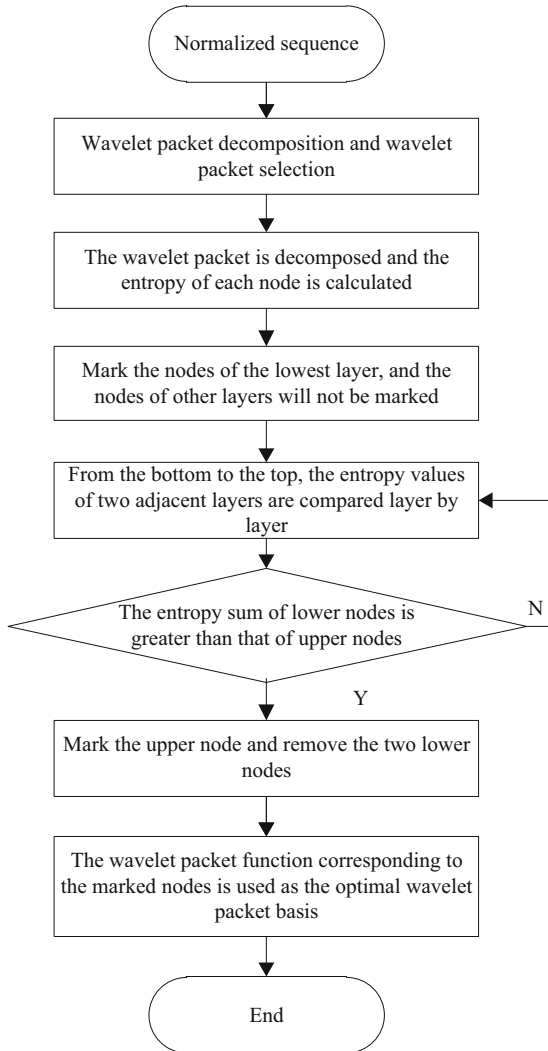
Fig. 3. The optimal wavelet packet base search process

Threshold quantizes the optimal wavelet packet basis decomposition coefficient, and obtains the oscillation signal containing the vibration information of the rotating machine after the noise is eliminated. So far, the filtering of the vibration signal of the rotating machinery is completed.

2.4 Extract Local Time Domain Feature Components of Vibration Signal

For the rotating machinery vibration signal after decomposition and denoising, a time-frequency period window is set to extract the signal video feature components. First calculate the time series data in the cyclic stationary process:

$$x(t) = x(t + T) \tag{7}$$

Among them, T is the signal period. Take the local smoothing window function and apply it to the time series data, then the instantaneous correlation function $J(t)$ of $x(t)$ is:

$$J(t) = x\left(t + \frac{\xi}{2}\right)x^*\left(t - \frac{\xi}{2}\right) \tag{8}$$

Among them, ξ is the instantaneous correlation coefficient, and $*$ is the conjugate [9]. According to the product of the free correlation function and the periodic function, in the stationary process of the period T, the instantaneous correlation function $J(t)$ of $x(t)$ is also the instantaneous correlation function of the time domain t. Expand the instantaneous correlation function to obtain the Fourier series, and determine the frequency of the vibration signal. The period and the function $J(t)$ have a strong autocorrelation. Taking $\frac{1}{T}$ as the fundamental frequency and taking the frequency value to an integer multiple, its value represents the cyclic frequency. The cycle frequency number represents the energy of the cycle frequency. After the machine fails, the cycle frequency number is used as an important diagnostic basis.

Clarify the time-frequency domain on the Fourier series curve. When using each of the time-frequency domains, groupings of discrete components of cyclic frequency are formed. Different components have different sensitivity to faults. The more sensitive discrete components of cyclic frequency are useful in diagnosing faults. Time has a very good relationship characteristic. During the diagnosis process of rotating machinery, it is not necessary to know the entire distribution of the signal in the time and frequency domain, only to check it. For the discrete components of the insensitive cycle frequency number, the Fourier series is expanded under a single cycle, and with the help of the time-frequency domain local windowing method, the Fourier series and the signal frequency have a harmonic relationship, that is, there is a quadrature in the $[0, 2T]$ interval. The relationship, using the orthogonality in the harmonic function interval, calculates the window function of the vibration signal distribution in the time domain and the frequency domain, finds the harmonic frequency that can be used for detection, after a series of simplification and integration, the detection The time-frequency distribution at the characteristic frequency component extracts the local time-frequency characteristic component of the vibration signal, and obtains the important information of the local time-domain characteristic component. So far, the extraction of the local time-domain feature components of the vibration signal is completed.

2.5   Analyze the Failure of Rotating Machinery Based on Machine Learning

The extracted local time domain features of the vibration signal are input into the artificial neural network of machine learning to diagnose the failure mode of rotating machinery.

Firstly, by means of graphs, real-time tracking of the changes in the fault parameters of the rotating machinery to obtain fault samples, using the artificial intelligence algorithm of the neural network to number and classify the rotating machinery faults, and then through the data collector to collect the vibration signals of the running state of the rotating machinery, Take the characteristic components of the vibration signal at each time and frequency as the characteristic quantity of the fault of the rotating machinery, so as to obtain the characteristic information described by the value of the vibration signal, and obtain the learning sample input by the neural network. A single hidden layer neural network is selected, and the local time-domain characteristics of the extracted vibration signal and the fault diagnosis mode of rotating machinery are respectively used as the input and output of the network [10]. Determine the neural network structure, initialize the number of nodes in each layer, obtain the output values of the hidden layer and the output layer, adjust the connection weights of the input layer to the hidden layer, and the output threshold of the hidden layer and output layer, calculate the actual output value and the training sample The deviation. The specific neural network structure is shown in Fig. 4.
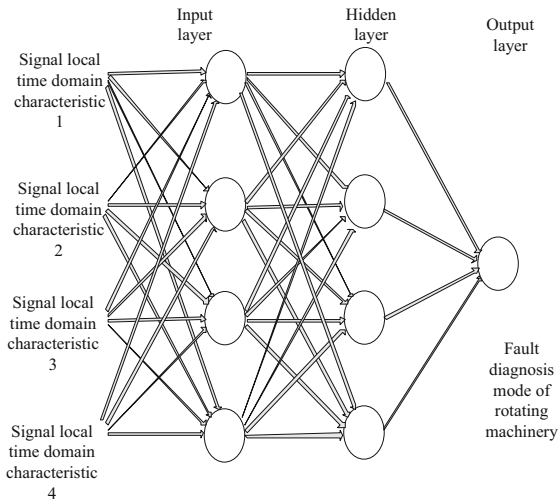
Fig. 4. Neural network structure of rotating machinery fault diagnosis

As shown in Fig. 4, when the network error converges to the minimum, the corresponding hidden layer node number is 4, and the rotating machinery fault diagnosis mode is used as an output node of the neural network. Calculate the input U_t of neuron τ in the middle layer of the network, the formula is:

$$U_\tau = \sum_{y=1}^{R} Z_{y\tau} P_y - \alpha_\tau \quad \tau = 1, 2, \ldots, Q \tag{9}$$

Among them, P_y is the fault feature vector input by the input layer neuron y, α_τ is the output threshold value of the middle layer neuron τ to each unit, and R and Q are the number of input layer and middle layer neurons. Use the transfer function to calculate the output of each unit in the middle layer and the response of each unit in the output layer, randomly select the next learning sample to input into the network, and repeat the above process until the neural network iteration error is less than the preset value [11]. Use the linear function to activate the input layer and output layer of the neural network, and then use the Sigmoid function to activate the hidden layer, and normalize the data before input to make the data at the same level, eliminate the influence of dimensions, and avoid the Sigmoid function the saturation area of the training data [12]. Update the number of training multiple times, reduce the error between the output value and the training sample, realize the neural network training of the device fault characteristics, compare the output value with the test sample, output the diagnosis result of the rotating machine, identify the fault diagnosis mode of the rotating machine, and It is applied to the rotating machinery tested. So far, the fault recognition of rotating machinery based on machine learning has been completed.

3   Experiment and Analysis

The time-frequency analysis method of the vibration signal distribution of rotating machinery designed in this paper based on machine learning and EMD decomposition is compared with two sets of traditional time-frequency analysis methods of the vibration signal distribution of rotating machinery. Use three sets of methods to perform time-frequency analysis on rotating machinery vibration signals to diagnose rotating machinery faults.

3.1   Experiment Preparation

The test data is collected from a single-stage gearbox failure test bench, and the failure type of the failed gear is tooth surface spalling. The number of gear teeth on the input shaft of the gearbox is 20, the number of gear teeth on the output shaft is 21, the sampling frequency is 20000 Hz, and the number of spectrum analysis points is 409600. The speed of the input shaft is 1000 r/min, the speed of the output shaft is 952.2 r/min, and the speed of the input shaft is 16. 67 Hz, and its meshing frequency is 333.3 Hz. The operating parameters of rotating machinery are shown in Table 1.

Table 1. Rotating machinery operating parameter settings

Parameter	Numerical value	Parameter	Numerical value
Number of reactive power compensation equipment put into operation	1 station	Active power correlation coefficient	865 units
Reactive power correlation coefficient	0.47	Active power distribution ratio	0.62
Reactive power distribution ratio	3.28	Power Factor	4.22
total capacity	11850 kVA	Compensation point reactive power compensation capacity	0.55–0.85
Rated voltage of compensation device	10 kV	Rated power	800 kVA
Reactive power compensation efficiency	96%	Load factor	2500 kW
Short-circuit capacity	450 MV.A	Distance between compensation point and center	80%

The time-domain waveform and frequency spectrum of the gear fault vibration signal are shown in Fig. 5.
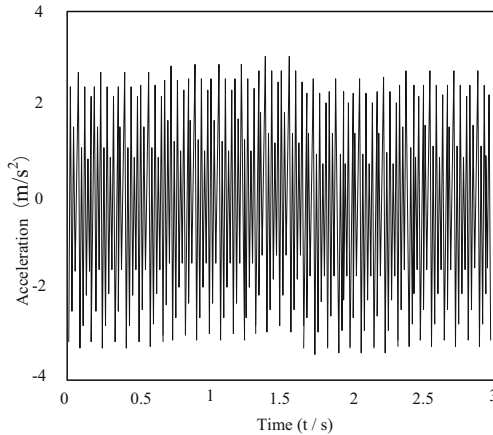
Fig. 5. Time-domain waveform diagram of gear failure
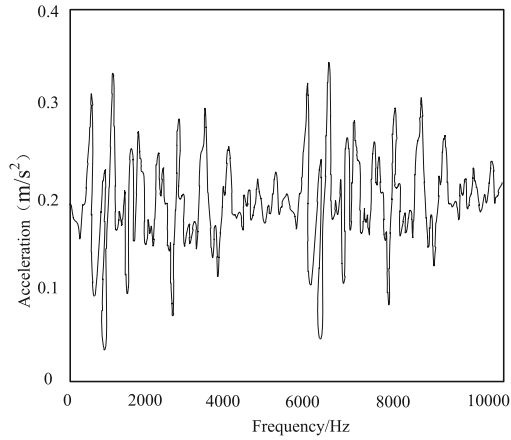
Fig. 6. Frequency spectrum of gear failure

It can be seen from Fig. 5 and Fig. 6 that the time-domain waveform of the time-domain waveform diagram is more complicated, and the frequency components of the spectrogram are also more complicated, it is difficult to distinguish the specific characteristics of the signal, and it is difficult to distinguish useful information.

3.2   Experimental Result

The First Set of Experimental Results
Record the three methods to extract the local time-domain features of the vibration signal, and the experimental comparison results are shown in Fig. 7.
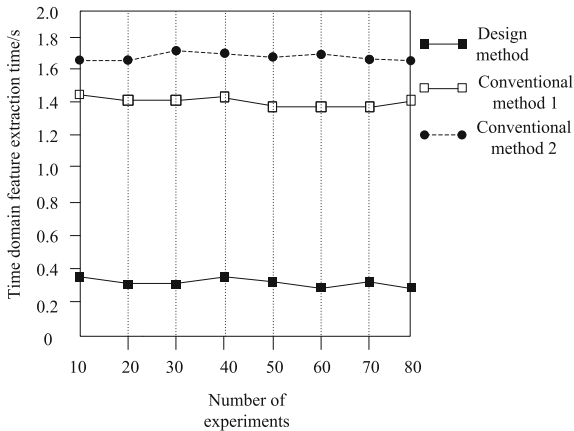
Fig. 7. Comparison results of time domain feature extraction time

It can be seen from Fig. 7 that the average extraction time of the signal time domain features of the method in this paper is 0.37 s, and the average extraction time of conventional method 1 and conventional method 2 are 1.42 s and 1.67 s respectively. Compared with the two conventional methods, the average extraction time of this method is Respectively reduced by 1.05 s and 1.30 s.

The Second Set of Experimental Results
On the basis of the first set of experiments, the vibration signal collection period was changed to test the time-domain characteristics of the signals extracted by the three sets of methods and the degree of similarity to the vibration signals of rotating machinery. The cross-correlation coefficient is used to measure the effect of signal extraction in the time domain. The larger the correlation coefficient, the more similar the extracted time domain features and the vibration signal. The experimental comparison results are shown in Table 2.

Table 2. Comparison of correlation coefficients of extracted signal time domain features

Signal acquisition period/s	Method correlation coefficient	Conventional Method 1 Correlation Coefficient	Conventional Method 2 Correlation Coefficient
0.02	0.792	0.734	0.701
0.04	0.801	0.746	0.708
0.06	0.813	0.752	0.714
0.08	0.838	0.771	0.723
0.1	0.849	0.782	0.735
0.12	0.851	0.789	0.741
0.14	0.872	0.794	0.749
0.16	0.889	0.805	0.758
0.18	0.891	0.816	0.763
0.2	0.905	0.837	0.771

As shown in Table 2, the correlation coefficient of the method in this paper is always higher than the two conventional methods, with an average correlation coefficient of 0.850. The average correlation coefficients of conventional method 1 and conventional method 2 are 0.783 and 0.736, respectively. It can be seen that the correlation coefficients of the method in this paper are increased by 0.067 and 0.112 respectively.

The Third Set of Experimental Results
After the three sets of methods extract the time-domain characteristics of the vibration signal, identify the fault diagnosis type of the rotating machinery, and calculate the fault diagnosis accuracy rate. The results are shown in Table 3.

Table 3. Results of the third group of experiments (%)

Signal acquisition cycle/s	A group recognition accuracy rate	B group recognition accuracy rate	C group recognition accuracy rate
0.02	98.7	90.4	86.4
0.04	97.7	91.6	87.9
0.06	98.1	89.9	85.3
0.08	97.5	90.6	86.5
0.1	97.4	91.9	87.5
0.12	96.9	92.4	87.6
0.14	100	90.7	86.9
0.16	99.2	89.7	86.3
0.18	97.1	90.1	86.1
0.2	98.2	92.8	88.5

As shown in Table 3, for the tooth surface peeling failure of rotating machinery, the average recognition accuracy of this method is 98.1%, and the recognition accuracy of conventional method 1 and conventional method 2 are 91.0% and 86.9%, respectively. It can be seen that this method recognizes The accuracy rate has been increased by 7.1% and 11.2% respectively.

In summary, the method in this paper shortens the time-domain feature extraction time of the vibration signal compared with the conventional method, and the extracted time-domain feature is more similar to the vibration signal, which improves the recognition accuracy of the fault type of the rotating machinery.

4   Conclusion

This study gives full play to the technical advantages of machine learning and EMD decomposition, improves the extraction effect of time-domain characteristics of vibration signals, and improves the fault identification accuracy of rotating machinery, making the average accuracy of the analysis results of time-frequency distribution of vibration signals of rotating machinery by the method in this paper reach 98.1%. However, there are still some shortcomings in this study. Therefore, in the future research, it will be considered to introduce the improved BITD algorithm to further improve the extraction efficiency of time domain features.

References

1. Hao, G.C., et al.: Time-frequency analysis of non-stationary signal based on NDSST. Geomat. Inform. Sci. Wuhan Univ. **44**(6), 941–948 (2019)
2. Shao, K.W., et al.: A time-frequency distribution method based on parameter adaptation and its application. Intell. Build. Constr. Mach. **1**(1), 60–63 (2019)

3. Hao, G.C., et al.: Time-frequency analysis of BGabor-NSPWVD algorithm with strong robustness and high sharpening concentration. Acta Automat. Sin. **45**(3), 566–576 (2019)
4. Yu, K., et al.: General parameterized synchrosqueezing transform and its application in rotating machinery vibration signal. J. Mech. Eng. **55**(11), 149–159 (2019)
5. Chen, H.K., et al.: Time-frequency analysis of laser Doppler radar vibration signals. Chin. J. Opt. **13**(5), 1014–1022 (2020)
6. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mobile Netw. Appl. **24**(1), 5–17 (2018). https://doi.org/10.1007/s11036-018-1134-8
7. Zhang, Z.W., Feng, Z.P., Chen, X.W.: Acoustic signal analysis of the resonance frequency region for planetary gearbox fault diagnosis based on high-order synchrosqueezing transform. Chin. J. Eng. **42**(8), 1048–1054 (2020)
8. Huang, C.J., et al.: Full vector BEMD method for fault feature extraction of rotating machinery. J. Vib. Shock **38**(9), 94–99+132 (2019)
9. Yu, B., Han, Y.B.: Design of vibration signal acquisition system based on STM32. Electron. Meas. Technol. **43**(4), 148–152 (2020)
10. Song, H.S., Shen, W., Zhou, H.: Time-frequency domain feature extraction of mixed signals based on blind source separation. Comput. Simul. **36**(4), 465–468 (2019)
11. Shi, W.W., et al.: Design of rotating machinery vibration test node based on WIFI real-time transmission. J. Sichuan Ordnance **40**(11), 92–96 (2019)
12. Gu, X.D., et al.: Vibration signal analysis based on wavelet packet transform and time-energy wavelet spectrum. Aerospace Control Appl. **46**(4), 29–33 (2020)

Synchronous Monitoring Method of Multi-manipulator Trajectory Signals Based on Machine Learning

Xiao-zheng Wan[1]([✉]), Song Zhang[2], Ji-ming Zhang[1], Hui Chai[1], and Huan-yu Zhao[1]

[1] Institute of Oceanographic Instrumentation, Qilu University of Technology (Shandong Academy of Sciences), Qingdao 266061, China
wxz18661817997@163.com

[2] College of New Energy, China University of Petroleum, Qingdao 266580, China

Abstract. The traditional trajectory signal monitoring method uses a combination of sensors and mathematical models to detect the trajectory signal of the manipulator, which is susceptible to interference from uncertain factors, resulting in low monitoring accuracy and poor real-time performance. In response to the above problems, this research designed a method for synchronous monitoring of multi-manipulator trajectory signals based on machine learning. After the multi-manipulator dynamic model is established, the coordinate system of the trajectory acquisition equipment is calibrated according to the principle of binocular vision. After preprocessing the trajectory image of the manipulator, the CLDNN model is used to realize the synchronous monitoring of the trajectory signal. The simulation experiment results show that the monitoring method shortens the monitoring time by about 62.5%, and has higher monitoring accuracy, which proves that its performance is better.

Keywords: Machine learning · Multiple manipulators · Manipulator trajectory · Signal synchronization · Signal monitoring

1 Introduction

A mechanical arm is an automatic operating device that can imitate certain actions of a human arm to grasp, carry objects or operate tools in a fixed program, and it is an important branch of industrial robots. The robotic arm is an important equipment in automated production, with the advantages of high safety and long working hours. After more than half a century of development, the control accuracy of industrial manipulators has been continuously improved, its types have become more and more abundant, and the application fields have also continued to expand.

There is a strong coupling between multiple manipulators, so although there are independent controllers for each mobile manipulator, each mobile manipulator should not be controlled separately, but should be considered as a whole [1]. It is an effective way to improve the control precision of multi-manipulator to monitor the error between its

moving trajectory and the theoretical trajectory of control instruction when controlling multi-manipulator.

Traditional monitoring method based on mathematical model is used to establish a mathematical model of the moving trajectory of the multi-manipulator, and the signals are monitored by comparing the difference between the feedback of the manipulator and the theoretical value of the model [2]. However, because the robotic arm system is a system with multiple inputs and multiple outputs, and has strong nonlinearity and strong coupling, these characteristics and the interference existing outside the system make the mathematical model of the robotic arm must have some uncertain factors, which is difficult Ensure the accuracy of the monitoring method. Some scholars also proposed a sensor-based tracking signal monitoring method for multi-manipulator. In this method, a variety of sensors are installed on the manipulator arm, and the tracking signals of the manipulator arm are monitored by analyzing the signals sent back by the sensors [3]. The accuracy of this monitoring method is not only limited by the accuracy of the sensor, but also the sensor will be interfered in the complex working environment of the robotic arm, and the real-time monitoring is difficult to guarantee.

With the development of computer vision, machine learning and sensor technology, various intelligent robot systems relying on computer vision, machine learning and advanced sensors are constantly coming out. The use of machine learning can improve the monitoring efficiency of the monitoring method, and can monitor multiple objects at the same time [4]. According to the above analysis content, in order to improve the real-time performance of trajectory signal monitoring, this article will take the multi-mobile manipulator as the research object, and study the method of synchronous monitoring of multi-manipulator trajectory signal based on machine learning.

2   Method Design

2.1   Multi-manipulator Motion Analysis and Dynamic Modeling

In order to be able to monitor the trajectory signals of multiple manipulators simultaneously, a mathematical model of the multi-mobile manipulator system must be established first. Usually, a fixed coordinate system is defined on each link of the robot arm, and two adjacent coordinate systems can represent each other, and their relative positions are generally described by the transformation matrix between the two adjacent links. The connecting rods are numbered starting from the fixed base of the robotic arm. Generally, the coordinate system at the fixed base is O_o, and so on, so the coordinate system on the connecting rod i is O_i.

For the robotic arm, the coordinate system of all links can be established in sequence according to the following steps [5]:

Find out the corresponding joint axis of each link and extend it in the direction of its axis.

Find the common perpendicular between the joint axis $i - 1$ and the joint axis i or the intersection point of the joint axis $i - 1$ and the joint axis i, and use the intersection point of the common perpendicular line and the joint axis or the intersection point of the joint axis $i - 1$ and the joint axis i as the connection The origin of the rod coordinate system O_i is used to establish the coordinate system.

Specify the direction of the Z_i axis as the direction of the i axis.

Specify that the X_i axis is along the direction of the common perpendicular between the joint axis $i-1$ and the joint axis i, and the joint axis $i-1$ points to the joint axis i. If the joint axis $i-1$ and the joint axis i intersect, the X_i axis is specified to be perpendicular to the joint axis $i-1$ And the plane where the joint axis i lies.

The Y_i axis direction is determined by the right-hand rule based on the X_i axis and Z_i axis directions.

In special cases, when the first joint variable is 0, the coordinate system O_o and the coordinate system O_1 coincide. For the coordinate system O_n, the origin and the direction of the X_n axis can be selected arbitrarily. When selecting, the connecting rod parameter is usually selected as 0.

It can be seen from the establishment rules of the coordinate system that when the axis Z_i and the axis of the joint i coincide, there are two choices for the direction of Z_i: the axial direction and the reverse extension direction. You can choose one of them without affecting the final result. When two joint axes intersect, it is impossible to determine which joint axis is in front and which is behind. The direction of the X_i axis cannot be defined according to the rules. You can choose between two possibilities, and then determine the direction of the y axis based on the right-hand rule based on the previous determination., When the joint axis $i-1$ is parallel to the joint axis i, the origin position of the coordinate system O_i can be arbitrarily selected.

The relationship between the linear velocity of the end of the robotic arm in Cartesian space and its joint angular velocity in the joint space is expressed by the Jacobian matrix.

The hypothesis of the expected trajectory at the end of the robotic arm is shown in Eq. (1):

$$x = x(q) \tag{1}$$

In the formula, x represents the trajectory of the end of the robotic arm in Cartesian space, and q represents the rotation angle of each joint of the robotic arm in the joint space. Take the derivation operation on both sides of formula (1) to time respectively to obtain the relationship between q and x as shown in formula (2):

$$\dot{x} = J(q)\dot{q} \tag{2}$$

In the formula, $\dot{x}$ is the speed of the end of the manipulator in Cartesian space; $\dot{q}$ is the angular velocity of the joint rotation of the manipulator in the joint space; $J(q)$ is the Jacobian matrix of the manipulator. The elements in row i and column j of the matrix are:

$$J_{ij}(q) = \frac{\partial x_i(q)}{\partial q_j} \tag{3}$$

The multi-manipulator dynamic model studied in this paper is obtained on the basis of the Lagrangian function method. The kinetic energy K of the system is defined as the sum of the translational kinetic energy and the rotational kinetic energy of each link, and its expression is as follows:

$$K = \sum_{i=1}^{n} \left(\frac{1}{2} I_i \omega_i^2 + \frac{1}{2} m_i \dot{r}_i^2 \right) \tag{4}$$

The potential energy P is defined as the sum of the gravitational potential energy of each connecting rod and its expression is as formula (5):

$$P = \sum_{i=1}^{n} m_i g r_{iy} \tag{5}$$

In the above formula, I_i is the central inertia tensor of the connecting rod; ω_i is the angular velocity of the connecting rod in the inertial coordinate system; $\dot{r}$ is the linear velocity of the connecting rod in the inertial coordinate system; r_{iy} is the direction along the y-axis in the inertial coordinate system Weight.

Bring the total kinetic energy and total potential energy equation of the robot arm link into the Lagrangian function, and then obtain the partial derivative of it, and finally the dynamic equation of the robot arm can be obtained as shown in the following formula [6]:

$$D(q)\ddot{q} + B(q, \dot{q})\dot{q} + G(q) = u \tag{6}$$

In the formula, $D(q)$ is the inertia matrix of the multi-manipulator, $B(q, \dot{q})$ is the centrifugal force and Ge force matrix of the multi-manipulator, $G(q)$ is the gravity term matrix, q is the current joint angle vector of the manipulator, and u is the controller output in the controlled manipulator. The driving torque on each joint. After the above-mentioned multi-manipulator motion trajectory model is established, the visual acquisition equipment that collects the motion trajectory of the mechanical arm is calibrated.

2.2   Trajectory Acquisition Equipment Dual Target Positioning

The pinhole imaging model of the camera actually uses the above geometric relationship constraints to simulate the image formation process, that is, to establish the coordinate conversion relationship between the two-dimensional image collected by the binocular camera and the three-dimensional objective object. The ideal imaging model of the camera is the pinhole imaging model, which does not consider the distortion of the camera in reality. There are four coordinates in this model: the world coordinate system, the camera coordinate system, the image coordinate system and the pixel coordinate system [7].

In an ideal camera imaging model, the default pixel plane is parallel to the image plane, and the pixel coordinate system is known, that is, the coordinates of point R in the two-dimensional digital image, and then the points of pixel $R(u, v)$ and world coordinates are derived in turn The derived relationship between $R_w(x_w, y_w, z_w)$. Due to the manufacturing process of the camera, it cannot guarantee that the imaging plane of the camera is absolutely perpendicular to the optical axis of the camera. From the imaging situation of the camera under ideal conditions, the mathematical derivation of the generalized pinhole imaging model can be derived as follows:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{z_w} \begin{bmatrix} f_x & -f_x/\tan\theta & u_0 & 0 \\ 0 & f_y/\sin\theta & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R_{3\times3} & T_{3\times1} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{7}$$

In the formula, θ is the angle between the imaging plane and the optical axis of the camera; f_x, f_y, u_0, and v_0 are collectively referred to as the camera internal parameters; $R_{3\times3}$ is the camera internal parameter matrix; $T_{3\times1}$ is the camera's projection transformation matrix. After calibrating the multi-mechanical ratio trajectory image acquisition device, the acquired trajectory image is processed.

2.3 Robotic Arm Trajectory Image Processing

According to the brightness component of the pixel in the robot arm trajectory image, the conversion relationship between RGB and YUV color space is constructed as shown in the following formula [8].

$$Y(i, j) = \alpha_1 R(i, j) + \alpha_2 G(i, j) + \alpha_3 B(i, j) \tag{8}$$

In the formula, (i, j) is the position of any pixel in the image; α_1, α_2, and α_3 correspond to the weights of the three color components of R, G, and B in the RGB color space; $R(i, j), G(i, j)$, and $B(i, j)$ are the brightness values of the three components of the image; $Y(i, j)$ is the grayscale value of the image after grayscale conversion. After converting the image gray level, a Gaussian low-pass filter is used to process image noise.

For image denoising processing of size $M \times N$, the transfer function of the two-dimensional Gaussian low-pass filter is as follows:

$$\begin{cases} H(u, v) = e^{-\frac{D(u,v)^2}{2\sigma^2}} \\ D(u, v) = \left[\left(u - \frac{M}{2}\right)^2 + \left(v - \frac{N}{2}\right)^2 \right]^{\frac{1}{2}} \end{cases} \tag{9}$$

In the formula, (u, v) is the image pixel of the input Gaussian low-pass filter; $H(u, v)$ is the filter transfer function; $D(u, v)$ is the distance from the center of the frequency rectangle; σ is the variance of the image gray value. Image denoising processing will result in blurring of the edges of the image. Sobel operator is used to detect the edges of the image. The calculation formula of Sobel operator is as follows [9]:

$$\nabla f = \left[\frac{\partial f}{\partial x} \ \frac{\partial f}{\partial y} \right]^T \tag{10}$$

$$mag(\nabla f) = \left[\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2 \right]^{\frac{1}{2}} \tag{11}$$

$$\partial(x, y) = \arctan\left(\frac{\partial f}{\partial x} \ \frac{\partial f}{\partial y} \right) \tag{12}$$

In the formula, ∇f is the image gradient; $mag(\nabla f)$ is the image gradient amplitude; $\partial(x, y)$ is the image direction angle.

After image preprocessing, the image is transformed from the spatial domain to the fuzzy domain according to the following formula:

$$\mu_{ij} = \tan\left(\frac{x_{ij} - x_{\min}}{x_{\max} - x_{\min}}\right) \tag{13}$$

$$\begin{cases} C = \gamma \frac{|\mu_{ij} - \overline{\mu}_{ij}|}{|\mu_{ij} + \overline{\mu}_{ij}|} \\ C' = \varphi(C) \end{cases} \tag{14}$$

In the formula, x_{ij} is the gray value of the image; μ_{ij} is the membership value of the fuzzy domain of the image; C is the local image contrast, which is also the fuzzy enhancement operator; $\varphi(\ )$ is the inverse transform function of the fuzzy domain; γ is the fuzzy parameter of the fuzzy enhancement. Generally, in order to obtain high-definition images of the robotic arm trajectory, the camera needs to continuously shoot multiple high-definition images. In order to improve the efficiency of track signal monitoring, the image is compressed.

2.4   Image Compression Processing

After processing the robot arm trajectory image, the PLOT compression algorithm is used to compress the image signal to reduce the image processing dimension. The essence of the PLOT compression algorithm is to find all the main trends in the collected signal, and use the segmented linear trend to replace the original collected data to achieve the best possible data compression effect. For the prediction time series object y_t, suppose its model is a discrete time signal sampled at equal intervals as shown in Eq. (15) [10].

$$y_t = \mu_t + \varepsilon_t \tag{15}$$

In the formula, μ_t is the expected value of the process at sampling time t, ε_t is the error between the expected and actual values, and $[t_{j-1}, t_j]$ represents the sampling time period. Assuming that ε_t is a random variable and obeys an independent normal distribution $N(0, \sigma^2)$, the trend of the process mean μ_t can be approximated by a piecewise linear expression. If the trend of the time series changes at time t_{j-1}, then in the j th sampling time period $[t_{j-1}, t_j]$, the specific form of the linear expression of the analog μ_t is shown in Eq. (16).

$$\mu_t = \mu_{t_{j-1}} + \delta_{t_{j-1}} + \beta_j(t - t_{j-1}) \tag{16}$$

In the formula, $\delta_{t_{j-1}}$ is the change of the process expectation value at time t_{j-1}, and β_j is the change slope of the process expectation value during the sampling period of $[t_{j-1}, t_j]$.

The core idea of the PLOT data compression algorithm is to simulate the linear function corresponding to all sampling time period $[t_{j-1}, t_j]$ as accurately as possible, not only to detect the linear change trend, but also to identify the linear trend change point. The time t_j, t_j is the starting point of the next sampling time period. time. In the process of piecewise linear estimation, it is necessary to detect and process the abnormal point data in the sampled signal at the same time to reduce the influence of abnormal points on the linear trend estimation. The linear function in the sampling time period is obtained by linear fitting all data points in the current time period using the least square algorithm, and the new sampling point must be re-fitted in the same way after the new sampling point appears. At the same time, the least squares estimate the variance of the fitting process through the mean square error. The judgment of the segment interval point is to use the linear function in the current time period to predict the next time sampling point, and then judge whether the new data point falls within the prediction interval. If it falls within it, it will not be processed. It is only used to refit the data trend. If it falls outside the prediction interval, it means that the data trend has changed or is an abnormal point at the next moment, and then processed accordingly. Use the neural network structure in machine learning to realize the synchronous monitoring of multi-manipulator trajectory signals.

2.5 Realize the Simultaneous Monitoring of Multiple Robotic Arms Trajectory Signals

This paper constructs a CLDNN model accordingly. The structure of the CLDNN network is to connect several layers of CNN after the input layer to extract local features. The output of the CNN is poured into several layers of LSTM units to reduce time domain changes. The output of the last layer of LSTM is input to the fully connected layer, with the purpose of adding features The space is mapped to an output layer that is easier to classify. After the convolutional layer, the BN layer is deployed for normalization to avoid over-fitting. The fully connected layer connects all features and transmits the output value to the softmax classifier.

The following cross entropy loss function is selected as the loss function of the model:

$$J(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \ln \hat{y}_i - (1-y) \ln(1 - \hat{y}_i) \right] \tag{17}$$

The original cross entropy is used to indicate the degree of similarity between two distributions. Here it is used to indicate the degree of similarity between the distribution of the true value y and the distribution of the network predicted value $\hat{y}$. The smaller the cross entropy, the more similar the two.

Considering that the depth of the network has an extremely important impact on the classification efficiency of the network itself, after the number of network layers continues to increase, the method of backpropagation will cause the gradient of the parameters to gradually become diffuse during the convergence process, and even disappear directly, resulting in The parameters do not converge for a long time. Let the network fit the residual function $F(x) = H(x) - x$ instead of fitting $H(x)$, so that the original function becomes $F(x) + x$. In this article, the residual structure block is defined as:

$$y = F(x, \{W_i\}) + x \tag{18}$$

In the formula, x and y respectively represent the input and output vectors of the structure block, and the function $F(x, \{W_i\})$ represents the residual mapping that needs to be learned. With the minimum output of the residual block as the goal, the convolutional network is trained with sample data to determine the parameters of the neural network.

The processed multi-manipulator trajectory image is output to the convolutional neural network, and the trajectory in the image is matched with the preset mechanical arm motion trajectory under the action of the activation function in the hidden layer. Determine the error between the actual robot trajectory and the theory, so as to realize the monitoring of the trajectory signal. So far, the research on the synchronous monitoring method of multi-manipulator trajectory signals based on machine learning has been completed.

3   Test Experiment

3.1   Experiment Content

This experiment is a comparative experiment. The synchronous monitoring method of multi-manipulator trajectory signals based on machine learning studied above is used as the experimental group, and the traditional trajectory signal monitoring method is used as the comparison group. By comparing the error and response time of the two signal monitoring methods, the reliability of the monitoring method studied above is verified.

3.2   Experimental Environment

The experimental platform includes Denso VP6242G and Quanser's open structure control module. The control module has the functions of all industrial systems and serves as an interface for interaction with the QUARC software. The closed loop frequency of the controller is 4kHz. Users can adjust the gain of the controller through the QUARC software, or design their own controller in the simulation environment, and directly use the completely open commands to bypass the internal PID controller. The experimental platform realizes the rotation of each joint through the servo motors at the joints and is equipped with corresponding measurement sensors. The measurement data controller and the QUARC software interact to realize the reproduction of the end trajectory of the manipulator.

3.3 Experimental Result

The comparison results of the two trajectory signal monitoring methods to monitor the trajectory error of the multi-manipulator are shown in Fig. 1.
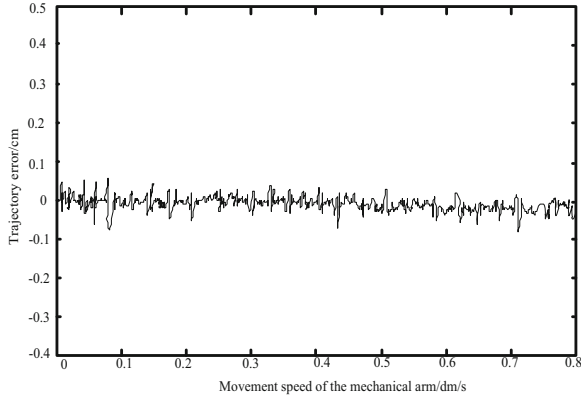
(a) Experimental group method monitoring error
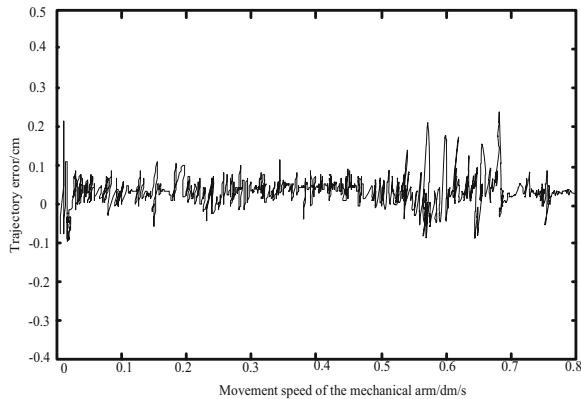
(b) Comparison group method monitoring error

Fig. 1. Comparison of monitoring errors

From the analysis of Fig. 1, it can be seen that compared with the error change curve of the contrast group, the monitoring error of the method in the experimental group for the trajectory of the multi-manipulator is significantly smaller. Although the monitoring errors of the two groups were within the allowable range, the fluctuation of the monitoring error curve of the contrast group method was significantly more severe than that of the experimental group. Therefore, the monitoring performance of the proposed method is better.

When the two methods are monitoring the trajectory of the same parameter manipulator, the response time of the monitoring method is shown in Table 1.

Table 1. Comparison of response time of monitoring methods/ms

Serial number	Experimental group method	Comparison group method
1	7.32	17.93
2	6.97	18.66
3	6.94	18.74
4	6.99	17.88
5	7.02	19.63
6	6.96	19.71
7	7.23	19.85
8	7.04	17.82
9	7.02	19.34
10	6.86	17.89

By analyzing the data in Table 1, it can be seen that the response time of the method in the experimental group in monitoring the trajectory signal of the manipulator is less than that of the method in the contrast group. After averaging the data in the table, it was found that the average response time of the experimental group method was 7.035 ms, and the average response time of the contrast group method was 18.745 ms. The above results indicate that the trajectory signal monitoring method studied in this paper has higher monitoring accuracy, real-time performance and better performance.

4 Concluding Remarks

Industrial manipulator is an industrial automation equipment which includes machinery, communication, automatic control, computer and other disciplines. It plays an important role in high-end manufacturing industry. In order to improve the control accuracy of the manipulator, this paper studies a multi-manipulator trajectory signal synchronous monitoring method based on machine learning in view of the problems existing in the traditional trajectory signal monitoring methods. Based on the establishment of dynamic model of multi-manipulator, this method calibrates the coordinate system of trajectory acquisition equipment by using binocular vision principle. Then, after preprocessing the trajectory image of the manipulator, the CLDNN model is used to realize the synchronous monitoring of the trajectory signal. This study also proves that the method has better performance through experiments.

References

1. Zhu, Z., et al.: Obstacle avoidance path planning of space redundant manipulator based on a collision detection algorithm. J. Northwest. Polytech. Univ. **38**(01), 183–190 (2020)

2. Ren, W., Jiang, M.: Research on trajectory planning algorithm of 6-DOF manipulator. J. Anhui Polytech. Univ. **34**(06), 52–59 (2019)
3. Qian, Q., Zhang, A., Sun, Y.: Event triggered adaptive robust trajectory tracking control for multi-joint manipulators. Acta Armamentarii **40**(08), 1732–1739 (2019)
4. Xu, Y., et al.: Application of improved LMD algorithm in signal detection of power quality disturbance in microgrid. Power Syst. Technol. **43**(01), 332–341 (2019)
5. Li, Y., et al.: Research on model of active tiles in data centers based on machine learning. Comput. Simul. **36**(12), 180–185+248 (2019)
6. Chen, H., Li, L.: Trajectory planning of time optimal manipulator based on complex method. J. Mech. Transm. **43**(03), 72–75+94 (2019)
7. Ma, X., Jiang, X.: Analysis for the generalization ability of machine learning based full reference image quality metrics. J. Commun. Univ. China Sci. Technol. **26**(04), 42–49 (2019)
8. Zhou, X., He, X., Zheng, C.L Radio signal recognition based on image deep learning. J. Commun. **40**(07), 114–125 (2019)
9. Liu, S., et al.: Parallel fractal compression method for big video data. Complexity **2018**, 2016976 (2018)
10. Liu, S., Fu, W., He, L., Zhou, J., Ma, M.: Distribution of primary additional errors in fractal encoding method. Multimed. Tools Appl. **76**(4), 5787–5802 (2014). https://doi.org/10.1007/s11042-014-2408-1

Big Data Stepwise Regression Correction Method for Forearm Wrong Posture in Track and Field

Yong-ming Chen[1] and Cai-xu Xu[2(✉)]

[1] Guangzhou Huali College, Guangzhou 511325, China
[2] School of Electronics and Information Engineering, WuZhou University,
WuZhou 543002, China
`xuccx32@outlook.com`

Abstract. The current large data correction method for track and field forearm wrong posture has not considered the forearm posture detection results and restricts the arm posture detection, resulting in the wrong forearm posture and the low accuracy of the big data for correcting forearm posture. For this reason, a track and field forearm is proposed Stepwise regression correction method for wrong posture big data. Make full use of the information detection characteristics of the physical factors of the stepwise regression, and assume the time series of the forearm posture to detect the motion posture of the forearm in the track and field; calculate the misjudgment probability of detecting the forearm motion posture, form the misjudgment probability matrix, and analyze it according to the matrix. As a result, the forearm posture big data misjudgment probability matrix constraint is obtained; the stepwise regression method is used to realize the big data detection and classification of the forearm wrong posture, judge whether the forearm posture is wrong, calculate the peak value of the forearm posture error, and correct the big data of wrong forearm posture. The experimental results show that the error posture big data correction accuracy of this method is higher.

Keyword: Stepwise regression · Track and field sports · Wrong forearm posture · Correction

1 Introduction

Track and field sports are one of the main contents of current sports. Correct and beautiful track and field movements can greatly improve students' strength, agility, flexibility and other qualities, physical coordination, sense of rhythm, and shaping a beautiful body shape. The benefits of [1]. Therefore, track and field sports are an important part of sports. However, due to the mutual influence and mutual restriction of personal habits, physical fitness and other factors, in the process of track and field sports, it is inevitable that various redundant or wrong actions will occur. If the wrong action cannot be found in the budding state in time To be corrected, the wrong power stereotype will be formed over time, which will not only affect the mastery of the correct movement technique

of track and field sports, but also may lead to injury accidents [2]. Therefore, in track and field sports, it is particularly important to check and correct wrong movements. Therefore, both at home and abroad are actively studying ways to correct track and field wrong postures. In China, there are many documents on track and field analysis and error correction, but most of them are divided into two types. One is to analyze and correct only a certain movement of track and field sports, but does not consider the connection with the overall movement, and cannot grasp the root cause of the error. The second is the listing and description of various wrong actions, not only did not analyze and discuss the cause of the error and corrective methods in detail, but also did not sort and analyze the wrong actions [3]. In foreign countries, for the correction of wrong actions in track and field exercises, it is necessary to choose a reasonable error correction method in accordance with the law of formation of track and field skills [4]. Relevant scholars have made some progress in this issue.

In reference [2], a simulation of manipulator grasping attitude control based on synovial error feedback is proposed. The multi-sensor fusion information acquisition method is used to measure and fuse the manipulator attitude parameter information, and the constraint parameter model and control objective function of manipulator attitude control are established. The experimental results show that the method has high control accuracy in the robot arm attitude control, and improves the robot arm grip. But the error capture time is not good.In reference [3], a general dual robot arm posture detection method based on three-dimensional positioning is proposed. The robot positioning accuracy of the target workpiece under a single camera is measured by three-dimensional positioning method, and the relevant features in the image information are interacted, which can quickly and efficiently separate the target object from the background. This method has efficient recognition ability, However, the positioning error of this method is obviously large.

In view of the above problems, this paper introduces the stepwise regression method to correct the big data of forearm wrong posture in track and field. From the psychological factors, environmental factors, physical factors and other aspects of technical error correction are analyzed and studied effectively, which can effectively improve the correction effect of forearm error posture in track and field.

2 Big Data Correction Method of Wrong Forearm Posture in Track and Field

2.1 Big Data Detection of Forearm Posture Based on Stepwise Regression

In track and field sports, the forearm posture time series $Y(t)$, (where t represents time), is composed of forearms, elbows and wrists. In order to make full use of the big data of forearm posture, the information on the change law of the factors generated in track and field sports is gradually used. Formula regression method separates its period term and trend term, and predicts the forearm motion posture [5]. Therefore, the separated random term can be regarded as a stationary time series. Let the periodic term be $T(t)$, the trend term is $P(t)$, the random term is $R(t)$, the number of dominant factors of the periodic term is I, the order of the detection model is P, and the number of periodic

terms is The number is i, the regression coefficients of the periodic term are A_0 and A_i, the number of dominant test periodic waves is k, the length is l, and the test periodic waves of length l are $f_{il}(t)$, the regression coefficients of the trend term are B_0 and B_i, and the constructor of the time t is $g_i(t)$, the regression coefficient of the random term is ϕ_i, and the mean value of the random term is μ, then:

$$Y(t) = T(t) + P(t) + R(t)$$

$$T(t) = A_0 + \sum_{i=1}^{I} A_i \sum_{l=1}^{k} a_{il} f_{il}(t)$$

$$P(t) = B_0 + \sum_{i=1}^{} B_i g_i(t) \tag{1}$$

$$R(t) = \mu + \sum_{i=1}^{P} \phi_i [R(t-i) - \mu]$$

In the formula, a_{il} represents the coefficient of the test periodic wave $f_{il}(t)$ [6]. At this time, substituting the forearm posture time series $\{Y_1(t), Y_2(t), \cdots, Y_n(t)\}$ (where n is the number of forearm posture factors) obtained by Eq. (1) into the stepwise regression to obtain the predicted value of the forearm posture factor $Q(t)$, then:

$$Q(t) = b_0 + \sum_{i=1}^{n} b_i Y_i(t) \tag{2}$$

In the formula, b_0 and b_i are time series coefficients. At this time, according to the formula (2), the forearm posture is recognized and the forearm posture is recognized to determine whether there is an error in the big data of the forearm movement posture during the track and field exercise.

2.2 Determine the Constraint Conditions for Correcting the Big Data Misjudgment of the Forearm Posture

Misjudgment Probability Matrix
In the process of detecting and recognizing the forearm posture, due to problems such as the angle and direction of the forearm movement, there will be a certain probability of misjudgment when recognizing whether there is an error in the forearm posture and correcting it [7]. Therefore, based on formula (2), the detected forearm motion posture in track and field is calculated and the probability of misjudgment exists when correcting the wrong posture of the forearm. Let i represent the correct detection of the forearm posture, j represent the large data detection error of the forearm posture, use p_{ij} to represent the probability that the detection result is i and the actual result is j, c_{ij} represents the actual result is i, and the actual result is the number of j. Then there are:

$$p_{ij} = \frac{c_{ij}}{\sum_{j=0}^{3} c_{ij}} \tag{3}$$

In the formula, the values of i and j are 00, 01, 02, 03. At this time, using formula (3), after taking all values of i and j, the misjudgment probability matrix can be established, as shown in Table 1.

Table 1. Misjudgment probability matrix

Actual class	Forecast			
	00	01	02	03
00	0.731	0.003	0.006	0.008
01	0.128	0.797	0.002	0.07
02	0.030	0.109	0.980	0.05
03	0.111	0.091	0.012	0.872

In the misjudgment probability matrix shown in Table 1, when the predicted value calculated by Eq. (3) is 00, the probability of forearm posture detection error 00, 01, 02, 03, 04 is 0.731, 0.128, 0.030, 0.111. The misjudgment probability matrix established at this time lays the foundation for the error correction algorithm.

Forearm Posture Big Data Misjudgment Probability Matrix Constraint
In the process of correcting the wrong forearm posture, there will be 04, 05, and 06 types of forearm posture recognition results with a correct rate of more than 96%. Therefore, on the basis of the misjudgment probability matrix, the forearm posture misjudgment probability matrix is constrained. Based on the probability value of the misrecognition

Table 2. Correspondence between i, j and the new category

Test results	Ranges	New category
00	[1,73]	00
	[73,86]	01
	[86,89]	02
	[89,100]	03
01	[1,80]	01
	[80,91]	02
	[91,100]	03
02	[1,98]	02
	[98,100]	03
03	[1, 7]	01
	[7, 13]	02
	[13,100]	03

probability matrix, 100 integers are divided into different areas from 1 to 100. At this time, the correspondence between the three is shown in Table 2.

At this time, suppose that in track and field sports, the forearm movement posture identified in (2) is Type m, and $m \in \{0, 1, 2, 3\}$. And use x_m to represent the random integer whose recognition result is m, and the value range of x_m is in $[1, 2, \cdots, 100]$. Suppose the new category after the calculation process in Table 2 is y_m and $y_m \in \{0, 1, 2, 3\}$. At this time, the piecewise function x_m is established according to Table 2, then:

$$f(x_m) = \begin{cases} 00, 1 \leq x_{00} \leq 73 \\ 01, 74 \leq x_{00} \leq 86, 1 \leq x_{01} \leq 80, 1 \leq x_{03} \leq 7 \\ 02, 87 \leq x_{00} \leq 89, 81 \leq x_{01} \leq 91, 1 \leq x_{02} \leq 98, 8 \leq x_{03} \leq 13 \\ 09, 90 \leq x_{00} \leq 100, 92 \leq x_{01} \leq 100, 99 \leq x_{02} \leq 100, 14 \leq x_{03} \leq 100 \end{cases}$$

$$(4)$$

At this time, (4) is the constraint condition of the misjudgment probability matrix. Combining the above content, the constraint flow of the forearm posture misjudgment probability matrix can be established, as shown in Fig. 1.
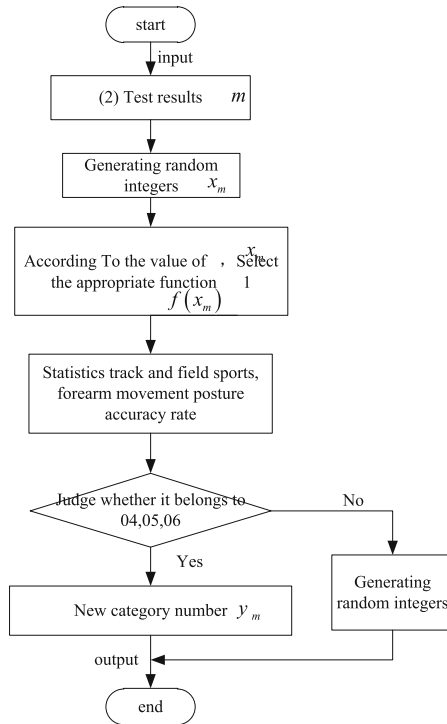
Fig. 1. Flow chart of the constraint of the probability matrix

At this time, under the above constraints, the wrong posture of the forearm movement in the track and field movement detected by the formula (2) is corrected.

2.3 Big Data Correction Algorithm for Wrong Forearm Posture

For the existing forearm movement posture in the track and field detected by Eq. (2), the output result of Fig. 1 is mapped to the [0, 1] interval, that is, the classification result is expressed in the form of probability. Suppose the last layer of fully connected data of the forearm posture of track and field is extracted, the one-dimensional matrix V represents this set of data, V_i represents the value of the i element in V, and the value range of i is determined by the number of tags in the model [8]. The big data of the forearm posture of track and field sports proposed this time has a total of n tags, so the value of i ranges from 1 to n. W_i represents the weight parameter in the maximum flexibility. Suppose the maximum flexibility is S_i and Z_i, then the maximum flexibility is S_i and Z_i as:

$$S_i = \frac{e^{Z_i}}{\sum_{i=1}^{n} e^{Z_i}}, i = 1, 2, \cdots, n$$

$$Z_i = \sum V_i * W_i$$

(5)

In the formula, e represents an increasing function. Since the classification result of the maximum flexible maximum is to select the class with the maximum probability, that is, the i maximum S_i corresponds. It can be seen from the formula (5) that S_i and Z_i have a corresponding relationship. Because the value on the denominator is constant, and e represents an increasing function, the size of the Z_i value determines the final classification result, and the i corresponding to the largest Z_i is the final classification result of the forearm posture. as shown in picture 2.
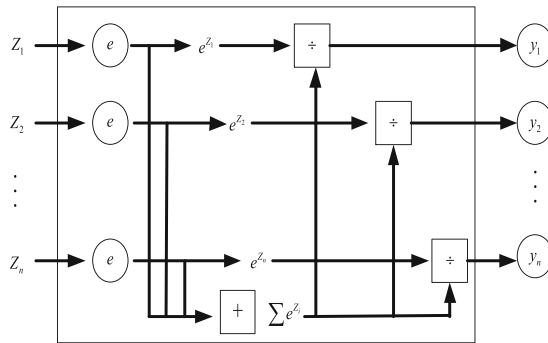
Fig. 2. Final classification of forearm posture big data

Under the condition of correcting the probability constraint of forearm posture misjudgment, the same forearm motion posture data set is divided into correct class and wrong class. Based on the above analysis, the value of Z depends on V. Therefore,

before classifying the forearm motion posture data set, the characteristic value V of the n fully connected layer is used to express the distribution law with a curve fitting function.

Observing the curve distribution of the two recognition results of the same type of gesture, it is found that there is a common feature dimension in V that has a large impact factor on the recognition result, so it is defined as a misrecognition feature matrix. The reason why the big data of the same type of forearm movement posture error based on the stepwise regression recognition process is that the value of the misrecognition feature matrix is too small, that is, the process of detecting the misrecognition feature matrix is to explore the error mechanism of the misrecognition gesture [9]. Therefore, suppose that the correct m forearm pictures identified by stepwise regression constitute set A, and the u forearm pictures identified incorrectly by stepwise regression constitute set B. And extract the eigenvaluesof the n fully connected layer, save it to the matrix V, and define the total number of features as h. When inputting any track and field sports, the forearm motion postures k and s, where $k \in A$, $s \in B$; the eigenvalues of the n fully connected layer, extract k and s respectively, and save them in the matrix V; respectively calculate k and s and input them to the flexible The value on Z of the maximum value function; arrange the data from small to large, and use a curve to describe the changing trend of Z, find out the characteristics that change drastically and the corresponding original dimension, and form the dimension of the set C; start traversing from C_1, Until C_h ends the loop, the frequency and frequency of each dimension are counted; the frequency is greater than 90% and the dimension is saved to the matrix Q, and the misidentified feature matrix Q is output, and the end.

Based on Eq. (5) and the classification results of the forearm motion posture in Fig. 2, in order to retain the position information of the forearm features and the relationship between the feature points, the data of each row in the matrix is represented by three-dimensional coordinates. Then there are:

$$z = griddata(x, y, z, X, Y, V) \tag{6}$$

Formula represents the big data fitting process of forearm motion posture [10–13]. (6) In the formula, $z = f(x, y)$, the surface is fitted with an irregular data vector x, y, z, and $griddata()$ means that the interpolation of the surface z at the point (X, Y) will be returned, and the surface always passes through these data points x, y, z. The input parameter (X, Y) is usually a regular grid. X, Y is a matrix of one-dimensional constant row vectors.

At this point, you can correct the wrong forearm posture in track and field, that is, enter the stepwise regression recognition category number; extract the current forearm motion posture and the characteristic value in the track and field; calculate the current forearm motion posture, the three-dimensional in the track and field Curved surface peak; when the three-dimensional curved surface peaks are within the normal swing range of the forearm; when the three-dimensional curved surface peaks are within the normal swing range of the forearm; output the original forearm swing posture. When the peak value of the three-dimensional curved surface is not in the normal swing range of the forearm, the above calculation process is used to correct the forearm movement posture and output the correction result of the forearm error posture. At this point, the correction of the wrong posture of the forearm is completed.

3   Experimental Demonstration Analysis

To verify the method for correcting the wrong forearm posture of track and field sports, two commonly used large two-dimensional forearm posture databases of track and field athletes were selected as the test objects of this experimental method to correct the wrong forearm posture of the track and field athletes. The proposed method for correcting forearm wrong posture in track and field sports is regarded as experimental group A, and the two methods for correcting forearm wrong posture in track and field sports mentioned in the introduction are regarded as experimental group B and experimental group C, respectively. The actual data of the track and field athlete's wrong forearm posture movement data set is known. The three methods for correcting forearm posture in track and field sports are used to detect the forearm posture of the selected track and field athletes this time, and the detected wrong posture is corrected. The three methods for correcting forearm posture in track and field sports are compared. When detecting the forearm posture, The accuracy of the detection result; the error generated from the correct posture when correcting the posture of the forearm.

3.1   Experiment Preparation

In order to fully measure the performance of the proposed algorithm, two commonly used large-scale two-dimensional track and field athletes' forearm posture movement databases were selected as the test objects of this experimental method. The first is the BU-3DFE data released by American researchers, which is currently the most comprehensive database containing forearm posture movements. In addition, the original scan data in it also contains other very complex interferences, such as clothes, hair, expressions, hand gestures, etc. It obtains a variety of forearm movement postures generated by 100 participants in track and field. Each forearm movement posture is divided into 4 different degrees. Therefore, each individual has 25 corresponding three-dimensional forearm movements. Posture, a total of 2500 two-dimensional forearm motion posture images. The second database is the GavabDB three-dimensional forearm movement posture database, which contains 61 objects in the same track and field sports and 412 three-dimensional forearm movement posture pictures. In the two databases, under the same track and field sports, there are different forearm motion postures, which include various interferences other than the forearm motion, such as clothes, hair, and shoulders. For this experiment, a computer with Linux PC, Intel Dual 2.3 GHz processor, 2G memory and GeForce 8600 graphics card was selected.

The parameters of this exercise posture correction experiment include the number of scale spaces n, the forearm movement feature threshold T, and the forearm movement area feature threshold R. For the selection of the values of these three parameters, the GavabDB database is used to determine. When the number of scale spaces n is less than 10, as shown in Fig. 3, many relatively sharp parts will correspond to a relatively large forearm motion posture, which will result in many forearm candidate points, that is, the distinguishing ability of features is not strong enough at this time. However, when the number of scale spaces gradually increases, the original three-dimensional surface will gradually become smoother, and the features of the forearm part will become larger than other areas, that is, the ability to distinguish features of the forearm becomes more and

more obvious at this time. Generally speaking, when the number of scale spaces n is greater than 40, as shown in a in Fig. 3, the forearm movement characteristic threshold T gradually becomes relatively stable. Here we tested the accuracy of forearm motion detection in different scale spaces, and the results are shown in Fig. 3 a.

Based on the above analysis of the number of scale spaces n, the number of scale spaces n is selected as 40. The selection of the forearm movement characteristic threshold T is based on the analysis of the number n of the scale space as described above. When the number of scale spaces n is about 40, the forearm area gradually corresponds to the maximum threshold T, so the forearm features of all points are sorted from large to small, and the first N_T points are selected for the next clustering process. The selection of the N_T value depends on the resolution of the forearm motion posture, that is, the number of discrete forearm posture points. When the forearm posture has about 15,000 points, a stable and accurate result will be obtained when N_T is selected between 30 and 50. The relationship between the accurate detection rate and N_T is shown in b in Fig. 3. The feature threshold R of the forearm movement area will be selected in a similar way to the forearm movement feature threshold T. When the value of N_R is generally between 100 and 150, good posture correction results can be obtained. The c in Fig. 3 shows the relationship between the average angle error of the attitude correction and N_R. To sum up, it is the value of the three parameters selected in this experiment: the number of scale spaces n, the forearm movement feature threshold T, and the forearm movement area feature threshold R.

3.2   The First Set of Experimental Results

In the two three-dimensional motion forearm posture data of GavabDB and BU-3DFE, forearm posture detection and segmentation are performed. Based on the determination of this experiment, the three parameters of the experimental scale space number n, forearm movement characteristic threshold T and forearm movement area characteristic threshold R are determined, and the result of forearm posture detection is verified by artificial visual observation. Among them, the GavabDB database contains 412 forearm movement postures, and the BU-3DFE database contains 2500 forearm movement postures. The detection results of the wrong forearm posture are shown in Table 3.

It can be seen from Table 3 that the experimental group A has extremely high accuracy in detecting the forearm posture in the GavabDB database, and there is only one error point in the number of forearm movement postures detected. For the BU-3DFE database, because the database, The number of forearm postures is huge, but the number of errors in the detected forearm postures is only 30; and the experimental group B has a higher accuracy rate of forearm posture detection in the GavabDB database than that of the BU-3DFE database. The number of detection data is affected, but the number of detected forearm motion postures and the number of errors are much higher than that of experimental group A; experimental group C's accuracy of forearm posture detection in the GavabDB database is lower than that of the BU-3DFE database. Compared with experimental group B, it is more suitable for multiple database detection, but the number of detected forearm movement postures and the number of errors are much higher than experimental group A. It can be seen that the method for correcting forearm posture
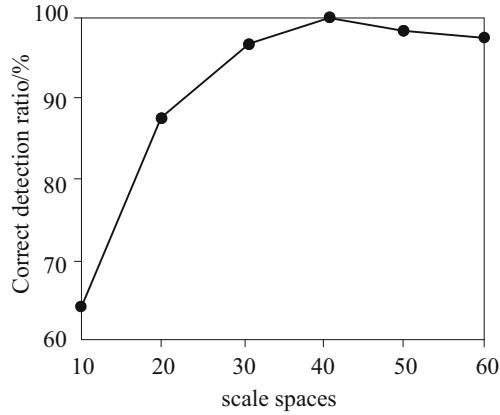
(a) Number of scale spaces
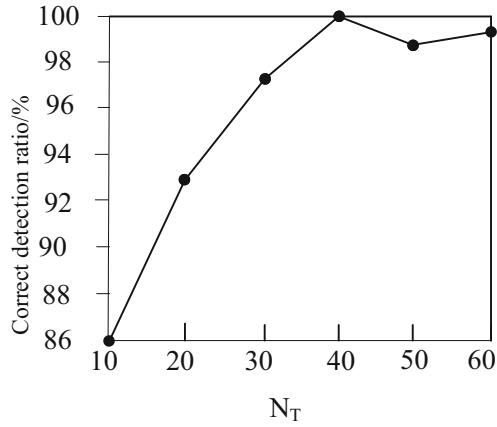
(b)Forearm motion characteristic threshold
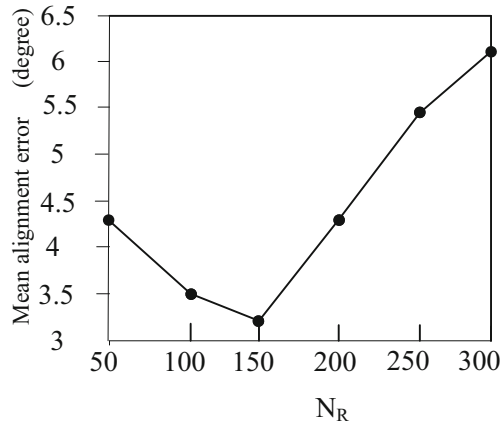
(c) Characteristic threshold of forearm motion area

Fig. 3. Determination of different parameters

Table 3. Accurate detection rate on the two databases

Method	Database	Number of detections	Accurate detection rate
Experiment Group A	GavabDB	411	99.8%
	BU-3DFE	2470	98.8%
Experiment Group B	GavabDB	369	89.6%
	BU-3DFE	2107	84.28%
Experiment Group C	GavabDB	351	85.0%
	BU-3DFE	2205	88.2%

errors in track and field sports in this study can more accurately detect forearm posture errors in track and field sports.

3.3   The Second Set of Experimental Results

On the basis of the first set of experiments, conduct the second set of experiments. Using the first set of experiments, the wrong posture in the forearm movement was detected, and three wrong posture correction methods were used to correct the wrong posture of the forearm. A total of 2881 successfully detected three-dimensional forearm postures from the two data are used to measure the error. It is found that the average error of elbow detection is 2.31mm, and the average angle error of forearm line positioning is 3.52°. The angular errors in the x-axis and y-axis directions of the forearm posture correction are 2.41° and 2.68°, respectively. At this time, according to the above detection results, three sets of correction methods are calculated to correct the angle of the wrong forearm posture and the error between the correct forearm posture angle. The corresponding error distribution diagram is shown in Fig. 4.

According to the analysis of Fig. 4, when the current number of arms is two, the forearm positioning error of experimental group A is 120 mm, that of experimental group B is 760 mm, and that of experimental group C is 660 mm; When the current number of arms is 8, the forearm positioning error of experimental group A is 60 mm, the forearm positioning error of experimental group B is 590 mm, and the forearm positioning error of experimental group C is 250 mm. Among them, experimental group A is the method of this paper, which shows that the forearm positioning error of this method is significantly smaller than other methods.It can be seen from Fig. 4 that the angle error of the wrong forearm posture generated in the two databases in the experiment group B and experiment group C is significantly higher than that in the experiment group A after correction. It can be seen that the method for correcting the wrong forearm posture in track and field sports this time has a higher correction rate for the wrong posture of the forearm.

Based on the above two sets of experiments, it can be seen that the large data correction method of forearm wrong posture in track and field sports in this study can quickly detect the large data of wrong forearm posture in track and field sports. After the big data of wrong posture is corrected, and The accuracy of correction is high.
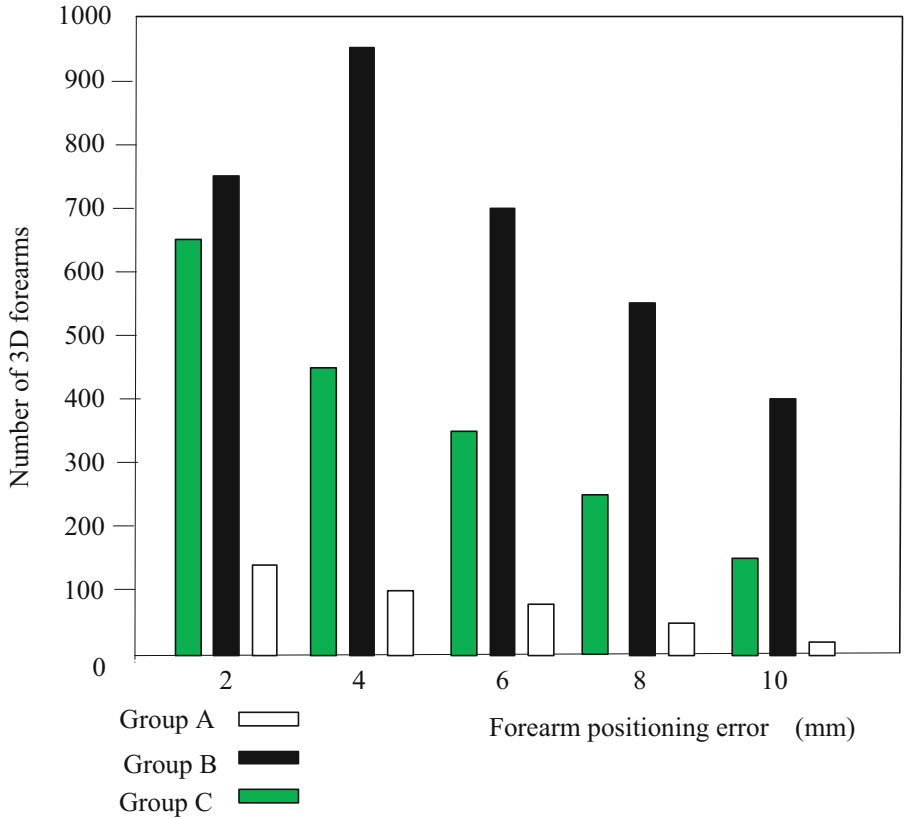
Fig. 4. Corresponding error distribution diagram

4 Concluding Remarks

Research on the big data correction method of forearm wrong posture in track and field, and make full use of the information detection characteristics of physical factors of stepwise regression to predict the posture of forearm in track and field. However, the large data correction method of forearm wrong posture in track and field in this study has not considered the presentation form of forearm movement posture. In different presentation forms, the big data of forearm wrong posture will affect the correction process. Therefore, in future research, the presentation form of forearm posture in track and field sports should be considered, and the big data correction method of forearm wrong posture should be analyzed in depth.

References

1. Lv, Z.: Simulation of grasping attitude control of robotic arm based on synovial error feedback. Int. J. Ind. Syst. Eng. **1**(1), 1 (2020)

2. Hsu, S.C., Li, T.H., Chan, H.Y., et al.: General dual wiimote 3D localization scheme: sensitivity analysis and its application on robot arm attitude detection. International Conference on Industrial Application Engineering, pp. 43–49 (2019)
3. Yi, H.A.O., Feifei, H.A.N., Shuai, Y.A.N.G., et al.: Analysis of the influence of AEB to the dummy motion posture in frontal impact. Automobile Parts **3**, 21–25 (2020)
4. Xiansheng, L., Shangrong, H.: Research on robot motion attitude measurement based on laser technology. Laser Journal **40**(3), 192–195 (2019)
5. Jingwei, C.A.O., Baoquan, Z.H.U.: Vehicle motion attitude measurement based on MEMS gyroscope and accelerometer. J. Chongqing Inst. Technol. **32**(4), 48–54 (2018)
6. Peng, C., Hai, Z., Dehong, G., et al.: Research on attitude control technology for subsurface buoy motion under near surface condition. J. Projectiles, Rockets, Missiles Guidance **38**(3), 164–167 (2018)
7. Liang, J.I.A., Xuewei, D.O.N.G.: Human motion gesture recognition based on BP neural network. Comput. Inf. Technol. **27**(6), 21–23 (2019)
8. Wen, R.E.N.: Research on motion aided training system based on attitude estimation. Electronic Design Eng. **27**(18), 149–152 (2019)
9. Fan, C., Qinghua, Z.: Study on sports injury attitude acquisition method based on 3D image analysis. Modern Electron. Technique **41**(4), 48–51 (2018)
10. Fang, W.U., Shiyong, W.A.N.G.: Experimental study on in-depth treatment of coking wastewater based on quadratic polynomial stepwise regression algorithm. Shanghai Chem. Ind. **45**(3), 20–24 (2020)
11. Zhu, J., Zhu, J., Wan, X., et al.: Object detection and localization in 3D environment by fusing raw fisheye image and attitude data. J. Vis. Commun. Image Represent. **59**(2), 128–139 (2019)
12. Xu, X., Sun, Y., Tian, X.: An attitude compensation method based on neural network using data from MEMS MARG sensors. In: 2019 IEEE 8th Data Driven Control and Learning Systems Conference (DDCLS). IEEE, (2019)
13. Gang, C.A., Yw, A., Yw, A., et al.: Detumbling strategy based on friction control of dual-arm space robot for capturing tumbling target. Chin. J. Aeronaut. **33**(3), 1093–1106 (2020)

Intelligent Sharing Technology of Mobile Medical Dynamic Data Based on Internet of Things

Hai-bo Zhang[1(✉)], Xiu-juan Duan[2], and Jian-mei Sun[1]

[1] Dalian University of Science and Technology, Dalian 116011, China
jhjsdf89@163.com
[2] Financial Department, College of Humanities and Information, Changchun University of Technology, Changchun 130122, China

Abstract. In order to solve the problem of long sharing time and low efficiency caused by a large number of false data in cloud computing, the intelligent sharing technology of mobile medical dynamic data based on Internet of Things is proposed. This paper constructs a data cache model based on diagonal priority, allocates bandwidth priority to high priority data, and avoids data transmission interruption in Internet of Things. P2P algorithm is used to get the request queue length of the service node through message interface in the interactive process between the request node and the service node. Establish an intelligent sharing communication protocol for the internet of things, improve subjective and objective factors of mobile medical institutions, build a sharing constraint mechanism, and ensure that mobile medical institutions at different ends receive complete data smoothly. The experimental results show that the shortest sharing time is 2.1s, the highest sharing efficiency is 99%, and the technology has a good sharing effect.

Keyword: Internet of Things · Mobile healthcare · Dynamic data · Intelligent sharing · P2P scheduling

1 Introduction

Mobile medical care refers to the collection, reception and pretreatment of physiological parameters such as ECG, blood pressure, pulse wave and heart rate of residents on an intelligent mobile platform by using Bluetooth wireless technology to connect various household medical electronic devices [1]. Combined with the diagnosis of doctors, it can make up the defect of medical monitoring equipment. SSL and TLS are adopted to realize the security strategy of the system, and the HL7 message exchange standard is integrated with the existing medical information system to provide data support for collaborative diagnosis among hospitals [2]. Using mobile medicine, health centers can provide remote, real-time health monitoring and rapid and timely assistance in the event of an emergency. However, there are a number of security and privacy issues that need to be addressed before sharing personal health information between users. In MHSN, the

patient's PHI is privacy-sensitive, and even if the patient's identity is kept private, only patients with the same condition can view the shared information. However, because of the randomness of patient movement, this network becomes a network of opportunities [3]. Therefore, the main security and privacy challenges in current mobile medical social networks are how to achieve the secret disease matching and the security and privacy protection in the communication process. Intelligent sharing of mobile medical dynamic data based on cloud computing, with infrastructure as a service, platform as a service and software as a service, enables some specific application functions to be encapsulated as services to users through their own dynamic extension of resources and fault-tolerant management functions. [4] With the wide application of cloud computing, the security problems caused by cloud computing have become increasingly prominent, and become a key factor restricting the further development and implementation of cloud computing [5]. Specifically, because in cloud computing environment, due to hardware, software and human uncertainties, cloud servers will inevitably erroneously modify, or even erase the data of users. In the process of data use, due to the characteristics of cloud computing outsourcing, the whole process of data retrieval and retrieval is performed in the cloud. The cloud server can control the whole retrieval process and determine the final results, and the data consumer will not be able to directly verify the integrity and correctness of the data returned by the cloud server. Therefore, the intelligent sharing technology of mobile medical dynamic data based on Internet of Things is proposed.

Using centralized P2P data scheduling, a data caching model based on diagonal priority is constructed, data is allocated according to the priority level of bandwidth, and data transmission is completed in wireless cellular network by using P2P technology. Based on the Internet of Things, an intelligent sharing communication protocol of Internet of Things is constructed under constraint mechanism, which makes data communication smooth. The intelligent sharing of mobile medical dynamic data based on Internet of Things is completed.

2   Prioritization of P2P Intelligent Data Sharing Based on Internet of Things

Because of the variety of wireless channel, high loss rate of nodes and different network speed, and the small number of nodes, the quality of service of streaming media in mobile medical service is poor. In the Internet of Things, P2P network is the core architecture of data transmission, which is used by many online video and large websites. But because of the characteristics of mobile network, the traditional P2P technology can not be directly applied to mobile medical streaming media transmission. Centralized data scheduling of P2P in fixed network is the core technology. It calculates the bandwidth, delay and packet loss rate of every node in the network, and selects some nodes from neighbor set to transmit data. In order to improve the precision of mobile medical dynamic data sharing, data sharing priority is designed based on the data cache model.

2.1 Construction of Data Cache Model Based on Twill Priority

The cache model is divided into two parts in block coordinates: the decision scheduling area and the scarcity scheduling area. The scheduling algorithm of the decision scheduling area assigns an urgency degree to each block of data, which is determined by layer coordinates and time coordinates. The closer to the playback time, the lower the number of layers, the higher the urgency, and the system gives priority to dispatching the data with the higher urgency; in the scarce dispatching area, bandwidth is allocated to the scarce data first, that is, the data blocks containing the small number of nodes of the data, with the aim of reducing the probability of frequent requests to the server due to node departures, which improves the absence of scarce resources due to high node leave rates in mobile networks [6–9].

The two regions have different scheduling algorithms, but the network bandwidth is uniformly allocated according to the priority of the entire cache, that is, all data blocks in the cache are periodically scanned and the bandwidth is allocated first to the higher-priority data [10, 11]. The build based on the twill priority data cache model is shown in Fig. 1.
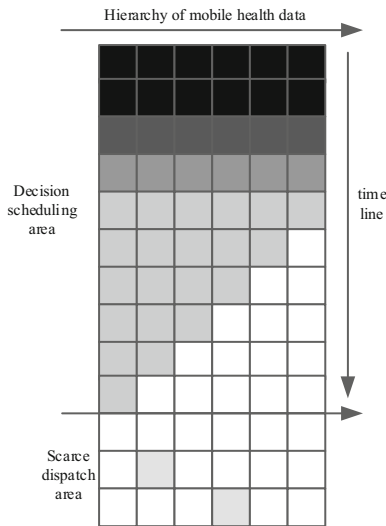
Fig. 1. Data cache model based on diagonal priority

As shown in Fig. 1, the two priority sizes can overlap, so the allocation of bandwidth is not limited to the region. The darker the color, the higher the priority, the lighter the color, the lower the priority. In the decision and scheduling area, the trend is gradually decreasing in the right direction, similar to the sub-twill. Cyclic loop computation, with each loop computing the priority of the currently available bandwidth, the decision scheduling area and the scarcity scheduling area for each data, then allocating the available bandwidth to the higher priority data, which is then allocated to the next scheduling phase until the scheduling terminates [12–14].

(1) Decision-making and scheduling area

Because of such factors as limited and drastically changing bandwidth, bad channel, and high node loss rate in mobile networks, dynamic data cannot be guaranteed to reach all at playback time [15]. Because this area is close to the playback time, the distance of playback time and the number of layers are used as the metrics of the data priority. The priority of the data in the area is allocated in the way of twill, near the playpoint, the priority of the low level data is high, far away from the playpoint, the priority of the high level data is low [16]. The strategy not only reduces the probability of avoiding data transmission interruption in the IOT, but also satisfies the definition requirement in the mobile environment, that is, to balance the data transmission fluency in the IOT [17]. The priority of the scarce data in the scheduling area is higher than that of the high-level data in the partial decision scheduling area.

(2) Scarce dispatch areas

The scheduling principle of this area is to allocate the bandwidth to the scarce data, which refers to the data blocks containing fewer nodes in the network. The aim of scarcity priority scheduling is to minimize frequent access to servers due to node departures and to reduce server load [18]. This algorithm improves the utilization of bandwidth and increases the system throughput. The principle of scarcity scheduling is that the lower the number of nodes in a region, the higher the priority. Compared with the decision scheduling area, the priority of this area will be higher than the high-level data of partial decision scheduling area [19].

2.2 Data Scheduling Algorithm Based on P2P Technology

Mobile P2P technology is the use of P2P technology in wireless cellular network for data transmission, it is P2P and mobile communication technology interdisciplinary. The current 4G technology makes the mobile P2P get the rapid development, it is widely used in the cellular network streaming media transmission and mobile network data interaction. P2P network is a kind of application layer protocol, which realizes distributed peer-to-peer relationship among peers and resources are shared by all peers and the peers can transmit data to each other. The node of P2P network is not only the data requester, but also the data provider. Peer to peer transmission is implemented between nodes, and each node acts as a server.

A P2P network is a distributed architecture that differs from a traditional client/server (Client/Server, C/S). As the center of C/S structure, the central server not only provides data resources to all nodes, but also has the basic information of nodes and networks in the system. In the traditional model, the client can only request data from the server and receive data from the server, and the server can not only receive, process and forward data, but also control the overall network situation. But the server is easy to become the network limit, when the node number increases gradually, the server's computation and the network bandwidth increase gradually, when achieves the bottleneck network efficiency to reduce suddenly. If a server is compromised by a virus or attack, the entire network faces disaster. At the same time, due to the limited single-point processing capacity of the server, the huge client requests are often not processed in time, so the network resource utilization is low, and P2P network does not have this problem.

The algorithm realizes the data communication between the request node and the service node through the interaction between the request node and the service node. The requesting node stores data such as the number of caches and historical bandwidth in the local information database, and obtains the request queue length of the service node through the message interface [20]. The peer communication in the P2P algorithm is shown in Fig. 2.
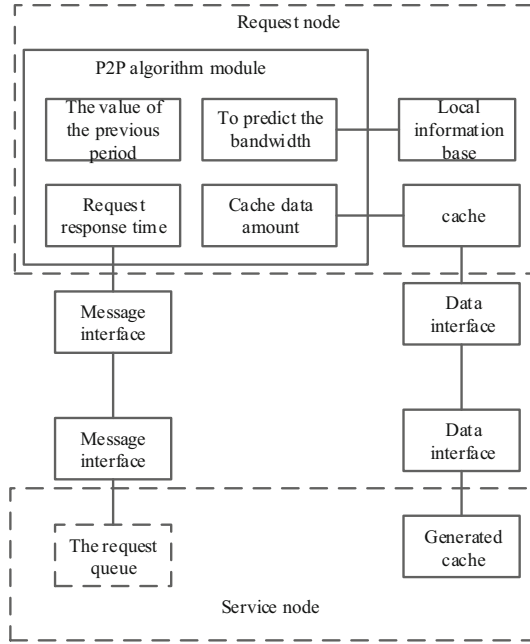
Fig. 2. Peer communication in P2P algorithm

As can be seen from Fig. 2, the algorithm dynamically adjusts the number of requests and the object of the request node by defining four variables Q[n], T[n], D[n], and L[n] to get the current network service status. Q[n] represents the number of data blocks requested from node i by node k for the nth time, and the calculation formula is:

$$Q[n] = Q[n-1] + \frac{\alpha}{m}D[n] + \frac{\beta}{m}S[n] + \delta\frac{U}{L}T[n] \tag{1}$$

As shown in formula (1), Q[n] is determined by the number of requests for data blocks at the previous time, the estimated bandwidth of the node, the amount of data in the cache and the response time of the node, α, β, δ being constant; m the size of the set of service nodes [21].

S[n] represents the data size of the buffer for the nth request from node k, calculated by:

$$S[n] = \frac{S_m - S[n-1]}{S_m} \tag{2}$$

In formula (2), S_m is the threshold for the cache size, and if the cache size at the previous moment was greater than this value, $S[n]$ is negative, reducing the number of data requests for Q[n] and vice versa. Proportional to $S[n]$ and Q[n], that is, the fewer the number of service nodes, the greater the impact of $S[n]$ on Q[n], raising the priority of scarce resources, increasing system throughput.

T[n] represents the response time of the nth request, including queuing delay and sending delay. The calculation formula is as follows:

$$T[n] = 1 - T[n-1] \tag{3}$$

As shown in formula (3), the current request response time is inversely proportional to the response time of the previous cycle. $T[n]$ is inversely proportional to the ring, and proportional to the queue length of the $L[n]$-node k request service. That is, the greater the uplink bandwidth, the greater the impact of $T[n]$ on $T[n]$, and the higher the uplink bandwidth utilization; Increasing the queue length of the request service reduces the value of Q[n] by $T[n]$, so the system can reduce the probability of service queue congestion.

$D[n]$ Represents the predicted bandwidth of node k nth direction node j, the calculation formula is:

$$D[n] = \sum_{t=1}^{n} \lambda_i D[n-1] \tag{4}$$

In formula (4), $D[n]$ records the information sent by the service node to the requesting node in the most recent m cycles, and estimates the current bandwidth based on the amount sent, thus predicting the amount requested from the service node. As shown in formula (4), n is a positive integer, and the closer the period, the greater the impact of bandwidth prediction, which follows the characteristics of network bandwidth changes. $D[n]$ is proportional to m, and the smaller the number of service nodes, the greater the impact of $D[n]$ on Q[n].

3   Research on Intelligent Sharing Technology Based on Internet of Things

3.1   Establishment of Data Sharing Mechanism Based on Internet of Things

The mobile medical dynamic data sharing based on Internet of things mainly follows the intelligent sharing communication protocol, as shown in Fig. 3.

As can be seen from Fig. 3, the real-time sharing of mobile medical dynamic data based on the Internet of Things meets the following conditions: the sender and the receiver have the same purposes, and the sharing results can achieve the desired effect during the whole data sharing process [22].

Therefore, in order to improve the subjective and objective factors of mobile medical institutions, effective management and shared constraint mechanism should be established, as shown in Fig. 4.
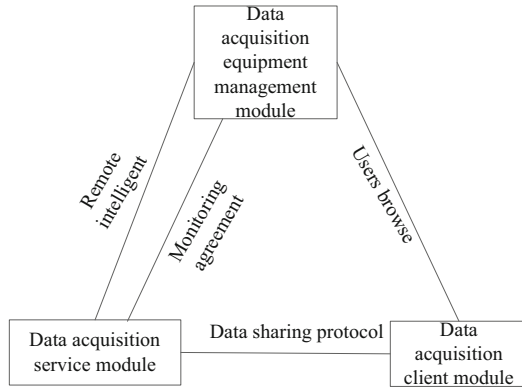
Fig. 3. Networked intelligent shared communication protocol
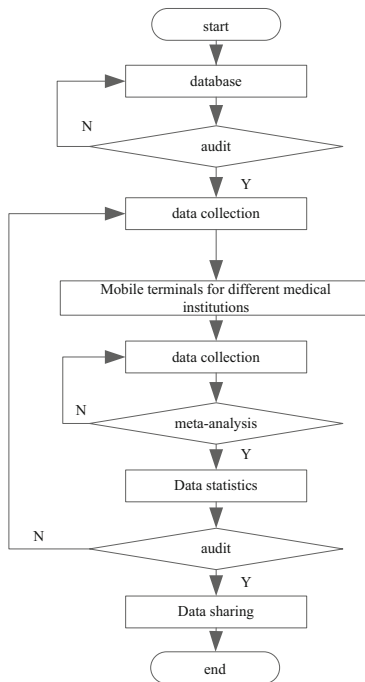
Fig. 4. Data sharing constraints based on Internet of Things

Under the constraint mechanism shown in Fig. 4, the dynamic transmission of mobile medical data based on Internet of Things needs to go through three steps: data integration, data sending and data receiving. Therefore, a constraint mechanism is needed to promote the effective sharing of data within the main transformer. Through the data sharing constraint mechanism, the mobile medical institutions based on the Internet of Things

can be urged to analyze from the common cognitive perspective, carry out reasonable transmission, and through accurate data transmission [23], the complete data can be successfully received by different mobile medical institutions, and then the dynamic data sharing of mobile medical services based on the Internet of Things can be realized.

Because of the large amount of mobile medical dynamic data based on Internet of Things, the sharing mode should be selected. For the complementary judgment matrix under different modes, in order to obtain the optimal shared results, the following steps are required:

Step 1: Set thresholds and iteration parameters before data is shared;

Step 2: According to the complementary judgment matrix to obtain the minimum deviation and the optimal weight vector;

Step 3: according to the following formula to obtain complementary judgment matrix consistency index;

Step 4: Output complementary judgment matrix that is the optimal weight vector, according to the vector of data security sharing mode to sort, in order to get the best sharing results.

3.2   Shared Process Design Based on Internet of Things

In this paper, a medical information sharing system is proposed, which can realize the information sharing by using the cloud to store the patient's diagnosis and treatment data, and allocating and managing the related authority of the institution. The sharing process based on the Internet of Things is shown in Fig. 5.

Under the constraint mechanism shown in Fig. 4, the dynamic transmission of mobile medical data based on Internet of Things needs to go through three steps: data integration, data sending and data receiving. Therefore, a constraint mechanism is needed to promote the effective sharing of data within the main transformer. Through the data sharing constraint mechanism, the mobile medical institutions based on the Internet of Things can be urged to analyze from the common cognitive perspective, carry out reasonable transmission, and through accurate data transmission, the complete data can be successfully received by different mobile medical institutions, and then the dynamic data sharing of mobile medical services based on the Internet of Things can be realized.

Because of the large amount of mobile medical dynamic data based on Internet of Things, the sharing mode should be selected. For the complementary judgment matrix under different modes, in order to obtain the optimal shared results, the following steps are required:

Step 1: Set thresholds and iteration parameters before data is shared;

Step 2: According to the complementary judgment matrix to obtain the minimum deviation and the optimal weight vector;

Step 3: according to the following formula to obtain complementary judgment matrix consistency index;

Step 4: Output complementary judgment matrix that is the optimal weight vector, according to the vector of data security sharing mode to sort, in order to get the best sharing results.

Fig. 5. Data sharing constraints based on Internet of Things

3.3   Shared Process Design Based on Internet of Things

In this paper, a medical information sharing system is proposed, which can realize the information sharing by using the cloud to store the patient's diagnosis and treatment data, and allocating and managing the related authority of the institution. The sharing process based on the Internet of Things is shown in Fig. 6.

The detailed steps are:

Step 1: The terminal device sends a document request for obtaining a specific identity to the server. The request includes: identity identification information and document name;

Step 2: The server searches the database for documents matching the request. The technical solution provides a complete way of data sharing. No matter what system the operating system of the terminal device belongs to, it can automatically generate the corresponding document structure and list and display them, which makes it convenient for users or doctors to query historical medical data. Specifically, because the documents are generated by the terminal according to the document number, and the information data is filled by the terminal device itself according to the received data content in the
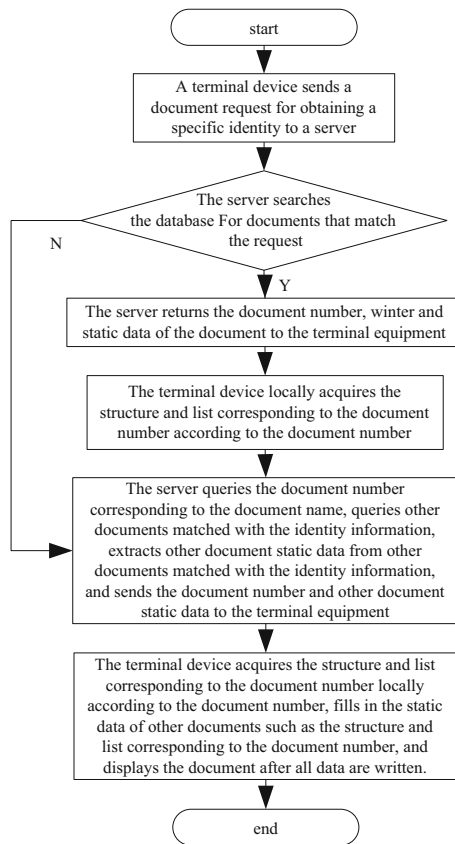
Fig. 6. Sharing process based on Internet of Things

corresponding document, so it will not be incompatible with the system. In addition, the locally generated documents also reduce the problem that the documents do not match with the terminal device. For example, the document interface displayed on the PC is displayed on the mobile phone, and it is easy to cause the problem of mismatch due to the size of the display screen, while the locally generated documents only extract the local documents corresponding to the document number according to the document number, so it will not cause such problem, and the system does not need to distinguish the application system of the terminal device and uniformly issue the document number. In addition, the static data of other documents can be invoked from other documents that match the identity information, even if no matching document can be queried in the above way. For static data, because of its stability and the characteristics of only matching the identity information, it can become the static information of other documents instead of the static information of this document, so that the operators such as doctors do not have to re-enter the user's static data during this operation, thus saving doctors' time and improving their work efficiency.

Step 3: The server returns the document number, static data and dynamic data to the terminal device;

Step 4: The terminal device acquires the structure and list corresponding to the document number locally according to the document number; fills in the static data and dynamic data of the document into the structure and list; for example, there are many rows of data in the dynamic data of the document, the server extracts the description of each row of records according to the structure of many rows of records, forms a list of row of descriptions, sends the return data ID into a list of row of descriptions to the terminal device, and the terminal device writes the list of row of descriptions into the document structure and list line by line, and displays the document when all data are written;

Step 5: The server inquires about and obtains the document number corresponding to the document name, inquires about other documents matching the identity information, extracts other static data of other documents matching the identity information, and sends the document number and other static data of documents to the terminal device;

Step 6: The terminal device obtains the structure and list corresponding to the document number locally according to the document number, fills in the static data of other documents into the structure and list corresponding to the document number, and displays the document after all data is written.

4   Experiment

In order to ensure the reliability of the experiment, we set up a fixed experimental device to verify the rationality of the research on the intelligent sharing technology of mobile medical dynamic data.

4.1   Construction of Network Open Source Simulation Platform

OverSim is an open source P2P network simulation framework based on OMNeT++ (Objective Modular Network Testbed in C++). It supports many P2P protocols, including not only structured overlay network but also unstructured overlay network. OverSim platform has three layers, which are application layer, coverage layer and foundation layer from top to bottom.

(1)   Application layer

The application layer is a P2P application development system for developers. The platform provides a number of perfect application APP instances, which send detection information periodically, obtain the data of the protocol bottom layer through the basic layer or the overlay layer, and record the information delay and hops in the overlay layer. The application layer provides the perfect system simulation index for developers to use together.

This level provides a data sharing system for communication terminals in a medical system, as shown in Fig. 7.
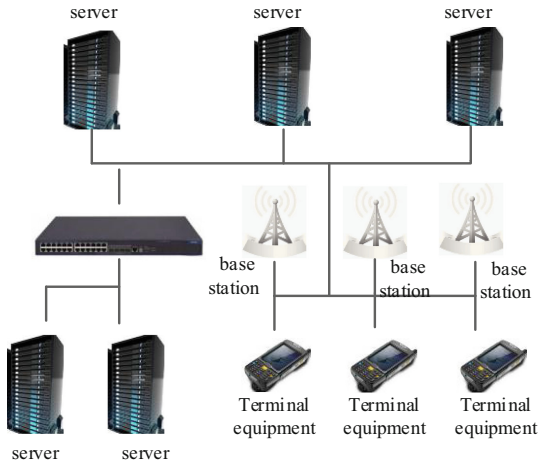
Fig. 7. Data sharing system

As can be seen from Fig. 6, the system is composed of two parts, the terminal device and the server. The terminal device is used for sending a document request to the server for obtaining a specific identity, including identity information and document name, and is used for obtaining the structure and list corresponding to the document number locally according to the document number. The static data and dynamic data of the document are filled in the structure and list, and the document is displayed when all the data is written; the server is used to retrieve whether there is a document matching the request in the database, such as query for a document matching the request, and return the document number, static data and dynamic data of the document to the terminal device. The server is used for querying documents without the requested match, querying for the document number corresponding to the document name, querying for other documents matching the identity information, extracting static data of other documents in other documents matching the identity information, and sending the document number and static data of other documents to the terminal device.

(2) Overburden

The overlay layer is based on the basic layer, which provides the support of P2P protocol for the application layer. This layer has the following four functions:

First, information processing of statistical data in overburden layer. The layer provides an RPC (Remote Procedure Call) interface to handle packet timeout and packet retransmission caused by packet loss. The interface can also count the data sent, received, forwarded and lost by nodes.

Second, visualization of topology structure;

Thirdly, the function of node lookup is provided, which is to find the corresponding node according to the node behavior in the topology.

Fourthly, the layer provides API protocol support for the application layer. Each layer protocol must provide at least one kind of keyword -based routing interface. For the distributed hash table based protocol, the same service can be provided to

realize the communication between the two layers. For the protocol that does not support distributed hash table, the key word based routing interface can also be used to realize the communication.

(3) Basic layer

This layer has three network models: iNET(integrated Network Enhanced Telemetry), simple and singleHost. The base layer provides routing table, IP address assignment and MAC communication for network simulation. It is the bottom layer of OverSim module.

4.2 Experimental Results and Analysis

The sharing effect is mainly from the sharing time and the sharing efficiency two angles to carry on the experimental verification analysis.

Shared Time Experiment Analysis

Based on the generation stage data, operation and maintenance stage data and scrapping stage data, the training set boundary decision graph is analyzed, and the result is shown in Fig. 8.
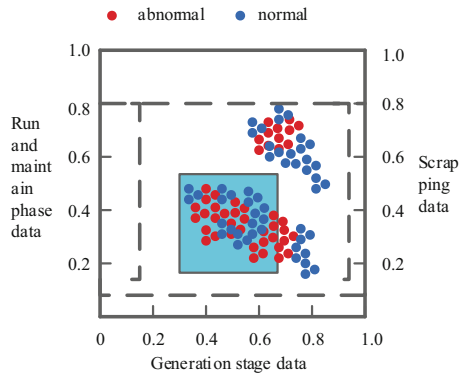
Fig. 8. Training set boundary decision graph

As can be seen from Fig. 8, the data in the boundary of the training set is the normal operation of mobile medical institutions. When the monitored data is beyond the boundary, the data can be identified as normal data and need maintenance. Cloud -based and IoT -based approaches are used to compare anomaly and normal data sharing, and the results are shown in Fig. 9.

As can be seen from Fig. 8, the abnormal data: when the data amount is 0.21×104 bit, the sharing time consumed by the traditional method is 20 s, while when the data amount is 0.62×104 bit, the sharing time consumed by the method studied is 18 s; when the data amount is 0.21×104 bit, the sharing time consumed by the method studied is 24 s, and when the data amount is 0.42×104 bit, the sharing time consumed by the method studied is 20 s. Once the shared data is found to be abnormal using the method
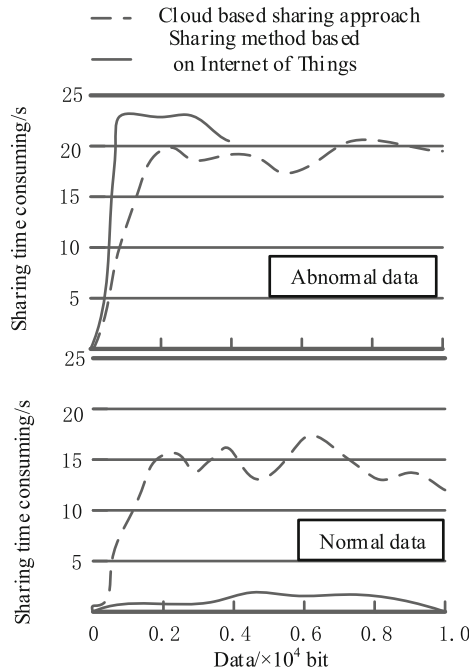
Fig. 9. Comparison of time-consuming results between the two methods

studied, the sharing is stopped immediately. But the cloud computing takes a long time to share the abnormal data. Normal data: when the data quantity is 0.21×104 bit, the sharing time consumed by the traditional method is 15 s, while when the data quantity is 0.62×104 bit, the sharing time consumed by the method studied is 19 s; when the data quantity is 0.21×104 bit, the sharing time consumed by the method studied is 2.1 s, and when the data quantity is 0.42×104, the sharing time consumed by the method studied is 3.0 s.

Experimental Analysis of Sharing Efficiency

(1) Comparative analysis of sharing efficiency under data interference

①Creating false information
    Create false information to cause some interference to data sharing, based on this situation, respectively, using cloud -based and IoT -based approach to analyze the efficiency of data sharing, the result is shown in Fig. 10.
    As can be seen from Fig. 9: when the data volume is 10×104 bit, the sharing efficiency of the method based on cloud computing is 43%, and the sharing efficiency of the method based on Internet of things is 88%; when the data volume is 30×104 bit, the sharing efficiency of the method based on cloud computing is
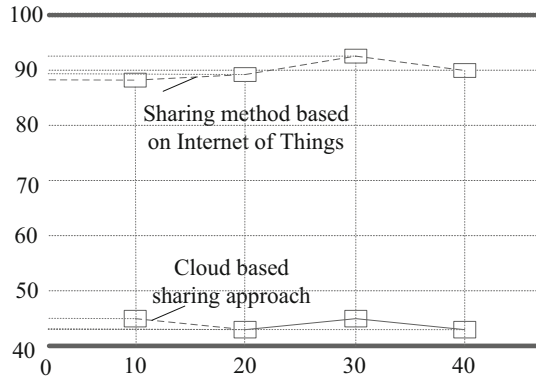
Fig. 10. Comparison of sharing efficiency of two technologies in case of creating false information

43%, and the sharing efficiency of the method based on Internet of things is 94%. Therefore, the method based on Internet of things is more efficient.

②Multiple attack points

Because the shared system is composed of a large number of data collected and processed, it is easy to be attacked by multiple attack points, resulting in more data interference. Based on this situation, the sharing efficiency of the two technologies is compared and the result is shown in Fig. 11.
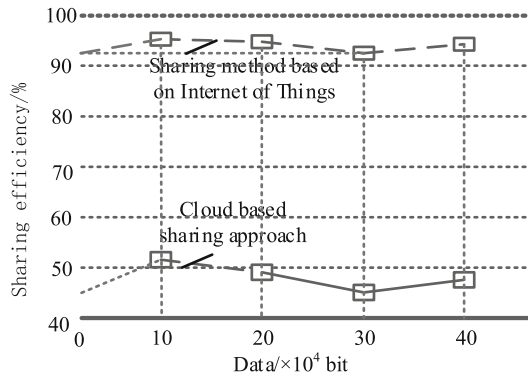
Fig. 11. Comparison of the sharing efficiency of the two technologies for multiple attack points

As can be seen in Fig. 11, both sharing efficiencies are on the rise compared with disinformation creation, and the efficiency of data sharing using the IoT -based sharing approach is consistently above 90%. When the data volume is 10×104bit, the sharing efficiency based on cloud computing method is 52%, and the sharing efficiency based on IoT method is 95%; when the data volume is 30×104bit, the sharing efficiency based on cloud computing method is 45%, and the sharing efficiency based on IoT method is 92%,

The results show that the sharing method based on Internet of Things has higher sharing efficiency in the case of data interference.

(2) Comparative analysis of sharing efficiency under normal circumstances
     In order to further verify the rationality of this method, the interference data are eliminated, as shown in Fig. 12.
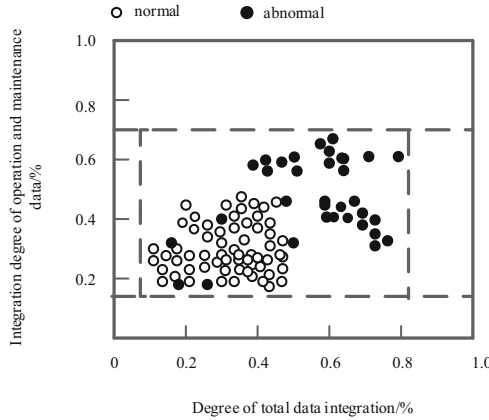
Fig. 12. Eliminating interference data

As shown in Fig. 11, the data works fine within the range of the host training set. If the monitored data is abnormal, it indicates an attack. If the monitoring data is correct, it needs to be corrected. On this basis, the sharing efficiency of the two sharing technologies is compared and analyzed, as shown in Table 1.

Table 1. Comparison and analysis of sharing efficiency of two technologies under elimination of interference data

Mobile medical data number	V2	V3
#1	82%	99%
#2	82%	99%
#3	81%	98%
#4	78%	98%
#5	78%	98%
#6	75%	97%
#7	75%	97%
#8	72%	98%
#9	70%	98%

From the Table 1, we can see that the efficiency of the two technologies is higher than 70% in the case of no interference data, which has good sharing efficiency.

5   Concluding Remarks

Due to some shortcomings of mobile medical network, such as high dynamic of nodes and poor channel quality, the data scheduling algorithm in P2P can achieve good results in mobile medical data intelligent sharing. Combined with the new data scheduling algorithm to overcome these shortcomings effectively in the mobile network, the experimental results show that the sharing time and sharing efficiency has higher advantages.

With the further development of mobile medical care, the intelligent sharing technology of mobile medical dynamic data based on Internet of Things will develop in the direction of integration. This technology will integrate all kinds of scene applications, so the data scheduling algorithm needs to be switched according to different applications. Since the proposed scheduling algorithm is based on the priority caching model, the future work will focus on the P2P algorithm, so that the series of algorithms can adapt to the needs of the integrated platform. Because this algorithm uses priority caching model, it has certain research value and research foundation.

It is concluded that the intelligent sharing technology of mobile medical dynamic data is suitable for the development of mobile medical platform, intelligence and diversification, and has far-reaching application value. In addition, the follow-up research can focus on the mobile medical dynamic data sharing mechanism, and improve the quality of mobile medical data transmission from many aspects.

References

1. Cheng, H., Shi, Y., Wu, L., et al.: An intelligent scheme for big data recovery in internet of things based on multi-attribute assistance and extremely randomized trees. Inf. Sci. **557**(4), 66–83 (2021)
2. Toh, S.: Analytic and data sharing options in real-world multidatabase studies of comparative effectiveness and safety of medical products. Clin. Pharmacol. Ther. **107**(4), 834–842 (2019)
3. Cole, C.L., Soumitra, S., Sarah, R., et al.: Ten principles for data sharing and commercialization. J. Am. Med. Inform. Assoc. **3**(3), 646–649 (2020)
4. Oliveira, J.L., Trifan, A., Silva, L.B.: EMIF Catalogue: A collaborative platform for sharing and reusing biomedical data. Int. J. Med. Inform. **126**, 35–45 (2019)
5. Elizabeth, F., Malcolm, O., Lamiece, H., et al.: Should free-text data in electronic medical records be shared for research? a citizens' jury study in the UK. J. Med. Ethics **46**(6), 367–377 (2020)
6. Dubovitskaya, A., et al.: Applications of blockchain technology for data-sharing in oncology: results from a systematic literature review. Oncology **98**(6), 1–9 (2019)
7. Chandrasekaran, R., Sankaranarayanan, B., Pendergrass, J.: Unfulfilled promises of health information exchange: what inhibits ambulatory clinics from electronically sharing health information? Int. J. Med. Informatics **149**(1), 104418 (2021)
8. Bahr, B., Cunningham, W.J., Dittrich, B., et al.: Data on sharing data. Nat. Phys. **15**(8), 724–725 (2019)

9. Saito, M.A., Bertrand, E.M., Duffy, M.E., et al.: Progress and challenges in ocean metapro-teomics and proposed best practices for data sharing. J. Proteome Res. **18**(4), 1461–1476 (2019)

10. Du, M., Chen, Q., Chen, J., et al.: An optimized consortium blockchain for medical information sharing. IEEE Trans. Eng. Manag. **68**(6), 1677–1689 (2020)

11. Fylan, F., Fylan, B.: Co-creating social licence for sharing health and care data. Int. J. Med. Informatics **149**(4), 104439 (2021)

12. Dexheimer, J.W., Greiner, M.V., Beal, S.J., et al.: Sharing personal health record data elements in protective custody: youth and stakeholder perspectives. J. Am. Med. Inform. Assoc. **26**(8–9), 8–9 (2019)

13. Kalkman, S., Delden, J.V., Banerjee, A., et al.: Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. J. Med. Ethics, medethics-2019–105651 (2019)

14. Shortreed, S.M., Cook, A.J., Yates, C.R., et al.: Challenges and opportunities for using big health care data to advance medical science and public health. Am. J. Epidemiol. **188**(5), 851–861 (2019)

15. Shu, D., Yoshida, K., Fireman, B.H., et al.: Inverse probability weighted Cox model in multi-site studies without sharing individual-level data. Stat. Methods Med. Res. **29**(6), 096228021986974 (2019)

16. Spector-Bagdady, K., Hutchinson, R., Kaleba, E.O., et al.: Sharing health data and biospec-imens with industry — a principle-driven, practical approach. N. Engl. J. Med. **382**(22), 2072–2075 (2020)

17. Timothy, B., Yan, Y., Thomas, S., et al.: Piloting a model-to-data approach to enable predictive analytics in health care through patient mortality prediction. J. Am. Med. Inform. Assoc. **27**(9), 1393–1400 (2020)

18. Sheehan, M., Friesen, P., Balmer, A., et al.: Trust, trustworthiness and sharing patient data for research. Journal of Medical Ethics, (3):medethics-2019–106048 (2020)

19. Kuo, Y.F., Raji, M.A., Lin, Y.L., et al.: Use of medicare data to identify team-based primary care. Med. Care **57**(11), 905–912 (2019)

20. Fritz, Z., et al.: Patient engagement or information overload: patient and physician views on sharing the medical record in the acute setting. Clin. Med. (Lond.) **19**(5), 386–391 (2019)

21. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mobile Networks Appl. **24**(1), 5–17 (2018). https://doi.org/10.1007/s11036-018-1134-8

22. Liu, S., Bai, W., Liu, G., et al.: Parallel fractal compression method for big video data. Complexity **2018**, 2016976 (2018)

23. Liu, S., He, T., Dai, J.: A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese. Mobile Networks Appl. **26**(5), 1891–1903 (2021). https://doi.org/10.1007/s11036-020-01725-x

Multi Channel Data Encryption Transmission Algorithm of Medical Internet of Things Based on Improved MQTT Protocol

Hai-bo Zhang[1]([✉]), Xiu-juan Duan[2], and Jian-mei Sun[1]

[1] Dalian University of Science and Technology, Dalian 116011, China
jhjsdf89@163.com
[2] Financial Department, College of Humanities and Information, Changchun University of Technology, Changchun 130122, China

Abstract. Since the concept of the Internet of things was put forward, it has attracted the attention of all countries in the world, and has become a technology developed by all countries and organizations in the world. With the development of Internet of things technology and application, information sharing is realized between things on the Internet of things, so the safe transmission of data becomes particularly important. The mainstream MQTT (Message Queuing Telemetry Transport) protocol has obvious advantages in the application of Internet of things network communication, but the protocol has a very big defect, its data transmission default is not encrypted, unable to guarantee the security of data. To solve this problem, this paper analyzes the security of the Internet of things, improves the MQTT protocol, and obtains a new multi-channel data encryption transmission algorithm of the Internet of things.

Keyword: Internet of Things · Multi channel data · MQTT protocol · Encrypted transmission

1 Introduction

The core of the Internet of things is to realize the information sharing between things, and data transmission, as the core technology to realize the communication between things, naturally becomes the research hotspot of the Internet of things technology. As a key element in the process of data transmission, data transmission protocol has been concerned by researchers from all walks of life in the development of Internet of things technology [1, 2]. With the continuous development of the research work of Internet of things in recent years, the big companies in the IT industry and related research institutions have developed network communication protocols suitable for the Internet of things environment. Among them, MQTT protocol has obvious advantages in the application of Internet of things network communication. It optimizes the application situation of Internet of things equipment resources and unstable network environment, which makes it low cost, high reliability and can effectively reduce the flow and power

consumption of terminal devices. At present, the application research of MQTT protocol at home and abroad mainly focuses on the message transmission of Internet of things, but MQTT protocol has a very big defect. Its data transmission is not encrypted by default and the data security can not be guaranteed. Therefore, based on the improved MQTT protocol, a new encryption algorithm is proposed for the multi-channel data of the medical Internet of things, so as to realize the data security transmission of the Internet of things.

2   Security Analysis of Internet of Things

In recent years, the Internet of things has made a breakthrough, and many applications based on the Internet of things have gradually emerged, bringing convenience to people's lives. At the same time, the security problems of the Internet of things are gradually highlighted. Whether the security problem of the Internet of things can be effectively solved is directly related to whether the Internet of things can be applied on a large scale. This paper first introduces the hierarchical model of the Internet of things, and then analyzes the security problems of the Internet of things and the security mechanisms that can be used [3, 4].

"Overview of the Internet of things" standard puts forward the reference model of the Internet of things. From bottom to top, it can be divided into four layers, namely perception layer, network layer, service layer, application support layer (processing layer) and application layer. Each layer has corresponding security and management functions. The reference model of Internet of things is shown in Fig. 1.
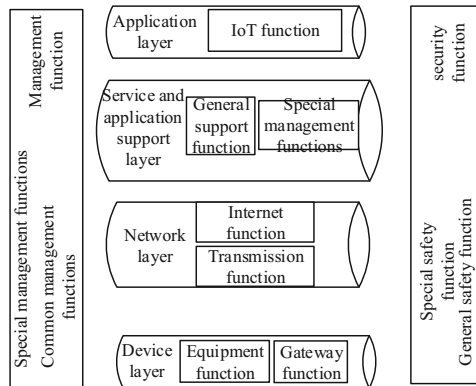
Fig. 1.   Reference model of Internet of things

According to Fig. 1, we can see the security problem analysis and security mechanism of each layer of the Internet of things. The main function of the perception layer is to realize comprehensive perception through the sensor, and gather the collected data through the access gateway. At the same time, it can also receive the instructions from the

control end to control the goods. Because the sensing node is usually in an unsupervised, open and complex environment, the single function and limited processing capacity of the node make it impossible to take complex security measures, and the information collected by the node is transmitted through the wireless network. Therefore, nodes are vulnerable to physical damage, impersonation attack and denial of service attack, and sensing information is easy to be eavesdropped, tampered and replayed in the process of transmission. The above problems can be solved by cryptographic technology, high-speed cryptographic chip, PKI public key infrastructure, information method security management platform and other security technologies, but it is difficult to find a lightweight, secure and general security mechanism [5–7]. The network layer mainly realizes the goods information transmission and transmission management. In addition to the security problems of the existing communication network, the network layer also has its particularity.

The Internet of things is a kind of network which is integrated with various heterogeneous networks. There are many ways to access the core network. Therefore, it involves a large number of heterogeneous networks interconnection and cross domain security encryption. Because of the large number of nodes in the network, a large number of data transmission at the same time can easily lead to network congestion and denial of service. These problems can be guaranteed by firewall, secure routing, virtual private network and other security technologies. However, the existing security mechanism is designed from the perspective of human to human communication, which can not be directly used for the communication between machines. Using the existing security mechanism may split the logical relationship between machines in the Internet of things [8–10]. The processing layer is mainly to intelligently analyze and process the information transmitted from the network layer to form a variety of Internet of things applications, and establish an efficient, reliable and reliable business support platform for the application layer [11, 12]. Its security problem mainly comes from the intelligent data processing process, because a large number of terminals in the Internet of things produce massive data transmission in the network. Therefore, only by using intelligent processing technology can these data be identified and analyzed in time, otherwise it may lead to network connection interruption and data loss. However, if the intelligent processing technology is out of control or exploited by attackers, it may make the intelligence become low-energy and the automatic processing fail, resulting in catastrophic damage. To solve these problems, we can use attack detection, virus prevention, content analysis, access control and other technologies to protect. Application layer is a level of information interaction with users directly, which provides users with diversified applications. As the application of Internet of things involves many aspects of users' lives, once the information is disclosed, the personal privacy, property, trade secrets and other information of users may be violated. Therefore, we need a sound identity encryption and access control mechanism to isolate illegal users or unauthorized users from accessing the service, so as to ensure the user's privacy information and application security [13, 14].

From the above analysis of the functions and security problems of the reference model of the Internet of things, it can be seen that the layers of the Internet of things cannot directly correspond to the levels in the Internet OSI reference model. For example, the main function of the perception layer is the overall perception of information. In

this process, it may involve all seven layer protocols of the Internet. Therefore, the traditional security protocols, such as IPSec Protocol in OSI network layer and ssl/tls protocol in transmission layer, cannot be directly applied to the Internet of things to realize the end-to-end data security transmission. The technologies that the security mechanisms available at the above layers depend on are essentially encryption and encryption technologies [15, 16]. Therefore, this paper introduces the MQTT protocol to realize the encryption of multi-channel data of Internet of things.

3   Multi Channel Data Encryption Transmission Algorithm in Internet of Things

3.1   MQTT Protocol

MQTT protocol is a message transmission protocol based on message publishing / topic subscription model, which uses client / server architecture to communicate. The design follows the principle of simple, open and lightweight, and is suitable for the application environment of Internet of things with unstable network environment, low bandwidth and limited equipment resources. The main features include.

(1) The topic based message subscription / publication mode can easily realize one to many message distribution and transmission, and decouple the communication publisher and subscriber;
(2) When MQTT message is transmitted in the network, its load content is shielded and can be used to transmit various types of messages.
(3) Provides three levels of message delivery quality: at least once, at least once, and exactly once. Message publishers and subscribers can make messages reach the destination on demand according to the actual transmission needs.
(4) The overhead of the protocol is small, the fixed header is only 2 bytes, the protocol exchange is minimized, and the network traffic is saved.
(5) Last will (also known as Testament) mechanism is provided. When the client is disconnected abnormally, the server will inform the relevant terminal.

MQTT publishes and subscribes messages based on topics, and uses topics to establish message transmission channels between publishers and subscribers. Topic refers to a label attached to an application message. MQTT consists of MQTT message proxy server and MQTT client. Figure 2 shows the structure of MQTT.

According to Fig. 2, the MQTT client refers to the program or device using MQTT, which can be either a publisher or a subscriber. Publishers can publish application messages to the message broker, and subscribers can subscribe to the message broker to receive interested application messages or unsubscribe. As the intermediary between sending message client and request subscription message client, MQTT message proxy server receives connection request from client and application message published by client, processes subscription and unsubscribe request of client, and forwards application message to qualified subscribed client [17, 18].
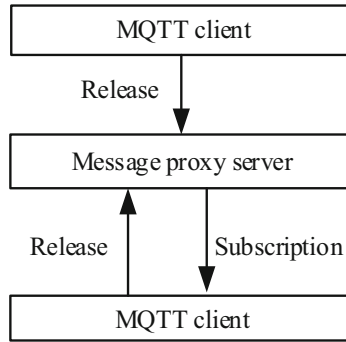
Fig. 2. MQTT structure

3.2 Encryption Transmission Algorithm

Encryption technology refers to the use of cryptographic algorithm to encrypt the data plaintext and then transmit it to the destination, and then restore the ciphertext to plaintext by corresponding means. Symmetric encryption algorithm is also known as the traditional encryption algorithm or single key encryption algorithm. Its encryption / decryption key is the same, or the decryption key can be derived from the encryption key. It requires the sender and receiver to negotiate a key before secure communication. The security of symmetric encryption algorithm depends on the key, and revealing the key means that anyone can encrypt / decrypt the message. The encryption and decryption of symmetric encryption algorithm are expressed as: encryption algorithm $C = EK (m)$, decryption algorithm $m = DK (c) K$. The typical symmetric encryption method is shown in Fig. 3.
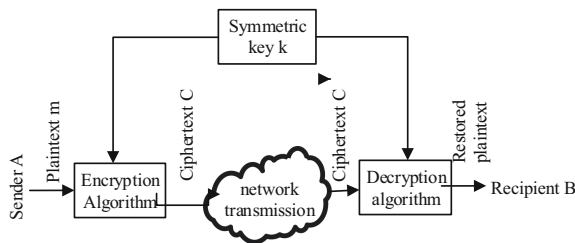
Fig. 3. Symmetrical cipher

According to Fig. 3, the advantages of symmetric encryption algorithm are fast algorithm speed, low requirements for physical devices and high encryption efficiency. The disadvantage is that both sides of the communication use the same key, the security is not high, and it is difficult to solve the problem of signature encryption and non repudiation.

And the key must be distributed through the secure channel. With the expansion of the network scale and the increase of the key quantity, the security management, distribution and transmission of the key are very difficult. This paper uses Rijndael encryption method, which is a new generation of data encryption standard designed by the National Institute of standards and technology to replace des. The standard has been widely used to replace des which used 56 bit key. AES is a symmetric encryption algorithm based on packet, which requires 128 bits of packet length. The key length can be 128 bits, 192 bits or 256 bits according to the actual application requirements. Comparatively speaking, the 128 bit key of AES is 1021 times stronger than the 56 bit key of Des. The algorithm mainly includes round number, round change and key expansion [19, 20]. The number of rounds represents the number of rounds to transform an input packet. The relationship between the number of rounds and the key length is shown in Table 1.

Table 1. Relationship between number of rounds and key length

AES type	Key length/byte	Group size/byte	Number of turns/time
AES-128	4	4	10
AES-192	6	4	12
AES-256	8	4	14

According to Table 1, each round of transformation in the encryption process consists of four different transformations: S-box transformation (nonlinear transformation), row transformation, column transformation and key addition layer (XOR operation with extended key). But the last round transformation does not include the column transformation. Each round transformation requires an extended key with the same length as the input packet. Because the length of the external input key is limited, it is necessary to expand it into a longer bit string to generate the encryption key of each round.

The specific steps of encryption are as follows.

(1) Copy the input group to the 4×4 byte State data group.
(2) XOR the first round key of the state.
(3) The state data set is replaced by nonlinear S-box.
(4) Transform the state.
(5) Column blending operation on State.
(6) XOR the next wheel key with the state (key adding layer), and repeat step 3 until all wheel transformations are completed.;
(7) Copy the State to the output matrix.

Due to the limited processing capacity of IOT terminals, complex encryption algorithms cannot be used. But the Internet of things environment is complex. To ensure the security of communication, simple encryption algorithms can not be achieved. To solve

this problem, this paper combines symmetric encryption algorithm with multi-channel transmission. The advantages of simple calculation and high efficiency are obtained by using symmetric encryption algorithm. As well as the multi-channel transmission data fragmentation makes it more difficult to eavesdrop on the complete ciphertext and crack and recover the plaintext, which reduces the complexity of operation while ensuring the security. After the two sides negotiate the correct key KAB, the sender will use the key KAB to encrypt the original message symmetrically. Then the encrypted message is divided. And add the corresponding identification (session number, fast identification) and message password, and then multiplex. After the message is sent to the receiver, the message is reorganized and decrypted according to the identification of the data block to get the initial message plaintext.

Multi-channel data encryption is a form of encryption, which can make people still get the result of ciphertext when they operate ciphertext, and the result of ciphertext decryption is the same as that of plaintext operation. For people, this technology is correct when retrieving and comparing data, but it is not necessary to decrypt data in the whole process. The previous multi-channel data encryption technology starts from the process of data coding and compression. There are many parameters in the process of data coding, such as I-frame data, P-frame data, And frame macro block, DCT coefficient and other related parameters. Through the analysis of the coding process, the parameters which have a great influence in the coding process are encrypted to achieve the total content of encrypted data and ensure the security effect. The encryption technology of dynamic data is similar to multimedia encryption technology, which is also based on the analysis of encrypted data. Because in many applications need to use uncompressed data, this kind of data without complex coding, so can not use multimedia encryption. However, compared with the general compressed data, the length of the data is much more than that of the latter. Therefore, in order to distinguish the different importance, then the application of dynamic encryption technology is more meaningful.

This paper uses the homomorphic public key encryption function to analyze and select the encrypted data transmitted and stored between nodes in the blockchain, so as to ensure the security of dynamic data in transmission and storage. The concept of blockchain is added to the existing dynamic data encryption method, and the dynamic data is stored distributed by using nodes in the blockchain. In each blockchain node, it is relatively high to allocate independent computing units to deal with homomorphic encryption or decryption. First, dynamic data encryption request is made on a blockchain network. After a series of operations, the request is responded and implemented.

Dynamic data encryption refers to the homomorphic encryption calculation of the collected original data. Multi channel data encryption is a way to calculate data encryption. The data encryption is realized by algebraic calculation and algebra operation on the same plaintext. The process of homomorphic encryption is: set the same station encryption process as JK, known homomorphic encryption process is composed of generating key, data encryption, decryption and data evaluation, set the generated key as KG, encryption process as ENC, decryption process as Dec, evaluation as Eval. Then we can get the formula.

$$JK = (KG, Enc, Dec, Eval) \tag{1}$$

Let the public key AK generate the corresponding security parameter o in the private key BK. AK is used to encrypt plaintext and BK to decrypt ciphertext. Set plaintext Z ∈ Q_n, where n is an integer. Homomorphic encryption of Z can be expressed as.

$$W_{cd}(v_1 + v_2) = W_{cd}(v_1) \oplus W_{cd}(v_2) W_{cd}(fv_1) = f \odot W_{cd}(v_1) \tag{2}$$

After the formula for data encryption transmission, after the transmission of data confidentiality process. If the decryption process is set to Dec, the formula can be obtained.

$$u = Dec(E, AK) \tag{3}$$

Finally, the decryption results are evaluated to complete the calculation process. The specific formula is as follows.

$$P = Eval(T, i, E) \tag{4}$$

Through the above process, the results of homomorphic encryption of data are substituted into the block structure, so as to realize the data security transmission. This method encrypts and stores dynamic data homomorphically based on blockchain technology, and points out that the operation of dynamic data on blockchain is permanently saved. These dynamic data are stored in a node participating in encryption operation by the blockchain, and all nodes constitute a distributed database method of dynamic data. In other words, any node that is damaged can be verified by hash node. At the same time, it can be stored in the database and interpreted to realize the encryption operation.

The data processing process described above is shown in Fig. 4.
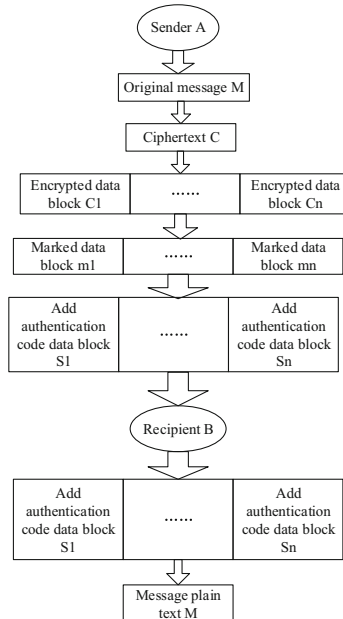
Fig. 4. Data processing process of both sender and receiver

According to Fig. 4, the biggest difference between the data encryption method used in this paper and the traditional encryption method is the use of multi-channel transmission technology. The encrypted message is divided and transmitted through multiple paths at the same time. The conventional encryption method is to encrypt the whole plaintext and transmit the whole plaintext through a single path, so as long as the attacker intercepts the data on the path, it will be easier to crack the message plaintext because of the large amount of information. The encryption and multiplex method mentioned in the section make the eavesdropper intercept a certain block of information from it is small and difficult to break the original. Moreover, the difficulty of eavesdropping is greatly increased by multiplex transmission. Compared with the traditional one-way transmission encryption method, the encryption and decryption process of the method are not different. Only the packet segmentation, reorganization and signature verification are added. This is acceptable compared with the high-intensity encryption and decryption algorithm.

4   Experiment and Result Analysis

In order to analyze the experimental performance directly, the simulation experiment in Matlab environment, the original data are transmitted, and the visualization experimental data results are obtained by using different algorithms. In addition, in order to standardize the experiment, this paper integrates C language program in the process of implementing the algorithm, and improves the function of computer language expression. In addition, the experimental conditions need to be set. The length of encrypted bit sequence is 1200, the time of medical sensitive information sampling is 200 s, and the training set size is 24. The author obtains medical information from medical database. In this experiment, 2170 data were extracted from the above data set by healthdata data set (healthdata.gov), adni. long. usc. edu, drive data set (www.isi.uu.nl/research/databases/drive/download.php), oasis data set (www.oasisbrains.ORG).

4.1   Comparison of Anti Attack Performance of Encryption

In order to verify the effectiveness of the method, the comparative experiment was set. The reference [3] method (a secure framework for authentication and encryption using improved ECC for IOT based medical sensor data) is used as the traditional method. The experimental results are shown in Table 2.

The experimental results show that, compared with the general data encryption, the general data encryption aims at the security of data storage and transmission, that is, the data owner must encrypt the data before continuing to operate. Moreover, it is impossible for the user without the key to decrypt the data and obtain the basic information about the original data. Only the key user can decrypt the data. In the whole process, the user can not decrypt the data, but can only transmit and store the data. If the user forcibly decrypts the data, it may lead to decryption error or decryption failure. Homomorphic encryption focuses on the security of data processing, mainly providing the encryption processing function of data, which means that non key users will not disclose data information when decrypting data. At the same time, the owner of the key can decrypt the data and get the result of the same state encryption.

Table 2. Experimental results of encryption delay

Iterations	Method of this paper/s	Traditional method/s
100	0.921	0.812
200	0.943	0.824
300	0.957	0.835
400	0.975	0.857
500	0.986	0.861

4.2 Comparison of Encryption Delay

For further encryption, the experimental results are shown in Table 3.

Table 3. Experimental results of encryption delay

Experiment times/time	Encryption delay/s	
	Traditional method	Method of this paper
1	1.5	0.3
2	1.8	0.2
3	1.7	0.4
4	1.6	0.3
5	1.4	0.5
6	1.3	0.4
7	1.4	0.5
8	1.6	0.3
9	1.7	0.4
10	1.5	0.2

According to the above experimental results, the encryption method in this paper works better. In this paper, the design of the power Internet of things encryption communication encryption method uses the AIE mode, the power Internet of things communication information encryption method to reduce the impact of network connectivity. Encrypting the communication data between the wireless network devices of power Internet of things makes the encryption method more defensive.

In addition, the design of encryption key reduces the network connectivity and the complexity of the key of the power Internet of things, and the memory occupied by the encryption method of encryption communication is significantly improved due to the

reduction of the complexity of the key. Therefore, it greatly reduces the power consumption of Internet of things encryption communication encryption methods, and improves the work efficiency of network information encryption. The power circuit, communicator and encryptor are designed in this paper. The power circuit provides working voltage for the method. The design of communicator and encryptor enhances the efficiency of the Internet of things communication and the security of network data transmission. Therefore, the design of the whole method is relatively simple and easy to implement. The encryption design of network information makes the network information more sharing, and network users can share legitimate network information resources. When encrypting communication information, the design of encryption key makes it more integrated and reliable. The energy of power Internet of things is greater, and the memory capacity of network information node stored by encryption method is significantly increased. Thus, the network communication defense ability of the encryption method is further improved.

5   Conclusion

Internet of things security as the key factor of Internet of things applications can really popularize, must cause high attention. However, security issues include many aspects. This paper only studies the security issues of end-to-end communication on the Internet of things. In view of the high security requirements of information transmission such as privacy and confidentiality in the end-to-end communication process of the Internet of things and the limited processing capacity of the terminal, this paper proposes an end-to-end secure communication method based on multi-channel transmission, which can ensure the security and reduce the computational complexity as much as possible. This paper analyzes the security problems existing in the Internet of things, proposes the combination of encryption and encryption technology with multi-path transmission, and designs an end-to-end multi-channel secure communication method. And the security of the algorithm is analyzed. Finally, the accuracy of the method in the medical Internet of things data security transmission is verified by experiments, which proves the superiority of the method.

References

1. Tan, A., Wang, S., Xin, N., et al.: A multi-channel transmission scheme in green Internet of Things for underground mining safety warning. IEEE Access **8**, 775–788 (2020)
2. Liu, X., Yang, X., Luo, Y., et al.: Verifiable multi-keyword search encryption scheme with anonymous key generation for medical Internet of Things. IEEE Internet Things J. (99):1 (2021)
3. Khan, M.A., Quasim, M.T., Alghamdi, N.S., et al.: A secure framework for authentication and encryption using improved ECC for IoT-based medical sensor data. IEEE Access, PP(99):1 (2020)
4. Attarian, R., Hashemi, S.: An anonymity communication protocol for security and privacy of clients in IoT-based mobile health transactions. Comput. Netw. **190**(2), 107976 (2021)

5. Ahmad, S., Kim, D.H.: A multi-device multi-tasks management and orchestration architecture for the design of enterprise IoT applications. Futur. Gener. Comput. Syst. **106**(May), 482–500 (2020)
6. Kreuzer, D., Munz, M.: Deep convolutional and LSTM networks on multi-channel time series data for gait phase recognition. Sensors **21**(3), 789 (2021)
7. Rustagi, A., Shukla, M., Samuel, F., et al.: Data Analysis and interpretation in IoT-based systems for critical medical services and healthcare applications. Wireless Pers. Commun. **2**, 1–16 (2021)
8. Cao, Z., Zhou, P., Li, R., et al.: Multi-agent deep reinforcement learning for joint multi-channel access and task offloading of mobile edge computing in industry 4.0. IEEE Internet of Things Journal, PP(99):1 (2020)
9. Chen, X., Liu, A., Zhao, M.J.: High-mobility multi-modal sensing for IoT network via MIMO aircomp: a mixed-timescale optimization approach. IEEE Communications Letters, PP(99):1 (2020)
10. Tsai, K.L., Leu, F.Y., You, I., et al.: Low-power aes data encryption architecture for a LoRaWAN. IEEE Access, **7**, 146348–146357 (2019)
11. Jiang, L., et al.: Toward practical privacy-preserving processing over encrypted data in IoT: an assistive healthcare use case. IEEE Internet of Things Journal, **6**(6), 10177–10190 (2019)
12. Mahdi, M., Salim, T.A., Kalid, H.N.: Secure patient data transmission using information hiding system and medical IoT. Technol. Reports of Kansai Univ. **62**(8), 4572-4585 (2020)
13. Huang K.: Secure efficient revocable large universe multi-authority attribute-based encryption for cloud-aided IoT. IEEE Access, PP(99):1 (2021)
14. Zeng, P., Zhang, Z., Lu, R., et al.: Efficient policy-hiding and large universe attribute-based encryption with public traceability for internet of medical things. IEEE Internet of Things Journal, PP(99):1 (2021)
15. Besher, K.M., Subah, Z., Ali, M.Z.: IoT Sensor initiated healthcare data security. IEEE Sensors Journal, PP(99):1 (2020)
16. Hui, L.A., Tao, J.: A ciphertext-policy attribute-based encryption scheme with public verification for an IoT-fog-cloud architecture - sciencedirect. Procedia Comput. Sci. **174**, 243–251 (2020)
17. Liu, S., Bai, W., Liu, G., et al.: Parallel fractal compression method for big video data. Complexity **2018**, 2016976 (2018)
18. Cheng, X., Zhang, Z., Chen, F., et al.: Secure Identity Authentication of Community Medical Internet of things. IEEE Access, PP(99):1 (2019)
19. Privacy-preserving aware data transmission for IoT-based e-health. Computer networks, **162**(Oct.24), 106866.1–106866.13 (2019)
20. Lu, Y., Li, J., Zhang, Y.: Privacy-preserving and pairing-free multirecipient certificateless encryption with keyword search for cloud-assisted IIoT. IEEE Internet of Things J. **7**(4), 2553–2562 (2020)
21. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mobile Networks Appl. **24**(1), 5–17 (2018). https://doi.org/10.1007/s11036-018-1134-8
22. Lin, H.Y., Hung, Y.M.: An improved proxy re-encryption scheme for IoT-based data outsourcing services in clouds. Sensors **21**(1), 67 (2020)
23. Liu, S., Fu, W., He, L., Zhou, J., Ma, M.: Distribution of primary additional errors in fractal encoding method. Multimedia Tools and Applications **76**(4), 5787–5802 (2014). https://doi.org/10.1007/s11042-014-2408-1

Design of Innovation and Entrepreneurship Effect Evaluation System for College Students Based on MOA Model

Dan Zhao[1](✉) and Huan-wei Liang[2]

[1] Changchun Institute of Finance and Economics, Changchun 130000, China
zhaodan45632@163.com
[2] Guangzhou Huali College, Guangzhou 511325, China

Abstract. With the change of social background, the environment of students' innovation and entrepreneurship is also changing. The traditional evaluation system can not meet the needs of higher vocational education. In order to better meet the current needs of higher education, improve the adaptability of College Students' innovation and entrepreneurship effect evaluation system, and achieve a good evaluation of students' Entrepreneurship effect, this paper designs the innovation and entrepreneurship effect evaluation system based on MOA model. In the hardware design, the CPU is determined according to the system functional requirements, and with the support of the CPU, the serial communication circuit is designed to realize the functions of data uploading and downloading. At this point, the system design is complete. The test results show that the designed evaluation system has high concurrency and anti-pressure ability, which shows that the system has good adaptability to heterogeneous data.

Keywords: MOA model · College students · Innovation and entrepreneurship · Effect evaluation

1 Introduction

Vocational education "take the service as the objective, take the employment as the guidance", the innovation and entrepreneurship education is its connotation. As an important part of the vocational education system, higher education should constantly improve the quality of innovation and entrepreneurship education and do a good job in innovation and entrepreneurship education. To achieve this goal, we must speed up the construction of the evaluation system and mode of innovation and entrepreneurship education, speed up its popularization and application, find out the problems existing in innovation and entrepreneurship education in time, and take effective measures to solve them [1].

Early on, UNESCO (United Nations Educational, Scientific and Cultural Organization) proposed that "a degree is no longer equivalent to a job" and that graduates should no longer be confined to job seekers, but should also become successful entrepreneurs and job creators. Innovation and entrepreneurship education originated in the US, and the Babson College, which offers a bachelor's degree in entrepreneurship education, gradually established a more comprehensive innovation and entrepreneurship education system [2]. The United Kingdom, Germany, Japan, Canada and other developed countries have also set off an upsurge of innovation and entrepreneurship education, taking the implementation of innovation and entrepreneurship education as a strategic plan to promote national economic and social development, and gradually forming an innovation and entrepreneurship curriculum system and teaching system with national characteristics and school characteristics. Although the innovation and entrepreneurship education in our country started late, it has become an important way to promote the innovation-driven development in our country and narrow the gap with the technological innovation and development in developed countries. Therefore, we must improve the quality of innovation and entrepreneurship in higher vocational colleges and build a corresponding evaluation system [3].

In addition, vocational education aims at service and employment- oriented. From this level, we can say that innovation and entrepreneurship education is the connotation of vocational education. However, the research and practice of vocational education evaluation in the past mostly focused on moral education evaluation, curriculum evaluation, skill evaluation, professional ability evaluation and so on. Therefore, it is necessary to design an innovation and entrepreneurship effect evaluation system for college students in order to provide a complete evaluation tool for the government and colleges themselves [4].

Although there are many mature research results in the design of college students' innovation and entrepreneurship evaluation system, such as the integration of social support environment, school investment and student performance. In addition, there are some people from the process of action, environmental factors and reliability and validity of the balance of college students to achieve innovation and entrepreneurship evaluation. However, with the change of social background, the environment of college students' innovation and entrepreneurship is also changing. The traditional evaluation system has been unable to meet the needs of higher vocational education. In order to better meet the current needs of higher education, improve the adaptability of the evaluation system of College Students' innovation and entrepreneurship effect, and realize the good evaluation of students' Entrepreneurship effect, this paper introduces the MOA model into this field, and designs the evaluation system of innovation and entrepreneurship effect based on the MOA (Motivation; Opportunity; Ability) model. In the hardware design, the CPU (Central Processing Unit) is determined according to the system functional requirements, and with the support of CPU, the system can realize the evaluation of students' Entrepreneurship effect, The serial communication circuit is designed to realize the function of data upload and download. So far, the system design has been completed, which is expected to provide some help to solve the above problems.

2 Hardware Design of Innovation and Entrepreneurship Effect Evaluation System for College Students Based on MOA Model

2.1 Selection and Design of System CPU

The system CPU needs to provide support for the function realization of the whole system, the most important is to support the serial communication device data acquisition, and data stored in the storage module through the DEBUS bus. According to the work to be done by the CPU, consider the following aspects of CPU selection [5]:

In the system debugging stage, needs to have the input output design, in order to take the different debugging method and the observation debugging result, therefore the CPU must have the certain number I/O port. To simplify the hardware, the output module can be a serial LCD module. Up to four I/O ports are required here. In order to observe the debugging process, you also need to access three to four LED, in this part of the need for about four I/O port. The input part is controlled by keys. The keys do not need a matrix keyboard, only need to set 3 to 4 adjustment keys. Here also need 3 to 4 I/O ports. LCD module, debug button and LED in the final integration of the system, can also be used as a formal output display, functional indicators and function keys. So these ports are initially used for debugging, and later will become part of the system, can not be excluded, a total of at least 12 I/O port CPU. In addition, two I/O ports are required to acquire EEPROM data using the I2C bus. In addition, because of the need of "conflict detection", the flag line of conflict detection needs to occupy 1 I/O port. Finally, it is best to reserve some I/O ports for future expansion of system functionality. So, at this point, I/O ports need to be at least 15 N each.

Considering the convenience of system debugging and maintenance, CPU should have the function of ISP online programming. Online programming capabilities are not always in use, so I/O ports used for online programming can be reused for other functions [6]. However, considering the stability of online programming and the inconvenience of reusing the software, 3 I/O ports are specially designed for online programming. At this point, the CPU's I/O port is at least 17 N.

Nowadays, the function of CPU is developing rapidly. In order to simplify the external devices, many CPUs integrate some common functions, such as A/D, D/A, voice, floating point, watchdog, etc. Since the design system does not require these functions, there is no need to choose a CPU with additional functions to increase costs, and the performance of the CPU can be seen from its own parameters, such as the capacity of the internal ROM and RAM, the number and bits of the internal timer, the number of internal interrupt sources, the size of the internal stack, the frequency at which the CPU runs, the data width of the CPU universal register, etc. The first metric to be considered here is the general register data width of the CPU. And if there is no complex and precise large-scale operation, the general 8-bit single-chip can meet the needs of field data acquisition [7].

According to the above considerations, the AT89S51 MCU produced by Atmel Company is selected as the CPU of the lower computer [8].

The AT89S51 comes in four package configurations. Choose the most popular PDIP (Plastic Dual Inline Package) package, and compared to PLCC and TQFB package, the price is cheaper, circuit wiring is also more simple [9]. As the core part of CPU, it needs to integrate serial communication interface, DEBUS bus cascade interface, ISP

online download interface, serial LCD (Liquid Crystal Display) display interface, key debugging interface and indicator LED (Light Emitting Diode) interface. Make better use of port resources. The main pin definition is shown in Table 1.

Table 1. Main pin definition of AT89S51

Pin	Functional features and definitions	Pin	Functional features and definitions
P1.0	CS_ LCD (+): serial LCD chip selection terminal	P1.1	SCLK(System clock): serial LCD clock line
P1.2	SDAT: serial LCD clock line	P1.5	Mosi: ISP online programming data line
P1.6	Miso: ISP (Internet Service Provider) online programming data line	P1.7	SCK (Studiecentrum voor Kernenergie): ISP online programming clock line
P3.0	RXD (receive external data): Asynchronous serial transmission output	P3.1	TXD (transmit external data): Asynchronous serial transmission output
P0.0	Idle: DEBUS bus status flag line	P0.5	ResetSon: Wake up connection
P0.6	IIC-SDAT: I^2C data line	P0.7	IIC-SCLK: I^2C clock line
RST	ResetSon:System reset, lower computer reset control	P2.5	LEDcontrol port
P2.4	LED control port	P2.3	LED control port
P2.2	Key control port	P2.1	Key control port
P2.0	Key control port		

Under the support of CPU, the system's internal upload, download, information sharing and other functions to provide support.

2.2   Serial Communication Hardware Design

Considering the needs of users, the serial communication hardware is designed to provide users with data upload, download, transmission and sharing functions [10].

In the system hardware design using RS-232 serial communication, PC using the RS-232C standard logic level, and AT8S51 using the TTL (Transistor-Transistor Logic) level, level conversion must be carried out. The exchange of levels falls broadly into two categories. One is the use of discrete components, one is the use of chips [11]. Circuit diagrams of discrete components have a lot of information to follow, some of which are quite simple. But this method needs to calculate the electrical index of discrete components according to the serial port of a specific PC. The choice of integrated chip is low cost, less peripheral components, more stable, can simplify the circuit board. Max232 chips are finally selected in the design based on the power supply requirements (SV), power consumption requirements (general), circuit board area requirements (smaller), price (cheapest) and other factors [12].

There are two pairs of drivers/receivers inside the MAX232. One pair, Tin and Rout, is connected to the TXD and RXD pins of the MCU respectively. The other end is connected to the 9-pin interface, which is connected to the serial interface of the PC. External connection of 4 capacitors, with the internal double charge pump to achieve level conversion. According to the function of online download, the hardware circuit of ISP online download is designed.

The concept of the ISP (In System Programming) was initially proposed by Lattice Corporation of the United States, which can directly program the internal memory of the CPU under the condition that the single-chip computer system operates without power. And after the end of the program, the SCM(Software Configuration Management) can restart the operation [13]. Because ISP makes the development process of embedded system more efficient, Atmel company also adds ISP online download function to its mainstream product AT89CSX series SCM. The ISP functional pins of the AT89S51 are P1.5 (MOSI: MASTER OUTPUT SLAVE INPUT), P1.6 (MISO: MASTER INPUT SLAVE OUTPUT), P1.7 (SCK), and their RST(RESET) reset pins are also attached [14].

In order to realize the downloading function of AT89SSX series single chip microcomputer, it needs the cooperation of hardware and application software [15]. The ISP interface for the underlying host is shown in Fig. 1.
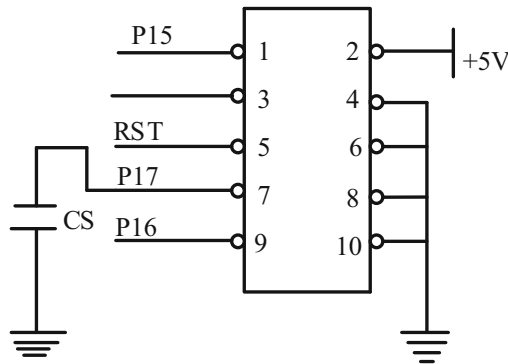
Fig. 1. ISP download line interface

According to the above, the DEBUS bus already contains the I2C bus and the status flag line Idle. In order to increase the function of waking up the lower computer, the control line of "ResetSon" is added [16]. The line is directly connected to the RST line of the lower computer. When the lower computer breaks down, it can be restarted by the ResetSon control line. The interface of DEBUS is designed in the form of "cascade", and the Vcc power line and GND (GrouND) bottom line are added to achieve the goal of "common power supply" and "common ground". At this point, the DEBUS hardware architecture is complete, totaling six lines: I2C's data line IIC-SDAT, I2C's clock line IIC-SCLK, Status Marker Idle, Machine Control Line ResetSon, Bottom Line GND, and Power Line VCCCo. DEBUS Bus interface portions are shown in Fig. 2.
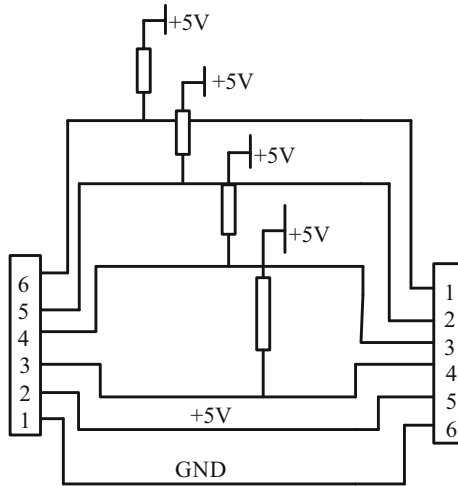
Fig. 2. Bus interface circuit

J8 is the cascade terminal, which is the access terminal of each lower computer. Each module of the lower computer has this cascade interface, and all the lower computers can be connected in parallel on the DEBUS bus through the interface. Of the 4 I/O ports of AT89S51, P0 port has the largest IOL, i. e., the strongest drop-down capability, so P0 port is chosen here as the terminal of DEBUS bus. The P0 port has no built-in pullup resistor, so additional pullup resistors are required. At this point, the system hardware design is completed [17].

3 Software Design of Innovation and Entrepreneurship Effect Evaluation System for College Students Based on MOA Model

3.1 Storage Mode of Design Effect Evaluation Resources

The evaluation resources of innovation and entrepreneurship often have many forms, such as documents, pictures, videos and so on. Most of these massive digital resources exist in the form of unstructured, and can not be effectively stored with a unified data structure, which easily leads to resource sharing difficulties. In order to achieve the goal of quality innovation and entrepreneurship effect evaluation resource sharing, the first problem to be solved is the description resource [18]. The evaluation of innovation and entrepreneurship effect is to define a standardized and feasible way to describe and package the evaluation resources of innovation and entrepreneurship effect, Realize the unified storage and retrieval of resources. This is the basis of effective resource sharing, so it is necessary to unify and standardize the representation of resources [19].

Defining a consistent data resource description framework through metadata standards is an effective way to solve this problem. Metadata is the description of the data

structure and content characteristics of the original data resources. The purpose of browsing metadata is to access the original data resources, i.e. innovation and entrepreneurship effect evaluation resources. The so-called metadata standard is a set of rules that describe the specific objects of innovation and entrepreneurship effect evaluation resources.

According to the CELTS-42 standard "Metadata Specification of Basic Education Resource", the data element set of Basic Education Resource includes 22 elements, using the language code specification defined in RFC 1766. The specification allows users to extend metadata elements according to the needs of the system, but must conform to the format and technical specifications defined by the specification elements [20]. According to this specification, the storage mode of system resources is designed as follows:

The resource itself and resource description information of innovation and entrepreneurship effect evaluation should be named with the same file name, and stored in the designated folder of HDFS distributed file system through Hadoop. The resource description information is generated by resource name and resource content configuration file in a fixed format, which is used to generate the index file of retrieval resources [21].

The metadata information of the resource is stored as a record in the MySQL database table, while the location of the resource in HDFS (Hadoop Distributed File System) is recorded and used as a way to download and obtain the resource. Metadata information is used to allow users to view the properties of a resource when it is displayed [22].

Users on the system page can view the metadata information of the resource and the content information of the resource itself (displayed on the page in the form of a preview) and obtain the corresponding resource in HDFS through the storage location field of the resource in the database table when the resource needs to be obtained. The metadata description table ResourceInfo is designed as shown in Table 2.

Table 2. ResourceInfo metadata description table

Field name	Field type	Explain
ResourceID	Integer	Resource ID
Title	Varchar (50)	Resource title
UserID	Integer	Resource owner ID
Description	TinyText	Resource description
TypeID	Integer	Resource type
FormatID	Integer	Resource file format size
FileSize	Integer	Resource file size
UploadTime	Date	Resource upload time
Value	Integer	Resource value set by uploader
Download Count	Integer	Resource downloads
ScanCount	Integer	Resource views
Identify	Varchar (255)	Resource index identifier

In the process of evaluation, the system fetches the related information from the resource database, and realizes the effect evaluation of college students' innovation and entrepreneurship [23].

3.2 Selection of Evaluation Indicators Based on MOA Model

The MOA model is composed of motivation, opportunity and ability, and their inter-relation and interaction promote the occurrence of certain behaviors. Motivation is the college students' willingness to participate in innovation and entrepreneurship activities, and opportunity is the effective component that is conducive to behavior in the external environment. It has positive effect on innovation and entrepreneurship.

Collect the data of college students in the process of innovation and entrepreneurship, process the data by SPSS, analyze the basic situation of the sample, and get the final analysis data by reverse recalculation. Using the concept model to verify and analyze the effect of innovation and entrepreneurship activities of college students, to select the appropriate evaluation indicators. After model verification and calculation, the results are shown in Table 3.

It can be seen from the data in Table 3 that each measurement index meets the evaluation requirements. When the index is known, the weight of each index is calculated.

3.3 Evaluation of the Effects of Innovation and Entrepreneurship

The geometric average values of each row of the judgment matrix are calculated by using the product square root method. The formula is as follows:

$$\bar{c}_i = \left(\prod_{j=1}^{n} a_{ij} \right)^{\frac{1}{n}} \quad i, j = 1, 2, \ldots, n \tag{1}$$

In formula (1), a_{ij} represents the elements in row i and column j of the original judgment matrix, n represents the number of indicators, and c_i represents the geometric average value in row i of the original judgment matrix. The geometric mean of each row is normalized to get the feature vector:

$$c_i = \frac{\bar{c}_i}{\sum_{j=1}^{n} \bar{c}_i} \tag{2}$$

In formula (2), c_i represents the weight of the i-th indicator, n represents the number of indicators, and $\bar{c}_i$ represents the geometric average value of the i-th row of the original judgment matrix. The weight coefficient of the first level index is calculated. On this basis, the maximum eigenvalue of the judgment matrix is calculated:

$$\lambda_{max} = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{j=1}^{n} a_{ij} c_i}{c_i} \tag{3}$$

Table 3. Factors load, reliability, validity and normal test results of measurement indexes

First level indicators	Secondary indicators	Factor load	Reliability coefficient	Mean variance extraction
Innovation and Entrepreneurship Environment	External support environment	.702	.493	.426
	School implementation environment	−.393	.154	
	School Entrepreneurship	.795	.632	
	Construction of teaching staff	.940	.884	
Investment in innovation and Entrepreneurship	Current situation of investment	.909	.601	.770
	Construction of practice platform	.775	.501	
	Curriculum system design	.708	.545	
Innovation and entrepreneurship process	Service guidance support	.738	.582	.523
	Student participation process	.763	.741	
Achievements of innovation and Entrepreneurship	social influence	.861	.584	.636
	Educational achievements	.764	.826	

The weights of single-level indicators can be obtained through the above calculation. Based on the above results, an evaluation model is established:

$$Y = \sum_{i=1}^{n} X_i c_i \tag{4}$$

In formula (4), X_i represents the score value of the evaluation index and c_i represents the comprehensive weight of the index. According to the calculation results of formula (4), the effect of innovation and Entrepreneurship of college students is judged. The larger the value is, the better the effect is. On the contrary, the worse the effect is. So

far, the evaluation system of College Students' innovation and entrepreneurship effect based on MOA model has been designed.

4   Performance Test of Innovation and Entrepreneurship Effect Evaluation System for College Students Based on MOA Model

4.1   Build Test Environment

The system is designed on the basis of multi-layer architecture. The function of each layer is relatively independent. There is a standard interface between each layer. It can not only ensure the scalability, flexibility and openness of the system, but also facilitate the system access and management. In order to verify the application performance of the system, the experimental environment is set up. The hardware environment is shown in Table 4.

Table 4. Hardware environment

Hardware type	Purpose	To configure	Number	Operating environment
web application server	Support the retrieval, analysis and other front-end applications	IBM-X3650 M4 CPU:Inter Xeon 4C E5–2609 Memory:ECC DDR3 128G	2	Windows 2010
Interface application server	Middle layer encapsulation service of retrieval interface	IBM-X3650 M4 CPU: Inter Xeon 4C E5–2609 Memory: ECC DDR3 128G	1	Centeos
Load balancing server	Realize load optimization and resource allocation in the process of big data loading and big user concurrency	IBM-X3650 M4 CPU: Inter Xeon 4C E5–2609 Memory: ECC DDR3 128G	1	Centeos

Under the support of the above environment, the test results evaluate the performance of the system. System testing is a process used to test the integrity, correctness and security of software. It is an indispensable part of software development. In order to ensure the quality and reliability of the evaluation system of College Students' innovation and entrepreneurship effect, functional testing and performance testing are carried out on the system. The function test is mainly to test whether each functional module of the system meets the requirements, and the performance test is mainly comparative test. Two traditional evaluation systems are introduced in the test to verify the actual level of each system under the same test conditions, including load stress test and concurrent test.

4.2   System Function Test

The function test of the evaluation system for college students' innovation and entrepreneurship effect adopts the method of unit test, which runs through all stages of the system development, that is, every functional module is developed and tested to ensure its accuracy. Test cases include normal input, boundary input and abnormal input. The output of each module of the test system meets the requirements under different conditions, and the system is improved according to the test cases.

The test of evaluation data acquisition and management module mainly includes the upload of test data, the information editing, modification and deletion of uploaded data. The test contents are shown in Table 5.

Table 5. Test of data upload and management module

Test content	Input	Expected results	Actual output
Resource upload	Select a file and confirm the upload	Prompt that the upload is successful, call to the resource information editing page	In line with expectations
Edit resource information	Enter the editing information and confirm the editing	Prompt to edit successfully, jump to the uploaded resource list	In line with expectations
Modify resource information	Enter the modification information and confirm the modification	If the modification is successful, skip to the uploaded resource list	In line with expectations
Delete resource	Select a data and click delete	Prompt to delete successfully	In line with expectations

For the test of effect evaluation module, the main function is to test evaluation resources and view other evaluation content. The test contents and results are shown in Table 6.

Table 6. Effect evaluation module test

Test content	Input	Expected results	Actual output
View comments	Click the resource icon	The resource details page displays all the user's comments	In line with expectations
Evaluation resources	Evaluation content	Evaluate success	In line with expectations

Evaluation information classification management module test, the administrator to view the evaluation information classification, add, delete, modify information classification functions. The test contents are shown in Table 7.

Table 7. Testing of the evaluation information resource category management module

Test content	input	Expected results	Actual output
Get resource category	Click manage resource category	Get resource type	In line with expectations
Add resource category	Enter the resource name and confirm to add	Increase success	In line with expectations
Modify resource category	Select the category, enter the modification name and confirm the modification	Modified successfully	In line with expectations
Delete resource category	Select category and confirm deletion	Successfully deleted	In line with expectations

Through the above functional tests, we can see that the system functions are able to meet the normal implementation, there is no abnormal situation. On this basis, comparative experiments can be carried out to further verify the actual level of the evaluation system.

4.3   System Load Capacity Testing and Analysis

Load test mainly tests whether the load pressure of the system meets the requirement under the condition of high concurrent access. Web Bench is used to complete the load stress test of the system server. Web Benceh is a website stress testing tool developed by Lionbridge. It can be used to test the server stress of ASP, PHP, JAVA, CGI, etc. It can also be used to test the load capacity of SSL secure websites. It can simulate up to 30,000 concurrent links to test the load capacity of websites. The traditional effect evaluation systems 1 and 2 are compared with the systems mentioned in the documents [7] and documents [8]. The test results are shown in Table 8.

Through the data in Table 8, it can be seen that the traditional effect evaluation system 1 increased the number of request failures significantly with the increase of test time, and reached 415 request failure links at 50 s, indicating that the system has poor load capacity and can not meet the needs of continuous request access. In the traditional evaluation system 2, when dealing with 1000 requests, with the increase of test time, the rate of request failure is also very obvious. 526 request failure links are reached in 50 s, which indicates that the system load capacity is poor and it is difficult to meet the needs of continuous request access. In the face of a large number of requests, the design system can ensure the success of all requests in the 50 s, indicating that the system can meet the continuous concurrent access in a certain period of time. The results show that

Table 8. Load test results of different effect evaluation systems

	Number of requests	Test time	Number of requests failed
Traditional effect evaluation system 1	1000	10	11
	1000	20	65
	1000	30	124
	1000	40	206
	1000	50	415
Traditional effect evaluation system 2	1000	10	24
	1000	20	92
	1000	30	187
	1000	40	294
	1000	50	526
The evaluation system proposed in this paper	1000	10	0
	1000	20	0
	1000	30	0
	1000	40	0
	1000	50	0

the evaluation system of College Students' innovation and entrepreneurship effect has good pressure resistance ability.

4.4 System Concurrency Test Results and Analysis

The system uses Apache JMeter to test concurrency. The concurrency test runs a stress test for about 60 s with different number of threads and different loops.

Observing the data in Table 9, we can see that the system access error rate of the traditional effect evaluation system 1 increases gradually when the thread number setting increases in turn, and reaches more than 50% when it increases to 1000, which is basically unbearable, and the throughput drops obviously, which can not maintain the high application level in the high thread. Traditional evaluation system 2 faces the situation that the number of threads is increasing continuously, but it also has the situation that the system access rate is increasing. When it increases to 1000, the system error rate reaches more than 60%, the system can not run normally, and the throughput drops more obviously. This shows that it is difficult for the system to keep good performance when the number of threads increases. However, with the increase of the number of threads, the error rate of design system access remains at a very low level of 0.5%, and the throughput does not decline significantly, and it always remains at a high level. Load performance test results show that the system based on MOA model has good adaptability and can be effectively applied to multi-party collaborative environment.

Table 9. System concurrency test eesults

	Number of threads	System access error rate%	Throughput (KB/sec)
Traditional effect evaluation system 1	200	0.95	107.16
	400	7.45	73.13
	600	12.17	64.14
	800	26.47	58.92
	1000	51.42	43.61
Traditional effect evaluation system 2	200	0.57	119.62
	400	9.26	94.32
	600	22.54	61.55
	800	43.36	54.23
	1000	62.34	41.45
The evaluation system proposed in this paper	200	0.26	358.5
	400	0.30	307.6
	600	0.51	277.8
	800	0.60	243.6
	1000	0.63	215.4

5   Concluding Remarks

In this paper, with the support of the original research data and literature materials, the MOA model-based evaluation system of college students' innovation and entrepreneurship effect is designed, which provides more possibilities for college students' innovation and entrepreneurship education. After the system design is completed, the high adaptability of the evaluation system is verified through a large number of comparative experiments, and the problems existing in the traditional evaluation system are solved. However, due to the limitation of time and technology, the system still has some shortcomings. The basic education of innovation and entrepreneurship involves a wide range, so it is possible to choose the evaluation index. Therefore, in the follow-up research, we will further improve the evaluation indicators and content, promote the improvement of innovation and entrepreneurship education level of college students, so as to achieve the purpose of carrying out satisfactory vocational education.

References

1. Zhang, Y.: Assisted teaching quality evaluation model based on active learning support vector machine. Mod. Electron. Tech. 42(07), 112–114 (2019)

2. Zhang, J., Wang, Y., Yu, L.: Teaching and assessment practice of the course of theoretical mechanics uniting four types of classrooms. Mech. Eng. **41**(02), 210–215 (2019)
3. Yu, J., Feng, Q., Ding, X.: A new exploration of academic evaluation system for college students based on CIPP and CDIO model. Heilongjiang Res. High. Educ. **37**(03), 52–56 (2019)
4. Wei, D., Wei, X.: Design of network experimental teaching evaluation system based on data mining. Mod. Electron. Tech. **43**(03), 142–145+149 (2020)
5. Nan, G., Hou, Z., Liu, K.: Construction of evaluation model of college students'entrepreneurial opportunity based on entropy weight method. J. Hefei Univ. Technol. (Nat. Sci.) **43**(12), 1713–1717 (2020)
6. Xu, Z., Yang, L., Luan, S.: Research on evaluation of transformation ability and efficiency of scientific and technological achievements in universities based on relay innovation. Sci. Technol. Manage. Res. **39**(24), 8–14 (2019)
7. Dang, J., Wang, D., Jing, T.: Design of entrepreneurship education evaluation system based on digital badge technology. E-educat. Res. 41(09), 75–80+101 (2020)
8. Dong, Q., Dai, G.: Evaluating system of learning effect based on fuzzy comprehensive evaluation. Math. Pract. Theo. **50**(05), 282–291 (2020)
9. Han, S., Li, R.: Kansei engineering based evaluation for the designation of products on electronic commence website. J. Univ. Shanghai Sci. Technol. 41(01), 97–102 (2019)
10. Zhang, J., Li, J., Chen, H.: Research on the performance evaluation and influencing factors of scientific research innovation in universities across the Yangtze River delta: based on DEA-Malmquist-Tobit Model. Sci. Technol. Manage. Res. 40(09), 80–87 (2020)
11. Yan, X., Tian, J.: Evaluation model of college students' entrepreneurship success rate based on grey relational theory. E3S Web Conf. 235(25), 03019 (2021)
12. Xin, M., Wang, Y., Fan, L., et al.: On the construction of evaluation system for undergraduate college students' ability of innovation and entrepreneurship. J. Heihe Univ. **10**(09), 105–107 (2019)
13. Yuan, L., Sun, Z., Sun, Z., et al.: Evaluation of college students' innovation and entrepreneurship ability in maker space. Theo. Pract. Innov. Entrepreneurship **2**(02), 20–22 (2019)
14. Zhang, F., Xi, L.: An evaluation model for the innovation and entrepreneurship thinking ability of college students based on neural network. Int. J. Emerg. Technol. Learn. (iJET) **16**(2), 188 (2021)
15. Liu, R., Guo, J., Daytime charming: Research on the current situation and countermeasures of college students' innovation and entrepreneurship curriculum system—taking Xi'an University of engineering as an example. Educ. Forum 000(014), 122–124 (2019)
16. Xin, Y., Wang, C., Dong, Y., et al.: Management and entrepreneurship management mechanism of college students based on support vector machine algorithm. Comput. Intel. IP(IP), 1–13 (2020)
17. Chen, Y.: College English Teaching quality evaluation system based on information fusion and optimized RBF neural network decision algorithm. J. Sens. **2021**(5), 1–9 (2021)
18. Qi, Y., Liang, T., Chang, Y.: Evaluation of college students' innovation and entrepreneurial ability for the science and technology service industry. Int. J. Emerg. Technol. Learn. (iJET) **16**(5), 228 (2021)
19. Wang, K., Yan, C.: An evaluation model for the cultivation and improvement of the innovation ability of college students. Int. J. Emerg. Technol. Learn. (iJET) **15**(17), 181 (2020)
20. Li, N.: A fuzzy evaluation model of college English teaching quality based on analytic hierarchy process. Int. J. Emerg. Technol. Learn. (iJET) **16**(2), 17 (2021)
21. Lei, H., Qin, L.: Effect analysis of the field management innovation of SERVQUAL model evaluation in pro-motion of outpatient service upgrade. Pract. J. Clin. Med. **016**(004), 233–235 (2019)

22. Gupta, H., Barua, M.K.: Evaluation of Manufacturing Organizations Ability to Overcome Internal Barriers to Green Innovations (2021)
23. Soltanzadeh, J., Elyasi, M., Ghaderifar, E., et al.: Evaluation of the effect of R&D subsidies on Iranian firms' innovative behavior: reconceptualizing behavioral additionality. J. Sci. Technol. Policy Manage. **11**(1), 17–48 (2019)

Design of Human Resource Distance Education System Based on Internet of Things Technology

Huan-wei Liang[1] and Dan Zhao[2(✉)]

[1] Guangzhou Huali College, Guangzhou 511325, China
[2] Changchun Institute of Finance and Economics, Changchun 130000, China
`zhaodan45632@163.com`

Abstract. When the traditional distance education system deals with parallel tasks, there are too many network packets running in the system, which leads to the slow running of the system. This paper designs a human resource distance education system based on the Internet of things technology. In the hardware part, the structure of teaching processor is designed to form the circuit structure of human resource data acquisition and balance the data acquisition. In the software part, the Internet of things technology is used to build the system architecture, control the actual number of network packets, construct the function logic, and finally realize the function of human resources remote teaching. In order to prepare the software and hardware environment of the teaching system, two kinds of traditional teaching systems and the designed teaching system are applied to carry out experiments. The results show that the designed distance teaching system has the least number of network packets and the fastest running speed.

Keywords: Internet of Things technology · Human resources · Distance education system · Network package

1 Introduction

With the sustained and rapid development of the social economy, today's world has entered the era of rapid progress of information technology. The strong impact of information technology on traditional education has made great changes in the current education work in terms of teaching methods and teaching compositions. To do a good job of education modernization and improve the quality of teaching work, we need to enhance the informatization of teaching process. According to the relevant investigation, the focus of distance education supported by physical network technology is to deal with the interaction among teaching work, resource reserve and quality management. With the strong innovation of the current computer network technology, especially the realization of the personalized interaction function led by the physical network technology, the operation efficiency and service quality of the emerging system which provides various technical services for education work have been rapidly improved, and the effect

of teaching interaction between teachers and students in the network situation is more significant, These are in line with people's expectations of the development of modern education [1, 2].

In related research, reference [3] proposed a problem-solving path with different difficulty levels in various disciplines by supplementing the existing learning tracking algorithm. The data set obtained by solving the path of the learner problem, and the optimal problem solving path through the recurrent neural network is suggested through the path data set. Reference [4] proposed a method of dynamically combining adaptive online learning courses based on learner activities, learning goals and instructional design strategies using Q-learning algorithm. The latter is based on the learner's behavior, and based on the learner's positive or negative feedback, it provides the necessary course content to achieve the learning goal.

Considering that experiments play an important role in the realization of the overall education goal, traditional education is restricted by the limitations of the region and experimental conditions, and it fails to show the teaching effect of the experiment effectively [5]. Therefore, it is necessary to construct a virtual teaching system with outstanding interactive effect, timely feedback and diverse display effects. Therefore, this paper designs a human resources distance education system based on the Internet of Things technology. In order to improve the operating efficiency of the system, the hardware part of the teaching processor structure is designed to form the human resources data acquisition circuit structure, which can balance the data acquisition. The software part uses the Internet of Things technology to build the system architecture and controls the actual number of network packets in operation to increase the operating speed of the system.

2 Hardware Design of Human Resources Distance Teaching System

2.1 Structure Design of Teaching Processor

The load-bearing device of the distance teaching system is mainly computer. So when designing the hardware of the distance teaching system, it is mainly designed for the computer carrying the teaching system [6]. Firstly, the processor required by the system is designed. The processor adopts S3C2410 as ARM processor, DSP processor selects processor of TMS320C6713B, and connects corresponding memory to form processor structure, as shown in Fig. 1.

According to the processor structure shown in Fig. 1, an FPGA is used to connect the interface between the arm processor and the external memory of the DSP processor, and to connect with the interface processor in the crossbar network. The HPI interface of the DSP processor is connected with the general interface of the arm, which helps to load the internal program of the processor.

In order to reduce the design difficulty of the simulation teaching system, cy7c1041dv33-iobgi static memory is selected. The SDRAM is used to control its storage capacity of 128MB, storage depth of 6 m and bit width of 32bit. Set the working voltage of the processor to ± 3.3 V. The maximum clock rate of the chip is 142 MHz,
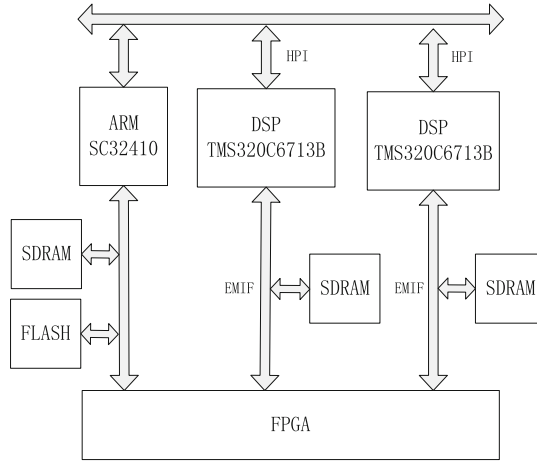
Fig. 1. Processor architecture

the data reading cycle is 5, and the reading time is 6 ns. The processor uses two FLASH memories, S29GL032M 10TAIR10 and no structure. Set the storage capacity to 2 m × 8 bit, control its working voltage to +3 V, debug the memory interface, and form the interface structure as shown in Fig. 2.
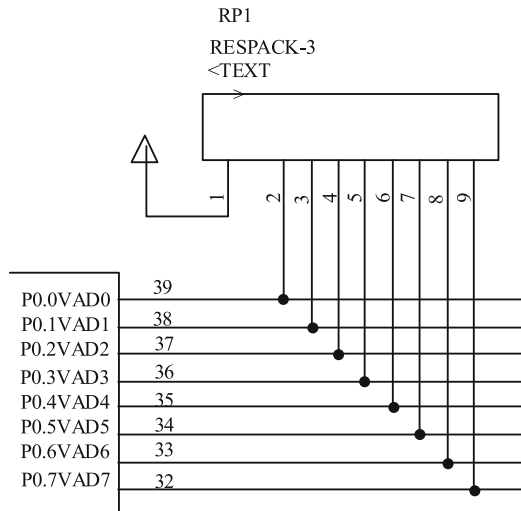
Fig. 2. Shows the interface structure

In the interface structure shown in Fig. 2, the working current for controlling read operation is 15 mA, and that for erasing and writing is 30 ma. Two FLASH memories with the above parameters are used, and one is used to store the code of distance teaching practice simulation teaching in the processor [7]. The other is used to store the

compensation calibration parameters of the setting processor. A storage capacity of 256 K is adopted × Two 8-bit SRAM control the data exchange inside the processor. All the above memory parts are connected to FPGA, and the control of memory is realized finally.

Fiber optic gyroscope (FOG) and flexible accelerometer (FAA) are used as the inertial components of the hardware. Rs-421 serial interface is used to realize the digital output of the processor and receive the data of FAA. I/F conversion circuit is used to convert the electrical signal in the processor into frequency signal. FPGA adopts EP3SL200F1152I3N with 65 nm process, and uses ALMs of 60 K in its internal to connect 4 GPIO in serial. One DC/DC power converter LTM4616 is used in power supply to provide two channels of current for processor [8]. Because the current requirement of the power supply in the processor is not high, the LDO conversion chip LTC1584-3.3 is selected to provide voltage and current for the hardware.

The teaching system takes the single-chip microcomputer as the core processing unit, combines with resistance and capacitance and other devices, and takes the single-chip microcomputer as the smallest processing unit. The final control core circuit composition block diagram is shown in Fig. 3
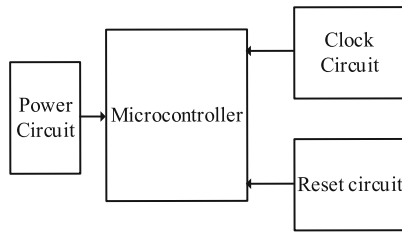
Fig. 3. Minimum system composition block diagram of STC89C52 single chip microcomputer

As shown in Fig. 3, the left side of MCU is mainly connected with power supply circuit, and the right side is connected with clock circuit and reset circuit. In the actual work of MCU, the time of accessing the memory from ROM is defined as a machine cycle, and the internal oscillator of MCU is used to store the access data of a machine cycle [9]. The XTAL1 and XTAL2 ports of the oscillator are used as the input/output
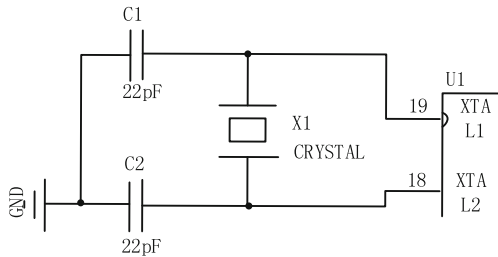
Fig. 4. Clock circuit diagram of single chip microcomputer

ports of the oscillator. The XTAL1 port of the oscillator is connected with a quartz crystal by internal and external clock mode. Then a capacitor is carried on the outside to form a parallel resonant circuit, which makes the internal oscillation circuit produce self-excited oscillation [10]. Finally, the connection diagram of the clock circuit is drawn. As shown in Fig. 4.

In order to prevent the single-chip microcomputer from being disturbed by the environment when it is running, the teaching system will fail. The reset circuit of the single chip microcomputer system is adjusted to the level switch reset mode, so that the teaching system is in the short circuit state when the power is turned on, and the reset pin is connected to the high level [11]. When the power supply is stable, the reset pin is grounded through resistance, so that the capacitor can isolate the DC level. The reset circuit of MCU is designed as shown in Fig. 5.
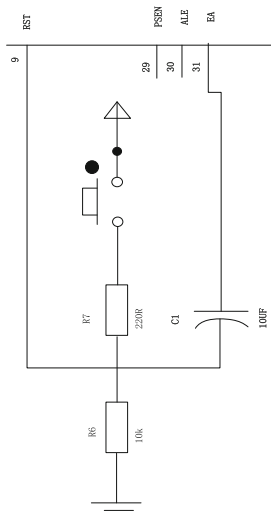
Fig. 5. Single chip reset circuit diagram

The power circuit uses USB to directly provide 20 V DC, in order to ensure that the program reading starts from the internal storage [12]. Connect the EA pin in the figure above to the high level to complete the design of the system processor.

2.2 Structure Design of Human Resource Data Acquisition Circuit

In order to ensure the smooth operation of the virtual scene simulated by human resources, a driving circuit is designed, which adopts the single power circuit structure, as shown in Fig. 6.

According to the structure shown in Fig. 6, when the power switch in the circuit is on, the supply voltage VCC helps the current of winding inductance L rise. When the current rises to the threshold, the system will be limited by the resistance R and control the current in a safe range [13]. In order to ensure the control of the controller by a single
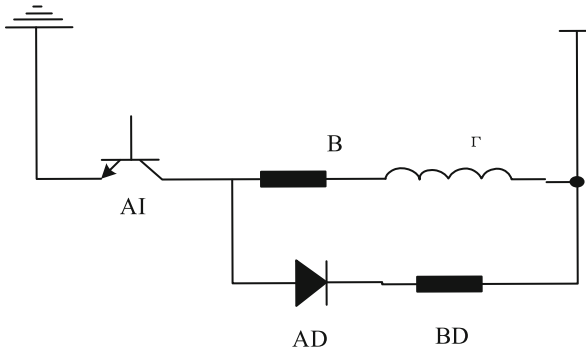
Fig. 6. Single power supply circuit

circuit, a power amplifier chip is connected to the circuit in Fig. 6. The chip selects ULN2003 as the driving amplifier, configures its pins, and sets pins 1–6 as the input pins to control the instruction of the teaching system. Pin 9–16 is output pin, external processor and power supply part [14]. Pin 7 is COM pin and control pin 8 is grounded. After the pin function is configured, a freewheeling diode is built into the chip, and the internal structure of the chip is shown in Fig. 7.
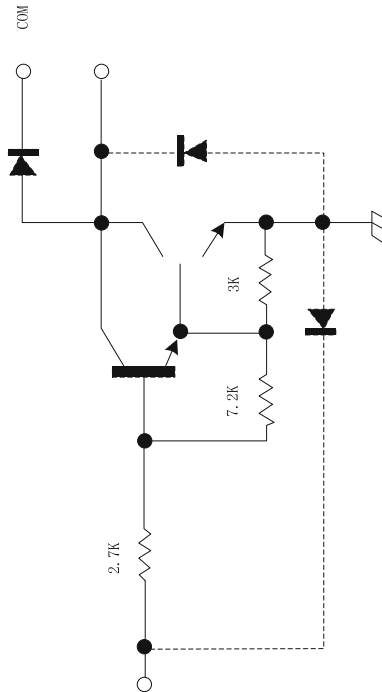
Fig. 7. Schematic diagram of internal structure of chip

As shown in the internal structure shown in Fig. 7, when the freewheeling diode chip output pin output low level. When the diode is not conducting, the output chip pin outputs high level. The high-level output value is too large, leading to the teaching system rate [15]. The hardware design of human resource simulation teaching system is completed.

3   Software Design of Human Resource Distance Education System

3.1   Using Internet of Things Technology to Build System Architecture

In the use of Internet of things technology to build a distance learning system, using B/S mode, using browser on direct access mode, in the front-end using htmlcss. JavaScript and other technologies to build a platform, in which HTML. CSS to achieve the display of the page, JavaScript to achieve the front-end business logic, including front-end data verification [16]. At the same time, we use Node. j s as our server-side language to quickly realize the construction of network services and application platform. The main function module on the server side directly supports the Internet of things browser and realizes the two-way communication with the client and the server side. WebRTC realizes the real-time transmission function of audio and video, and uses Express as the development framework of Node js [17] to ensure the smooth development of the server side. According to the needs of different users, the attributes of information-based teaching resources are demarcated. After contacting the entities of teaching resources with E-R diagram, they are converted to the selected DBMS record type. The sub pattern is used as the interface between the application and the resource database, and the data at the interface is collected and integrated into a data set A. A data transfer function is constructed

$$A(s) = \frac{\omega^2}{s^2 + Q} \tag{1}$$

In formula (1), s represents the data transmission time, Q represents the data transmission amount, and ω represents the buffer parameter. According to the above transmission process [18], it is assumed that the hardware structure has the same sensitivity to each incoming data, so an attribute parameter r is set, and the quantitative relationship of the parameters can be expressed as follows:

$$r^2 = (1 - e_{11})\left(1 + \frac{e_{11}(a_{11} + e_{21})}{2}\right) \tag{2}$$

In formula (2), e_{11} and e_{21} represent the data transmission amount at different times, and a_{11} represents the sensitivity parameter. Under the control of the attribute parameter, a shared signal delay parameter is set to form an attribute mode

$$\begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = P\begin{pmatrix} u_k \\ v_k \end{pmatrix} + E_2\begin{pmatrix} 1 & j \\ 0 & \kappa \end{pmatrix} \tag{3}$$

In formula (3), u_k represents data stability parameter, v_k represents data transmission speed, j represents delay parameter, κ represents hardware structure sensitivity to

data, k represents signal transmission time, and P represents teaching resource attribute parameter. Under the control of different signal transmission time, in order to unify the data format of teaching resource data, the above processing sensitivity parameters and delay parameters are normalized

$$G = f \frac{T}{\kappa j} \tag{4}$$

In formula (4), f represents the working frequency of the hardware element and T represents the working period. According to the change of the above calculation formula, the normalized frequency of attribute parameters changes as shown in Fig. 8
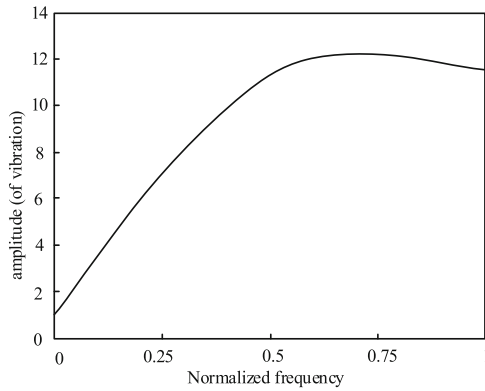
Fig. 8. Normalized change of attribute parameters

Under the normalized change of attribute parameters as shown in Fig. 8, the minimum amplitude frequency at different time points is controlled as the attribute division point, and the final system architecture is built.

3.2 Realizing the Function of Human Resource Distance Teaching

The front desk part focuses on the function composition of the students and teachers platform, registration module and login module. The front desk part is responsible for the teacher platform, student platform, registration function and login function. Among them, the function of the student platform is to effectively process the personal information, resource display, recent courses and other contents related to students [19]. Especially, given the student classroom, it can make students realize the existence of online video viewing, course browsing, group interaction discussion, etc., which is the key point of students to complete efficient learning. The teacher platform can handle the introduction and maintenance of each course, provide students with the necessary resources for effective teaching, enrich the knowledge reserve of the whole system, and the realization logic formed is shown in Fig. 9
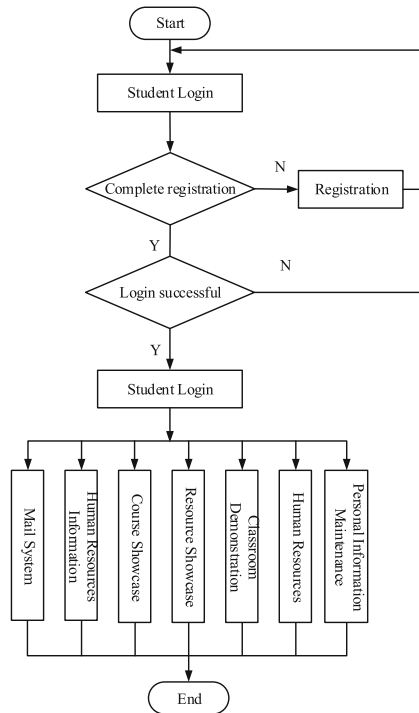
Fig. 9. The logical structure of the construction

Under the logic structure shown in Fig. 9, referring to the above logic situation, it can be made clear that students can perform the following operations in the platform when they formally log in to the system. First, the mail system in the station. This part can complete the mail receiving and sending, as well as the announcement and address book display operation. Students can use email to achieve daily mail delivery and other work. In the announcement, students can see the important announcement notes recently given. The address book is the email record between students and related personnel, which makes it easier to find the contact person. Second, the introduction of human resources, where students can refer to the details of their intended teachers [19]. Third, the latest course is displayed. There are recently introduced curriculum materials collected here, and students can check them in real time to determine the contents of the courses that are interested in in-depth study. Fourth, the latest resources display. Here, we collect the latest information, and students can check them in real time to determine the content that they are interested in in in-depth study. Fifth, students' classroom display. This will allow students to better complete the relevant work. Learning class can let students use it to realize online video access, course browsing, group interaction discussion and other operations.

4 System Test

4.1 Test Preparation

The main function of the system is online teaching management, the main function is teaching resource management, teaching program management, curriculum and training program management, curriculum learning and other functions. In addition, the online teaching system needs to meet the user access of multiple teaching bases, so the system test should ensure the correct operation of the system in terms of function and performance. Functional testing is mainly about requirement verification, which aims to ensure whether the final system implementation is consistent with the initial requirement analysis, whether it can meet the needs of users, and whether the requirement implementation is correct. The purpose of performance test is to verify whether the performance of the system can meet the requirements of multiple users accessing the system at the same time and whether the system runs normally. The test environment is as follows:

Table 1. Test environment parameters prepared

Composition of testing machine	Parameter name	Parameter details
Operator	CPU	Intel(R)Xeon(R)CPU ES-2430 0@2.20 GHz
	Memory capacity	64 G
	Hard disk capacity	1000 G
	Operating system	CentOS release 6.2 (Final)
	Web server	Node.js v0.10.26
	Database	MySQL5.6
Service machine	CPU	AMD A8-3870 APU with Radeon(tm) HD Graphics 3.00 GHz
	Memory capacity	18G
	Hard disk capacity	500G
	Operating system	Windows 7 旗舰版Service Pack 1
	Browser version	IE9, IE10, IE11, Firefox, Chrome, 360 browser

Under the parameters of operation machine and service machine prepared in Table 1, function test needs to cover the requirements analysis, so the design of requirement point and function test case is very complicated. Each requirement needs a large number of test cases to verify. As an essential part, testing accounts for more than 70% of a software life cycle. It shows that testing is very important in the whole process of software development, and software testing fully reflects the reliability of software. Through the testing process, we find out the neglected errors in the process of software development. In the process of software development, every link of the system should be tested periodically, and the problems should be found and repaired immediately. First of all, unit test is needed to test whether the components of the system are complete. This step is the test of the most basic module, which is usually completed by the programmer independently of the tester in the process of system development. The complete testing process needs to be completed by checking the syntax of the original code, finding data exceptions and detecting the basic algorithm of the program. Here, we need to build a test case and import it into the module under test to test the correctness of the module in the running state. The test environment is shown in Fig. 10.
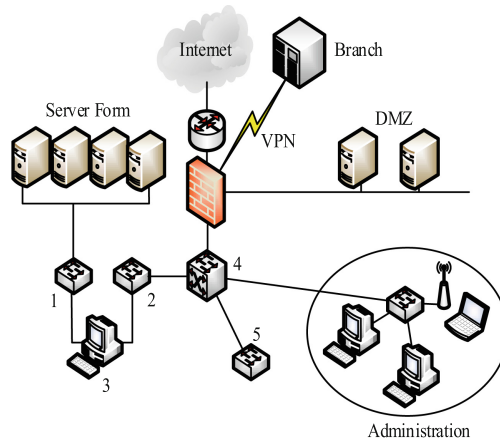
Fig. 10. System test environment

In the test environment shown in Fig. 10, software testers need to use the software as a user, test the input of the software through various situations, and check whether the output of the software is normal. The problems found in the process of trial operation are continuously corrected and improved, which are divided into the following three stages. First of all, test the software source code, according to the expected effect of online teaching system components, run the program code and feedback. At the same time, check the software source code, find the errors in the system module and correct them in time. Then, test the function of the system module. Use a variety of accounts of people with different identities to log in to the system, and perform the functions of reading, searching, adjusting, deleting, etc. to verify whether the software can achieve the expected effect, and whether the input information is correctly processed and applied. The third is to carry out open trial operation of the platform, collect and summarize

the problems encountered by teachers and students in the trial process, and make corrections and improvements. Delimit the use case number of human resource distance education system, and correspond to different operation processes in the number. The corresponding code and operation process are shown in Table 2.

Table 2. Set operation process and code

Operation name	Operation steps	Process description
CZMC-01	Fill in the account	Jump to page after correct input
CZMC-02	Fill in the password	If the password is filled in incorrectly, the corresponding prompt will be given
CZMC-03	Select login	Display management information
CZMC-04	Click on the question type	Navigate to the optional module
CZMC-05	Entry title	Jump to submit page
CZMC-06	Set the correct answer	Mark the score and give the score
CZMC-07	Input analysis	Jump after correct submission
CZMC-08	Select submit	Jump to the next interface
CZMC-09	Show add title content	Module classification
CZMC-10	Right key	Constructing the evaluation process
CZMC-11	Scoring criteria	Show the end after completing the selection

In the set operation process and compilation, prepare two traditional teaching systems and the designed teaching system for testing, and compare the performance of three teaching systems. The knowledge base used in the experiment in this article is based on a limited field. Here, the relevant courses such as "Comprehensive Experiment of Human Resource Management" in university undergraduate teaching are selected as the source of the knowledge base, and there are more than a thousand questions in the question bank.

4.2   Results and Analysis

Based on the above experimental preparation, the operation of the three kinds of teaching system is controlled, and the online number of the three kinds of distance teaching system is 50–1000. Taking 50 people as a functional test group, corresponding to different test items. The response time of the system is taken as the experimental index. The shorter the response time, the greater the throughput per unit time of the system and the more stable the operation. The response time of online teaching system is shown in Table 3.

Table 3. Response time of three distance education systems

Number of people online at the same time/person	Response time/S		
	Traditional online education system 1	Traditional online education system 2	The design of online education system
50	8.25	5.01	2.12
100	8.57	5.49	2.12
150	8.59	5.41	2.13
200	8.57	5.59	2.97
250	8.84	5.44	2.95
300	8.92	5.39	2.81
350	8.46	4.84	2.51
400	8.31	5.23	2.08
450	8.72	4.72	2.66
500	8.03	5.46	2.53
550	8.36	4.25	2.05
600	8.99	4.02	2.59
650	8.28	5.77	2.54
700	8.85	4.85	2.15
750	8.48	5.16	2.62
800	8.23	4.47	2.89
850	8.84	4.65	2.85
900	8.81	4.27	2.67
950	8.37	5.72	2.43
1000	8.92	5.15	2.53

According to the experimental results shown in the table above, the response time of traditional distance education system 1 is about 8.5 s, and the actual response time of the system is the longest. The response time of traditional distance education system 2 is about 5 s, but the actual response time of the system is longer. The response time of the designed teaching system is about 2.4 s. Compared with the two traditional teaching systems, the response time of the designed teaching system is the shortest.

Keep the above-mentioned experimental environment unchanged, the number of control system interface debugging is 20 times, call the task manager of the hosting system hosting server, calculate and count the call success rate of the three teaching systems, the higher the success rate of interface call, the system interface connection effect The better, the experimental results are shown in Table 4:

Table 4. Call success rate of three teaching systems

Debugging times	Call success rate/%		
	Traditional teaching system 1	Traditional teaching system 2	The designed teaching system
1	28.3	60.7	84.5
2	25.3	40.9	82.7
3	27.2	57.4	81.9
4	28.2	43.7	87.4
5	29.6	53.8	82.8
6	27.4	41.7	80.9
7	20.9	47.5	80.7
8	27.6	53.1	89.4
9	27.7	58.9	85.1
10	25.9	41.1	80.9
11	24.7	50.7	87.2
12	26.1	48.8	80.9
13	23.9	55.3	86.6
14	26.9	55.2	84.6
15	21.3	54.4	83.9
16	21.8	55.3	88.3
17	21.5	60.2	81.5
18	24.3	43.1	83.5
19	28.2	52.5	85.5
20	25.4	43.4	89.7

According to the debugging results shown in Table 4, the call success rate of traditional teaching system 1 is about 25%, and the actual debugging success rate of serial port is low. The success rate of traditional teaching system 2 serial port debugging is about 50%, but the actual success rate of debugging serial port is high. Compared with the two traditional teaching systems, the success rate of the designed teaching system is the highest.

In the above experimental environment, the test number of 500 people used in the three teaching systems is controlled, and the concurrent test number of distance teaching is set to be 10–200 groups. According to the set concurrent test group, the network packets used by the teaching system are counted and calculated. The experimental results are shown in Table 5.

Control three kinds of teaching system to deal with the teaching system with 500 people online at the same time, after setting different number of concurrent tests. According to the experimental results shown in Table 5, the average number of network packets

Table 5. The number of network packets used by the three teaching systems

Number of concurrent tests	Number of network packets		
	Traditional teaching system 1	Traditional teaching system 2	The designed teaching system
10	301	182	71
20	253	181	82
30	264	166	88
40	273	180	90
50	266	193	94
60	266	173	98
70	315	170	102
80	225	152	106
90	337	189	113
100	263	196	116
110	237	197	124
120	299	167	125
130	280	191	128
140	243	162	134
150	378	160	137
160	225	164	142
170	215	188	150
180	331	152	150
190	379	178	151
200	388	194	151

required for the normal operation of the traditional teaching system 1 is about 300, and the actual operation of the teaching system uses a large number of network packets, which is prone to network congestion. The average number of network packets required for the normal operation of traditional teaching system 2 is about 160, while the number of network packets required for the actual operation is less. Compared with the two traditional teaching systems, the designed teaching system uses the most network packets in processing.

5 Conclusion

The progress of network information technology accelerates the information circulation and improves the efficiency of online teaching circulation. In the past, the traditional teaching method was to teach through the topic, resulting in a waste of human resources. However, with the gradual popularization of online teaching, the whole teaching process

is electronized to reduce the waste of resources, improve the teaching quality and efficiency, and promote the accumulation and sharing of academic achievements. Based on the consideration of users, using physical network technology as the basis, the design of human resources distance teaching system can improve the shortcomings of traditional teaching system, and provide theoretical support for the design of teaching system in the future.

References

1. Casolla, G., Cuomo, S., Di Cola, V.S., et al.: Exploring unsupervised learning techniques for the Internet of Things. IEEE Trans. Industr. Inf. **16**(4), 2621–2628 (2019)
2. Hou, R., Kong, Y.Q., Cai, B., et al.: Unstructured big data analysis algorithm and simulation of Internet of Things based on machine learning. Neural Comput. Appl. **32**(10), 5399–5407 (2020)
3. Min, Y.A.: Problem solving path algorithm in distance education environment. J. Korea Soc. Comp. Inform. **26**(6), 55–61 (2021)
4. Boussakssou, M., Hssina, B., Erittali, M.: Towards an adaptive E-learning system based on Q-learning algorithm. Procedia Comput. Sci. **170**, 1198–1203 (2020)
5. Aljawarneh, S.A., Vangipuram, R.: GARUDA: Gaussian dissimilarity measure for feature representation and anomaly detection in Internet of Things. J. Supercomput. **76**(6), 4376–4413 (2020)
6. Wang, P., Xue, F., Li, H., et al.: A multi-objective DV-Hop localization algorithm based on NSGA-II in internet of things. Mathematics **7**(2), 184 (2019)
7. Liu, S., Sun, G., Fu, W. (eds.): eLEOT 2020. LNICSSITE, vol. 339. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63952-5
8. Rico-Bautista, D., Medina-Cárdenas, Y., Guerrero, C.D.: Smart university: a review from the educational and technological view of Internet of Things. In: Rocha, Á., Ferrás, C., Paredes, M. (eds.) Information Technology and Systems: Proceedings of ICITS 2019, pp. 427–440. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-11890-7_42
9. Liu, S., Li, Z., Zhang, Y., et al.: Introduction of key problems in long-distance learning and training. Mob. Netw. Appl. **24**(1), 1–4 (2019)
10. Gao, P., Li, J., Liu, S.: An introduction to key technology in artificial intelligence and big data driven e-learning and e-education. Mob. Netw. Appl. 1–4 (2021)
11. Xia, Y.: Resource scheduling for piano teaching system of Internet of Things based on mobile edge computing. Comput. Commun. **158**, 73–84 (2020)
12. Zhou, X., Li, X., Su, N.: Design and internet of things development of network teaching resource base system for educational technology. J. Phys. Conf. Ser. **1769**(1), 012005 (2021)
13. Rao, A.R., Clarke, D., Bhdiyadra, M., et al.: Development of an embedded system course to teach the Internet-of-Things. In: 2018 IEEE Integrated STEM Education Conference (ISEC), pp. 154–160. IEEE (2018)
14. Akbar, M.A., Rashid, M.M.: Technology based learning system in Internet of Things (IoT) education. In: 2018 7th International Conference on Computer and Communication Engineering (ICCCE), pp. 192–197. IEEE (2018)
15. Savov, T., Terzieva, V., Todorova, K.: Computer vision and Internet of Things: attention system in educational context. In: Proceedings of the 19th International Conference on Computer Systems and Technologies, pp. 171–177 (2018)
16. Vlasov, A.I., Yudin, A.V., Shakhnov, V.A., et al.: Design methods of teaching the development of internet of things components with considering predictive maintenance on the basis of mechatronic devices. Int. J. Appl. Eng. Res. **12**(20), 9390–9396 (2017)

17. Huang, Y., Jin, X.: Innovative college English teaching modes based on Big Data. Kuram ve Uygulamada Egitim Bilimleri **18**(6), 3428–3434 (2018)
18. Zhang, J.: Design and analysis of intelligent teaching system in colleges and universities based on big data application. In: 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), pp. 431–434. IEEE (2021)
19. Solé-Beteta, X., Navarro, J., Vernet, D., et al.: Automatic tutoring system to support cross-disciplinary training in Big Data. J. Supercomput. **77**, 1818–1852 (2021)

Intelligent Encrypted Storage Method for Medical Health Database Based on Internet of Things

Qing-bang Zeng[1]($\boxtimes$) and Wen-da Xie[2]

[1] College of Information Engineering, Jiangmen Polytechnic, Jiangmen 529000, China
honda455@163.com
[2] Computer Engineering Technical College (Artificial Intelligence College), Guangdong Polytechnic of Science and Technology, Zhuhai 519090, China

Abstract. By analyzing the process of encrypting and decrypting, the traditional method of encrypting and storing medical health database is designed, and the data in database is transformed slowly, which leads to high cost and long time of encrypting. The intelligent encryption of medical health database is divided into three parts: encryption algorithm, decryption algorithm and matching algorithm, and the key family, key cycle and key management mode of the three parts are determined to design the intelligent encryption algorithm for medical health database; the relationship between the encryption and decryption process and user read-write operation is analyzed to design the database encryption and decryption processing mode and transform database data; the cloud computing technology in the Internet of Things technology is adopted to design the database encryption scheme based on the Internet of Things technology to realize the intelligent encryption algorithm for medical health database. The experimental results show that the size of encrypted and unencrypted data is 7% and 4%, and the average ratio of key management time to data processing time is 0.274% and 0.298%, respectively, in one operation cycle, with less space overhead, faster database encryption speed and better encryption performance.

Keywords: Internet of Things · Medical treatment · Health data · Intelligent encryption · Storage method

1 Introduction

With the rapid development of the Internet of things, cloud computing has become the main supporting technology of the Internet of things due to its strong computing power, convenient access, low cost and other advantages, which is the development direction and research focus of the IT field at present [1]. At present, cloud computing has entered a period of rapid development, and various big data applications based on cloud computing are gradually enriched. Users' awareness of adopting cloud services to process various computing, storage and other tasks has increased. Many governments, educational and

medical institutions and large enterprise groups take the initiative to consider integrating their business into the cloud. In the era of rapid expansion of information, cloud storage in the storage and management of large data applications has a very obvious advantage.

Medical cloud combines cloud computing, cloud storage and big data processing technology to improve the construction of hospital intelligent information technology to a higher level. First of all, cloud storage and cloud computing are introduced into the construction of medical cloud. Through unified deployment and centralized management, the investment of hardware equipment for hospital information management is reduced [2], the utilization ratio of storage resources is improved, and the operation and maintenance costs of medical information system are reduced. Secondly, due to the popularization of cloud computing in health care system and major hospitals, a large number of clinical data, patient information and so on are rapidly expanding to form medical big data. Finally, medical big data itself has high application and research value, so it is necessary to extract valuable and meaningful data from hospital information platform data to construct medical big database.

Traditional personal computer and server architecture can not meet the requirements of medical big data processing [3]. Cloud storage technology provides an efficient management mode for medical big data processing. In the application of medical cloud system, users are most concerned about the data security, especially the sensitive medical information data. In medical cloud system, in order to protect the data security in sensitive medical database, the common method is to encrypt the data in the database and store the encrypted medical data to cloud server. However, the data encrypted by traditional encryption algorithm has lost a lot of data processing functions. Therefore, how to encrypt the data of medical database efficiently is one of the urgent problems in the application of medical cloud.

In reference [4], an intelligent database encryption method based on AES algorithm is proposed. The method comprises the following steps: constructing a security key for encrypting sensitive database information, adopting an arithmetic coding design scheme to carry out adaptive feature classification and vector quantization coding design of sensitive database storage information, carrying out cyclic code reconstruction and block encryption design of sensitive database storage data encryption under advanced encryption standard protocol (AES), constructing encryption and decryption keys, and carrying out feature block recombination on plaintext decoded by intelligent encryption cyclic code of sensitive database, The random sampling method is used to reconstruct the sample space of the information stored in sensitive database, and the multi-layer encryption design of highly sensitive database storage is realized by combining the key negotiation and elliptic linear coding method. In the literature [5], a decentralized blockchain information management scheme is proposed to realize the safe storage of medical data. The scheme adopts improved PBFT consensus algorithm and optimized Hash encryption algorithm, and stores medical data safely and effectively in distributed database, ensuring the integrity and traceability of medical data. At the same time, a brand-new data interaction system is designed to prevent the direct interaction between the third party and the database, so as to prevent the untrustworthy third party from maliciously destroying the medical data and ensure the security of the data. Finally, the access control and Lucene retrieval mechanism are used to protect patients' privacy

and realize fast retrieval of medical data. Experimental analysis shows that compared with algorithms such as Proof of Work, POW) and Delegated Proof of Stake, DPOS), the improved PBFT consensus algorithm provides better stability and throughput for medical blockchain system. Compared with the common database interaction, the data interaction system effectively prevents the direct operation of the database, and has better security and tamper resistance.

The most common method of database encryption is based on a single record. The basic idea of this method is that each record is encrypted into ciphertext and stored in the database file, and the smallest granularity of decryption is generally records. Each foreign large-scale database manufacturers have more or less provided the solution of database encryption, partly solved the problem of encrypting confidential data. However, due to various limitations, the solutions provided by various database vendors can not achieve the effect of multi-user sharing key and encrypting data, and it is very troublesome to migrate ciphertext data between different DBMSs. Because the application of database and computer network in our country is later than abroad for many years on the whole, so the need of database encryption software in our country is very urgent. Through the use of peripheral encryption software, the data in database is encrypted, and a database encryption system can be designed. The encrypted tables and fields can be chosen by users freely, and the encryption process can be transparent to the application.

Therefore, this paper proposes to use the Internet of Things technology to design an intelligent encryption storage method for medical and health database. Through collusion mode, the key family, key period and key management mode of intelligent encryption algorithm of medical database are determined. According to the relationship between encryption and user operation, the database encryption and decryption processing mode is designed, the conversion of database data is realized, and the database encryption scheme using Internet of Things technology is realized. It can provide theoretical support for the safe storage of medical data, ensure the secrecy of medical data, and has important practical significance for the safety of patient privacy.

2   Intelligent Encrypted Storage Method for Medical Health Database Based on Internet of Things

2.1   Design Intelligent Encryption Algorithm for Health Care Database

The design of the health database intelligent encryption algorithm, which is divided into encryption algorithm, decryption algorithm and matching algorithm in three parts. Among them, encryption algorithm is to encrypt plaintext into ciphertext, according to the type of plaintext is divided into ordinary string encryption and pattern string encryption algorithm; decryption algorithm is to restore the ciphertext data in the database to plaintext data [6], and can be based on the check algorithm to infer whether ciphertext is tampered with; matching algorithm is to match the characters in the pattern string with the characters in the ordinary string stored in the database one by one, so as to find out ciphertext data containing keywords. In addition, we also need to determine the key family, key cycle and key management.

Encryption Algorithm

The design of intelligent encryption algorithms for health care databases will use a collusion pattern, which often contains ordinary characters and special characters, such as letters a, b, c, d, and special characters such as *, /, and ?, +, {, }, etc. At the same time, numbers can be either ordinary characters or special characters. A special character in a pattern string has a special meaning, which generally indicates that the preceding ordinary character occurs more than once, such as a indicates that a occurs one or more times. For a pattern string can not be simply encrypted into ciphertext, because the pattern string of special characters encrypted into ciphertext, will lead to special characters lost its meaning, then the server's regular parsing engine can not parse the meaning of the expression [7]. For example, if the pattern string a is encrypted into ciphertext, the server cannot know what it means, but only encrypts a, and if the + does not, the corresponding ciphertext is repeated one or more times for the server. So, for ordinary characters can use encryption technology, to hide its specific meaning. However, special characters should retain their specific meaning and can contain ordinary characters that are encrypted or not encrypted using encryption algorithms. The result is an ambiguity in the ciphertext string. For example, suppose the ciphertext character a is encrypted and becomes a sign, then the ciphertext string ++ can mean either the ciphertext string aa or the ciphertext string a +, whereas aa and a + have completely different meanings in a regular expression.To solve this problem, a flag flag is introduced to distinguish between a character that is a plain ciphertext character data or an unencrypted special character data. The form of the flag prefix code is combined with each ciphertext character, meaning that each clear text character in the pattern string is processed (encrypted or reserved) and the flag bit is added before that character. The length of flag is one byte (8 bits).

After the introduction of flag bit and special character encoding, the design of mode string encryption is as follows: mode string encryption granularity is a single character, for the processing of a single character, first determine its character category, according to the character category set flag value [8]. The characters shall be encrypted. If they are ordinary characters, they shall be encrypted into ciphertext characters by AES algorithm. If they are special characters (key characters in regular expression), they shall be encoded according to the encoding method shown in Table 1.

Table 1. Special character encoding table

Character	Code	Character	Code
^	0x0000	*	0x0001
$	0x0002	#	0x0003
+	0x0004	?	0x0005
{	0x0006	}	0x0007
[	0x0008	]	0x0009
.	0x000A	,	0x000B
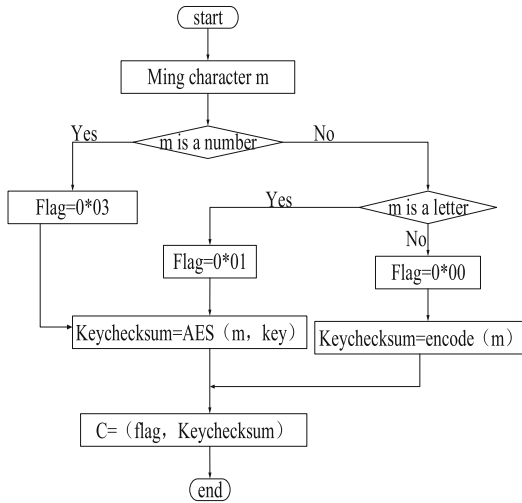
Fig. 1. Flow chart of modal string encryption algorithm

The encrypted characters are called trap gates, which are recorded as keychekcsum, and the flag flag is then stitched into the trap keychecksum to form the final ciphertext data, as shown in Fig. 1. The ciphertext pattern string can be obtained by processing each character in the pattern string, and then the ciphertext data satisfying the regular expression can be queried by ciphertext retrieval algorithm.

Decryption Algorithm
Decryption of ciphertext data is the inverse operation of encryption algorithm. The ciphertext character is stored as ({c} checksum), so decrypting c gets the corresponding plaintext character. Firstly, C is decrypted by AES algorithm. The plain-text character m can then be calculated by the exclusive OR operation of the random numbers r and c produced by the random number generator [9]. In order to verify whether the decrypted data has been tampered with, the AES algorithm can be used to encrypt m to get the trap keychecksum. Through the irreversible hash algorithm MD5, c and key are calculated to get the check code. Finally, the comparison between cs and ckecksum is equal. If not, the ciphertext data c is tampered with. Because the key in the pattern string is different from the key in the normal string, an attacker cannot change the value of checksum even if the normal string key is compromised.

Matching Algorithm
The matching granularity of ciphertext matching algorithm is single character, and the receiving object is ciphertext pattern string. Each ciphertext character in a ciphertext pattern string is composed of ({flag} keychecksum). First, the meaning of the ciphertext pattern string is parsed by analyzing whether each keychecksum is an ordinary character encrypted or a special character encoded using flag flags [10]. Then ciphertext matching algorithm is used to match ciphertext characters one by one. For ciphertext matching

algorithms, the storage format of ciphertext characters is ({o} oheoksum) and the character format of ciphertext pattern strings is ({flag} keychecksum), so the value of ({c} keychecksum) is computed as checksum 'through the irreversible hash algorithm MD5, comparing whether checksum is equal to a checksum, and if equality proves that the two ciphertext characters correspond to the same plaintext character, matching can proceed downward until a destination string that satisfies the meaning of the pattern string is found.

Key Family

Key family A is used to encrypt ID number and Key family B is used to encrypt hospital information. The key family B is used to encrypt the same group of data [11]. Each key in the family of keys has its own specific encryption object, but at the same time, only one key should be used for encryption, while the others may be used to decrypt data, and some keys may have expired, as shown in Fig. 2.
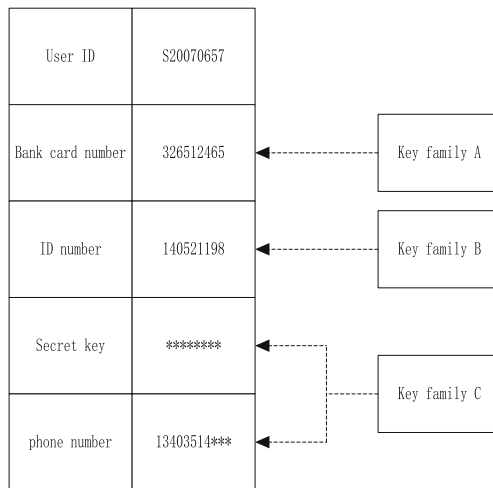
Fig. 2. Key family example

Consider assigning a key family to each column in the database that needs to be encrypted. If there are three columns in a table that need to be encrypted, three key families should be set for encryption [12]. As an example of the key family shown in Fig. 2, it is assumed that ID number, bank card number, key and mobile phone number are the data that need to be encrypted and protected. Key family A encrypts bank card number, Key family B encrypts ID number, Key family C encrypts key and mobile phone number.

However, if each key family encrypts one column, all keys are fetched for decryption when a row of data is read, which degrades system performance. If only one key family is used for all columns, the impact on system performance can be reduced to some extent, although the extent depends on how well the key engine handles the request. In fact,

the performance gains from encrypting multiple columns with the same key family are limited, so it is not worth sacrificing security.

Key Cycle

Key cycle, that is, the key from the creation, waiting, enabled, expired, deactivated until the destruction of the entire process. One key cannot be used all the time, because the more and longer data encrypted with the same key, the more likely an attacker is to crack the key and the less secure it is, so the key needs to be changed periodically [13]. During the replacement process, all data encrypted with the old key needs to be decrypted, and then encrypted with the new key, the old key will be discarded, and will be destroyed later. The whole process of the key cycle is shown in Fig. 3.
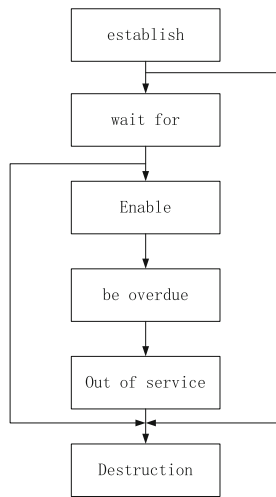
Fig. 3. Key cycle

A key is created as required, and is assigned a key family and activation time. Before the activation time arrives, it is in a waiting state, when the activation time is reached, it is enabled, active, and is used by the encryption system to encrypt and decrypt data until it is disabled by the new key and destroyed by the system after a certain time. Keys that are expired will never be used again. Deactivation key should be deleted as soon as possible [14], even if the key is no longer used, there are security risks, a little improper handling, it may cause data leakage.

When the old key is replaced by the new key, all data encrypted by the expired key is decrypted and then encrypted with an active key. When all the data is encrypted, the expired key is marked as inactive and may be deleted very quickly [15]. There is also a risk that a large amount of data can be decrypted and encrypted in a very short time. Key replacement is a very resource-consuming process. Frequent key changes may result

in errors and outweigh the benefits. Therefore, the time for key replacement should be reasonably set according to the actual needs of the data system.

Key Management Method

Key management plays an important role in database encryption technology. Even if the encryption intensity is high and the key algorithm is complex, once the key is compromised, the data in the database is not safe and the encryption is meaningless. Key distribution management is a difficult problem, which involves the whole process of key creation, wait, enable, expire, deactivate and destroy.

There are two kinds of keys in database encryption system: user key and data key. User key is mainly used to query and modify data. A user may provide his personal information (user name, secret key, etc.) to the database management system, and after the verification and confirmation of the legitimate authority of the user, the user may conduct corresponding operations. Depending on the permissions, database users are divided into ordinary users and database administrators. Ordinary users may have different permissions, such as VIP users who generally have more permissions than free users. Database administrator has the most authority to deal with both plaintext data and ciphertext data. For the sake of security, there may be many administrators managing a database. Even so, administrators are often a hidden danger of database security that can not be ignored.

Data keys fall into two categories: master keys and working keys. The master key is used for encryption and is changed periodically. The security of the master key is very important to the whole database system, because once it is compromised, all the data will be in an insecure state. Work keys are used to encrypt tables or data items, and each table or data item requires a different work key.

Keys are often generated randomly through mathematical functions, there is no law, so it is impossible to predict the statistical analysis of the key attacks is impossible, so the key itself without security risks. The key generating function shall meet the following conditions: the probability of the same key of different data items is extremely low; even if some information of the data items is obtained, it is impossible to infer the information of other data items from ciphertext; the key of one data item cannot infer other keys; and the key generating speed shall be as fast as possible.

The key storage is divided into: the master key storage, the work key storage, user key storage. The storage of the master key is suggested to use some hardware encryption devices with high encryption strength and inviolability. If it is implemented by software, the master key generator should be kept out of all hidden troubles, and the administrator who manages the master key should be absolutely trusted. The master key should be changed periodically for security reasons.

The working key is usually encrypted by the master key and stored in the database data dictionary in the form of ciphertext. Because of its large number, it is generally not allowed to modify the operation directly.

Generally, the user key is stored in the database table, the user will save a copy, so in addition to ensuring database security, the user should also properly keep the key, it is best to use key protection card to access the database.

2.2   Database Encryption, Decryption and Conversion

In addition to the encryption algorithm, the encryption and decryption of the method is the basis of the whole operation process throughout the method. It includes both the encryption and decryption of user information and the encryption and decryption of data keys.

The decryption process is similar to encryption except that the parameters of the Cipher object are set slightly differently. It is important to note that the encryption and decryption processes cannot be performed simultaneously. If you want to do both encryption and decryption, you need two instances of the Cipher object. In fact, we can find that the whole system's encryption/decryption process is closely related to the user's read-write operations. The relationship between the encryption and decryption processes and user read-write operations is shown in Fig. 4.
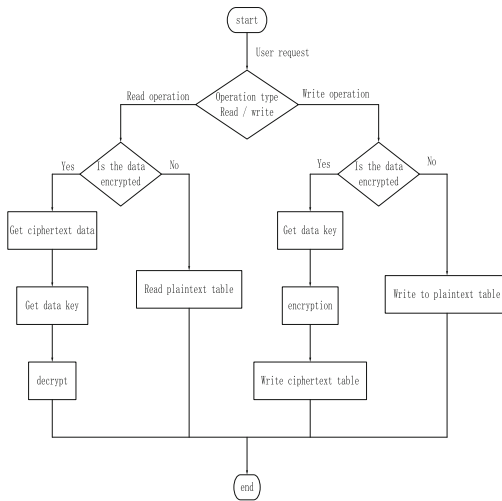
Fig. 4. Relationship between encryption and decryption and user read and write operations

As you can see in Fig. 4, when a user reads, the system determines whether the pre-read data is encrypted. If not encrypted, the system directly reads the plaintext data and returns. If it has been encrypted, the system will obtain the relevant data keys for decryption. When a user writes, the system also determines whether the pre-written data needs to be encrypted. If no encryption is required, the system writes the plaintext data directly and returns. If encrypted storage is required, the system will obtain the relevant data keys after encrypting the plaintext data.

In this study, in order to facilitate users to observe the changes of data before and after encryption, binary data is converted into hexadecimal string and stored in the database. Here is mainly for the character type of data encryption. The data conversion code is contained in the Utils class, which is used to convert byte arrays to hex strings. It contains two methods, {by} tes? The HexString () method is used to convert a byte array to a

hexadecimal string. Another hex String2Bytes () method, which is inversely related to the above method, converts a hexadecimal string to a byte array.

Because the computer directly processes the binary data the speed is quicker, therefore in the actual application, directly uses the binary form to encrypt the ciphertext data carries on the storage. Also, binary data saves more storage space than encoded hexadecimal data.

2.3 Intelligent Encrypted Storage of Medical Health Database Based on Internet of Things Technology

The cloud computing technology of Internet of things technology is adopted to realize the intelligent encryption algorithm of medical health database. The process is as follows: Firstly, the data users in cloud computing database are partitioned by sharding functions, and the data is distributed to different processors by the master node as the dispatcher. Each Mapper completes a portion of the final result. The Reduce function is responsible for consolidating some of the results of all Mappers, as shown in Fig. 5.
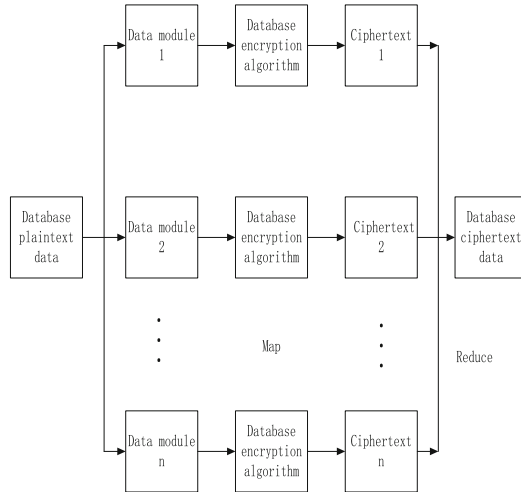
Fig. 5. Database encryption scheme based on Internet of Things

As can be seen from the analysis of Fig. 5, the implementation of intelligent encryption algorithm for health care database based on Internet of Things technology mainly includes collusion mode encryption stage and parallel encryption stage, which are described below:

Fragment function to realize the encryption stage (serial encryption stage): the initial file is divided into n fragments through one-time fragmentation, in which the file size is represented as L, the i fragments as i and the i fragment database data size is represented as L_i. In order to decide whether the file reading process is over or not, we use POS to represent the location of the real-time processed data in the initial file, and use i to store the data after the slicing through filebuffer, and i sliced data filebuffer to store the value.

The encryption results of the serial encryption phase are processed in parallel, and the Map function realizes the encryption phase as (the parallel encryption phase): any Map node is encrypted in its entirety, resulting in any block of ciphertext. Set the input of the Map function as key/value pair, the offset of the block data in the initial data is key, and the clear text representation of the block data is value.

The Reduce function implements the encryption phase as (parallel encryption phase): after the Map phase encryption processing is completed, the key value pairs of all nodes are counted and sorted by the value of the key as a weight, and the value is pieced together to get the final file.

A sorting algorithm is used to process the file assembled from the above values, and the encryption complexity calculation formula is obtained as follows:

$$O(t) = s \qquad (1)$$

In this formula, t represents time, O represents encryption complexity, and s represents the encryption process. At this point, assume that the constant of the encryption process is α and t_0, and at $t > t_0$, there are $O \leq \alpha s$, and assume that the complexity associated with sorting in the Reduce function process is $m\ x$ operations, and $\alpha > 0$, then use expression (2) to minimize the encryption complexity:

$$O_{\min} = \min \alpha ks O(t) \qquad (2)$$

In the formula, k represents the offset of the chunked data from the initial data.

Based on the above contents, the algorithm, data encryption, decryption and conversion process, as well as the two calculation formulas are substituted into the database encryption scheme diagram based on the Internet of Things technology shown in Fig. 5 to realize the intelligent encrypted storage of medical health database.

3   Experiment and Analysis

Two groups of traditional encrypted storage methods were selected, and the health database was chosen as the experimental object to verify the intelligent encrypted storage method of health database based on Internet of Things.

3.1   Experimental Preparation

According to the operating characteristics of the encryption storage algorithm selected in this experiment, three sets of encryption storage methods are set up on the computer, as shown in Table 2.

Table 2. Running environment

The environment	Configure the	Specification
Software	Operating system	Unbutu
	Operating system Type	The 64 - bit
	Agent server open source software	SNProxy
	Network card	Gigabit network
	Network card transmission rate within LAN	1Gbps
	Database	MYSQL 5.0.41
	Auxiliary Function Writing Language	Javascript
	Test the script	Python
	IDE	MyEclipse 7.5
Hardware	Server server	MongoDB
	Ser ver Version	MongoDB3.0.2
	Proxy server	SNProxy
	CPU of physical machines	Intel i5 7500
	CPU Memory	8GB
	CPU main frequency	2.9 GHz
	Mechanical hard disk capacity	500GB

Based on the running environment of encrypted storage method shown in Table 2, the encryption and decryption algorithms of the three encrypted storage methods are stored in the SNPROXY proxy server.

During the test, the client first connects to the MongoDB server via the SNProxy proxy server to perform the query and record the output. The client then connects directly to the MongoDB server to perform the query and record the output. Compare the two query results before and after, observe whether the first query results are clear and correct, and whether the second query results are ciphertext data. This verifies the performance of three cryptographic storage methods.

Through the method of comparison test, the time consumed by various basic operations (adding, deleting, modifying, querying) of MongoDB server database without using SNProxy proxy server and using SNProxy proxy server is measured respectively. The time is the time interval between the client initiates the request and receives the reply, so as to determine whether the response time of the whole system can be severely slowed down when the system performs the encryption and decryption calculation.

In the design of SNProxy proxy server, the key management mechanism is also introduced, and the additional time cost is introduced in the generation of encrypted sub-key and the management of online and offline user information.

3.2   Experimental Results

Results of the First Set of Experiments
According to the experimental process of this experimental design, three groups of encrypted storage methods are tested respectively. In the functional test process of this design, when inserting, key-value query, aggregation, range, fuzzy query, update, delete and other operations are carried out, the function test result is shown in Table 3.

Table 3.   Functional test results table of three cryptographic storage methods

Method	Type of operation	Implementation results	Output results	
			SNProxy	MongoDB
Research method	Insert operation	Correct	–	–
	Key Value Query	Correct	Plain text	Secret text
	Aggregation operations	Correct	Plain text	Secret text
	Scope operations	Correct	Plain text	Secret text
	Fuzzy query	Correct	Plain text	Secret text
	Update operations	Correct	–	–
	Delete operations	Correct	–	–
Traditional method 1	Insert operation	Correct	–	–
	Key Value Query	Correct	Plain text	Secret text
	Aggregation operations	Error	Plain text	Plain text
	Scope operations	Correct	Plain text	Secret text
	Fuzzy query	Error	Plain text	Secret text
	Update operations	Correct	–	–
	Delete operations	Error	Plain text	Secret text
Traditional Method 2	Insert operation	Error	Plain text	Secret text
	Key Value Query	Correct	Plain text	Secret text

(*continued*)

Table 3. (*continued*)

Method	Type of operation	Implementation results	Output results	
			SNProxy	MongoDB
	Aggregation operations	Error	Plain text	Secret text
	Scope operations	Correct	Plain text	Secret text
	Fuzzy query	Correct	Plain text	Secret text
	Update operations	Correct	–	–
	Delete operations	Correct	–	–

As can be seen from Table 3, in the process of testing the three groups of encrypted storage methods designed in this experiment, the output results of the traditional method 1 are all clear text data when the client uses SNProxy for query, but in the process of deletion, errors of encrypted storage operation occur, which leads to multiple clear text in the output results when the client directly connects to the MongoDB database for query again; the traditional method 2 uses SNProxy for query, the output results are all clear text data when the client directly connects to the MongoDB database for query again, but in the process of insertion, errors of encrypted storage operation occur, which leads to the existence of clear text in the output results when the client directly connects to the MongoDB database for query again; the research method uses SNProxy for query and the query results are correct, and when the client directly connects to the MongoDB database for query again, the output results are all clear text data. This means that SNProxy encrypts the data and enables the MongoDB database to perform a query on the ciphertext. Obviously, this research method has the better encryption performance.

Second Set of Experimental Results

All the experimental data selected in this experiment are normalized, and the experimental data size is controlled to one unit. The data size of the three methods after encryption is calculated, and the ratio of the data size before encryption to the data size before encryption is compared. The data size ratio before and after encryption is compared. The experimental results are shown in Fig. 6.

It can be seen from Fig. 6 that the data size ratio of the medical health database data encrypted and stored by traditional method 2 is the largest before and after encryption, which indicates that the data of the medical health database encrypted by traditional method 2 increases, resulting in large space overhead; The ratio of data size before and after encryption in traditional method 1 is 3% smaller than that in traditional method 2, which indicates that the data of medical health database encrypted by traditional method 1 increases and the space overhead is smaller than that of traditional method 2. The data size ratio before and after encryption is the smallest, which is 7% and 4% smaller
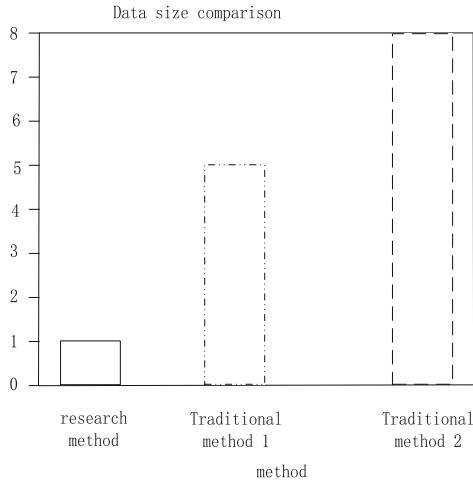
Fig. 6. Data size ratio before and after encryption

than that of the two traditional methods, indicating that the data of the medical health database encrypted by the research method increases and the space cost is smaller than that of the two traditional methods. Therefore, this research method has better encryption performance.

The Third Group of Experimental Results
Change the number of experimental data selected in the experiment, let three groups of encryption storage methods manage the data keys of medical health database in the same cycle, and process the data of medical health database, calculate the ratio of time
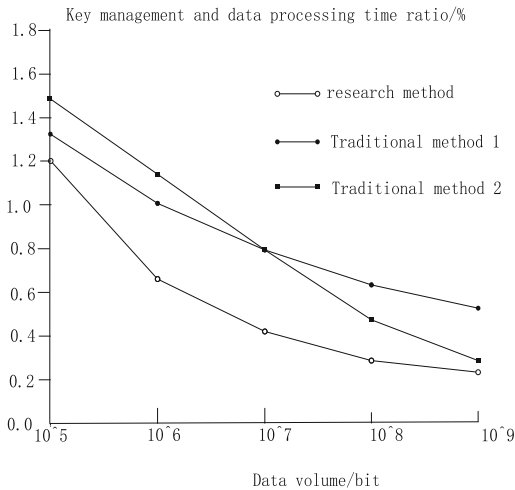
Fig. 7. Key management to data processing time ratio

consumed by key management and data processing in the same cycle, and compare the time ratio of key management and data processing in the same cycle. The experimental results are shown in Fig. 7.

As you can see in Fig. 7, the amount of time spent in key management is doubly reduced as the amount of data in the database increases by an order of magnitude, depending on the number of documents in the database. However, the average ratio of key management time to data processing time of traditional method 2 in one operation cycle is 0.832%, the average ratio of key management time to data processing time of traditional method 1 in one operation cycle is 0.856%, and the average ratio of key management time to data processing time of research method in one operation cycle is 0.558%. Compared with the two traditional methods, the average ratios of key management time to data processing time are 0.274% and 0.298%, respectively, so this method has better encryption performance.

According to the above three experimental results, this research method can meet the needs of data security and encrypted storage, and make the basic operation of database better supported. It has less space overhead and faster database encryption speed.This is because the method studied in this paper decomposes the intelligent encryption algorithm of medical database into key family, key period and key management mode with the help of Internet of Things technology, manages the key by using ciphertext matching algorithm, and increases the steps of data conversion, thus improving the security protection performance of database encryption and decryption.

4   Closing Remarks

To sum up, the research on the intelligent encrypted storage method of medical health database based on the Internet of Things makes full use of the technology of the Internet of Things to enhance the encryption effect of the encrypted medical health database, so that the time expenditure and data expansion rate of the encrypted medical health database are in an acceptable range and the cost of application system access is reduced. However, there are still some deficiencies in this study, in the future research, we still need to study the key management scheme to protect the data of online users of medical health database.

References

1. Shuaihang, Y., Jinrong, H.: Application of data encryption technology in computer network data security. J. Yanan Univ. (Natural Science Edition), **40**(1), 78–82 (2021)
2. Linxin, W., Shigang, L., Shulin, L., et al.: De privacy and data encryption of smart grid big data based on deep learning. Electronic Design Eng. **29**(3), 175–178,183 (2021)
3. Yiming, W.: Analysis of key matrix encryption method under one-way HASH function. Wuxian Hulian Keji **18**(1), 69–70 (2021)
4. Yahong, L.: Research on intelligent encryption method of sensitive database based on aes algorithm. J. Nanyang Inst. Technol. **011**(002), 31–35 (2019)
5. Hui, W.A.N.G., Yuxiang, L.I.U., Shunxiang, C.A.O., et al.: Medical data storage mechanism integrating blockchain technology. Comput. Sci. **4**, 285–291 (2020)

6. Wenbo, X.I.E., Yongzhuang, W.E.I., Zhenghong, L.I.U.: Parallel implementation and analysis of SKINNY encryption algorithm using CUDA. J. Comput. Appl. **41**(4), 1136–1141 (2021)
7. Minmin, X.I.E., Yong, W.A.N.G., Lin, Z.H.O.U.: IEC 61850 Communication message encryption algorithm based on SM2-SM3. Autom. Panorama **38**(1), 108–112 (2021)
8. Hailing, L.: Analysis and discussion of data encryption technology in computer network communication security. Telecom Power Technol. **37**(4), 222–223 (2020)
9. Zijie, Z.: The role and application of data encryption technology in computer network communication security. Telecom Power Technol. **37**(3), 205–206 (2020)
10. Guoqing, Z.: Data encryption algorithm. Comput. Eng. Softw. **41**(10), 129–131 (2020)
11. Xia, W.U.: Research on data storage encryption technology under the background of Big Data. China Comput. Commun. **32**(24), 136–138 (2020)
12. Junli, Y.A.N.G.: Research on user privacy data encryption protection system based on Big Data analysis. Microcomput. Appl. **36**(8), 65–67 (2020)
13. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mob. Netw. Appl. **24**(1), 5–17 (2018). https://doi.org/10.1007/s11036-018-1134-8
14. Shuai, L.I.U., Weiling, B.A.I., Gaocheng, L.I.U., et al.: Parallel fractal compression method for big video data. Complexity **2018**, 1–16 (2018)
15. Shuai, L., Tenghui, H., Jianhua, D.: A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese. Mob. Netw. Appl. 1–13 (2021)

Design of Single Camera 3D Laser Scanning System Based on Artificial Intelligence

Wen-da Xie[1] and Qing-bang Zeng[2(✉)]

[1] Computer Engineering Technical College (Artificial Intelligence College), Guangdong Polytechnic of Science and Technology, Zhuhai 519090, China
wowogogo22@yeah.net

[2] College of Information Engineering, Jiangmen Polytechnic, Jiangmen 529000, China
honda455@163.com

Abstract. At present, the binocular/multi-visual 3D laser scanning system needs to register and match the acquired images, and needs to calibrate and calibrate frequently, which is easy to cause errors. In the hardware design, we choose the model and parameter of camera, laser, stepper motor and its driver. In the software design, we use octree structure to index the 3D data tree structure, preprocess the scanning data, introduce artificial intelligence scanning algorithm, design the optimal hyperplane based on SVM to complete the noise recognition, and finally extract the scanning image features to complete the design of the whole scanning process. Experimental results show that, compared with the traditional system, the light stripe obtained by the design system is complete, and the ranging result is more close to the actual distance.

Keywords: Artificial intelligence · Monocular vision · 3D laser scanning · System design

1 Introduction

The 3D laser scanning system integrates the knowledge of optics, machinery, electronic control, image processing, computer graphics and many other fields into one, which can quickly obtain the 3D point cloud data of the object surface. 3D acquisition methods are generally divided into contact measurement and non-contact measurement. The basic principle of 3D laser scanning measurement is as follows: Laser stripes are formed when the laser is irradiated to the surface of the object to be measured. 3D information of laser stripes is extracted from the images taken by the cameras on the left and right sides. The laser generator and the cameras rotate synchronously around the object to obtain 3D coordinate data of the surface of the object to be measured quickly. At present, it is a hotspot in the field of computer vision to reconstruct 3D models of objects and environments by using computers. In the mid- nineties of the twentieth century, an advanced high-precision automatic stereo scanning technology - three-dimensional laser scanning technology appeared in people's vision. This new technology, also known as

"Scene Reproduction Technology", integrates optical, mechanical, electrical and computer technologies, and is mainly used to describe the shape, structure and color of space objects, so as to model and reconstruct the shape of objects with high accuracy.

As the main equipment of obtaining 3D information in reverse engineering, laser 3D scanning system is widely used in mechanical CAD system and NC machining, film and TV stunt making based on virtual reality, archaeological survey and virtual museum for the purpose of protecting cultural relics, costume design and 3D communication. Since the 1980s, with the rapid progress of information technology, laser 3-D scanning equipment has been developing towards the direction of low cost, high precision, high speed, easy operation and portability [1, 2]. At present, in order to ensure the scanning speed, the binocular/multi-visual 3D laser scanning system is usually used. Reference [3] proposed a virtual reconstruction system of building space structure based on laser 3D scanning, and designed the overall reconstruction system according to single-chip microcomputer mapping and 3D laser scanning technology. The process, preprocessing and registration method of obtaining building spatial structure data by laser 3D scanning are analyzed in detail. Three-dimensional reconstruction of architectural spatial structure model is carried out by using triangular mesh. The external space acquisition tool and interactive editing environment are provided for the operator to realize the virtual reconstruction of the architectural space structure. The experimental results show that the virtual reconstruction of building space structure can be realized by using this system. The 3D modeling results have high fitting degree with the actual results, short modeling time and strong applicability. Reference [4] proposed the pavement deformation analysis method based on 3D laser scanning point cloud, and used the point cloud data obtained by 3D laser scanning technology to operate the road surface. The road surface and digital model established qu2alitative and quantitative analysis of the pavement deformation, and then the overall deformation area and deformation of the road surface. The results show that the 3D laser scanning technology can make the road digital analysis more intuitive and simple, which is of great significance to the automatic development of road disease monitoring technology. In the process of using, the images acquired need to be registered and matched, which is easy to cause errors and need to be calibrated and calibrated frequently. Therefore, this paper designs a single camera 3D laser scanning system based on artificial intelligence, to improve the scanning accuracy. The innovation of the single camera 3D laser scanning system based on artificial intelligence is to use the artificial intelligence scanning algorithm to design the optimal hyperplane based on support vector machine to complete noise identification. The experimental results show that the designed system can obtain a more complete light strip, and the ranging result is closer to the actual distance, which is of great significance to the actual 3D laser scanning.

2 Design of Single Camera 3D Laser Scanning System Based on Artificial Intelligence

2.1 Hardware Design

In the process of system design, the structure of 3D laser scanning system is different because of different application environment and different principle of scanning and

imaging. The 3D laser scanning system designed in this paper generally consists of a high precision and high speed laser transmitter, a set of reflecting prisms and a built-in digital camera (which can obtain digital images of objects in front of the camera). Objects can be scanned in all directions by the rotation of the tilt, and a group of scanning point clouds can be obtained by processing and integrating the acquired data. Monocular vision is usually considered as an ideal pinhole projection model, which can describe the relationship between the coordinate of 3-D points and the relative 2-D points in the image plane. The system described in this paper makes full use of the geometry of pinhole projection model to construct an efficient 3D laser scanning system. The scanning system consists of a horizontal template, in order to obtain high precision data points, the instrument used must be of high precision. In the hardware design of the system, the hardware part is designed.

Select the Camera
The following factors shall be taken into consideration when selecting the type of camera: (1) The resolution of the camera shall not be lower than 680×480, and the frame rate shall not be lower than 30fps at this resolution; (2) The brightness, gray level, white balance and other parameters of the image may be adjusted, and the interference light may be filtered; and (3) The size of the picture taken by the camera shall meet certain requirements. Taking the above factors into consideration, the SONY Effio-E 4140+673 CCD camera is selected. The main technical parameters of the camera are listed in Table 1:

Table 1. Camera parameters

Name	Parameter	Company
Camera pixel	More than 380,000	/
Signal to noise ratio	48/58	dB
Color imaging	Colour to black	/
White balance mode	Automatic	/
Resolution	700	TVL
Focusing mode	Autofocus	/
Lens	DC lens	/
Others	Backlight compensation	/

The camera can be dynamic detection, manual digital noise reduction, backlight compensation, and the picture clear in line with the system requirements.

Selective Laser
When selecting a laser, the projection angle of the line laser should be considered first to ensure that the length of the laser line projecting to the object should be at least 800 mm when the projection distance is 600 mm. In addition, select a slightly more powerful

laser, so that the projection of the laser line will be more thin, improve the accuracy of modeling. And to choose stable light intensity, less scattering of the line laser [5, 6]. The optical fiber laser has been widely used in laser radar, laser ranging and imaging because of its high photoelectric conversion efficiency, good thermal and mechanical stability, compact structure and good output beam quality. Pulsed laser has higher peak power than CW laser and can transmit long distance. Therefore, pulsed laser is often used in the detection of long-range targets. For laser 3D imaging, the pulse width of laser pulse determines its radial resolution.

According to the above rules, the system selects the invisible infrared laser. It is required that the laser should meet the requirement of high quality in 110° sector and the laser intensity distribution should be uniform. Table 2 lists the main parameters of the lasers used in the system:

Table 2. Main parameters of line laser

Name	Parameter	Company
Output wavelength	1021	±nm
Operating voltage	4.3–5	VDC
Angle of line	110	°
Facula diameter	Adjustable	mm
Output power	100	mW
Working current	<180	mA
Facular focusing	Adjustable	/
Mode of operation	Continuous laser	/

Under the conditions of the above parameters, it is necessary to verify whether the length of the line laser meets the requirements. The calculation formula of linear laser is:

$$L_x = L_t \times \tan \frac{\theta}{2} \tag{1}$$

In the formula, L_x represents the length of the line laser that can be projected, L_t represents the projection distance of the laser, and θ represents the line angle degree. In this paper, the length of line laser is required to be more than 800 mm. Through algebraic calculation, it can be seen that the selected laser can meet the requirements of the system.

Selecting Step Motor and Its Driver

The objective of this paper is to build a 3D map model of the surrounding environment, so the camera and the line laser are required to rotate at 360°. Therefore, in the design of the 3D laser scanning system, the camera and the line laser are fixed into an adjustable whole by mechanical structure, and then fixed to the axis of the stepping motor. In the choice of stepper motor requires stepper motor to be able to rotate in 360° range, and motor speed

adjustable, in order to facilitate system debugging [7, 8]. The choice of motor driver must match the two-phase four-wire stepper motor, and the selected driver must also be two-phase output. For the convenience of motor adjustment, the choice of current subdivision and drive can be adjusted [9, 10]. In order to save costs, the laboratory has selected the existing stepper motor, model 57BYGH7630, the system choice of stepper motor driver for the DM542A subdivision type two-phase hybrid. Table 3 shows the main technical parameters of the stepper motor and driver:

Table 3. Parameters of stepping motor and its driver

Classification	Name	Parameter	Company
Stepping motor	Stepping angle	1.8	°
	Static moment	2.0	N·m
	Moment of inertia	480	$g \cdot cm^2$
	Lead wire	6	Root
	Phase voltage	3.0	V
	Phase current	3.0	A
Stepping motor driver	Input voltage	24–50	VDC
	Input current	<2	A
	working temperature	−10–48	°C
	Output current	1.0–4.2	A
	Consumption	80	W
	Weight	200	g

Subdivided two-phase hybrid stepper motor driver DM542A power supply mode for DC power supply, supply voltage of 16–50 V. It is suitable for stepper motor type: The outer diameter is 42–86 mm, 1850 V drive voltage, current less than 4.0 A two-phase hybrid. This driver uses the AC servo driver current loop for subdivision control, motor torque fluctuation is very small, low speed operation is very smooth, almost no vibration and noise. At high speed, the torque is much higher than that of other two-phase actuators [11, 12]. The actuator is widely used in high resolution equipment, such as packaging machinery, CNC machine tools, engraving machines, etc. In order to make the motor run more stably, the decelerating gear is equipped in the 3D laser scanning system. The deceleration gearbox model is 06AG 15, its deceleration ratio is 1:15. At this point completed the design of the hardware part.

2.2 Software Design

Scan Data Preprocessing

In the 3D laser scanning system, the pulse data sent by the laser equipment must be received first. With the support of the hardware structure designed in this paper, the

scanned image has a large amount of point cloud data, and the arrangement is chaotic. So we need to preprocess the data. First, the receiving channel is designed. In the receiving channel, the drift error can be corrected by measuring the amplitude of the pulse, and then a lookup table is used to correct the result [13]. The time of arrival of the pulse is generated by the rising edge of the pulse, so the large pulse does not affect the determination of the time of arrival of the pulse. The measurement of the pulse amplitude is achieved by using two identical peak detectors. The first peak detector is placed behind the preamplifier, and the second peak detector is placed behind the post-stage amplifier. This configuration maximizes the compensation range. For weak signals, the second peak detector is used to compensate, because at the small signal level, the compensation is sensitive to variations in the amplitude of the small signal [14, 15]. When the second peak detector is saturated, the amplitude information from the first peak detector is available because the first peak detector is still in the linear region. By the time the first peak detector is saturated, the signal is very large, but the arrival time of the pulse is insensitive to the large signal, so saturation is not a big problem in the early stage of data processing, the relationship between the discrete points can be established by building a point cloud spatial index. This article uses an octree structure to index the 3-D data tree. The example image is shown Fig. 1:
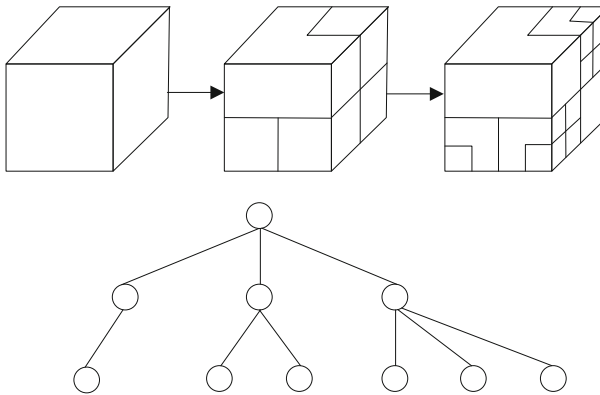
Fig. 1. Octree diagram

Octree is an extension of data structure based on binary tree of one-dimensional data and quadtree of two-dimensional data. Each node of octree is divided by a cube. This division method is based on the line segment of binary tree and the rectangle of quadtree. This stereo partitioning allows an octree to have up to eight child nodes corresponding to each partitioned cube node on the octal rectangle. When a node has no children, it means that the node can evenly represent the corresponding space without further subdivision [16, 17]. The storage point of 3-D point cloud belongs to dimensionless technology. That is to say, its state is a particle in space, and its corresponding volume can not be calculated [18]. When using octree to search, the nearest node is found, and the nearest neighbor is searched to approximate the three splitting planes.

After the relationship between discrete points is established, the noise of the original point cloud data should be eliminated. Based on the characteristics of vehicle radar point cloud data, it is necessary to reorganize the original data to compress the original data, which can be voxelized to achieve denoising. A cube in an octree can represent a voxel, and the cube can record the location of the voxel. The location information can be expressed as (r, c, h), and the calculation formula is as follows:

$$r = \frac{x - x_{\min}}{v}, \ c = \frac{y - y_{\min}}{v}, \ h = \frac{z - z_{\min}}{v} \tag{2}$$

The location information (r, c, h) of voxel respectively represents the distance between voxel and the smallest x, y, z and location point in all point clouds. $x_{\min}$, $y_{\min}$ and $z_{\min}$ are the minimum values of x, y and z. The number of points recorded in voxels will exist in the range or boundary of voxels. In this way, the relationship between point cloud and voxel is established to complete the scanning data preprocessing.

Introduction of Artificial Intelligence Scanning Algorithm

After getting the preprocessed scanning image, it is necessary to automatically diagnose the noise in the scanning process [19]. This paper uses a typical optimized dynamic time warping algorithm to express the relationship between the collected signal and the reference signal. In the process of diagnosis, assume that the reference signal is expressed as Q and the acquisition signal is expressed as C, then:

$$\begin{cases} Q = q_1, q_2, \ldots q_i, \ldots, q_n \\ C = c_1, c_2, \ldots c_j, \ldots, c_m \end{cases} \tag{3}$$

In the formula, n represents the length of Q, that is, the number of frames of the image, q_i represents the characteristic value of the i frame, and m represents the length and number of frames of C. When the length between the acquisition signal and the reference signal is the same, the distance between them can be calculated directly, and the dynamic time warping method is needed. Firstly, a matrix with dimension $n \times m$ is constructed. Any element in the matrix can represent the distance between q_i and c_i. In distance calculation, Euclidean distance is used as the calculation basis. While representing the distance, it can also reflect the similarity between two points. The dynamic time warping algorithm mainly finds Q and C through the optimal path:

$$W = w_1, w_2, \ldots, w_k, \max(m, n) \leq k < m + n - 1 \tag{4}$$

In the formula, W represents the optimal path, and the k element in W can be defined as $w_k = (i, j)_k$, which is constrained by boundary conditions, continuity and monotonicity. In order to minimize the structural risk, this paper uses SVM to map the image samples to the high-dimensional feature space, and constructs the noise data of the image into the optimal hyperplane by support vector, as shown in Fig. 2:
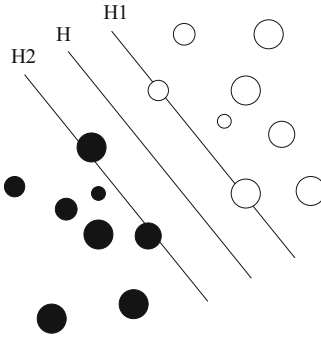
Fig. 2. Optimal hyperplane based on SVM

In the above picture, black circle and white circle represent two kinds of samples, which are divided into two kinds of noisy data and non-noisy data, and are used to construct the optimal hyperplane. In the process of constructing the optimal hyperplane, there are:

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}\|p\|^2 \\ s.t. \ y_i(p \cdot x + b) \geq 1, \ i = 1, 2, ..., l \end{cases} \tag{5}$$

In the formula, p is the normal vector and b is the threshold. Through the above formula, we can get the image stripe features of noise data and non noise data in different working modes, and treat these data as training data through support vector machine [20, 21]. After inputting the data to be tested into the support vector machine, it is compared with the training data samples. For the scanning system designed in this paper, a variety of indicators will be involved in the noise recognition process [22]. The image noise recognition in the scanning system can be represented by the multi indicator factor set U, that is, the geometric domain of evaluation indicators. Therefore, the noise recognition reference index set can be represented by V, Then the two sets can be expressed as follows:

$$U = \{u_1, u_2, \ldots, u_n\}, V = \{v_1, v_2, \ldots, v_n\} \tag{6}$$

For the weighted evaluation of noise recognition, it is necessary to establish the fuzzy vector A of weight allocation, that is, the index weight set. The fuzzy relationship of the product of the multi index factor set U and the noise recognition reference index set V can be expressed by R, which is in the form of fuzzy matrix. The membership function between the index and the evaluation level can be obtained, and the matrix multiplication of A and R can be obtained

$$T^* = \left(t_1^*, t_2^*, \ldots, t_n^*\right) \tag{7}$$

The final noise recognition result Q can be expressed as:

$$Q = T^* \times P' \tag{8}$$

In the formula, P' is transpose matrix, so as to complete the automatic diagnosis of scanning image noise.

Extraction of Scanning Image Features

In order to extract the information of 3D laser scanning target accurately and effectively, we must distinguish it from other objects according to the characteristics of the target. In some complex scenes, in order to achieve the precise scanning of the target, we need to filter out the objects except the target objects. In the process of clustering, eigenvectors are associated with each point, and in eigenvectors, geometric radiometric values are also included. Clustering is to filter out the low point cloud, and then set up the horizontal grid after projection. Mark the grid as 0 for the grid with point clouds and null for the grid without point clouds. In the neighborhood, the grid containing the point cloud will be grouped into the same cluster block and a new natural number will be re-assigned as a marker. The grid diagram after clustering is shown in Fig. 3:

Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null
Null	1	1	Null	Null	Null	Null	Null	Null	Null	5	5
Null	1	1	1	Null	Null	Null	4	Null	Null	5	5
Null	Null	Null	1	Null	Null	Null	4	Null	Null	5	5
Null	Null	Null	Null	Null	Null	4	4	4	Null	5	5
Null	2	2	Null	Null	4	4	4	Null	Null	5	5
Null	2	2	2	Null	Null	Null	4	Null	Null	5	5
Null	2	2	2	Null	Null	Null	Null	Null	Null	5	5
Null	Null	2	Null	Null	Null	Null	Null	Null	Null	5	5
Null	Null	Null	Null	3	3	3	Null	Null	Null	5	5
Null	Null	Null	3	3	3	3	Null	Null	5	5	5
Null	Null	Null	Null	3	3	Null	Null	Null	Null	Null	Null

Fig. 3. Cluster diagram of scanning target

After clustering, we can see some clustering blocks, but a clustering block does not represent the same kind of point cloud, in the subsequent feature extraction, we need to remove other objects together with the object. According to this characteristic, a plane can be intercepted at 1.3–1.5 m from the ground, and the point clouds of the upper and lower parts of the object can be projected horizontally and meshed. After the large clustering unit is obtained, the projection area of the upper and lower part of the point cloud is calculated, the projection span is calculated, and the 3D information of the object is extracted. The scan process is shown in Fig. 4:
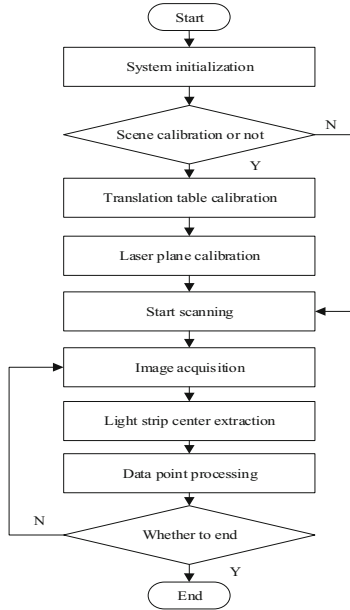
Fig. 4. Workflow of a scanning system

Before the system starts to run, it is necessary to complete the calibration of the camera, obtain the camera focus, imaging origin, distortion coefficient and other information. The initialization phase reads camera calibration data from an XML file that has been written, as well as the calibration data of the original translation table and the laser. The CCD image acquisition stage completes the automatic correction of the acquired image. In the phase of parameter acquisition, the laser plane calibration is performed and the translation velocity of the translation table is obtained. At the beginning of the scan phase, start the translation station to determine the location of the scan data points. In the scanning stage, the light stripe center is extracted and the 3D projection transformation is completed to obtain the point cloud data. When the light strip removes the scanning area, the system automatically terminates the scan, or manually terminates the scan. After scanning, it is necessary to correct the error. Due to the propagation of laser pulses in atmosphere or the interaction between laser pulses and target, the shape and amplitude of receiving pulses are different from those of transmitting pulses. Therefore, it is difficult to determine the time of arrival of the pulse, which has an effect on the ranging results. The error caused by the discrimination of the time reaching the pulse is called drift error. The function of time discrimination is to change the input analog pulse signal into logic signal by locating the pulse arrival time. Another is to extract the useful signal from the noise. At present, there are three kinds of identification methods: frontier identification method, zero-crossing comparison method, also known as high-pass capacitance identification and constant ratio timing method. The above principle methods have their own advantages and disadvantages, according to different applications to choose which method to use. In addition to drift errors due to changes in the

amplitude and shape of the received signal, there are also errors due to temperature and service time and measurement errors due to noise.

3   System Experimental Test

3.1   Build a Test Environment

According to the structure, hardware and software of the system, the image acquisition system, line laser and control turntable system are connected with the computer, and set up on the designed and assembled robot test prototype. Collect data on a dataset, which has the advantage that most conversations are public, which means that a lot of data is available through its API. A collection of open data sets contributed by data scientists involved in machine learning projects.The experimental environment was scanned by controlling the step motor to drive the camera and the laser, and the scanning process was filmed into the computer. After the video is processed by computer, the reconstructed 3D model is displayed. After the completion of the system to carry out system debugging, mainly divided into the following steps:

(1) Camera debugging: adjust the camera's parameters such as the angle of view, resolution, brightness, framing, saturation, contrast, backlight compensation, and white balance, turn off the camera's functions such as automatic gain control, automatic color compensation, and automatic weak light compensation, and then calibrate the camera; (2) step motor debugging: use the step motor controller to program and control the motor rotation, make the motor rotate uniformly, and collect the image of the scanned area; and (3) line laser debugging: control the angle between the camera spindle plane and the line laser plane to be controlled at 30. Between 600 and 600 to ensure measurement accuracy.

After the system is built and debugged, the scanning measurement can be started. In the process of scanning measurement, we should try our best to ensure the environment without light interference. When using 3D laser rangefinder to construct 3D model, the characteristic of point cloud data is different for different objects, so the method and process of data processing are also different. After the data acquisition, data processing, including: data acquisition, data smoothing, denoising, data model reconstruction and visualization. In the whole process, the camera calibration, system debugging is the premise of the measurement experiment, and is also the key to improve the accuracy of the system. The data processing process provides reliable point cloud data for 3D model reconstruction, which can ensure the accuracy and speed of the whole system. This system performance test runs under the Windows XP system, using VS2008, OpenCV and OpenGL's open source function library compilation. During the experiment, the scanning process is filmed as video, and then the video frames are extracted offline for processing, which can facilitate the debugging and save the experimental time.

3.2   Analysis and Comparison of Experimental Results

In the same test environment, the system designed in this paper and the traditional binocular vision scanning system and the multi-vision scanning system were used to carry out the experiment. The scanning photos obtained in this paper are shown in Fig. 5:

Fig. 5. Partial scan results from the scanning system designed in this paper

The results from a traditional binocular scanning system are shown in Fig. 6:

Fig. 6. Partial scan results from traditional binocular vision scanning system

The results from a traditional multi-vision scanning system are shown in Fig. 7:

Fig. 7. Partial scan results from a conventional multivisual scanning system

From the results of the above three scans, it can be seen that there are a lot of burrs and defects in the center of the light stripe obtained by the traditional binocular scanning system and the multi-visual scanning system, which will affect the modeling accuracy in the later calculation.

At the center of the light strip, the distance from the surface of the space object illuminated by the light strip to the laser can be calculated by using the principle of triangulation, and a number of test points are selected for comparison. The test distance of the three systems is shown in Table 4:

Table 4. Comparison of distance measurement results (Unit: mm)

Test point serial number	Single camera scanning system designed in this paper	Traditional binocular vision scanning system	Traditional multivisual scanning system
1	2712.363	2538.978	2606.603
2	2768.491	2331.808	2708.700
3	2148.405	2186.134	2419.489
4	2590.905	2484.065	2497.770
5	2424.949	2495.316	2429.288
6	2491.978	2441.131	2166.875
7	2484.598	2628.465	2983.623
8	2326.824	2532.241	2193.066
9	2730.697	2292.340	2790.358
10	3053.439	2473.615	2769.631

(*continued*)

Table 4. (*continued*)

Test point serial number	Single camera scanning system designed in this paper	Traditional binocular vision scanning system	Traditional multivisual scanning system
11	2702.654	2964.539	2495.410
12	2471.716	2966.710	2761.785
13	2282.009	3036.451	2537.333
14	2147.153	2231.956	2821.215
15	2204.823	2940.512	2950.758
16	2423.309	2293.090	2176.451
17	2512.986	2274.348	2887.923
18	3004.273	2896.337	2569.137
19	2134.994	2183.157	2798.103
20	2303.156	2153.667	2306.675

The fitting value of the system is 0.973, the fitting value of the traditional binocular vision scanning system is 0.914, and the fitting value of the traditional multi-vision scanning system is 0.908. From the above results, we can see that the designed 3D laser scanning system based on artificial intelligence has higher precision and reliability than the two traditional scanning systems in point ranging.

When using 3D laser scanning system for 3D modeling, there will be many error factors, such as the error of laser scanning system itself, the error caused by the reflection of the scanned object, the error caused by the external environment and so on. The errors of the system itself are mainly caused by the performance defects of the instrument itself, including the laser fringe error is too thick, the camera has radial and tangential distortion; In addition, the errors caused by different material processes on the surface of the scanned object mainly include the errors caused by the roughness and inclination of the surface of the scanned object. In addition, the error caused by the external environment is mainly caused by the light.

4   Conclusion

In this paper, a single camera 3D laser scanning system based on artificial intelligence is proposed. The system can solve the problem of registration and matching errors in binocular/multi-vision 3D laser scanning system. Experimental results show that, compared with the two traditional systems, the system designed in this paper has the highest precision in scanning imaging distance, ranging curve fitting accuracy of 0.973. The scanning system proposed in this paper can meet the requirements of general users.

References

1. Xugang, L., Yinfei, C., Haifeng, H.: Application status and existing problems of 3D laser scanning technique in mine surveying in China. Metal Mine (03), 3540 (2019)
2. Junjie, M., Zhilong, L., Chengming, L., et al.: Research on visual collaborative design of complex rock foundation based on 3D laser scanning technology. Building Struct. **49**(21), 124–128 (2019)
3. Meng, W., Cuiyan, W., Huihui, D.: Virtual reconstruction system based on laser three-dimensional scanning building space structure. Laser J. **40**(11), 170173 (2019)
4. Mengyuan, S., Da-guang, H., Jieming, G., et al.: Road pavement deformation analysis method based on 3 D laser scanning point cloud. Sci. Technol. Eng. **19**(24), 386–391 (2019)
5. Ping, W., Chenxin, W., Xuan, Z.: Design and application of educational artificial intelligence system based on automation method. China Educ. Technol. **06**, 7–15 (2020)
6. Dongbo, Z., Zhuolin, W., Lixue, J.: Application of 3D laser scanning technology in defect detection of external wall thermal insulation system. Construct. Technol. **49**(09), 20–23 (2020)
7. Hongyu, M., Xin, Z.: 3D laser scanning technology and BIM technology in ancient building protection surveying and mapping. Geotechn. Eng. Technique **33**(04), 222–225 (2019)
8. Jun, L., Tao, G.: Application of new mobile 3D laser scanning technology in ellipsometry of shield segments. Bull. Survey. Mapp. **12**, 159–162 (2019)
9. Fuyin, Y., Hongyu, C., Tiejun, L., et al.: Research on monitoring and early warning and management system of leakage water based on 3D laser scanning subway operation tunnel. Construct. Technol. **48**(23), 76–79+128 (2019)
10. Hongquan, X., Yuetao, C., Fang, Z., et al.: Application of precise point positioning to the piggyback mobile laser scanning system. Eng. Surveying Mapping **28**(04), 60–63 (2019)
11. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mob. Netw. Appl. **24**(1), 5–17 (2018). https://doi.org/10.1007/s11036-018-1134-8
12. Liu, S., Fu, W., He, L., Zhou, J., Ma, M.: Distribution of primary additional errors in fractal encoding method. Multimedia Tools Appl. **76**(4), 5787–5802 (2014). https://doi.org/10.1007/s11042-014-2408-1
13. Hudson, R., Faraj, F., Fotopoulos, G.: Review of close-range three-dimensional laser scanning of geological hand samples. Earth Sci. Rev. **210**(1), 103321 (2020)
14. Lin, Y.M., Song, C., Rutledge, G.C.: Direct three-dimensional visualization of membrane fouling by confocal laser scanning microscopy. ACS Appl. Mater. Interfaces. **11**(18), 159–164 (2019)
15. Kiyono, T., Asawa, T., Oshio, H.: Laser-scanning-based method for estimating the distribution of the convective-heat-transfer coefficient on full-scale building walls. Bound.-Layer Meteorol. **178**(3), 1–24 (2021)
16. Feng, P., Zou, Y., Hu, L., et al.: Use of 3D laser scanning on evaluating reduction of initial geometric imperfection of steel column with pre-stressed CFRP. Eng. Struct. **198**(1), 109527.1–109527.12 (2019)
17. Long, J., Xiong, W., Wei, C., et al.: Directional assembly of ZnO nanowires via three-dimensional laser direct writing. Nano Lett. **12**(12), 59–63 (2020)
18. Duan, D., Podlesnik, J., Scharf, I., et al.: Fine sand particles enable antlions to build pitfall traps with advanced three-dimensional geometry. J. Exp. Biol. **223**(Pt 15), jeb.224626 (2020)
19. Tu, D., Xiao, G., Zhang, X., et al.: Laser stripe matching algorithm with coplanar constraint in underwater laser scanning systems. Optical Eng. **58**(11), 114108.1–114108.9 (2019)
20. Denk, W., Strickler, J.H., Webb, W.W.: Two-photon laser scanning fluorescence microscopy. Science (New York, N.Y.), **248**(4951), 73–76 (2019)

21. Hasselbach, J., Bogatscher, S., Rembe, C.: Laser scanner module with large sending aperture and inherent high angular position accuracy for three-dimensional light detecting and ranging. Opt. Eng. **58**(8), 087101.1–087101.13 (2019)
22. Zhao, Y., Spingler, F.B., Patel, Y., et al.: Localized swelling inhomogeneity detection in lithium ion cells using multi-dimensional laser scanning. J. Electrochem. Soc. **166**(2), 27–34 (2019)

Design of Enterprise Intelligent Decision Support System Based on Data Mining

Qiu-ying Lv[1(✉)] and Yang Su[2]

[1] School of Business, Fuyang Normal University, Fuyang 236041, China
`crnjb1121@aliyun.com`
[2] School of Economics and Management, Yantai Institute of Technology, Yantai 264005, China

Abstract. Faced with the lack of data and redundant data, traditional decision support system is difficult to obtain accurate decision support index, which leads to the decline of system response and control performance. This paper studies the enterprise intelligent decision support system based on data mining. In the hardware design, the impedance conversion circuit, the signal transmitting and receiving FPGA interface circuit are designed to strengthen the software service. In the software design, the enterprise data design mode is set based on data mining, the decision indicators are selected according to the influencing factors, and the enterprise decision results are generated through the data analysis rules of the decision support system. The experimental data show that the response time of the proposed system is 0.0448 s and 0.0403 s lower than that of the two traditional systems in complex environment; When the data is missing or redundant, the control quality of the proposed system is 22.16% and 15.57% higher than that of the two traditional systems, respectively. Therefore, the decision support system based on data mining has better performance.

Keywords: Data mining · Enterprises · Intelligent decision making · Support system

1 Introduction

The process of enterprise development will involve a large amount of development information. At the same time, some large-scale enterprises need to collect a large amount of data to get the difference value in order to realize the data link comparison work, so as to provide a data support for the subsequent enterprise development and transformation work. For this reason, the relevant researchers will take the research content of reference [1] as a reference, combine the targeted design principles and overall architecture, take the enterprise development as the foundation, establish the intelligent decision support system belonging to the enterprise, and provide more accurate data for enterprise decision. Some scholars combine the research content of document [2], combine the "Internet plus" with decision support system, combined with the specific speeches of various experts, optimize some hardware and software of the system, and

strengthen the data acquisition ability of the system. Reference [3] proposes that the construction of operation decision support system can meet the data needs of different departments of the hospital for operation management, and improve the decision-making level and management efficiency of the hospital. Therefore, the relevant personnel take this reference as a reference, and combine Hadoop big data technology with system decision-making to provide more comprehensive statistical data for enterprises.

On the basis of the above design, data mining and enterprise intelligent decision support system are integrated to provide more reliable technical support for enterprise development and transformation. From the hardware design, the impedance conversion circuit, signal sending and receiving FPGA interface circuit are selected to strengthen the software service. Focus on the software design part, set the enterprise data design mode based on data mining, select the decision indicators according to the influencing factors, and generate the enterprise decision results through the data analysis rules of the decision support system. Finally, it is verified by comparative experiments.

2   Hardware Design of Enterprise Intelligent Decision Support System

2.1   Design Impedance Conversion Circuit

In the process of decision-making enterprise related data, affected by the external signal and its own equipment signal, it is easy to appear noise interference. In order to achieve the research purpose of intelligent and accurate decision-making, the oscilloscope is introduced into the design system for noise reduction. The input impedance of the oscilloscope is known to be 100 Ω. It can achieve the purpose of matching the source end with the load end. In order to effectively reduce the signal attenuation and reflection and achieve more accurate data decision-making, a new impedance conversion circuit is designed for the signal conditioning channel of oscilloscope, which can improve the input impedance of the later stage circuit, reduce the output impedance and reduce the load effect of the input signal. The impedance transformation circuit is the same as the two port network, and its input impedance is the total input impedance after the input port of the oscilloscope. Figure 1(a) is the design diagram of the impedance transformation circuit in the ideal state.

P1 in Fig. 1(a) represents the input resistance of the front stage of the oscilloscope; P2 represents the input impedance of the later stage variable gain amplifier. The input impedance of the circuit is very large and the output impedance is very small, so it is easy to appear the phenomenon of impedance mismatch, which affects the normal work of the system circuit when searching literature. Therefore, according to the requirements of the noise level, bandwidth and dynamic range of the oscilloscope, the impedance conversion circuit is designed by using integrated operational amplifier. The integrated operational amplifier is not easy to be affected by the distributed parameters, so it has better impedance conversion effect [4]. In this study, the integrated operational amplifier and voltage follower are built. According to the known equivalent input noise in Fig. 1(b),
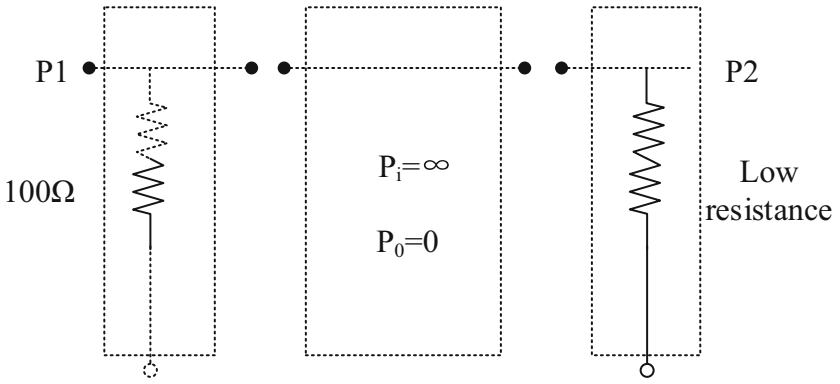
(a) Im pe dance transform circuit in the ideal state
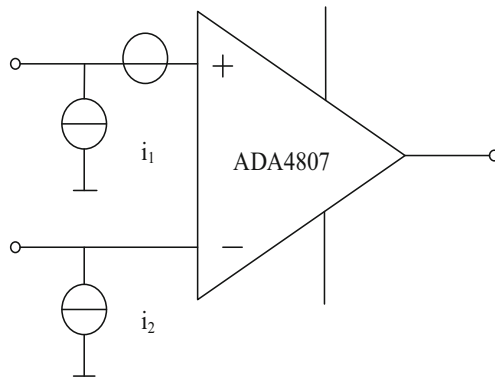
(b) Im pe dance - equivalent input noise of the transform circuit

Fig. 1. Impedance transformation circuit in the ideal state

the impedance conversion circuit required for retrieval is designed. The circuit is no longer built by using the peripheral resistance. Through the noise characteristics of the device itself, the influence of resistance thermal noise on the retrieval results is avoided.

2.2 Design of Interface Circuits FPGA Signal Transmission and Reception

The design of decision support system, using AD9787 converter to achieve signal transmission. The selected converter is a dual channel parallel structure, and the connection between the hardware and FPGA is established, and the communication transmission of data signal is controlled by the clock signal. The hardware mentioned here is field programmable gate array, and the design scheme of interface circuit between them is shown in Fig. 2 below.
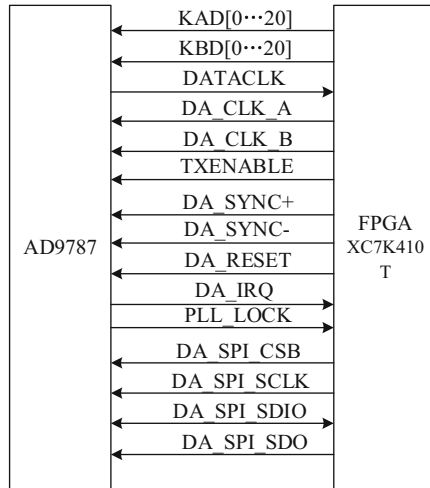
Fig. 2. FPGA Interface circuit AD9787 to the converter

When designing the interface circuit shown in Fig. 2 above, it is necessary to determine whether the I/O level of the interface is compatible in advance. According to the level division standard, the DA is divided into three levels_ CLK_ A and DA_ CLK_ B is set as the level standard of differential clock signal; Will DA_ SYNC+ and DA_ SYNC - set as the level standard of differential synchronization signal. The above ports are connected with FPGA, and two pairs of differential I/O interfaces are connected with BANK20 of XC7K410T, while other ports are connected with bank10. Using the above circuit, KAD [0… 20] and KBD [0… 20] correspond to a and B signals, using databack to synchronously feed back the clock signal of DAC, using TXENABLE and DA_ The reset port sets the enable signal and reset signal, while the da_ IRQ and PLL_ Lock port, respectively, indicates the interrupt signal in case of error and the indication signal of PLL locking. Finally, four groups of signals are used to represent the bus signals of SPI under the control of SPI protocol. Through the interface circuit designed above, the clock control and data communication control of the system are realized. When the signal receiving FPGA interface circuit is designed, the analog signal is collected by using dual channel parallel converter AD9650, and the converted signal is transmitted to FPGA in digital form. Figure 3 shows the interface circuit between FPGA and ad9650.

It is known that the DC power supply of the converter is 1.8 V, so it supports lvcmos-18 level standard. The bank32 of xc7k410t is used to connect the converter and FPGA. It is known that bank32 belongs to HP bank group and uses 1.8 V DC power supply to provide power for I/O, so the interface also supports 1.8 V DC power supply. It should be noted that the I/O interface of FPGA supports differential signal level, and the ADC_ CHA_ D[0…15]、ADC_ CHB_ D [0… 15] port, as a tool for 14 groups of differential lines to obtain ad9650 digital signal. When the circuit is used to transmit signals, the signals of all differential lines and different lines are transmitted alternately, and the configuration mode is determined through SPI interface to realize the task of register configuration [5]. So far, with the help of AD9787 and AD9650, the FPGA interface circuit of signal transmitting and receiving is designed.
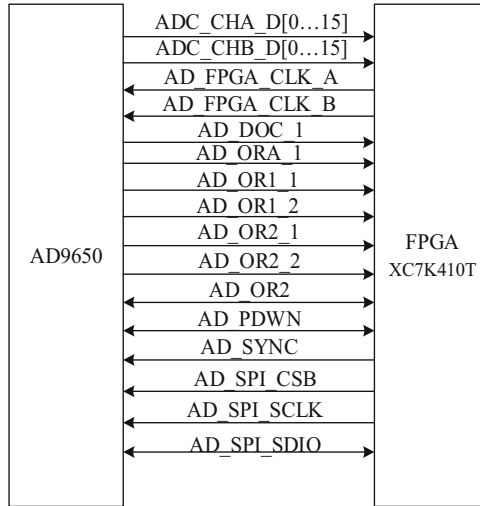
Fig. 3. FPGA Interface circuit AD9650 with converter

3 Software Design of Enterprise Intelligent Decision Support System Based on Data Mining

3.1 Set Enterprise Data Classification Mode Based on Data Mining

In data mining, clustering analysis is an advanced machine learning technology, which can calculate data sets with distance differences according to the target variables or attributes. So, the data mining technology is used to classify the data of enterprise monitoring matrix. The hierarchical clustering analysis method pre initializes similarity matrix and distance, then updates and merges the minimum set, and updates the new data set repeatedly and ends iteration after reaching the set number of sets. Set the distance to r, the distance between the a monitoring area and the b monitoring area is:

$$r = \sqrt{\sum_{m=1}^{n} (x_{am} - x_{bm})^2} \tag{1}$$

In the formula: x_{am} and x_{bm} represent the actual monitoring value of index m in the a and b monitoring areas respectively. It is known that there are linear characteristics among the decision-making data of enterprises, so all the decision-making historical data of enterprises in the region are normalized, and the data set composition analysis results are obtained

$$C_{ab} = \sum_{e} \mu_{ae} x_{eb} \tag{2}$$

In the formula: C_{ab} represents the composition of enterprise data; μ_{ae} is the influence degree of the change index on the composition data; e represents the total number of

changes in the constituent data [6]. According to the data mining method, training sample data, get the actual available parameters, find out the spatio-temporal association between different enterprise decision data, use the intelligent computing program of the model, monitor the changes of these data in a fixed period, find out the implicit relationship between indicators and other parameters in each time and space area, Complete the recognition of data changes in a specific area. The monitoring model designed in this paper can reasonably explain the important implicit relationship between data in the way of intelligent reasoning, and provide analysis basis for data change and invariance. According to the above calculation parameters and design concept, the data composition change recognition pattern is constructed based on data mining, as shown in Fig. 4 below. The model is used to monitor the data change characteristics contained in the system in real time and provide data sources for supporting programs [7].
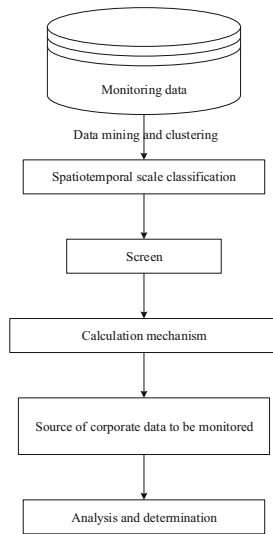
Fig. 4. Modelling of monitoring cell

Establish an effective connection between the identification mode shown in Fig. 4 and the central monitoring unit of the system, and control the data monitoring behavior of the system through this mode.

3.2   Selection of Decision-Making Indicators

According to the above design, the divided enterprise data will be obtained. According to the above results, the purpose and influencing factors of enterprise strategic decision are considered to select the decision-making index, analyze and confirm various enterprise information, and form a set that can support the basic strategic decision-making of various resources and further screen data. At present, there are many methods to set up decision-making indicators, including comprehensive analysis method, attribute category analysis

method, etc., comprehensive analysis and comparison. This design selects the excellent part of these methods, combines with the specific reality, and takes the previously used indicators that have an impact on the enterprise development strategy as a reference to reset new decision-making support indicators. There are different strong and weak correlations among some indicators, and some of the data will be affected by enterprise factors, as shown in Table 1 below.

Table 1. Factors affecting the enterprise decision - making indicators

Serial number	Secondary indicators	Serial number	Secondary indicators
k1	Number of employees	k11	Enterprise scale
k2	Resource quality	k12	Corporate culture
k3	Input training	k13	Enterprise Strategy
k4	Per capita income	k14	Competition among enterprises
k5	Labor cost	k15	Enterprise income
k6	Labour productivity	k16	Market Competition
k7	Employee satisfaction	k17	Market labor force
k8	Sector structure	k18	National policy
k9	Labour law	k19	Socioeconomic
k10	Other resources	k20	Scientific and technological productivity

It can be seen from the above Table 1 that the factors influencing the decision-making indicators include internal self influence and external environmental influence. Therefore, it is necessary to establish a preliminary selection set of decision-making indicators, which is represented by $D = \{d_1, d_2, \ldots, d_n\}$. Entropy is used to represent the index dispersion, and the original matrix is $F = (s_{ij}) \times n$, where s_{ij} represents the i-th candidate index and the j-th sample data. In this case, the entropy dispersion can be calculated by the following formula:

$$\gamma = \alpha_i - \varepsilon \sum_{i=1}^{n} \gamma s_{ij} \ln s_{ij} \tag{3}$$

In the formula: α_i represents the entropy value under the i alternative index; $\varepsilon = \frac{1}{\ln n}$. Represents decision reference data. The larger the γ value is, the greater the screening ability of the index is; When the γ value is too small, it indicates that the index is not accurate enough. At this time, the index has little ability to distinguish enterprise information, so it needs to be re selected [8].

3.3   Building Data Analysis Rules of DSS

The decision classification algorithm is used to construct the analysis rules of the support system, which makes the decision support system generate decision information related to the development situation of the enterprise quickly. The database table generated automatically by the known system is the object of decision classification algorithm processing, so it is assumed that the analysis system is:

$$Z=(U,V,G,H,f) \tag{4}$$

In the formula: $U = u_1, u_2, \ldots, u_n$ represents a nonempty finite set with n data; B is the set of condition classification attributes; K is the set of decision attributes; H is the set of attribute values, and there exists $H = \cup_{\beta \in U} H_\beta$, where β is the limiting parameter; In this function, $(V \cup G)$ constitutes the attribute set of the data table, and $(V \cap G) \neq \varnothing$. The decision classification algorithm is used to define the change characteristics of all kinds of enterprise decision data, and to deal with the related analysis problems of enterprise development situation. The following equations are the boundary equations of the data set obtained by the decision classification algorithm

$$\begin{cases} F^-(U) = \cup_{\beta \in U} \{[u]_F \cap U \neq \varnothing\} \\ F_-(U) = \cup_{\beta \in U} \{[u]_F \subseteq U\} \\ DN_F(U) = F^-(U) - F_-(U) \end{cases} \tag{5}$$

In the formula, $F^-(U)$ is the upper approximation of the non empty finite set U; $F_-(U)$ is the lower approximation of the non empty finite set U; $DN_F(U)$ is the boundary area with area D and the number of boundary N; $[u]_F$ represents the equivalent class on F. Under the application of decision classification algorithm, the approximate area of the system for the enterprise data set is shown in Fig. 5.
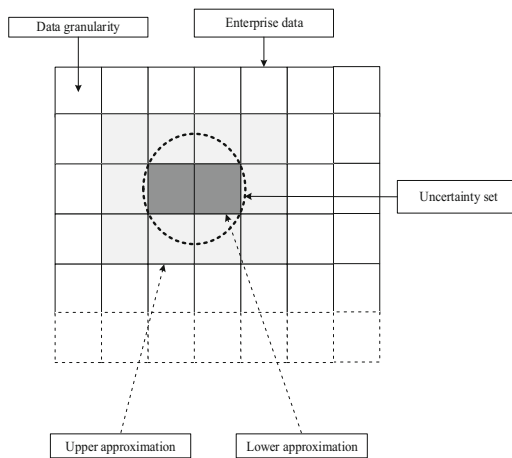

Fig. 5. Upper and lower limits of the data set

According to the upper and lower approximation, the concepts of positive domain, negative domain and boundary domain are introduced. The dark part in the graph represents the positive domain, which is the set of lower approximation; The white part is negative domain; The light color part is the boundary region. Suppose that an enterprise's information presents two states, represented by $z = (w, w')$, where w indicates that the enterprise has achieved good results in development; w' indicates that the development of the enterprise has not achieved good results [9–11]. Let P denote the degree of decision support, and $T = (t_P, t_N, t_D)$ denote the three action indexes of decision support system. All the information of the enterprise is divided into positive domain, negative domain and boundary domain as shown in Fig. 5 above. Let η_{ij} denote the action cost of the system, where $i \in (P, D, N)$ and $j \in (P, N)$, φ_{ij} are required to satisfy the following equations under normal conditions:

$$\begin{cases} \eta_{PP} \leq \eta_{DP} \leq \eta_{NP} \\ \eta_{NN} \leq \eta_{DN} \leq \eta_{PN} \end{cases} \quad (6)$$

Choose the action with the lowest cost, and get the analysis result of the enterprise development situation.

$$\begin{cases} \text{If } F(t_P|[u]) \leq F(t_D|[u]) \text{ and } F(t_P|[u]) \leq F(t_N|[u]), \text{ then } u \in F^-(U) \\ \text{If } F(t_D|[u]) \leq F(t_P|[u]) \text{ and } F(t_D|[u]) \leq F(t_N|[u]), \text{ then } u \in F\_(U) \\ \text{If } F(t_N|[u]) \leq F(t_P|[u]) \text{ and } F(t_N|[u]) \leq F(t_D|[u]), \text{ then } u \in DN_F(U) \end{cases} \quad (7)$$

A given decision classification algorithm requires the following formula to be true, namely $P(w|[u]) + P(w'|[u]) = 1$. according to the above formula, the system decision rules under the application of the algorithm can be obtained.

$$\begin{cases} \text{If } P(w|[u]) \geq \delta_1 \text{ and } P(w'|[u]) \geq \delta_3, \text{ then decide } F^-(U) \\ \text{If } P(w|[u]) \geq \delta_2 \text{ and } P(w'|[u]) \leq \delta_1, \text{ then decide } F\_(U) \\ \text{If } P(w|[u]) \geq \delta_3 \text{ and } P(w'|[u]) \geq \delta_2, \text{ then decide } DN_F(U) \end{cases} \quad (8)$$

Among them:

$$\begin{cases} \delta_1 = \dfrac{\eta_{PN} - \eta_{DN}}{(\eta_{PN} - \eta_{DN}) + (\eta_{DN} - \eta_{PP})} \\ \delta_2 = \dfrac{\eta_{PN} - \eta_{NN}}{(\eta_{PN} - \eta_{NN}) + (\eta_{NP} - \eta_{PP})} \\ \delta_3 = \dfrac{\eta_{DN} - \eta_{NN}}{(\eta_{DN} - \eta_{NN}) + (\eta_{NP} - \eta_{DP})} \end{cases} \quad (9)$$

In the above formula, δ_1, δ_2 and δ_3 respectively represent the decision support parameters corresponding to the three action indicators of the system. When $(\eta_{PN} - \eta_{DN})(\eta_{NP} - \eta_{DP}) > (\eta_{DP} - \eta_{PP})(\eta_{DN} - \eta_{NN})$ was established, there was $\delta_1 > \delta_2$, so $\delta_1 > \delta_2 > \delta_3$ was further deduced. The most simplified decision rules for the development situation of the enterprise obtained by the system are as follows:

$$\begin{cases} \text{If } P(w|[u]) \geq \delta_1, \text{ then } u \in F^-(U) \\ \text{If } \delta_1 < P(w|[u]) < \delta_2, \text{ then } u \in F\_(U) \\ \text{If } P(w|[u]) \leq \delta_3, \text{ then } u \in DN_F(U) \end{cases} \quad (10)$$

The decision classification algorithm sets the system analysis rules, and according to the above rules, generates an automatic pop-up statistical table of enterprise information.

3.4   Generate Enterprise Decision-Making Results

In order to improve the analysis results of the system, the real-time state of the data is determined based on Bayesian network. According to the characteristics of Bayesian network, the variational Bayesian inference is carried out. Using the criterion of minimizing singular distance, the virtual distribution $g(\lambda_t)$ is optimized to make the posterior probability density function $q(\lambda_t|R_{1:t})$. The singular distance between two functions is defined.

$$KLD[g||q] = \int g(\lambda_t) \ln\left(\frac{g(\lambda_t)}{q(\lambda_t|R_{1:t})}\right) d\lambda_t \tag{11}$$

In the formula, KLD is singular distance; λ_t is the total variable; $R_{1:t}$ is the posterior probability of the observation vector. Suppose that the virtual distribution corresponding to the total variable λ_t can be decomposed into the following form:

$$g(\lambda_t) = \prod_{m \in \theta_t} g(\lambda_t^m) = g(x_t)g(\varepsilon_t)g(v_t)g(\sigma_t)g(h_t) \tag{12}$$

In the formula, λ_t^m represents the m-th univariate component of the total variable λ_t; $g(\lambda_t^m)$ is the best approximation result of single variable λ_t^m; $g(x_t)$ is the moving target position; $g(\varepsilon_t)$ is the expected value of position; $g(v_t)$ is the moving speed; $g(\sigma_t)$ is the target transfer accuracy; $g(h_t)$ is the observation accuracy. Therefore, according to the above calculation results, Bayesian network is used to analyze the spatial transition state of the target data in the system, and then the dynamic law of the data change in the next cycle is predicted. It is known that there is mutual independence between the observation data, so according to the likelihood distribution corresponding to the observation information, the mobile tracking position is updated according to Bayesian network, and the position with deviation is corrected [12, 13]. According to the basic definition of likelihood distribution, the nonlinear observation results and observation accuracy are uncertain, and the posterior distribution $q(\lambda_t|R_{1:t})$ loses the ability of closed expression. Therefore, according to the multi-layer dynamic characteristics of Bayesian network, the posterior distribution $q(\lambda_t|R_{1:t})$ is adjusted to obtain the best variational approximation of a single variable.

$$f(\theta) = \exp\langle \ln g(\theta, h(\theta))\rangle_{q(\lambda_t|R_{1:t})} \tag{13}$$

In the formula: θ represents prior parameter; $h(\theta)$ is the Markov carpet of a single variable λ_t^m in Bayesian networks. According to the above calculation process, the moving target in the system database is located and tracked, and the real-time change law of the target is mastered. The data fusion is carried out through Bayesian network, which provides the operation basis for the decision support system to generate enterprise decision results. So far, the enterprise intelligent decision support system based on data mining is realized.

4   Experimental Study

4.1   Experimental Preparation

In order to verify that the design of the decision support system, can be truly applied to the development of enterprises, the establishment of simulation experiment test environment. Several PC servers are selected as application server and database server. The CPU model of the server is Intel Xeon 3.0 GHz * 4, with 32 g memory and 1 TB hard disk. It supports Windows Server 2018 operating system. Figure 6 shows two different test environments for this experiment.
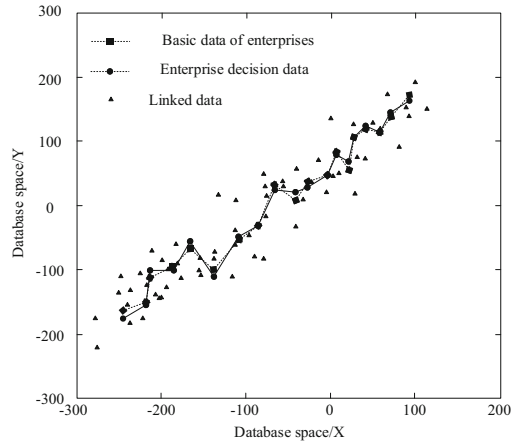
(a) Simple enterprise decision objectives
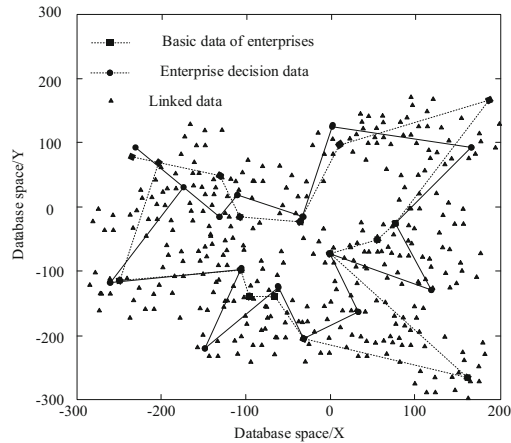
(b) Decision-making objectives of complex enterprises

Fig. 6.   Experimental test environment

In Fig. 6, (a) and (b) are respectively affected by different volumes of associated data, which have different degrees of impact on the decision support of the system. Taking the above test conditions as variables, two groups of traditional design decision support systems based on big data in reference [4] and artificial intelligence in reference [5] are selected as the control group, which are recorded as traditional system 1 and traditional system 2 respectively, and the performance of the three groups of systems is analyzed.

4.2   Response Time Test

As an important parameter, response time is one of the measurement standards of system quality, so when testing system quality, it is necessary to test the corresponding time. Three groups of systems are used to process the enterprise development data, and the response time of the system entering the information integration interface, information monitoring interface, information retrieval interface and information extraction interface is distinguished. Figure 7 below is the test result of system response time based on the test environment in Fig. 6 (a). Among them, the test times were 200 times each time, and 10 groups were tested.
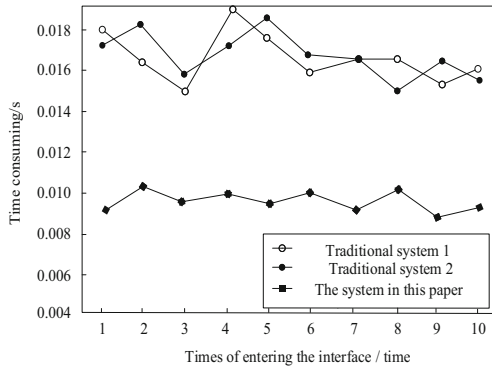
(a) Enter the information integration interface
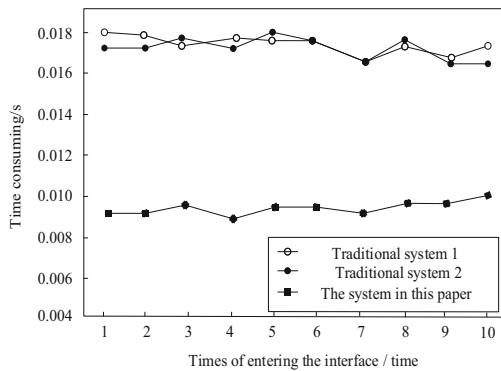
(b) Enter the information monitoring interface

Fig. 7. System response time test results in a simple environment
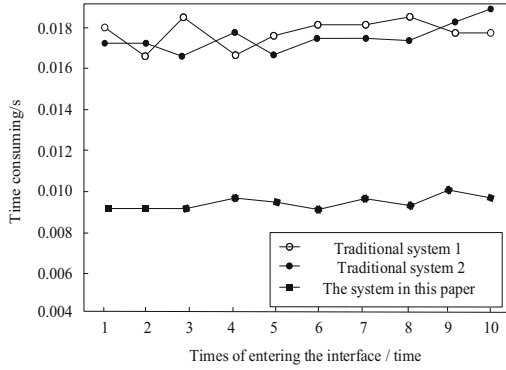
(c) Access to the information retrieval interface
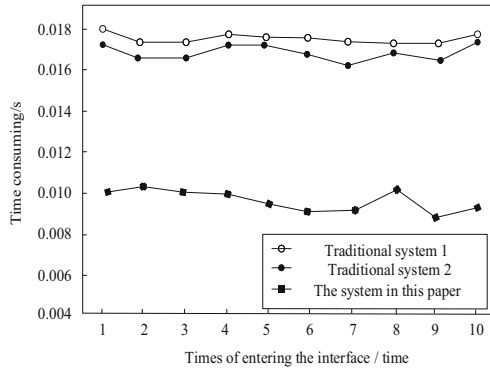
(d) Access to the information extraction interface

Fig. 7. continued

According to the four groups of test curves, it is obvious that the response time of the system in this paper is slightly lower than that of the traditional system. The total response time of the four groups of systems is exported and summarized in Table 2 below.

Table 2. Average response time of system in simple environment (s)

Test group	The system in this paper	Traditional system 1	Traditional system 2
a	0.0098	0.0167	0.0182
b	0.0086	0.0178	0.0177
c	0.0094	0.0178	0.0183
d	0.0107	0.0179	0.0172

a, b, c and d in Table 2 correspond to (a), (b) (c) and (d). According to the test results in Table 2, the response time of the system in this paper is obviously less than that of the two groups of traditional systems, but the difference between the three groups of systems is small and can be ignored. The main reason is that the system uses data mining technology to process and classify the response data before decision-making, which optimizes the transmission performance of the system. Based on the test conditions in Fig. 6 (b), the response time of three groups of systems in complex environment is compared, and the results are shown in Fig. 8.
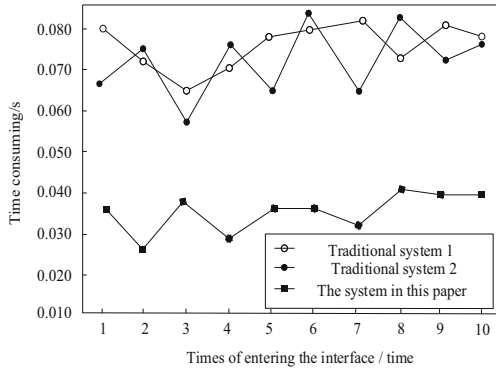
(a) Enter the information integration interface
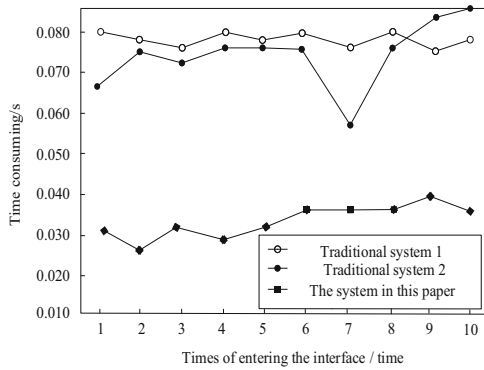
(b) Enter the information monitoring interface

Fig. 8. System response time test results in complex environment.
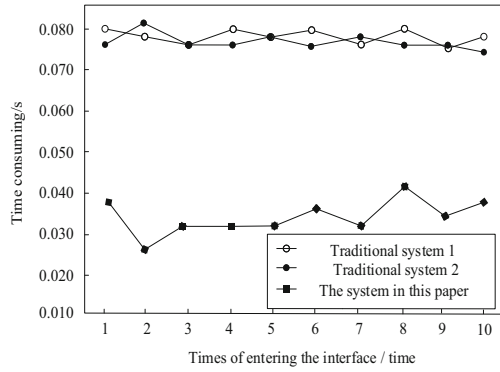
(c) Enter the information retrieval interface
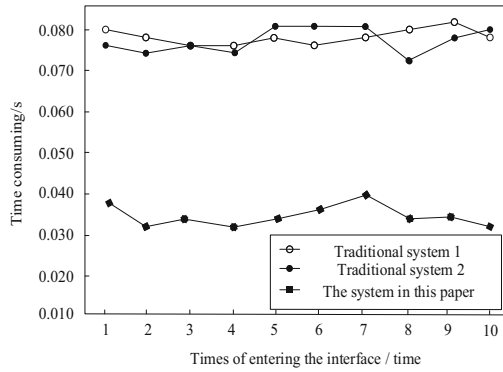
(d) Enter the information extraction interface

Fig. 8. continued

According to the test results of the second stage, although the response time of the system in this paper increases, it is far less than the two groups of traditional systems. The total response of the system is also exported and summarized in Table 3 below.

Table 3. Average response time of system in complex environment (s)

Test group	The system in this paper	Traditional system 1	Traditional system 2
a	0.0325	0.0715	0.0691
b	0.0322	0.0794	0.0699
c	0.0317	0.0792	0.0752
d	0.0331	0.0788	0.0767

According to the test results shown in Table 3, in the face of a complex system environment, the response time of the system in this paper is obviously slightly less than that of the two groups of traditional systems, but there is a great difference between the three groups of systems at this time. The average time consumption of the three groups of systems in the four groups is 0.0324 s, 0.0772 s and 0.0727s respectively. According to the above calculation results, the time consumption of the system in this paper is 0.0448 s and 0.0403 s shorter than that of the traditional system respectively, which shows that the response performance of the system in this paper is better. In this paper, the response time of the system is increased because the complex environment has a certain impact on the efficiency of data processing, but compared with the comparison system, the data mining technology only slows down the processing efficiency, and the overall response performance is still better than the comparison system.

4.3   Decision-Making Accuracy Tests

The first half year enterprise data of a company is selected as the basic background of the test, and three groups of systems are used to integrate and manage the information during this period, and the authenticity and integrity of the feedback data of the system are compared. The following Table 4 shows the missing and redundant data in the test task.

Table 4. Problems with data

Question number	Missing ratio	Redundancy ratio
1	4.2%	3.5%
2	8.3%	1.7%
3	16.7%	11.3%
4	7.45%	6.4%

According to the statistical data in Table 4, there are four types of data problems in the data of the enterprise. Three groups of systems are used to make data decisions under the test conditions shown in Table 4. The test results of the decision accuracy of the system are shown in Fig. 9 below.
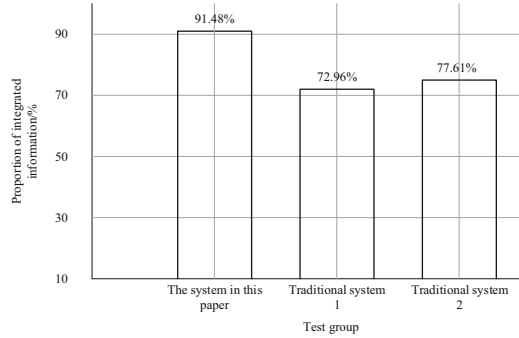
(a) Conditions 1
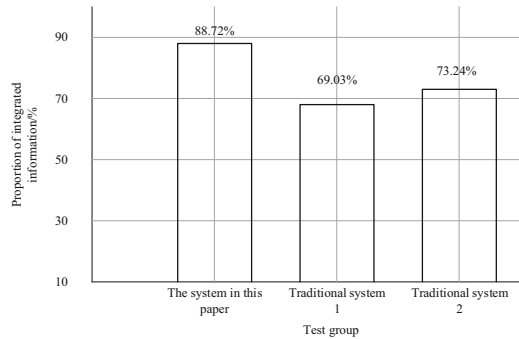
(b) Conditions 2

Fig. 9. Enterprise information integration control quality
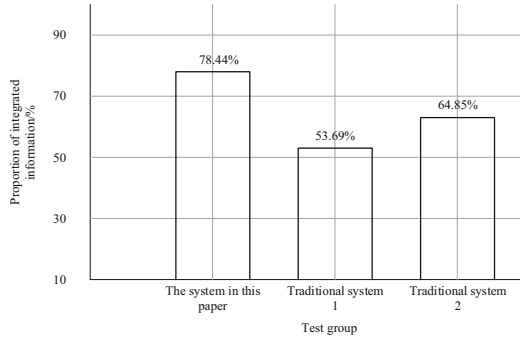
(c) Conditions 3
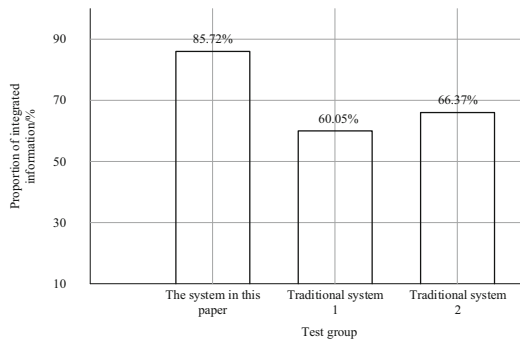
(d) Conditions 4

Fig. 9. continued

According to the test results shown in Fig. 9, when the data defects in the three groups of system applications are small, the system decision accuracy is high; When the system data defect is large, the system decision accuracy is low. The average decision accuracy of different systems was 86.09%, 63.93% and 70.52% respectively. According to the above calculation results, the decision accuracy of the design system based on data mining technology is 22.16% higher than that of the traditional system 1 and 15.57% higher than that of the traditional system 2.

5   Conclusion

Starting from the functional requirements of decision support system, this paper analyzes that enterprise management is a complex decision-making problem. It has the characteristics of dynamic, distributed, large amount of data and complex nature of data. In the face of sudden disasters and massive, complex, changeable and distributed data and information, it is difficult for decision makers to make timely decisions. The traditional decision support system can not meet the needs of modern enterprises because of its lack of function. Data mining technology is one of the effective ways to solve this problem.

Therefore, this paper designs an enterprise intelligent decision support system based on data mining. According to the above test results, the response time and control quality of the system are better than the traditional system, which provides more reliable data for enterprise decision-making. But in the process of designing the system, the calculation program involved is more complex, and the system is prone to collapse. In the future, it is necessary to simplify some software programs and test the operating pressure of the system to prevent the system from frequent crashes.

References

1. Gao, X., Lei, X., Wang, Y., et al.: Building of decision support system for primary health care service. J. Med. Intell. **40**(04), 24–27 (2019)
2. Wang, W., Xue, M., Wang, Y., et al.: Development of transportation planning and management support system: Highlight of the 24th urban transportation development forum in China. Urban Transp. China **18**(01), 102–113 (2020)
3. Wang, W., Huang, J., Guo, S.: Construction of hospital operation decision support system based on Big Data. China Digit. Med. **14**(02), 49–51 (2019)
4. Zhan, J., Li, T, Li, C.: Decision support system of adaptability evaluation for TBM selection based on artificial intelligence. J. China Coal Soc. **44**(10), 3258–3271 (2019)
5. Ding, Q., Li, J., Liu, N.: The applied research of data mining technology of aggregate information client-side: A case study of the user login and database on recommendable news. Stat. Inform. Forum **34**(02), 121–128 (2019)
6. Huayi, W., Rui, H., Lan, Y., et al.: Recent progress in taxi trajectory data mining. Acta Geodaetica et Cartographica Sinica **48**(11), 1341–1356 (2019)
7. Liu, K., Wang, H., Liu, Y, et al.: A quantitative method for evaluating management level of distribution network planned outages considering investment. Power Syst. Technol. **43**(07), 2282–2291 (2019)
8. Li, Y., Mu, F., Liang, J.: Construction and application of evaluation index system for wetland tourism resources. Econ. Geogr. **39**(01), 192–197 (2019)
9. Hui, Z, Di, M, Wei, W, et al.: Inter-operation mechanism for handle system and domain name system:implementation based on markup language describing protocol data unit. Appl. Res. Comput. **36**(01), 194–198 (2019)
10. Peng, H., Zhu, Z.: Analysis on human-machine collaborative decision-making for personalized adaptive learning. E-educ. Res. **40**(02), 12–20 (2019)
11. Gao, P., Li, J., Liu, S.: An introduction to key technology in artificial intelligence and big data driven e-Learning and e-Education. Mob. Netw. Appl. **26**(5), 2123–2126 (2021). https://doi.org/10.1007/s11036-021-01777-7
12. Liu, S., Liu, D., Srivastava, G., Połap, D., Woźniak, M.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. **7**(4), 1895–1917 (2020). https://doi.org/10.1007/s40747-020-00161-4
13. Liu, S., Liu, D., Muhammad, K., Ding, W.: Effective template update mechanism in visual tracking with background clutter. Neurocomputing (2020). https://doi.org/10.1016/j.neucom.2019.12.143

Recommendation Method of Nursing Teaching Resources in Geriatric Internal Medicine Based on Internet of Things Technology

Hua Fan[✉]

Medical and Nursing Branch, Panjin Vocational and Technical College, Panjin 124000, China
jwqddsf000@aliyun.com

Abstract. When many learning websites recommend course resources, the correlation between recommended resources and target resources is not very good, which leads to the low correlation between recommended courses and target resources. This paper designs a method of elderly nursing teaching resources recommendation based on Internet of things technology. Integrate interactive data and model data to browse course feature data. The correlation between recommended resources and target resources is compared, and the correlation between browsing function data and recommended resources is calculated. According to the degree of feature association, the weight of recommended resources is calculated. Through the weight set sorting, with the help of Internet of things technology to rearrange the recommended order of curriculum resources, complete the design of elderly nursing teaching resources recommendation method based on Internet of things technology. Through experimental analysis: the highest correlation between recommended courses and target resources is 72%. The results show that the method is more perfect.

Keywords: Internet of Things technology · Geriatric internal medicine nursing · Curriculum resource recommendation · Recommendation order

1 Introduction

Digital learning of Internet of things has become an important way of distance education and learning [1]. The network teaching resources are also growing rapidly. The overload of information brings many challenges to teaching organizers and learners. They have to spend a lot of time and energy to screen out the resources that meet their needs. And the Internet of things technology is an effective way to solve this problem [2]. It has been widely used in the business field and has achieved remarkable results by recommending the information and commodities of interest to users according to the characteristics of users' interest and historical behaviors. The pre oral recommendation system has also been applied to e-learning environment. It can help teaching organizers and learners to effectively reduce the burden of information overload and improve the efficiency of work learning. However, it is difficult to apply the recommendation methods

and systems directly to business field in e-learning, because the teaching and learning process is more complex than the business system, and is subject to the user's knowledge level and teaching The influence of the task of oral standard and the characteristics of the subject. With the development of Internet of things technology and the improvement of people's living standard [3].

In the related research, the common resource recommendation methods include personalized learning resource recommendation method based on 3D feature co domination and online learning resource recommendation method based on ontology and recurrent neural network. Among them, the personalized learning resource recommendation method based on Collaborative domination of three-dimensional features improves the matching relationship between learners and online learning resources, establishes the Personalized Learning Resource Recommendation Model Based on Collaborative domination of three-dimensional features, and describes it parametrically; Secondly, a binary particle swarm optimization algorithm based on fuzzy control of Gaussian membership function (fcbpso) is designed to solve the objective function of the recommended model; Finally, five groups of comparative experiments are conducted to verify the good performance of the proposed method. This method designs reasonable indicators in recommendation, but it has the problem of less resources. In the online learning resource recommendation method based on ontology and recurrent neural network, this method uses ontology method to model and represent the domain knowledge of learners and learning resources, and uses recurrent neural network to mine learners' learning mode to get the final resource recommendation list. This method has high recommendation accuracy, but poor linkage.

The cause and spectrum of diseases have changed greatly, which also led to the progress of medical science and technology and the change of medical mode. It also poses new challenges to nursing work. It requires that the nursing work should change the patients from the whole concept of biology, psychology and society, from the simple disease nursing mode to the "people-oriented" nursing mode; from the individual nursing of patients to group nursing; Let the patient produce the concept of changing from passive nursing to active disease prevention [4]. Of course, with the change of disease spectrum and aging of population, the cooperation and exchange in Internet of things technology and nursing are expanding day by day. Diversified community care services, such as family care, hospice care, day ward. These all put forward higher requirements for nursing, requiring the development of nursing work to specialization and specialization, and the nurses' theoretical study, self-study ability, active change of nursing service mode and the concept of nursing education.

2 The Design of the Recommended Method of Nursing Teaching Resources for the Elderly

2.1 The Collection and Browsing Course Characteristic Data of Internet of Things Technology

The data collection of course features mainly includes the collection of interactive data between online layer and users and the construction of model, including teacher behavior

record, teacher model, course model and resource model. Teacher behavior record is used to record various operations of teachers on the system platform. Based on the Internet of things technology, it provides data basis for personalized recommendation. According to the behavior objects of teachers, teachers' behavior records can be divided into resource behavior records, teaching task records and label behavior records [5]. The operation behavior of resources includes downloading, scoring, collecting and browsing; teaching task record refers to the teacher's teaching design and lesson preparation in the process of UI interaction; tag behavior record refers to the teacher's operation of recommended resource tags, including selection, modification, deletion and addition [6]. The course model is used to represent the relationship between courses and the content characteristics of each lesson [7]. It is composed of curriculum relationship and curriculum label based on Internet of things technology. Use resources related to the course. Describes the state sequence of a data set, each of which depends on the first and finite states. Therefore, the Internet of things technology is used to express the characteristic relationship between courses, that is, the course prepared by teachers in the next state is only related to the current position of the course. By using the state transformation relationship of Internet of things technology, teachers' position in the course structure can be predicted when they log in to the system again, so as to collect relevant resources more accurately, as shown in Fig. 1.
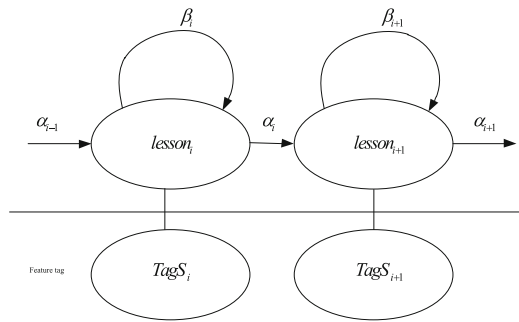
Fig. 1. Course characteristic data

In the figure, αi is the number of course nodes, which represents the probability from lessoni state to lessoni+ 1, βi represents the probability from lessonI to lessoni, $\alpha i + \beta i = 1$, $\alpha 0 = 0$. The curriculum label is used to represent the content characteristics of each class. It is represented by vector space model, and the formula is as follows:

$$TagS_i = \left\{ tag_k : w_k \middle| 1 \le k \le N_t, \sum_{k=1}^{N_t} w_k = 1 \right\} \tag{1}$$

where tagk is the k tag, Wk is its weight, and Nt is the number of tag sets. The initialization and adaptive adjustment of tag sets are completed by the course tag adjustment module in the offline layer. It mainly describes the resource owner's description of the resource, including resource title, resource keyword, resource description and resource type (such

as picture, video, audio, document, etc.); it represents the content characteristics of the resource, expressed by VSM. the formula is as follows:

$$TagS_i = \left\{ tag_k : w_k \middle| 1 \le k \le N_i, \sum_{k}^{N_i} w_k = 1 \right\} \tag{2}$$

Among them, Nt is the number of label sets, initialization and adjustment process, which is completed by the resource tag adjustment module of offline layer, and consists of static and dynamic curriculum attributes. Record the necessary information filled in by users during registration, such as user name, registration mailbox, etc., and record the position of teachers in the curriculum model and dynamically track the user's context. Analyze and process data of data layer, initialize and adjust the model, and calculate the relevance or similarity between curriculum and resources, resources and resources based on content and user behavior data, and provide data support for mixed recommendation algorithm of online layer [8]. The course label adjustment module initializes and adjusts the curriculum label in the curriculum model, which makes the curriculum label more accurately and comprehensively represent the content characteristics of the course. The module is composed of the crawler module and the text analysis module. The course name is used as the query word by Internet technology, simulating the search process and crawling the relevant web content. The paper analyzes the Internet of things technology module uses word frequency reverse file frequency algorithm to analyze the web page content obtained by the web crawler, or extract the content characteristics of teaching task record in the data layer, so as to realize the adjustment of course label [9]. In the course of course label initialization, the extracted content features are directly regarded as curriculum tags; in the course of course label adjustment, the content features are obtained by analyzing the teaching task records, and the common words and tags with low weight are removed by the method of unified training. Through the potential semantic analysis method, the curriculum label matrix is constructed, the eigenvector of the matrix is solved by matrix decomposition method, and the feature vector is analyzed by distance measurement method, and the adjacent neighbor of the label is solved to expand the curriculum label set and dynamically adjust the course label, so as to improve the course label, as shown in Fig. 2.

The resource label adjustment module initializes and adjusts the resource label in the clinical nursing resource model. Different from the adjustment of course label, due to the differences in the types of resources (including video, pictures, documents, audio, etc.), it is impossible to analyze the content of resources based on DF algorithm. Meanwhile, the description text of static attribute of clinical nursing is relatively short [10]; therefore, in the initialization process, the static attributes of resources are analyzed by using file frequency Internet of things technology, common words are filtered and keywords are extracted as label sets. The weight of the label obtained by nursing name, clinical key words and nursing description is different. The process of resource label adjustment is similar to the course label adjustment process. According to the teacher's feedback behavior to the resource label, the resource label set is optimized by statistics and potential semantic analysis. The course resource relevance training calculation module and resource similarity training module are based on the curriculum content characteristics (curriculum tags) and resource content features (resource tags), and calculate the similarity between curriculum and resources, resources and resources respectively according to
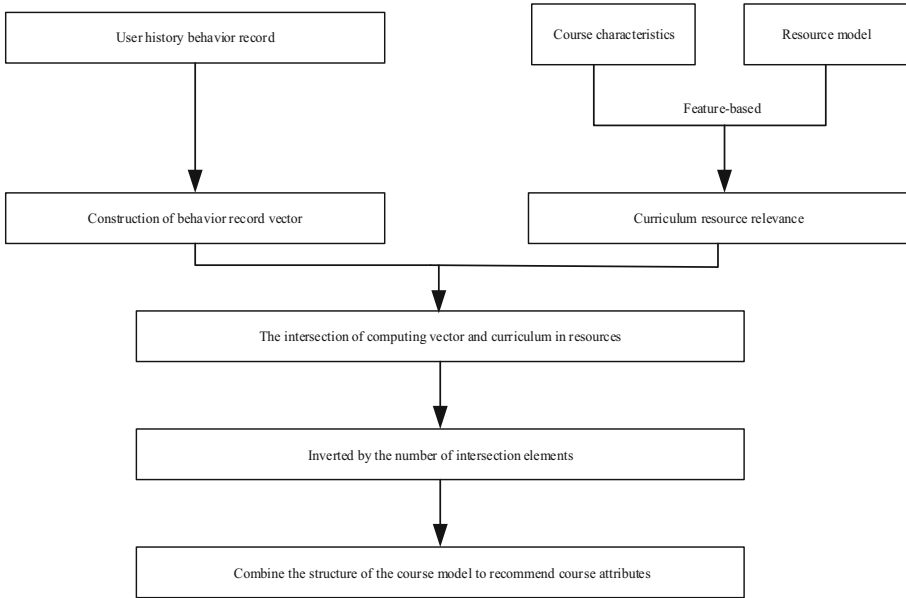
Fig. 2. Reasoning of course resource characteristics

similarity measurement and k-nearest neighbor algorithm, and establish the relationship between curriculum and resources, resources and resources. The resource association degree calculation module analyzes the teacher's resource behavior record and transforms it into the "shopping basket" problem. It transforms the resources of each login operation into the user's purchase record, mining by association rules, and training the correlation between resources. Teacher model reasoning, based on the teacher's behavior record of resources in the system, and combined with the results of the calculation of curriculum resource relevance, infers the current position of the context in the curriculum model, thus completes the collection of the whole browsing course feature data, and then forms the comparison between the feature data and the recommended resources, and then determines the correlation value between the two.

2.2 The Correlation Degree Between Browsing Feature Data and Recommended Resources

According to the obtained browsing course feature data, the correlation degree of feature data is compared, and the problem of course recommendation is abstracted and formalized as follows: given the learner learning data on the learning website of geriatric medical nursing teaching course, we get a set of M learners $L = \{l1, l2, …, lM\}$ and a set of N courses $C = \{c1, c2, …, cN\}$. The interaction matrix between learners and courses can be obtained from the records of course selection, which is recorded as $R \in RM \times N$. The interaction matrix $Y = \{yic \mid l \in L, c \in C\}$ is defined by the learner's course selection behavior:

$$y_{lc} = \begin{cases} 1, \text{Learners chose the course} \\ 0, \text{other} \end{cases} \tag{3}$$

ylc = 1 means that learner 1 has selected course c, otherwise ylc = 0. As a supplement to the recommendation information, we also have a knowledge map G = (E, R), E and R represent the set of entities and relationships in knowledge map G. Each data in knowledge map G is composed of triples (h, r, t). h, r and t represent the header entity, relationship and tail entity of the triples respectively, where h ∈ E, r ∈ R, t ∈ E. For example, triples mean that the invention of C language originated in BCPL. In the course recommendation scenario, a course c ∈ C consists of a word sequence, namely C = (W1, W2, W3,…), for example, "Python development simple Crawler" = (Python, development, simplicity, crawler). Each of these words may correspond to an entity e ∈ E in the knowledge map G. For example, Python and crawler in the course Python development simple crawler correspond to entities python (computer programming language) and web crawler, respectively. Based on the interaction matrix Y between learners and curriculum and curriculum knowledge map G. The goal of the recommendation algorithm is to estimate one on one course for the elderly medical care course learner who has not previously chosen. What is the possibility of choosing a course. That is, through the model, a prediction function is trained to turn y = F(l,c | θ, Y, G), where y represents the probability of selecting C from the course selected by the learner 1 of the elderly medical nursing course predicted by the model, and theta represents the parameters related to function F. According to the comparison of the target feature data and the resource feature data, the paper finds out whether there are other target feature data between them. According to the recommendation algorithm of knowledge map enhancement, the compactness between the two is calculated. The resource feature and target feature direction quantization module are shown in Fig. 3.
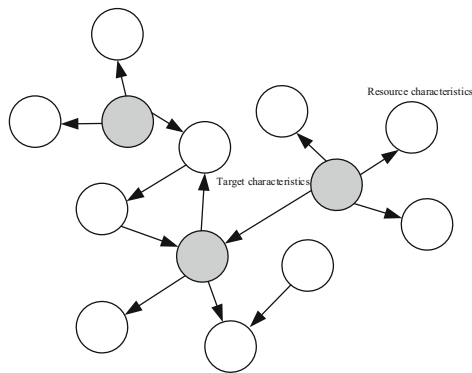
Fig. 3. Recommended data target feature orientation module

Based on this, the vector module compactness of target feature and resource feature is calculated. The sparse feature vectors of nursing course learners and courses in geriatric

medicine were used as input, and transformed into dense vector representation through full connection layer. The deep feature learning module takes the embedding of learners and courses as the input, as shown in Fig. 4.
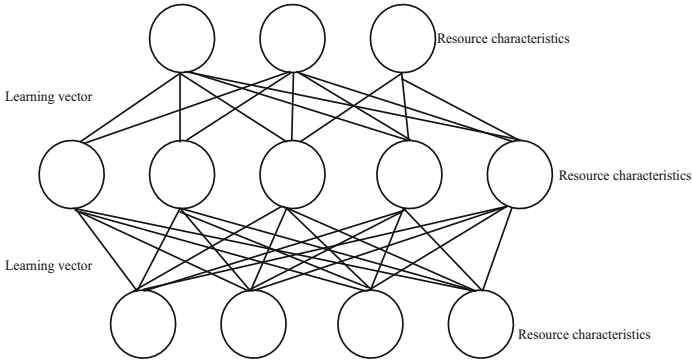
Fig. 4. Vectorized feature input

The deep perception mechanism is used to mine the characteristics of learners and courses, and the nonlinear characteristics of learners and courses are obtained. The knowledge map feature learning module can extract features based on the idea of multi-layer relationship diffusion. While learning the vector representation of entities and relationships in the knowledge map, we also learn the linear feature representation of learners and courses. Connect the prediction module, connect the vectors from the deep collaborative filtering module and the knowledge map feature learning module, and then normalize them to get the recommended prediction value. Thus, the correlation degree between browsing feature data and recommendation resources is calculated.

2.3 The Recommended Resources are Weighted According to the Feature Association Degree

The calculated correlation degree value is used as the activation coefficient in the MLP module to perform the weight assignment of the feature correlation degree. The number of MLP layers can be adjusted according to the experimental requirements, and the number of neurons is halved layer by layer. Use LM to denote the hidden vector of the learner in the MLP module, and CM to denote the hidden vector of the course in the MLP module. Each output layer of MLP is expressed as, the input of the module is also the learner vector and the course vector obtained by the input and vectorization module. Through the learning and training of this module, we can get the entity set contained in the hidden vector LR of the learner in the Ripple module and the hidden vector CRo of the course in the Ripple module as E1 = {e1, e2,…, eu}. The entity set E1 is regarded as the seed set of the knowledge graph. Just like throwing a stone into the water and splashing layers of ripples, the seed entity will spread outward along the correlation r in the knowledge map, forming a circle of "water waves". The set of triples involved in each wave of water, that is, the set of triples with a distance of k from the seed set is called the set of water waves bang (k = 1, 2, 3,…, H). The water wave set will interact with

the course word segmentation entity for multiple iterations, so as to obtain the learner's degree of love (or interest preference) for the course. Combining all these preference values can get the user's characteristic LR. The relationship weight Pi from course name to course participle can be calculated by the following formula:

$$P_i = \left(C_R^T R_i h_i\right) = \frac{\exp\left(C_R^T R_i h_i\right)}{\sum_{(h,r,e)\in Co_i} \exp\left(C_R^T R h\right)} \tag{4}$$

Among them, i is the data set, c is the feature data of the course, R is the recommended learning feature, h is the feature interaction matrix, and T is the parameter obey average value. According to this, the process of assigning the recommended resource weight to the feature relevance is completed.

2.4 Combine Recommendation Weights to Arrange the Recommendation Order of Course Resources

According to the weight value of the above-mentioned feature association degree, the recommended resources are arranged in order. In the knowledge graph feature learning module of the original model, the Internet of Things technology is used to carry out the feature ranking method of multi-layer relationship diffusion. The core idea is to use the multi-layer relationship that seed entities diffuse in the knowledge graph to find the user's interest characteristics. Especially for the related entities of the superimposed part of the layer caused by the "splashing ripples" of seeds, a phenomenon similar to "interference enhancement" in physics will occur. This strengthening phenomenon caused by multiple "hits" and the use of co-occurrence networks to find hot spots and preferences are shown in Fig. 5.
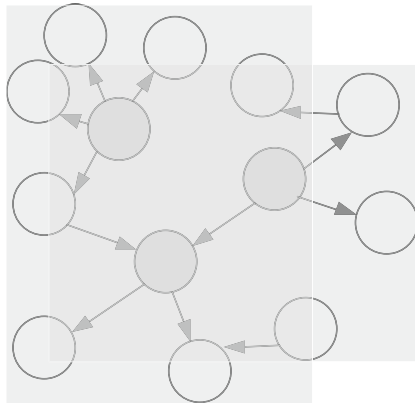
Fig. 5. The feature weights of recommended resources and courses are overlapped and added

The overlapping and shaded part in the figure is the co-reality network Co-net, which contains the course entity (the dark part) and the course word segmentation entity (the

light part). For the training of the co-reality network and the determination of the weights of the edges, we adopt the same method as the knowledge graph training. The common reality network information is analogous to the knowledge graph triplet information, and the relationship is assigned to the two types of "technical participle" and "non-technical participle", which can form (course name, technical participle, technical entity) and Course name, non-technical participle, non-technical entity) triples, the set of triples is called Col. Among them, the loss function of the algorithm includes the loss value of the learner-course interaction matrix, the loss value of the learning knowledge graph and the common reality network, and the regular term to prevent overfitting of training. In the specific training process, we used stochastic gradient descent (SGD) to optimize the parameters. At the same time, during the training process, we added negative samples to the learner's course selection records and knowledge graph data to improve the efficiency of training, and the number of negative samples and positive samples remained the same. By optimizing the loss function, the model will use the backpropagation algorithm to update the relevant parameters and characteristics of the model. So far, the design of the recommended method for the elderly internal medicine nursing teaching curriculum resources based on the Internet of Things technology is completed.

3   Experiment Analysis

Analyze the comparative experiments of traditional method 1, traditional method 2 and the recommendation method design of elderly internal medicine nursing teaching curriculum resources based on Internet of things technology, compare the relevance of recommended courses and target resources, so as to confirm the better recommendation method.

3.1   Experiment Preparation

The data crawling and cleaning in this research, and the realization of the model are all realized by writing code in Python language. The model uses Tensorflow as the basis of algorithm implementation. The main hardware and software configuration of the experiment is shown in the following table (Table 1):

Table 1. Basic values of experimental configuration

Experimental configuration	Parameters table
CPU	I52200U
RAM	8192MB
Existing	5400
System operating system	4182MB
Python version	3.4.6
Tensorflow	1.14

The course selection experiment data used in this article is taken from a certain geriatric internal medicine nursing course website, which includes learner selection record data, auxiliary information of all selected courses (course name, course category, course difficulty, etc.), and learner-related information. The final crawled data set situation is shown in Table 2:

Table 2. Basic parameters of experimental data

Data set	Group A	Group B	Group C
Number of courses	943	833	941
Number of learners	73773	52415	67453
Number of interactions	4116285	3015389	443792
Course auxiliary information	Category, name, instructor	Name, teacher	Classification name
Auxiliary information for learners	Name, study time, online time, search preference	Study time, online time	Search preferences, study time

According to the accuracy estimation of the scoring matrix of the experimental data parameters, predict the relevance of the recommendation task and the target resource, and use Accuracy to evaluate the final effect of the model recommendation. Accuracy represents the percentage of the correct number of samples in the test set that the model classifies to the total number of samples. Generally speaking, it is the ratio of the number of samples that classify positive samples into positive samples and negative samples into negative samples to the total samples. In the prediction task, the value of the correlation degree can be calculated by the following formula: target correlation degree recommended resource feature data set/target resource feature data set.

3.2   Experimental Result

Figure 6 shows the correlation between the recommended courses and target resources of the traditional method 1, traditional method 2 and the Internet of Things-based elderly internal medicine nursing teaching curriculum resource recommendation method.

According to the data in the figure, the correlation degree between the recommended courses and target resources in the traditional method 1 is up to 63%. As the number of recommended courses increases, the correlation degree gradually decreases, and the correlation degree between the recommended courses and target resources is as low as 33%. The traditional method 2 recommends courses and the target resources have a maximum relevance of 65%, and the minimum decrease of relevance is 30%. Based on the Internet of Things technology, the relevance of the recommended courses and target resources of the recommended methods for the elderly internal medicine nursing teaching curriculum resources is 72%, and the relevance is reduced to 51%. Therefore, it is better to recommend the resources of nursing teaching courses for the elderly based on the Internet of Things technology.
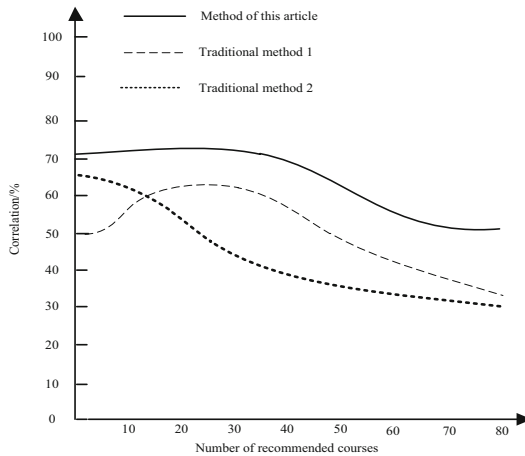
Fig. 6. The degree of relevance between recommended courses and target resources

4   Conclusion

Through the research of this article, the relevance of recommended courses for geriatric internal medicine nursing has been improved, and the accuracy of recommended course resources has been improved. In the future, in the learning part of the common reality network, in order to better integrate the common reality network with the original knowledge map, it should use the hit frequency or reference other models to train and generate, so that the weight of the target feature can be optimized, and the accuracy of the recommendation can be improved. rate.

References

1. Zhmud, V., Liapidevskiy, A., Avrmachuk, V., et al.: Analysis of barriers to the development of Industrial Internet of Things technology and ways to overcome them. IOP Conf. Ser. Mater. Sci. Eng. **1019**(1), 120–125 (2021)
2. Khan, M., Imtiaz, S., Parvaiz, G.S., et al.: Integration of Internet-of-Things with blockchain technology to enhance humanitarian logistics performance. IEEE Access **01**(12), 1 (2021)
3. Nandish, B.M., Pushparajesh, V.: Review of internet of things: distributed power in smart grid. IOP Conf. Ser. Mater. Sci. Eng. **1055**(1), 1213–1214 (2021)
4. Jiao, L., Sui, Y., Yang, G., et al.: The construction of the evaluation system of nurses' post-training and the application of the system in 25 grade-A general hospitals in China. Nurs. Open **8**(1), 45–52 (2021)
5. Atkinson, M., Lindberg, J., Zanotti-Cavazzoni, S., et al.: 333: Multidisciplinary collaboration leads to increased nursing satisfaction. Crit. Care Med. **49**(1), 155 (2021)

6. Li, H.J., Zhang, Z., Zhang, P.W.: Personalized learning resource recommendation method based on three-dimensional feature cooperative domination. Comput. Sci. **46**(01), 461–467 (2019)
7. Laura, R.N.: An es1cape room simulation focused on renal-impairment for prelicensure nursing students. Teach. Learn. Nurs. **16**(1), 95–99 (2021)
8. Gcawu, S.N., Rooyen, R.M.V., Jordan, P., et al.: Clinical teaching practices of nurse educators at a public college of nursing in South Africa: A survey study. Nurse Educ. Pract. **50**(12), 102935–102940 (2021)
9. Chen, J., Lu, Y., Shang, F., et al.: A fuzzy matrix factor recommendation method with forgetting function and user features. Appl. Soft Comput. **100**(1), 106–115 (2021)
10. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mob. Netw. Appl. **24**(1), 5–17 (2018). https://doi.org/10.1007/s11036-018-1134-8

Recognition of Human Abnormal Behavior in Static Image of Intelligent Monitoring System Based on Neural Network Algorithm

Hai-jing Zhou[✉]

Chongqing Vocational Institute of Tourism, Chongqing 409099, China
`zhouhaijing554@yeah.net`

Abstract. In recent years, human abnormal behavior recognition in static images is a hot topic in the vision field of intelligent monitoring system. In order to achieve the goal of accurate recognition of human abnormal behavior in static images, a method of human abnormal behavior recognition in static images of intelligent monitoring system based on neural network algorithm is proposed. The pedestrian detection algorithm based on neural network algorithm for filtering channel features is used to locate each target human body, collect feature data and denoise. Then the appearance model of each human structure feature frame is constructed, and finally the neural network algorithm is used for behavior classification and recognition. The experimental results show that the human abnormal behavior recognition method based on neural network algorithm in the static image of intelligent monitoring system can recognize the behavior of multiple human bodies in a single frame image. This method can provide multi category classification of human abnormal behavior at the same time, the experimental effect is obvious, and the accuracy is high.

Keywords: Neural Network · Intelligent monitoring · Static image · Behavior recognition

1 Introduction

Human abnormal behavior recognition in static images has a wide range of application prospects and potential economic value in security monitoring, advanced human-computer interaction, video conference, behavior based video retrieval and medical diagnosis. It is a research hotspot in the field of computer vision. Video intelligent monitoring system is one of its most important applications [1]. At present, the detection, tracking and recognition of static target in intelligent monitoring system is still in the primary stage, and the behavior analysis of static target is relatively simple. However, due to the interference factors, the recognition effect is still difficult to meet the current research objectives. Based on this, this paper discusses the common problems of human abnormal behavior recognition in the monitoring scene, and defines a small amount of behavior on the scene as abnormal behavior. Based on this, this paper studies the human abnormal

behavior recognition method in video surveillance, and expounds the research significance, development status and application fields of human abnormal behavior detection and recognition method [2]. In the field of computer vision, crowd abnormal behavior detection technology can be widely used in video surveillance, intelligent video analysis, group behavior recognition and other fields. Jiang Jun et al. [3] proposed a crowd abnormal behavior detection method based on pulse line flow convolutional neural network using deep residual network as the backbone network, extracting the apparent characteristics of crowd behavior through the input of the spatial network, and using the pulse line Stream extracts the movement characteristics of crowd behaviors to further improve the accuracy of crowd abnormal behavior detection. Obtain significant spatial characteristics of crowd behavior through convolutional neural network, and obtain temporal characteristics of crowd behavior through pulse flow combined with convolutional neural network; finally, the output of the two networks is averaged to complete the detection of abnormal crowd behavior. In the process of criminal investigation, Wu Songwei [4] and others used visual-based behavior analysis and understanding technology to discover and identify the abnormal behavior of potential perpetrators in a timely manner, which has extremely important practical significance for crime early warning. By designing the hardware and software system framework of the video surveillance abnormal behavior detection system, researching video-based pedestrian detection methods and abnormal behavior detection and recognition methods, and completing the implementation and development of related algorithms on the system, thereby improving the accuracy of the system and efficiently detecting abnormal behaviors ability. Zhang Xin and Qi Hua [5] studied and analyzed the complex and diverse problems of abnormal human behavior characteristics in view of the situation where the classification of human behavior in the surveillance video in public places is greatly affected by the background. Based on the YOLOv4 network model, a network human body abnormal behavior detection model is established. The experimental data results are obtained through K-Means clustering, and the processed data samples are sent to the three models for training. The test results show that it can meet the requirements of high accuracy and real-time. The advantages and disadvantages of the existing methods are summarized and analyzed, and some key technical methods are studied and improved. The main contents of this paper are static target detection, static target tracking, human behavior feature extraction and human behavior recognition methods. In static target detection, the parameters of hybrid Gaussian model are initialized by neural network method, which avoids the influence of static target on the initialization of Gaussian model [6]. The recognition method based on neural network can adaptively select the number of recognition points, and improve the accuracy and efficiency of static target recognition.

2 Recognition of Human Abnormal Behavior in Static Image of Intelligent Monitoring System

2.1 Collection of Abnormal Behavior Characteristics of Human Body in Static Image

Aiming at the problem of human abnormal behavior recognition in monitoring scene, a method of human abnormal behavior recognition based on neural network is proposed.

Establish the relationship between human static characteristics and behavior cognition, describe the middle characteristic behavior of human behavior, and realize the characteristic discrimination of human abnormal behavior [7]. In order to ensure the effect of feature recognition, firstly, the feature parameters of human abnormal behavior in static images are collected, and the acquisition steps are optimized, as shown in Fig. 1
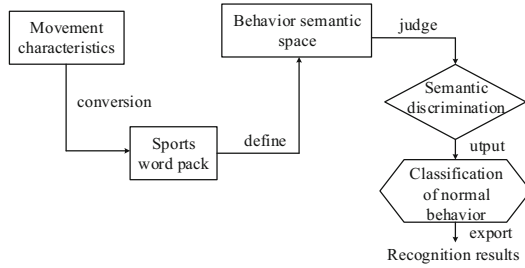
Fig. 1. Human behavior characteristics collection process of static image

In the process of feature recognition, it is necessary to use time-space interest points as low-level static features. The same kind of human behavior has similar concept distribution. The static distribution structure can be established to divide the human behavior categories and judge the specific behavior feature categories [8]. In the process of feature collection, in order to ensure the invariance of behavior features, better reflect the static information, and realize the comprehensive complementary of intelligent monitoring system features and global features, it is the key to realize the recognition of human abnormal behavior in static images. Therefore, it is necessary to further optimize the feature data processing steps based on neural network theory [9]. The platform of human behavior action capture is composed of a reference module of attitude heading and a computer, which can capture the body, upper and lower arms and legs. When the platform starts to work, the attitude and heading reference module uploads the attitude data of human action capture at 30 Hz frequency through the wireless network. The data is received by the computer, and the data sent by the attitude and heading reference module is received by the address query mode [6]. In order to reduce the volume and electromagnetic interference, it is necessary to enhance the platform independence and use the extended neural network for real-time data acquisition. After de-noising processing, data fusion is carried out to capture attitude angle, angular velocity and acceleration in real time. Then, the data is sent to the upper computer through the wireless network module, and the data is processed by the upper computer program, which is transformed into the signal form [7].

The neural network is used to collect the original data, and the relative attitude angle is obtained through the conversion relationship between Euler angle and quaternion [8]. The absolute attitude angle can be calculated by collecting the acceleration and using the direction cosine transformation of the magnetic field between the geographic coordinate system and the neural network coordinate system. When using the neural network to calculate the attitude, the data is divergent and cannot be used for a long time. However,

the accelerometer is vulnerable to vibration and its dynamic performance is poor. Therefore, it is necessary to fuse the data generated by the accelerometer to obtain a stable and accurate attitude angle. According to the vibration intensity, the accelerometer filter gain is updated in real time to ensure the dynamic response characteristics of the attitude and heading reference module and eliminate the steady-state error quickly. According to the three-axis acceleration value processed by neural network, the neural network is used to calculate the tilt angle roll and pitch angle pitch, and the yaw angle yaw is calculated according to the tilt compensation magnetometer value. According to the state vector, the high-precision attitude angle is calculated by using Euler angle conversion matrix. Through the tilt angle can capture human walking, running behavior signal, pitch angle can capture human static behavior signal, yaw angle can capture human jumping, squatting behavior signal. In fact, the complexity of the collected images is much greater than that of the data set used in this paper, including the influence of weather, light, angle of view and occlusion. It is a challenging task to mine practical multi person abnormal behavior recognition methods, which can better reduce the complexity of the algorithm, improve the running speed and so on. The processing of human abnormal behavior features in static images of intelligent monitoring system generally includes the following processes: static target detection, target tracking, feature selection and static representation, behavior recognition and understanding. Figure 2 is the whole process of human abnormal behavior processing in the static image of an intelligent monitoring system.
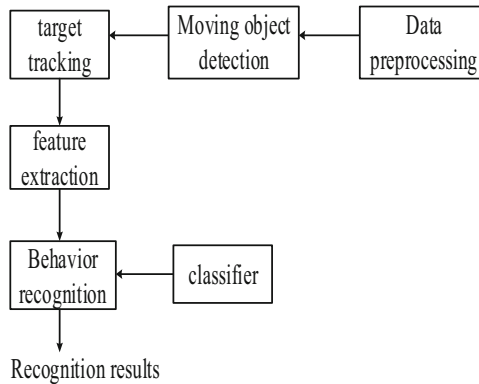
Fig. 2. Steps of human abnormal behavior processing in static image

In the process of human abnormal behavior recognition in static images, there is diversity due to the change of video content. It is necessary to capture the video behavior features based on the collected results and abstract them into data to provide further description of intermediate behavior features for video description [10]. Because in the process of human abnormal behavior recognition in static images, the recognition content of human behavior features is combined in an unknowable way, so it is necessary to use neural network for human behavior derivation and recognition [11]. Aiming at the problem of human complex behavior modeling with a long time span and a variety of basic actions, a collaborative clustering algorithm is used to model the correlation

between static words and human actions synchronously. At the same time, neural network is used to model the temporal correlation of human behavior, so as to improve the recognition accuracy of human complex behavior.

2.2 Optimization of Human Abnormal Behavior Recognition Algorithm in Static Image

Static object detection is to extract the change region from the background in the video image sequence, which is the basic and key part of surveillance video analysis, because the quality of static object segmentation directly affects the subsequent processing. In the process of image recognition, the image background will be interfered by the external environment, so the static target extraction becomes a more difficult work [12]. In practical applications, in order to better adapt to the static target detection algorithm of various environmental changes, combined with inter frame difference method, optical flow method, background subtraction method for denoising. By separating the changing region from the background image, the static target region and the background region can be segmented correctly. Multi path recognition of human motion features mainly involves image feature analysis, processing, feature extraction and pattern recognition. The general recognition process is shown in Fig. 3.
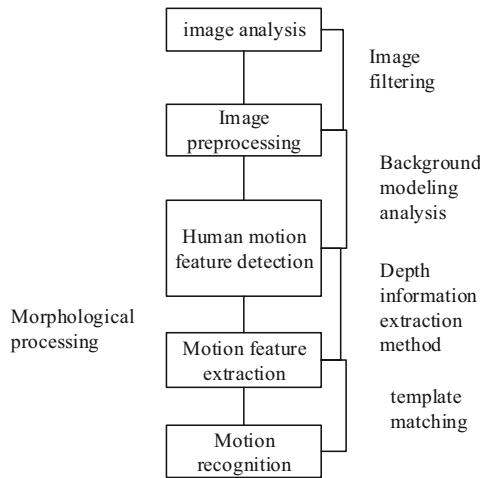
Fig. 3. General recognition process of human body features

Before motion feature detection, it is necessary to preprocess the acquired image information to filter the noise and provide accurate data for subsequent recognition. The human motion feature detection step is set to extract the related feature points from the complex human motion region. Under the occlusion condition, the clear human image can be obtained through morphological processing. In order to judge which type of motion the current motion belongs to, the appropriate human body representation method is selected and the useful motion information is extracted from the human body

sequence. Starting from the general recognition process, the multi-path recognition of human motion characteristics is carried out by using depth information.

In the process of abnormal behavior recognition, we need to introduce weighting coefficient to segment background and static target effectively. Chaotic particle swarm optimization is used to improve the iterative speed of image processing. Based on the commonly used description of posture and static features, the static features of the edge and limb shape of the video image are collected and judged to achieve de-noising [13, 14]. Combined with the neural network algorithm, the time-space feature of the image is detected, and the $I(x, y)$ structure coordinate of the image to be recognized is transformed into the image feature sequence $I(x, y, t)$. In order to detect features from image sequence $I(x, y, t)$, a separable linear filter is applied to construct the following response function R for a single fixed camera video

$$R = (I(x, y, t) \times h_{ev})^2 + (I \times g(x, y, \sigma) \times h_{od})^2 \tag{1}$$

Among formula (1):

$$g(x, y; \sigma) = \frac{e^{(-(x^2+y^2)/2\sigma_t^2 - t^2/2\tau_t^2)}}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}} - 1 \tag{2}$$

Then the smoothing kernel of the static image is as follows:

$$h_{ev}(t; \tau, \omega) = -R\cos(2\pi t\omega)e^{t^2/\mathrm{r}^2} \tag{3}$$

$$h_{od}(t; \tau, \omega) = R - \sin(2\pi t\omega)e^{t^2/\mathrm{r}^2} \tag{4}$$

If the feature weight of the static image is ω, e^{t^2/r^2} is the loss vector under the t-interference environment. τ is a still image smoothing coefficient, which has a great influence on the mesh performance. When the change of τ value is small, the error rate of human abnormal behavior recognition in static image is almost constant. Based on this, in the process of human abnormal behavior recognition in static images, when the sample data s is input as the original feature data of the model, the deviation value is B. Furthermore, the activation function is introduced based on the nonlinear calculation principle, and the normalized sample C is obtained by reducing the dimension of the merging layer.

$$C = Bsh_{od}(t; \tau, \omega) \cdot Xh_{ev}(t; \tau, \omega) \tag{5}$$

Assuming that the feature image sample to be recognized is p, the sample matrix is regularized. And record it in the feature image D. Then the standard sample vector and the standard learning sample vector are calculated

$$E = p \sum_{1}^{k} |D - CK|^2 \tag{6}$$

In formula (6), E represents the Euclidean distance between the sample to be classified and the learning sample, and the number of all kinds of samples is the same, which

is k. The Gaussian function of smoothing factor E = 0.5 is selected as the activation function to obtain the initial probability formula.

$$\Delta P = e^{-D/2a^2}/2VE \tag{7}$$

In the above algorithm, M is a video sequence with m, and the sample set is V, and it is recorded as $V = \{v_1, v_2, ..., v_M\}$. Using the difference between the current two frames in the video image sequence or between the current frame and a previous frame to extract the static region in the image for denoising. The principle is shown in Fig. 4.
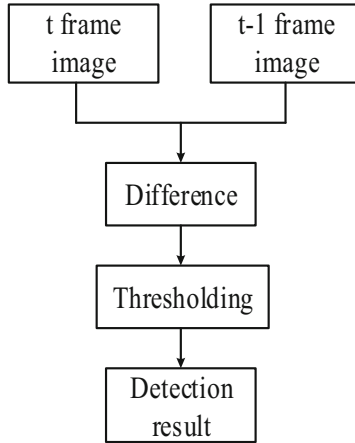
Fig. 4. Denoising of characteristic parameters of inter frame difference

Inter frame difference is the difference operation of the corresponding pixels between two frames. When the illumination condition in the scene is relatively stable, the difference value of the background part in the image sequence is zero. There is noise in the actual scene, and the difference between the regions of non-static objects may not be zero, which will affect the detection results of static objects. Based on this, we need to thresholding the difference results. By thresholding the detection results, the binary image with only static target is obtained, which can effectively suppress the noise interference. The background model is used to update the background image, and then the current frame and the background image are used to get the static target detection technology. The background subtraction method is also a special inter frame difference method. This method can get a more complete static target with fast operation speed, and can adapt to the real-time requirements of the system. When using this method to detect static targets, it is very important to get a more stable background model. Based on this, the background subtraction method is further optimized, and the specific steps are shown in Fig. 5.

```
┌─────────────────────┐
│   Video sequence    │
└─────────────────────┘
          │
          ▼
┌─────────────────────┐
│     Background      │
│     subtraction     │
└─────────────────────┘
          │
          ▼
┌─────────────────────┐
│    Thresholding     │
└─────────────────────┘
          │
          ▼
┌─────────────────────┐
│   Moving target     │
└─────────────────────┘
```
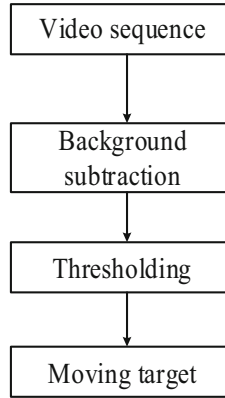
Fig. 5. Background subtraction processing steps

The background subtraction method is used to directly compare the gray value of the background image and the pixel of the static object entering the background image. Once the difference between the gray value of the pixel of the two images is greater than a set threshold, it indicates that there is an abnormal phenomenon. The mathematical formula can be expressed as follows:

$$F(x, y) = \begin{cases} 1, & |f_1(x, y) - f_2(x, y)| > S \\ 0 & other \end{cases} \quad (8)$$

In formula (8), $f_1(x, y)$ is the background image, $f_2(x, y)$ is the background image when the abnormal phenomenon occurs, $F(x, y)$ is the binary image on (x, y), and S is the initial threshold. A window with 512 sample points is used to segment the original signal, and its rectangular window is used to segment the original signal. After windowing, a single signal contains 1024 samples. If the sampling frequency of the signal is 100 Hz, the time span of the signal after windowing is 10.24 s, which is enough to complete a single action. The acceleration signal is processed by feature extraction to obtain an original feature matrix, which is trained to recognize the behavior signal.

The windowing method is used to process the processed human behavior signal. The sensor signal is extracted from the overlapped half window factory rectangular window. After the window factory is appropriate, a variety of features are extracted from a single window factory signal to form a feature vector to represent human behavior. Standard deviation is one of the statistical characteristics, which effectively reflects the degree of data dispersion of sensors. Because people are in static state, acceleration data are basically unchanged, and the standard value is 0. The acceleration data of people are constantly changing during the course of motion, and the standard value is greater than 0. The specific formula is:

$$a = \sqrt{\frac{1}{SF(x, y)} \sum_{i=1}^{S} (A_i - \overline{A})^2} \quad (9)$$

In formula (9), S is the number of samples and $\overline{A}$ is the average value of samples. According to the formula, the static and dynamic behaviors of human body can be distinguished. Skewness can measure the skew direction of acceleration sensor data distribution, and analyze the gravity direction diagram. The specific formula (10) is as follows.

$$P = \frac{S \sum_{i=1}^{s} (A_i - \overline{A})^3}{(S-1)(S-2)a^3} \tag{10}$$

According to formula (10), skewness can effectively distinguish the human body jumping and squatting behavior. Kurtosis change curve can directly reflect the change degree of all signals at the peak of data curve, which is an important statistical feature. The specific calculation formula (11) is as follows.

$$Q = \frac{\sum_{i=1}^{s} (A_i - \overline{A})^4 n_i}{Sa^4} \tag{11}$$

In formula (11), n_i is the sample interval. The formula can effectively distinguish human running behavior. The correlation coefficient can be used to measure the degree of linear correlation between variables. The specific calculation formula (12) is as follows.

$$W_{AB} = \frac{\sum_{i=1}^{s} (A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_{i=1}^{s} (A_i - \overline{A})^2} \sqrt{\sum_{i=1}^{s} (B_i - \overline{B})^2}} \tag{12}$$

In formula (12), A and B represent correlation coefficient variables respectively. According to the formula, combined with the correlation coefficient of gravity YZ, we can effectively distinguish between human walking and going upstairs. When the static target stays in a certain position for a long time, the pixels of the static target have a higher probability of occurrence in that position. Based on this, in the process of neural network updating, the target pixels will get a larger weight, so the target pixels will be misjudged as background pixels, and then the static target will be extracted. In this way, the static target will be missed and the hole will be formed. So the extraction of static target is not complete. Therefore, we need to use a morphological processing method to deal with the whole static target. The common algorithms in image processing include expansion, corrosion, open operation and close operation. The language of mathematical morphology is set theory, which uses special structural elements to process images. Common structural elements include 4-point neighborhood and 8-point neighborhood of 3×3. The definition of 3×3 neighborhood is shown in Fig. 6.
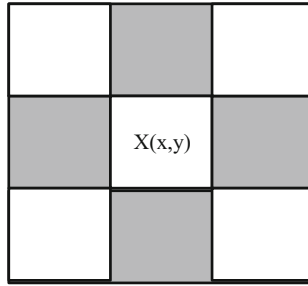
(a) 4-domain
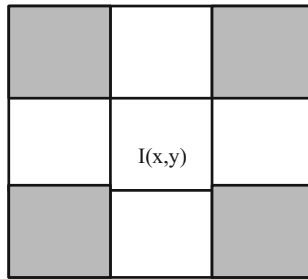
(b) \$field

Fig. 6. Definition of neighborhood background features of surveillance video

The main reasons for the change are the entry or exit of the monitoring target in the monitoring scene, the change of light and the influence of noise. In order to solve this problem, the initial threshold S is used as the threshold to distinguish the segmented object from the actual background. Set the pixel greater than or equal to the threshold as the object, and the pixel less than the threshold as the background. Based on this method, the feature data of static image can be extracted effectively, and the interference error in the process of image recognition can be eliminated effectively.

2.3   Realization of Abnormal Recognition of Static Portrait in Monitoring System

Static target detection and tracking is the research foundation of target behavior analysis and recognition in intelligent video surveillance system. Static object detection and tracking is to extract the static object in video sequence by image processing technology. Then, according to the characteristics of the static target, the static trajectory of the target is obtained. Finally, the behavior of the target can be analyzed and recognized according to the trajectory change and static characteristics of the target. Static object detection is the bottom module of computer vision, which aims to separate the static object from the background image. The effect of static detection will directly affect the accuracy of the following target tracking and behavior analysis. According to the current situation, it is very important to adopt the appropriate static target detection algorithm. The new video frame X_t and the B Gaussian distribution of the background are matched one by one. If

X$_t$ matches one of the B Gaussian distributions, it means that the pixel is the background, otherwise it is the foreground. Generally, the gray value of background pixel is assigned as 0, and the gray value of foreground pixel is assigned as 1. In this way, the static object can be extracted from the background. Based on this, the algorithm flow structure of neural network is optimized, as shown in Fig. 7.
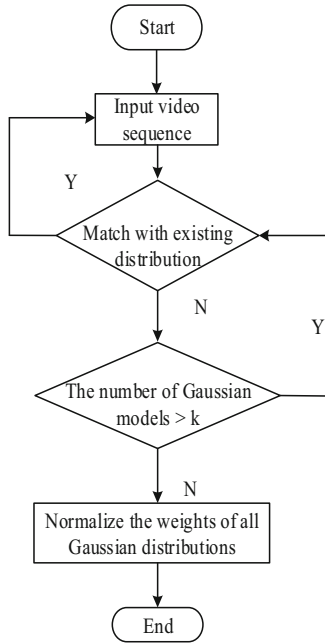
Fig. 7. Flow chart of portrait monitoring algorithm based on Neural Network

In the neural network, it is assumed that the probability of background pixels is high. Therefore, the Gaussian distribution with the highest probability of a pixel is used to represent the background model of the pixel. Human static object tracking is to find the position of the static human body in the current frame according to the position of the human body in the previous frame, and match the static human body in the two frames. If it can match, it means that the static human body in the two frames belongs to the same person. At present, the main reason for the slow progress of research on multi person abnormal behavior detection system is the limitation of universality and robustness. It is very difficult to obtain the behavior characteristics of multiple people and identify and classify them into complex scenes. The recognition method of human abnormal behavior in static image of intelligent monitoring system based on neural network algorithm mainly improves the speed and accuracy of the system from three aspects. The filter channel feature detector is used to calibrate the position of all human bodies in the image. It can reduce the influence of background on pose estimation and determine the position of some parts of the human body quickly and accurately. Through the neural network algorithm to obtain human static features and classification,

the accuracy of classification is guaranteed and the running speed is improved. In order to ensure the recognition effect of human abnormal behavior, region tracking method and active contour tracking method are used to recognize human abnormal static feature multi-path. The specific identification process is shown in Fig. 8.

The behavior characteristic signal acquired by the algorithm is analyzed
Signal space initialization

↓

Collecting discrete data
Form four-dimensional data

↓

The extended Kalman filter can be used to process the data space vector value

↓

Noise treatment

↓

Optimization estimation results

↓

Human behavior recognition based on trust level
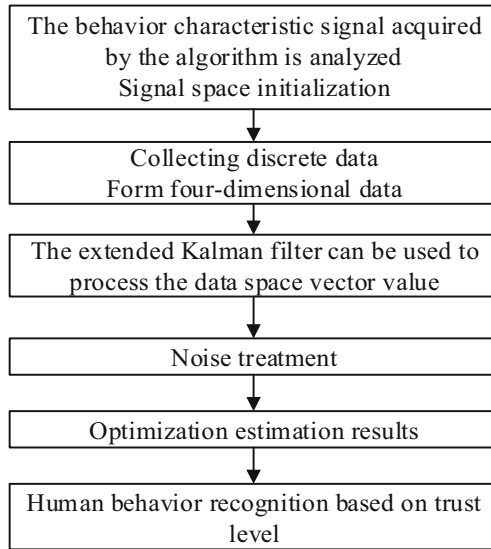
Fig. 8. Specific recognition process of portrait features

According to the identification process, the trust level of each component of the data space can be set. The optimal characteristic data is obtained by drying each grade. So as to better protect the effect of abnormal behavior recognition and improve the accuracy of human abnormal behavior recognition.

3   Analysis of Experimental Results

In order to verify the practical application effect of human abnormal behavior recognition in the static image of intelligent monitoring system based on neural network algorithm, the experimental detection is carried out. The experimental process is generated by the model and represented by behavior collection and display. Finally, the method described in the previous section is used to detect abnormal behaviors online and classify normal behaviors in real time. In this experiment, we use matlabr2019a platform to detect static objects in a video with two static human bodies in the behavior database of the Institute of automation, Chinese Academy of Sciences. Using 2.0 GHz 7 64 bit processor, 8 GB ram random access memory, Dx11 model graphics adapter. The experiment uses a host computer with the parameter configuration shown in Table 1.

Table 1. Experimental host configuration

Parameter	Index
built-in	USB3.0 bus
Graphics card	NAIDIA
processor	64 bit
CPU	IntelCore

Under the condition of Microsoft Kinect camera, the depth sensor is used to establish the experimental database. The specific data are shown in Table 2.

Table 2. Experimental data

Database	Resolving power	Remarks
MSR	320 * 2240	15fps; TXT file to the clearance node
RGBD	320 * 240	Indoor daily exercise; single person; whole body
LIRIS database	640 * 480	Indoor office action or interaction; single person or whole body
Chalearndatabase	640 * 480	16 daily behaviors; 15fps; TXT file joint points
Kitchendatabase	640 * 480	Camera installation; sports oriented
ReadingAct database	320 * 240	19 movements; 30 FPS; two cameras

In order to verify the practical application effect of this method, experimental detection is carried out. In order to ensure the accuracy of the experiment, the evaluation index of each behavior is standardized, and the detection accuracy, accuracy and recall rate of the four abnormal behaviors under this method are given. The specific results are shown in Table 3.

Table 3. Evaluation index of experimental results

This paper	Accuracy	Recall	Accuracy
Slap	0.900	0.888	0.932
Robbery	0.903	0.932	0.821
Shooting	0.903	0.865	0.923
Boxing	0.909	0.881	0.912

It can be seen from Table 3 that the method in this paper has a good ability to recognize abnormal human behaviors in four different types of static images. In terms of

abnormal feature recognition, the average recognition rate of the system in this paper is 90%, the image recall rate for abnormal behaviors is 87%, and the overall average effect of the system is 89%. This shows that the system in this paper has a good recognition rate for abnormal behavior detection. The result of human behavior is conducive to the stability of the detection system, so the static behavior data of the human body is collected according to the time step. Through the data measurement and analysis of the standard deviation, skewness, peak value and correlation coefficient of the X, Y, and Z axes. The results of the study are shown in Fig. 9.
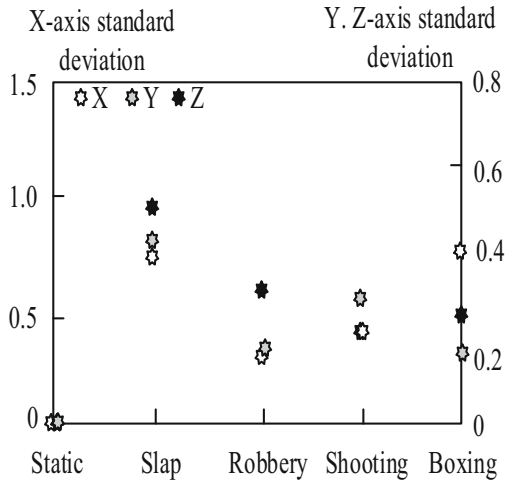
Fig. 9. Human behavior capture results

When human behavior is at rest, the standard deviation of X, y and Z axes is 0. When the human body is walking, the standard deviations of X, y and Z axes are 0.7, 0.45 and 0.5 respectively. When the human body is running, the standard deviations of X, y and Z axes are 0.3, 0.22 and 0.35 respectively. The standard deviations of X, y and Z axes are 0.4, 0.3 and 0.25 respectively when the human body is in the state of jumping up. When the human body is squatting, the standard deviations of X, y and Z axes are 0.8, 0.2 and 0.3 respectively. When the human body is in a static state, the component interval value does not change, so the study of human static abnormal behavior. By comparing the value range of various human behaviors, it can be seen that the traditional recognition method will mistakenly recognize the running state after walking faster. When the running speed is increased, it will be mistakenly identified as jumping state, if the jumping range is small, it will be mistakenly identified as squatting state, and the experimental results are inconsistent with the behavior habits.

The selected detector based on channel filtering features can effectively and accurately mark the position of each human body in the image, which can provide a strong guarantee for the subsequent pose estimation and improve the success rate of behavior recognition. In order to show the practical application effect of this method more clearly, the experiment was carried out by comparing it with the traditional recognition method, and 2–10 people's behavior images were randomly identified and tested. The experimental time is controlled in the range of 928 s. On this basis, a static real image is randomly selected to detect the recognition results. The specific detection results are shown in Fig. 10.
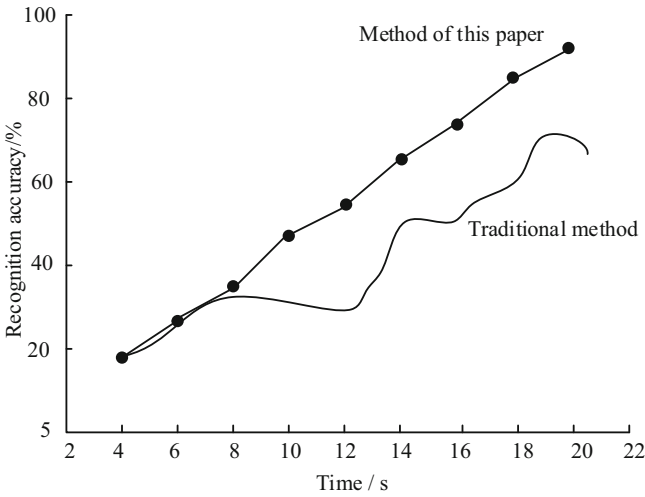
Fig. 10. Comparison test results

According to the test results in Fig. 10, the comparative test results of the traditional method and the method in this paper can be obtained. We can see that both the traditional method and the method in this paper have good recognition results in the range of less than 10 s, but as time continues to increase, the traditional recognition ability has eased, and the recognition accuracy is maintained at about 60%. However, with the increase of time in this paper, the recognition accuracy reaches 98%. It can be concluded that, compared with the traditional method, the human abnormal behavior recognition method based on neural network algorithm in this paper has a higher conversion rate in the actual application process, and can better meet the static human abnormal recognition target. In addition, the delays of abnormal behavior recognition under different methods are compared and recorded. The specific test results are shown in Table 4.

Table 4. Contrast detection of recognition delay between the two systems

Traditional recognition system				This paper introduces the recognition system			
Image acquisition	dynamic tracking	Dynamic identification	Accuracy%	Image acquisition	dynamic tracking	Dynamic identification	Accuracy%
—	S	S	59	—	S	S	69
S	—	S	86	S	—	S	94
S	S	—	70	S	S	—	91
—	—	S	55	—	—	S	79
—	S	—	60	—	S	—	71
S	—	—	40	S	—	—	52
S	S	S	100	S	S	S	100

From Table 4, we can get the image acquisition effect, image tracking ability, the traditional method under the dynamic recognition effect and the image recognition delay of the method in this paper. S represents the normal state of recognition. Through comparative analysis, it is found that there is a delay in the image recognition process. The image recognition delay rate of traditional methods is about 70% on average. In the static image of the intelligent surveillance system proposed in this paper, the neural network algorithm is used to recognize abnormal human behaviors. The method has high accuracy in the actual application process, the recognition accuracy rate is 89%, and the delay performance is 90%. Based on the above experimental results, the overall performance of the system algorithm in this paper is good, and the running speed has also been improved to a certain extent, which can better meet the requirements of abnormal behavior research on static human images.

4 Conclusion

The neural network algorithm can play a guiding role in the recognition of abnormal human behaviors in static images, and can recognize the behaviors of multiple humans on a single frame of images, and provide a multi-category classification method of abnormal human behaviors. Automatically learn human behavior models through video training sets without any manual annotations, and can apply the learned human behavior models to detect abnormal behaviors and recognize normal behaviors in real time. This paper builds the appearance model of the human structure feature framework, collects the abnormal behavior characteristics of the human body in static images, performs denoising processing according to the characteristic data, and uses the chaotic particle swarm search method to improve the iterative speed of image processing. Use neural network algorithm features to analyze the trajectory changes and static characteristics

of the target object, analyze and recognize the behavior of the target, and optimize and upgrade the recognition effect of abnormal human behavior in the static image according to the actual parameter results obtained, and improve the system design. This research is based on the recognition of abnormal human behavior in static images of the intelligent monitoring system based on neural network algorithm, which overcomes the shortcomings of accuracy and efficiency of traditional methods in the process of modeling complex human behavior. It alleviates the behavior type ambiguity caused by the lack of sufficient visual evidence in the real-time behavior recognition process of the system, so that the system can better complete the tasks of real-time abnormal behavior detection in the monitoring scene and online normal behavior recognition. It is conducive to optimizing the system update and upgrading, effectively improving the response speed of the system to recognize human behavior, reducing system memory usage, and improving the overall recognition effect of the system. In the actual application process of the system in this paper, the detection and recognition rate is 89%, the accuracy is 90%, and the delay is low, which shows that the method in this paper has higher accuracy, better stability, and more obvious effect.

References

1. Gorban, A.N., Mirkes, E.M., Tukin, I.Y.: How deep should be the depth of convolutional neural networks: a backyard dog case study. Cogn. Comput. **12**(1), 388–397 (2020)
2. Chen, S., Shen, Y., Yan, Y., et al.: Cholesky decomposition based metric learning for video-based human action recognition. IEEE Access (99), 1 (2020)
3. Jun, J., Zhuojun, Z., Mingliang, G., et al.: An abnormal crowd behavior detection method based on streak flow CNN.Adv. Eng. Sci. **52**(6), 215–222 (2020)
4. Songwei, W., Jun, L., Changjun, F., et al.: Design and implementation of video-based abnormal behavior detection system for criminal investigation. Comput. Era (9), 67–71,75 (2020)
5. Xin, Z., Hua, Q.: Research on human abnormal behavior detection algorithm based on YOLOv4. Comput. Digit. Eng. **49**(4), 791–796 (2021)
6. Al-Obaidi, S., Al-Khafaji, H., Abhayaratne, C.: Modeling temporal visual salience for human action recognition enabled visual anonymity preservation. IEEE Access **8**(9), 213806–213824 (2020)
7. Dhiman, C., Vishwakarma, D.K.: View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. IEEE Trans. Image Process. (99), 1 (2020)
8. Wang, Z., Li, N., Wu, T., et al.: Simulation of human ear recognition sound direction based on convolutional neural network. J. Intell. Syst. **30**(1), 209–223 (2020)
9. Veni, S., Anand, R., Mohan, D., et al.: Feature fusion in multimodal emotion recognition system for enhancement of human-machine interaction. IOP Conf. Ser. Mater. Sci. Eng. **1084**(1), 012004 (10p) (2021)

10. Liu, S., Bai, W., Liu, G., et al.: Parallel fractal compression method for big video data. Complexity **2018**, 2016976 (2018)
11. Liu, S., Fu, W., He, L., Zhou, J., Ma, M.: Distribution of primary additional errors in fractal encoding method. Multimedia Tools Appl. **76**(4), 5787–5802 (2014). https://doi.org/10.1007/s11042-014-2408-1
12. Liu, S., Ma, X., Wu, H., et al.: An end to end framework with adaptive spatio-temporal attention module for human action recognition. IEEE Access (99), 1 (2020)
13. Jegham, I., Khalifa, A.B., Alouani, I., et al.: Vision-based human action recognition: An overview and real world challenges. Digit. Investig. **32**(7), 200901 (2020)
14. Liu, S., Pan, Z., Cheng, X.: A novel fast fractal image compression method based on distance clustering in high dimensional sphere surface. Fractals **25**(4), 1740004 (2017)

Fall Behavior Recognition Algorithm in Video Surveillance Based on Feature and Deep Learning

Hai-jing Zhou(✉)

Chongqing Vocational Institute of Tourism, Chongqing 409099, China
`zhouhaijing554@yeah.net`

Abstract. Aiming at the problem of low recognition rate and accuracy rate of human fall behavior recognition algorithm in current video surveillance, the human behavior feature extraction module is relatively backward. To solve this problem, a fall behavior recognition algorithm based on feature and deep learning is designed. The video image preprocessing is completed by dilation and erosion. The covariance matrix of image features is constructed to extract the features of human fall behavior. The standard image database is constructed, and the deep learning algorithm and neural network are used to complete the human fall behavior recognition in video surveillance. So far, the design of human fall behavior recognition algorithm in video surveillance based on feature and deep learning is completed. Through the example test, the application effect of the feature algorithm is better than that of the traditional algorithm.

Keywords: Deep learning · Feature learning · Behavior recognition · Intelligent video

1 Introduction

With the rapid development of economy at home and abroad, people's life has undergone earth shaking changes, and the process of urbanization has accelerated significantly. However, a series of problems also follow. With the rapid growth of urban population, the number and proportion of left behind elderly and left behind children are increasing, and the health and safety problems of the elderly are becoming more and more prominent. Accidents or deaths occur in the elderly due to falls every year. In addition, with the increase of urban population, the urban environment is becoming more and more complex, and sudden violent incidents such as theft, mob fighting and terrorist attacks often occur, which has brought serious interference to social public order and brought great losses to the people. Governments of all countries pay more attention to the prevention of public security. As an important prevention tool, video surveillance has been widely used in public security, transportation, banking and other important departments [1]. At present, intelligent monitoring is mostly used in the family, through monitoring the activities of children and the elderly to carry out management, to avoid

the safety problems caused by physical discomfort falls. Through literature research, it can be seen that the analysis of abnormal behavior of moving human body is based on intelligent video monitoring, using computer vision technology, taking networking and intelligence as the development direction, to analyze and identify the human target behavior in the video sequence, so as to deal with and alarm the abnormal behavior in time. Greatly improve the ability of video surveillance, and save human and material resources in the monitoring work, at the same time, intelligent analysis can reduce the probability of missing and false positives [2].

At present, there are a lot of problems in human fall behavior recognition in video surveillance, which lead to abnormal recognition results and can not achieve the expected effect. In view of the drawbacks of human fall behavior recognition in traditional video surveillance, it is necessary to realize intelligent video surveillance to meet the needs of society and realize computer simulation of human brain. Realize the automatic analysis and understanding of the video monitoring screen, and timely alarm for the abnormal events, so as to facilitate the staff to deal with them in time, so as to minimize the loss [3]. In recent years, many experts and scholars have done in-depth research on the analysis and understanding of video information, and intelligent video monitoring technology has been further developed. Video understanding is to use computer to process video data and obtain effective information. There are two key elements of people and objects in video, and people as one of the objects of concern makes the behavior recognition of people in video surveillance become the focus of many application fields [4]. Abnormal behavior recognition is one of the specific subclasses, which is directly related to people's interests, so it is of greater significance for the study of human abnormal behavior. Therefore, in this study, feature learning and deep learning are applied to the human fall behavior recognition algorithm in traditional video surveillance, and the human fall behavior recognition algorithm in video surveillance based on feature and deep learning is designed. In order to obtain better recognition results based on the traditional recognition algorithm, and provide technical support for the development of video surveillance.

2   Design of Human Fall Behavior Recognition Algorithm in Video Surveillance

Before the beginning of this study, the problems existing in the use of traditional fall behavior recognition algorithm in video surveillance are analyzed. The deep learning algorithm and feature learning algorithm are applied to this study. The process of human fall behavior recognition in optimized video surveillance is shown in Fig. 1.

According to Fig. 1, the research content is set as follows:

This paper first introduces the research background and significance of this topic, and then introduces the research status in this field at home and abroad. And the existing methods are analyzed, and the research content and organization structure of this paper are introduced.
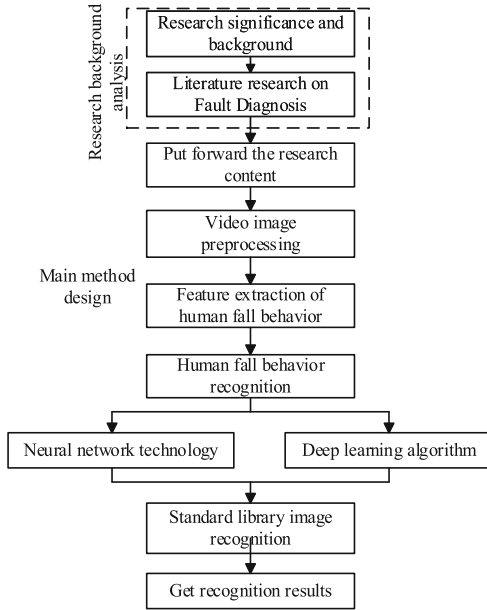
Fig. 1. Construction process of human fall behavior recognition algorithm in video surveillance

Related theory and technology. Firstly, the human behavior recognition algorithm is introduced to lay a theoretical foundation for the following research work. After that, it introduces the theory of deep learning and feature learning in detail, and then introduces the commonly used classification algorithms and deep learning tools.

Human behavior recognition. The calculation of characteristic covariance matrix is studied, and the covariance matrix is transformed into log Euclidean space vector. The bag of words model is improved for human behavior modeling, and the traditional framework is replaced by spectral clustering dictionary learning and local constraint linear coding. The spatial pyramid model is combined to pool features, and nonlinear support vector machine is used to classify human behavior.

Experimental demonstration and analysis. The corresponding example is set to verify the algorithm proposed in this paper, and this method is compared with the traditional method to determine the difference of the use effect of different methods.

Summary and prospect. The paper summarizes the work done in this paper, analyzes the shortcomings of the current algorithm, and puts forward the corresponding improvement direction.

2.1 Video Image Preprocessing

Digital image processing is the use of digital processing technology, image processing methods and technologies, such as noise removal, image enhancement, image restoration, image segmentation and so on [5]. In today's society, with the rapid development of information technology, the computer processing ability has been greatly improved, and the development of mathematics is also very rapid. The application demand of people in various fields of production and life is also growing. As the most intuitive way

to obtain and express information, image is still the focus of people's research. These factors promote the rapid development of digital image processing technology. In this study, we will mainly use morphology to preprocess the human body image in the video. According to the research goal, we will use expansion operation and corrosion operation to complete the preprocessing process.

Dilation operation is a processing method that combines the background pixels connected with the known target region into the target region by using structural elements, so that the boundary of the target region expands to the outside. The algorithm can fill the holes in the target area and eliminate the noise in the target area. Set the monitoring image set as X and the structural element as Y. mathematical morphology image processing uses Y to process X. If Y is used to dilate X, denoted as $X \oplus Y$ and $\oplus$ as dilation operator, then the formula of dilation operation (1) is as follows:

$$X \oplus Y = \left\{ a | (Y)_{ab} \cap A \neq \emptyset \right\} \tag{1}$$

The formula (1) shows that the expansion process is to first map the structural element Y itself to the origin, and then translate a. Using Y to expand X is the sum operation of all Y and X after translation a. the operation result is not an empty set, that is, there is at least one non-zero common element. According to this explanation, formula (1) can also be written in the form of formula (2)

$$X \oplus Y = \left| \left\{ a | (\overline{Y}) \cap (\overline{X}) \neq \emptyset \right\} \right. \tag{2}$$

Using formula (2), the expansion operation of video image is completed. Then, the image erosion operation is completed. The function of erosion operation is just opposite to that of expansion operation. It is a processing method that uses structural elements to eliminate some boundary points of the target area and make the boundary of the image target area shrink inward. This process can eliminate small and meaningless isolated noise points and segment two small connected objects. If Y is used to corrode X, denoted as $X \cdot$, and $\cdot$ is denoted as corrosion operator, then the formula of corrosion operation is as follows:

$$X \cdot Y = \{ a | (Y)_a \subseteq X \} \tag{3}$$

Formula (3) shows that the result of corrosion process can be interpreted as all the points a contained in X after the structural element Y is translated into a. After image preprocessing, the acquired image is enhanced. In this study, the representative smoothing filtering method of spatial domain enhancement method is mainly used [6]. The basic idea of smoothing filtering method is to weaken or eliminate the high-frequency component of the image, enhance the low-frequency component of the image, so as to remove the random noise in the image and smooth the image. According to the research results, the neighborhood average filtering method is selected to complete this part of processing.

Neighborhood average filtering method is also called mean filtering method, and its implementation process is: first, average a pixel and all pixels in its neighborhood. Then the average value is used to replace the value of the corresponding pixel in the output image to achieve smooth filtering. Set the original image to the 5×5 template type.

Using this model, the calculation process is completed. The specific formula (4) is as follows:

$$s(x, y) = \frac{1}{N} \sum_{k \in D} c(i, j) \tag{4}$$

In formula (4), it is assumed that D is the set of neighborhood pixels with (x, y) as the center, and N is the total number of pixels in the neighborhood. Each pixel is calculated according to formula (4). And the obtained gray value is assigned to the corresponding pixel to enhance the image. The above processing content is integrated to complete the video image preprocessing process.

2.2 Feature Extraction of Human Fall Behavior

In digital image processing, the covariance matrix can represent the image regional features, such as the regional covariance matrix [7]. It can also represent video sequences, such as spatiotemporal 3D feature covariance matrix. This paper mainly studies the application of spatiotemporal 3D feature covariance matrix in human behavior recognition algorithm. Suppose that $A(x, y, t)$ represents a continuous video sequence or a local region in the video sequence, (x, y, t) is the horizontal, vertical and temporal coordinates of the pixel, and $M := \{(x, y, t)^T x \in [1, O], y \in [1, R], t \in [1, T]\}$ pixel region is defined as. Where O is the region pixel width, R is the region pixel height, T is the region frame length, and $|M|$ is the total number of pixels in region M. The eigenvector of the pixel (x, y, t) is $j(x, y, t)$, let the covariance $W_M : \mathrm{cov}(J)$, where $J := j(M)$, its eigencovariance matrix is:

$$W_M : \mathrm{cov}(J) = \frac{1}{|M| - 1} \sum (j(x, y, t)^T - \alpha_J)(j(x, y, t)^T - \alpha_J)^T \tag{5}$$

In formula (5), $\alpha_J = \sum \frac{1}{|M|-1} j(x, y, t)$ is the average of eigenvectors. If the $j(x, y, t)$ dimension of the eigenvector is s, then the W_M dimension of the covariance matrix is $s*s$ and is symmetric. The gradient optical flow feature is added to the covariance matrix to improve the accuracy of feature acquisition. Gradient is a kind of image sharpening operator, which is used to enhance the edge of image and highlight the mutation of image gray level. The edge information of moving human body can be enhanced by weakening the area with slow gray change. Gradient is a sharpening filter based on the first order differential of the image to complete the analysis of human behavior. According to this setting, the above setting is optimized, and the image gradient is as follows:

$$grad(j(x, y, t)) = [J_x(x, y, t), J_y(x, y, t)]^2 \tag{6}$$

In formula (6), $J_x(x, y, t)$ and $J_y(x, y, t)$ are the first-order differential of the image in the x, y direction, and the gray image can be smoothed by using vectors $[-2, 0, +2]$ and $[-2, 0, +2]_T$ as templates respectively. Then the gradient amplitude $E(J(x, y, t))$ and $\beta(E(J(x, y, t)))$ directions of the image are as follows:

$$\begin{cases} E(J(x, y, t)) = \sqrt{J_x^2(x, y, t) + J_y^2(x, y, t)} \\ \beta(E(J(x, y, t))) = \tan^{-1} \dfrac{J_y(x, y, t)}{J_x(x, y, t)} \end{cases} \tag{7}$$

Optical flow is the velocity distribution of the image brightness mode, which can be generated by the relative motion between the target and the observer, so it can represent the spatial distribution of the target and the change of the corresponding distribution. In the past few decades, hundreds of scholars have been studying the calculation of optical flow, such as Lucas kanada sparse optical flow method and horn Schunck dense optical flow method [8]. In the research, horn Schunck method is used to calculate the optical flow, where $J(x, y, t)$ is the brightness of the pixel at (x, y, t), and $(g(x, y, t), h(x, y, t))$ is the corresponding optical flow vector. Assuming that the incident light is uniform and the reflected light is smooth, the brightness of a specific pixel is fixed

$$\frac{mJ}{mt} = \frac{\chi J}{\chi x} \frac{mx}{mt} + \frac{\chi J}{\chi y} \frac{my}{mt} + \frac{\chi J}{\chi t} = 0 \tag{8}$$

Let $g = \frac{mx}{mt}$, $h = \frac{my}{mt}$, then formula (8) can be written as $J_x g + J_y h + J_t = 0$, where J_x, J_y and J_t denote partial derivatives in x, y, t directions. Using the above formula, more reliable data of human body characteristics can be obtained. Due to the diversity of human behavior recognition, the background is complex and changeable. In order to ensure good recognition accuracy, behavior features must accurately describe human behavior, try not to be affected by external factors, and maintain strong robustness. Human contour is invariant to raster, color and texture in different scenes. Therefore, the human behavior extracted from the contour must have good robustness to these changes, and can well represent the human behavior.

2.3 Human Fall Behavior Recognition

According to the above design content, in this study, the deep learning algorithm hybrid neural network technology is used to complete the human fall behavior recognition process. Character behavior may occur in different areas of the video frame. The convolution characteristic graph contains the spatial structure of the input. Therefore, this study further proposes a trajectory attention map, which can locate the moving target area in the video on the convolution feature map. Attention map is widely used to locate the target area in video. It has the same length and width as convolution feature map. Convolution feature map and attention map in each frame are corresponding to each other. At present, attention graph is trained through end-to-end network, so it needs a lot of training data. And in a very complex environment is not so accurate [9, 10].

The trajectory attention graph proposed in this study is calculated by dense trajectories and improved dense trajectories. Dense trajectories can describe people's trajectories in complex environment, so these trajectories are always distributed in the moving target area of video. In order to ensure that the recognition algorithm in this study can get high accuracy and reliability analysis results. In this study, the improved dense trajectories will be used to analyze the human motion trajectory.

Improved dense trajectories is an improved version of dense trajectories, which can overcome the influence of camera motion in video. It assumes that the relationship between two adjacent frames can be described by a transformation matrix. In order to calculate the value of the transformation matrix, it first calculates the surf feature and optical flow feature in two adjacent frames. Then the RANSAC (RANdom SAmple

Consensus) algorithm is used to match these features. So the value of transformation matrix can be calculated. The first human motion track from the frame can be expressed as:

$$\left\{e_l^t, e_l^{t+1}, ..., e_l^{t+n-1}\right\} = \left\{(x^t, y^t)(x^{t+1}, y^{t+1}), ..., (x^{t+n-1}, y^{t+n-1})\right\} \tag{9}$$

In Eq. (9), n is the length of the trajectory, and all the improved dense trajectories in frame t can be expressed as. $\left\{e_1^t, e_2^t, ..., e_l^t\right\}$. Based on the dense trajectory points calculated from each frame, the trajectory attention map can be further calculated. The size of the pixel value in the track attention graph is determined by counting the number of track points in the receptive field corresponding to each pixel in the track attention graph. The track attention graph of a video can be expressed as $C \in R^{HWT}$, where H and W are the height and width of the track attention graph, and T is the time length of the video. $b_{i,j}^t$ is the value of position (i, j) in the track attention graph corresponding to the t th frame. The value of $b_{i,j}^t$ is calculated as follows.

$$b_{i,j}^t = N_{e^t \in \partial_{i,j}^t}(e^t) \tag{10}$$

In Eq. (10), $\partial_{i,j}^t$ is the receptive field corresponding to $b_{i,j}^t$ pixel in the trajectory attention map. N is the statistical function, which can count the number of trajectory points in the receptive field. Then the track attention map is normalized by Eq. (11), and the normalized track attention map can be expressed as.

$$\overline{b_{i,j}^t} = \frac{b_{i,j}^t}{\sum_{i,j} b_{i,j}^t} \tag{11}$$

After using formula (11) to process the human trajectory, the trajectory is used to recognize the human fall behavior. In order to make better use of the information of human falling behavior in video, the result of fusion of two stream model U corresponds to the input of LSTM layer at a time to obtain the information of long-term fighting behavior. LSTM network controls which information is retained and which information is discarded by setting input gate, forgetting gate, output gate and storage unit. At a certain time t, the calculation part is also updated according to the LSTM principle. First, the "forgetting gate" will control whether the last time state f_{t-1} can be transferred to the next layer, and the specific updates are as follows:

$$l_i = \wp\left(Q_i R_i + P_i h_{n-1} + v_i\right) \tag{12}$$

The "input gate" can be represented by the following functions:

$$I_i = \wp\left(Q_j R_i + P_i h_{n-1} + v_j\right) \tag{13}$$

LSTM uses "forgetting gate" and "input gate" to change the state of storage unit, and uses "output gate" to control the connection between long-term operation information and current motion information, that is, how much storage unit state C is output to the current output value of LSTM. Finally, the final output of LSTM is obtained by using the "output gate" and the state of storage unit.

$$\overline{\varepsilon_i} = \tan h(Q_{i1} R_i + P_c h_{n-1} + v_c) \tag{14}$$

$$\varepsilon_i = l_i \oplus \varepsilon_{i-1} + I_i \oplus \overline{\varepsilon_i} \tag{15}$$

$$z = \wp\left(Q_i R_i + P_c h_{n-1} + v_c\right) \tag{16}$$

$$h_i = z \tanh(\varepsilon_i) \tag{17}$$

where $\wp$ is the activation function of sigmoid. R is the characteristic matrix of the input at time t. Q_i, Q_j, Q_{i1} and Q_{j1} represent the weight matrix from input layer to input gate, forgetting gate, memory cell and output gate respectively. P_i, P_c, P_{c1} and P_{i1} represent the weight matrix from hidden layer to input gate, forgetting gate, storage unit and output gate respectively. v_i, v_c, v_{i1}, v_{c1} respectively represent the offset values of input gate, forgetting gate, storage unit and output gate. According to the above calculation process, the corresponding standard behavior library is constructed, and the secondary behavior library is used to set the human fall behavior recognition process as follows (Fig. 2):
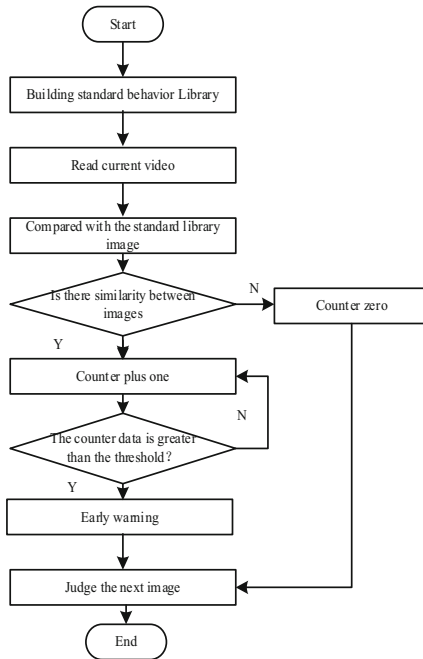
Fig. 2. Human fall behavior recognition process

This paper completes the preprocessing of the video image through the expansion operation and the erosion operation. The image feature covariance matrix was constructed to extract the fall behavior features of the human body, and then a standard image library was constructed on this basis. Use deep learning algorithms and neural networks to monitor human fall behavior. Since the number of images in the standard

library is limited, it is impossible to include all the images of standard behavior. In addition, people's fall behavior is a continuous process, so in this study, when multiple frames appear abnormal continuously. It is considered that the human body has fallen behavior, and then the alarm processing is carried out. By integrating the above settings, the design of human fall behavior recognition algorithm in video surveillance based on feature and deep learning is completed.

3 Example Analysis

3.1 Design of Test Platform

In this study, a human fall behavior recognition algorithm based on feature and deep learning is proposed to verify the effect of this method. In the test process, in order to ensure that the comparison process between this method and the traditional method 1 (reference method [2]) and traditional method 2 (reference method [3]) is not affected by external factors, the test platform is set as shown in the Table 1 shown.

Table 1. Test platform parameters

Use link	Parameter category	Parameter name	Specific model
Video image processing module	Hardware	CPU	Intel core i5 2520M
		CPU frequency	3.6 GHz
		Memory	8 GB
		Hard disk	500 GB*2
	Software	Development language	C language
		Image processing library	opencv
		Programming tools	Anaconda
Video image recognition module	Hardware	CPU	Intel core i5 2520M
		CPU frequency	3.6 GHz
		Memory	8 GB
		Hard disk	500 GB*2
	Software	Computing equipment	Python
		Operating system	Windows10

According to the content in Table 1, complete the construction of the test platform, and introduce the related content of the feature recognition algorithm and the traditional algorithm into the test platform, so as to provide a carrier for the comparative test of the two algorithms.

3.2   Test Data Set Settings

In this test, the traditional dynamic diagram is used as the experimental object. In the original dynamic image, the background area of dynamic image will appear motion blur. At the same time, the background area of the dynamic image contains a large number of vertical black edges. This is because the input video frame contains a large number of horizontal camera jitters, so the dynamic picture will mistakenly regard these background jitters as the motion of the target in the video. Thus, the motion information of the background is encoded into the dynamic image, which greatly reduces the quality of the dynamic image. Therefore, before using this part of the image, the preliminary processing is carried out, and the image types are divided according to the image types, which are used as the source of test data (Table 2).

Table 2.   Test image set

Test image set serial number	Number of test images	Contains the number of human fall images
X1	100	25
X2	150	30
X3	200	50
X4	200	120
X5	150	100
X6	100	75
X7	150	26
X8	300	34
X9	250	25
X10	200	30

The above test image data set is used as the test object in this test, and the feature recognition algorithm and traditional algorithm are used to complete the recognition process, and the test results are obtained according to the recognition results.

3.3   Test Plan Setting

In this test, we will complete the application comparison process between the feature recognition algorithm and the traditional algorithm. In order to obtain the test results with high reliability, we set the test index as three parts. They are: recognition rate of human fall behavior, recognition rate of human fall behavior and recognition response time under different clarity. According to the above test indicators, the application effect of the feature recognition algorithm and the traditional algorithm is analyzed to determine the difference between the different recognition algorithms. At the same time, the

design and development direction of the follow-up human body recognition algorithm in surveillance video are controlled according to the test results.

3.4   Analysis of Test Results

Table 3. Recognition rate of human fall behavior(%)

Test image set serial number	Feature recognition algorithm	Traditional method 1	Traditional method 2
X1	90.64	85.08	85.89
X2	90.45	86.05	85.12
X3	90.83	85.38	85.92
X4	91.87	85.89	85.07
X5	90.53	85.91	85.06
X6	90.15	85.84	85.04
X7	91.61	85.23	85.17
X8	90.64	85.12	85.49
X9	90.99	85.53	85.67
X10	91.22	85.32	85.76

The results show that the recognition effect of feature recognition algorithm and traditional algorithm is different for different types of surveillance video images. The feature recognition algorithm can recognize the preset image types effectively and has high recognition accuracy. Traditional algorithms can recognize some image types in the test with high precision, but they can not be effectively recognized for some of the more complex image types. The results of the above test can be determined that the recognition rate of the feature recognition algorithm proposed in this paper is higher and the effect of application is better (Tables 3, 4).

Table 4. Recognition rate of human falling behavior(%)

Test image set serial number	Feature recognition algorithm	Traditional method 1	Traditional method 2
X1	63.62	52.87	50.52
X2	62.51	50.17	53.41
X3	61.46	51.08	54.97
X4	64.47	54.47	51.74

(*continued*)

Table 4. (*continued*)

Test image set serial number	Feature recognition algorithm	Traditional method 1	Traditional method 2
X5	62.48	50.31	50.76
X6	63.11	54.15	50.96
X7	63.69	53.75	54.53
X8	62.77	53.58	53.56
X9	62.59	50.25	51.41
X10	64.12	52.31	50.01

The results of recognition rate obtained by different algorithms are analyzed and the corresponding results are obtained. According to the results, we can see that the recognition rate of different recognition algorithms is different for different types of video images. The feature recognition algorithm uses deep learning algorithm in the design process, and improves the accuracy of behavior recognition to a certain extent, and obtains the recognition accuracy that traditional algorithm can not achieve. In the follow-up research, this algorithm can be used to promote the intelligent development of video monitoring. At the same time, according to the design idea of feature recognition algorithm, the content of traditional algorithm is optimized to improve the application effect of the algorithm.

After analyzing Fig. 3, it can be found that the response time of the feature recognition algorithm is more stable under different image separation rates, and there is no problem of large fluctuation. Under different image resolutions, the recognition time of traditional algorithms is different, and the fluctuation is increased. According to the analysis results, it can be determined that the recognition effect and use experience of the feature recognition algorithm are better.

Through the analysis of all the test results obtained in this test, it can be found that the application effect of the feature recognition algorithm is better than the traditional algorithm, and this algorithm can be used in the future design to obtain more effective recognition results of human falling behavior.
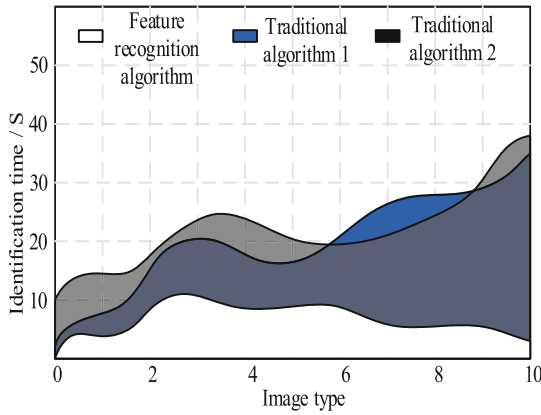
（a）High resolution image
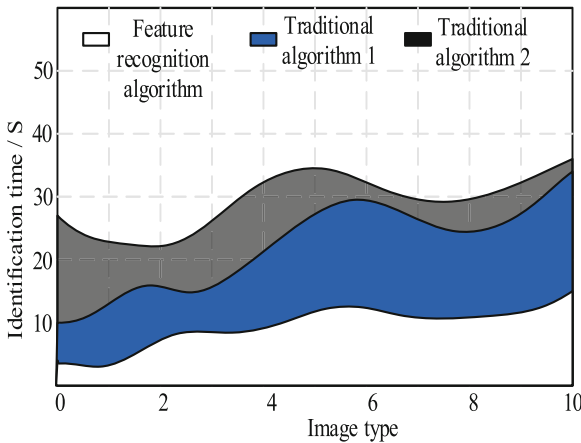
（b）Low resolution image

Fig. 3. Recognition response time under different definition

4 Discuss

With the development of science and technology, people pay more and more attention to public security. As an important preventive tool, video surveillance has been gradually popularized in people's daily life. At the same time, with the application of surveillance video in social life more and more popular, the intelligent video surveillance technology is getting more and more attention. People hope to realize the intelligent video surveillance, and hope that the computer can automatically identify abnormal behaviors such as falling, fighting and so on. With the development of computer vision and deep learning technology, this idea is gradually realized through the method of deep learning. This research mainly designs a human fall behavior recognition algorithm based on feature and deep learning, and makes the following work and contributions.

(1) This paper summarizes the main algorithms of human fall behavior recognition based on feature and deep learning. Firstly, according to the different types of input data, this paper classifies the current methods of human behavior recognition reasonably, reviews the development history and research status of human behavior recognition based on color video, discusses the shortcomings and defects of current human behavior recognition algorithms, and puts forward many corresponding improvement measures.

(2) A hybrid model based on feature and deep learning algorithm is proposed to recognize the fall behavior. After simple processing of the video data set, the hybrid model uses the sliding window and weight sharing of convolutional neural network to obtain the local features of the video sequence, which is used as the input of the next layer, and uses the timing to obtain the temporal features of the video data, They make full use of their respective advantages. In addition, the deep learning can automatically extract behavior features, which avoids the process of manual feature extraction. The paper analyzes the fall data of this paper by using traditional pattern recognition method, convolutional neural network, long-term memory network and this model respectively, and verifies that the method is feasible and effective in this paper based on the data set provided in this paper.

There are still some deficiencies in this study, which need to be improved.

(1) For the study of fall behavior, the data set used in the experiment is fixed and single person behavior. In addition, for the research of fall behavior, we use the single person behavior, without considering the group behavior, which has deviation from the actual situation. In the future, we should conduct in-depth research on fall behavior which is closer to the actual scene.

(2) In the study of behavior in video surveillance, when choosing neural network model, we can consider using deep network model such as Google net, VGg to study.

5   Conclusion

Now, with the development of information technology, video surveillance is more and more common, intelligent video surveillance has been more and more widely concerned, and the analysis of abnormal human behavior based on intelligent video surveillance has achieved certain results. But the moving human body is non rigid, has a high degree of freedom, and the scene of video surveillance is complex and changeable, so the research on human abnormal behavior is still in the analysis of some simple human behavior. Although the human fall behavior recognition method designed in this paper meets the needs of customers, its function is relatively single, and the reliability, scalability and security of the algorithm have not yet reached the standard, and the efficiency in the process of operation is not particularly high. Therefore, in the follow-up work, we hope to develop a more robust platform to make the proposed algorithm more competitive in the current intelligent monitoring platform.

References

1. She, X., Su, X.: Research on human fall behavior using CNN and LSTM-based hybrid model. Applicat. Res. Comput. **36**(12):3857–3859+3868 (2019)
2. Ding, Y., Pu, H., Zhang, J., et al.: Implementation of behavior recognition based on machine vision. In: 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). IEEE, pp. 1–5 (2019)
3. Chen, Y., Hu, R., Li, S., et al.: Recognition algorithm for human behavior in video based on combined features and SVM. J. Shenyang Univ. Technol. **42**(06), 665-669 (2020)
4. Kun, H., Zefan, H.: Falling behavior recognition method based on dynamic characteristics of human body posture. J. Hunan Univ. (Nat. Sci.) **47**(12), 69–76 (2020)
5. Huang, Y., Wan, C., Feng, H.: Multi-feature fusion human behavior recognition algorithm based on convolutional neural network and long short term memory neural network. Laser Optoelectron. Progress **56**(07), 243–249 (2019)
6. Liu, Y., Zhang, Y., Wang, C., et al.: Behavior recognition algorithm based on RGB-D and deep learning. Comput. Eng. Design **40**(06), 1747–1750+1762 (2019)
7. Ma, L., Pei, W., Zhu, Y., et al.: Fall action recognition based on deep learning. Comput. Sci. **46**(09), 106–112 (2019)
8. Fang, C., Jia, K., Liu, P.: Identification of taxi violation behavior based on surveillance video. Comput. Simulat. **37**(05), 326–331 (2020)
9. Zhou, Q., Xing, J., Yang, Q., et al.: Sequential image deep learning-based Wi-Fi human activity recognition method. J. Communicat. **41**(08), 43–54 (2020)
10. Nie, H., Xiong, X., Guo, Y., et al.: Video abnormal behavior identifying algorithm based on deep learning. Modern Electron. Techn. **43**(24), 110–112+116 (2020)
11. Liu, S., Liu, X., Wang, S., et al.: Fuzzy-aided solution for out-of-view challenge in visual tracking under IoT-assisted complex environment. Neural Comput. Appl. **33**(4), 1055–1065 (2021)
12. Liu, S., Liu, D., Srivastava, G., et al.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. 1–23 (2020)
13. Liu, S., Liu, D., Muhammad, K., et al.: Effective template update mechanism in visual tracking with background clutter. Neurocomputing (2020). https://doi.org/10.1016/j.neucom.2019.12.143

Pneumonia Detection Algorithm Based on Improved YOLOv3

Hailong Liu, Jinrong Cui$^{(\boxtimes)}$, and Chaoda Peng

College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China
tweety1028@163.com

Abstract. Pneumonia is a kind of disease caused by bacteria, viruses and other pathogens, which can seriously endanger human health, and has strong infectivity. Timely and accurate detection of pneumonia symptoms can not only make patients receive timely treatment, but also prevent the disease from spreading to others. This paper proposes an improved object detection algorithm YOLOv3-P which was based on YOLOv3. Using the idea of Path Aggregation Network (PANet) for reference, after feature fusion, the location information is enhanced through a bottom-up path, which makes full use of the feature information of each layer. And the better backbone network CSPDarkNet53 was used to replace DarkNet53 of YOLOv3, to better extract features from the pneumonia images. Experiments on the lung X-ray image data set provided by the North American Society of Radiology show that the average precision of the algorithm reaches 50.43%, which was improved compared with the YOLOv3 algorithm, and has good performance compared with other common object detection algorithms. YOLOv3-P can help doctors judge the location of pneumonia tissue faster and more accurately.

Keywords: Object detection · Pneumonia · Deep learning · YOLOv3

1 Introduction

Pneumonia is a respiratory disease caused by bacteria, viruses and other pathogens. It can lead to fever, cough, headache and other symptoms, and has a strong infectious. The number of people infected with pneumonia was increasing every year in the world. Pneumonia is a serious threat to human health. The diagnosis of pneumonia based on X-ray images is one of the most important methods for the diagnosis of pneumonia. But because the X-ray image is a black-and-white image, it is difficult for doctors to distinguish the diseased part from the normal part due to the lack of important information such as color and texture.

In recent years, machine learning has developed rapidly, and it has a wide range of applications in the field of computer vision [1–3]. In 2012, AlexNet [4] was proposed by Krizhevsky. AlexNet won the championship in the Imagenet image classification competition of the year. Convolutional neural network can learn deep feature information

from images, and it can be applied to object detection fields to improve the detection effect of the algorithm. R. Girshick et al. Proposed the first object detection algorithm R-CNN [5] based on convolutional neural network in 2014.

R. Joseph. proposed YOLOv1 [6] object detection algorithm in 2016. The detection accuracy of R-CNN series algorithm is very high, but its detection speed is slow. In order to solve the problem of real-time detection in object detection tasks, YOLOv1 was born. YOLOv1 algorithm directly inputs the image into the network and directly regresses the size and position of the target. However, compared with the two-stage object detection algorithm, its positioning accuracy is lower, especially for the very close targets and small-scale targets.

In 2017, R. Joseph improved based on YOLOv1 and proposed YOLOv2 [7] object detection algorithm. YOLOv2 absorbs the idea of Faster R-CNN algorithm and also adopts the method of the prior box. It uses the clustering method of K-means [8] to obtain the width and height of the prior box, to find a more suitable prior box.

In 2018, R. Joseph made some improvements on the basis of YOLOv2 and proposed YOLOv3 algorithm [9]. YOLOv3 uses three different proportions of feature maps to get the prediction box, which improves the detection effect of small targets, and proposes DarkNet53 feature extraction network.

With the development of the neural network, it is possible to use deep learning to detect and classify medical images. Setio [10] used multiple Convolutional Neural Networks (CNN) to identify pulmonary nodules, and fused the final results, and achieved good results. In reference [11], the de-noising self-coding method is used to extract the depth features of pulmonary nodules for classification, and its performance in fine-grained classification is better than that of traditional morphological and texture bottom feature learning methods.

In view of the current pneumonia detection algorithm prone to misdiagnosis and missed diagnosis, this paper improves the YOLOv3 detector, greatly improves its performance, reduces the phenomenon of misdiagnosis and missed diagnosis, so as to help doctors diagnose pneumonia faster and more accurately. X-ray images lack color and texture information, and it is difficult to distinguish the diseased part from the normal part. To solve this problem, this paper uses the idea of PANet [12] for reference, and further improves the Feature Pyramid Networks (FPN) [13] structure of YOLOv3. After the feature fusion of FPN, the positioning information is enhanced through the bottom-up path, making full use of the feature information of each layer. And use better backbone network CSPDarkNet53 [14] to replace the backbone network DarkNet53 of YOLOv3, so as to better extract features from the image.

The experimental results on the lung X-ray image data set provided by the Radiological Society of North America show that the detection accuracy of YOLOv3-P is greatly improved compared with the original algorithm YOLOv3, and it has better performance than several other commonly used target detection algorithms, which can help doctors judge the location of pneumonia tissue faster and more accurately.

2 Related Work

2.1 Two Stage Detectors

The first two stage detector based on deep learning is R-CNN, which uses CNN to extract features from images, but the network is complex and the training speed is slow. SPP-Net [15] can input images of any size, only need to do a convolution feature extraction to get the feature image, but it cannot achieve end-to-end detection. Fast R-CNN [16] uses RPN to generate suggestion boxes, which greatly speeds up the generation of suggestion boxes. Compared with other algorithms mentioned above, the speed of the algorithm is fast, but it has not achieved real-time detection yet.

2.2 Single Stage Detectors

Different from two stage detectors, single stage detectors do not need to generate suggestion boxes, only need to directly regress the type and location of the target, thus greatly speeding up the detection speed. YOLOv1 divides the input image into grids, and each grid cell is responsible for detecting the falling objects. SSD [14] significantly increases the detection effect by detecting targets of different sizes on feature maps of different scales.

Backbone + FPN   Head + Detection result
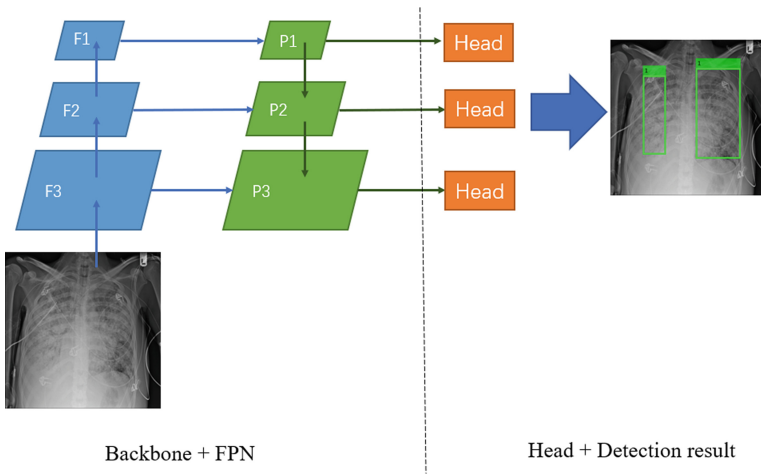
Fig. 1. Network structure of YOLOv3.

2.3 YOLOv3

YOLOv3 is an improved version of YOLOv2, and its backbone network is DarkNet53. YOLOv3 uses the idea of FPN for reference and fuses the feature images of the same size in the up-sampling stage. YOLOv3 network outputs three different scale feature maps, which are 13×13, 26×26, 52×52. Different size feature maps are used to detect different size objects. Figure 1 shows the network structure of YOLOv3.

3 Our Approach

In this paper, through the improvement of YOLOv3, YOLOv3-P is proposed. The algorithm first extracts features through the CSPDarkNet53 backbone network, and then obtains three feature maps with different scales through the PAN. Finally, each feature map is classified and regressed to get the final result. The network structure of the YOLOv3-P algorithm is shown in Fig. 2.
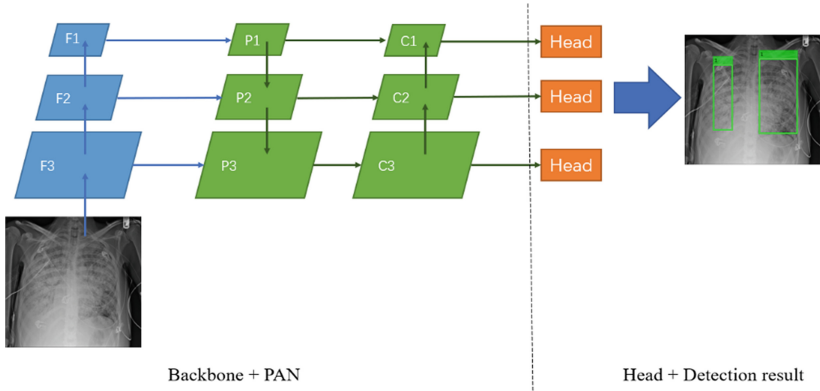
Fig. 2. Network structure of YOLOv3-P.

The training process of YOLOv3-P includes three stages: feature extraction, target location and loss calculation. The main network used in feature extraction is CSPDarkNet53 network. Object location is divided into two steps: classification and location regression. The loss calculation includes classification loss and location loss. After the lung images are input into the YOLOv3-P network, three feature maps F1, F2 and F3 with different sizes are output through the backbone network. Among them, F1 is 52 × 52, F2 is 26 × 26, F3 is 13 × 13. F3 gets P3 through 1 × 1 convolution reduction channel, P3 is up sampled to get a feature map of size 26 × 26 and fused with F2 to get P2, P2 is up sampled to get a feature map of size 52 × 52 and fused with F1 to get P1. After obtaining different sizes of feature maps P1, P2 and P3 through FPN, P1 passes through 1 × 1 convolution reduction channel to obtain feature map C1 with size of 52 × 52, C1 is down sampled to obtain feature map with size of 26 × 26, which is fused with P2 to obtain C2, C2 is down sampled to obtain feature map with size of 13 × 13, which is fused with P3 to obtain C3. At the same time, after each feature fusion, the fused feature image will be convoluted by 3 without changing the size and channel. After obtaining different size feature maps C1, C2 and C3, different size targets are predicted on each feature map.

3.1 PANet

The PANet was proposed in reference [17], which shortens the information path between low-level and high-level features through a bottom-up path. The specific method is, after

extracting the image features through the backbone network, first through the top-down path, the high-level features are fused by up-sampling and low-level features, and then through the bottom-up path, the whole feature layer is enhanced by using more accurate positioning information of low-level features, to get a better feature map for subsequent classification and regression.

3.2 CSPDarkNet53

CSPDarkNet53 is a backbone network based on DarkNet53, the backbone network of YOLOv3, and learning from the experience of CSPNet [18]. CSPNet solves the problem that network reasoning needs a lot of computation from the perspective of network architecture. Replacing DarkNet53 with CSPDarkNet53 can enhance CNN's learning ability, make the network lightweight while maintaining accuracy, and reduce the amount of computation and memory costs.

4 Experiments

4.1 Dataset

The experimental data set comes from the X-ray images of the lung provided by the Radiological Society of North America, including 4000 Gy-scale images with the size of 1024 pixels × 1024 pixels, 75% of which are used for training and the rest for testing. The input size of the network is 416 pixels × 416 pixels, and the gray image into RGB (Red, Green, Blue) image.

4.2 Training Details

In the experiment, the pre training model of DarkNet53 and CSPDarkNet53 on COCO data set is used as the feature extraction network, and Adam [19] is used to optimize the model. The experimental environment is: the deep learning framework is pytorch1.2 + cuda10.0, the GTX Tian x graphics card with 12 GB video memory, and the operating system is Ubuntu 18.04.

4.3 Metrics

In order to select the appropriate algorithm, this paper uses the average precision (AP) as the evaluation index. P-R curve is based on the two variables of precision (P) and recall (R). AP is obtained by calculating the area under P-R curve. And the Intersection over Union (IoU) is 0.5. The evaluation index of detection speed is frame per second (FPS).

4.4 Ablation Study

In order to verify the influence of different improved methods on the algorithm, this paper sets up three groups of experiments, which are: 1) YOLOv3; 2) YOLOv3 + PAN; 3) YOLOv3 + CSPDarkNet53 + PAN. The experimental results are shown in Table 1.

Table 1. Influence of different improvement methods on the performance of YOLOv3.

Method	AP/%	FPS
YOLOv3	44.21	40.96
YOLOv3 + PAN	46.47	39.12
YOLOv3 + PAN + CSPDarkNet53	50.43	34.12

As can be seen from Table 1, after replacing FPN in YOLOv3 with PAN, the AP of the algorithm is improved by 2.26%. On this basis, after replacing DarkNet53 with CSP-DarkNet53, the AP can continue to improve by 3.96%. This shows that the improvement of this algorithm to YOLOv3 can improve the AP of the algorithm.

4.5   Compared with Other Detectors

In order to verify the detection performance of YOLOv3-P, this paper compares YOLOv3-P with other object detection algorithms, and the experimental results are shown in Table 2.

Table 2. Performance comparison of different algorithms.

Method	AP/%	FPS
SSD	41.78	70.35
Faster R-CNN	46.12	9.47
YOLOv3	44.21	40.96
YOLOv3-P	50.43	34.12

As can be seen from Table 2, in the lung X-ray image data set, the AP of our algorithm YOLOv3-P is higher than other algorithms. The AP of SSD is lower than that of YOLOv3, and the AP of Faster R-CNN is higher than that of YOLOv3. The AP of YOLOv3-P is 8.65% and 4.31% higher than SSD and Faster R-CNN, respectively.

4.6   Detection Effect of Different Algorithms

Figure 3 shows the detection effect of different algorithms on four groups of pneumonia images, in which the blue box represents the real coordinate box, the green box represents the detection box, and the red box represents the detection error box.

As can be seen from Fig. 3, SSD algorithm has poor detection effect, large positioning error, and false detection and missing detection problems. The reason is that SSD algorithm does not fuse feature maps of different sizes, and positive and negative samples are unbalanced in the training process. The detection effect of YOLOv3 is better than that of SSD, because YOLOv3 algorithm combines different scale feature maps and

makes better use of the information of pneumonia image. Faster R-CNN algorithm uses region proposal network (RPN) to generate suggestion box, which makes the detection effect better and the problems of missed detection and false detection less. YOLOv3-P has the least false detection and missing detection problems, and the positioning accuracy is higher, because this algorithm replaces FPN with PAN, to make better use of the information of different sizes of feature maps. At the same time, it uses a more excellent backbone network CSPDarkNet53, so that the network can extract better feature information.
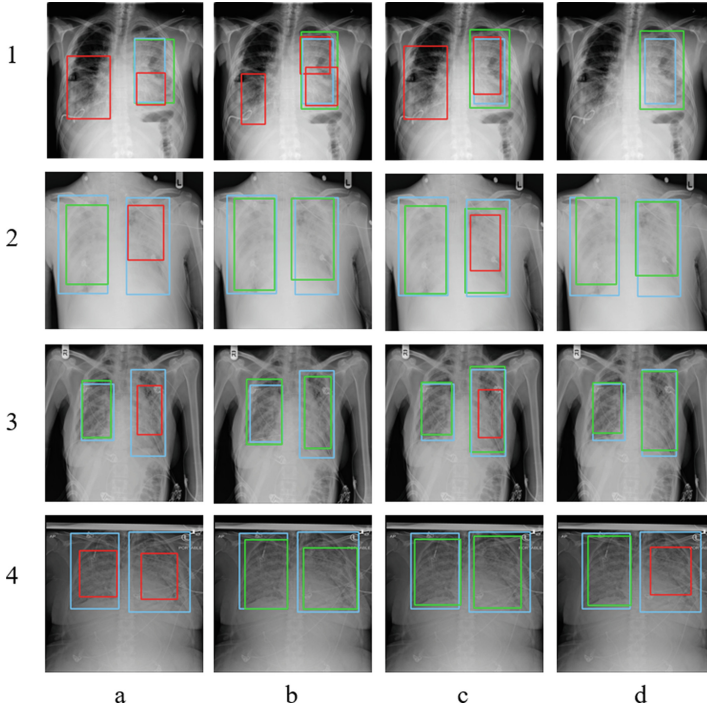
Fig. 3. Detection effect of different algorithms. Blue box: real coordinate box; Green box: detection box; Red box: detection error box. (a) SSD; (b) Faster R-CNN; (c) YOLOv3;(d) YOLOv3-P (Color figure online)

5   Conclusions

In view of the lack of color and texture information in the X-ray of the diseased part of pneumonia, the feature is not obvious and so on, this paper proposes the YOLOv3-P algorithm based on the improvement of the YOLOv3 algorithm. The algorithm uses a better feature extraction network CSPDarkNet53 to extract features better, and uses pan structure to make better use of feature information. The experimental results show that YOLOv3-P improves the detection accuracy significantly compared with the original algorithm, and has good performance compared with other classical target detection

algorithms. YOLOv3-P can help doctors judge the location of pneumonia tissue faster and more accurately.

References

1. Gao, G., Yang, J., Jing, X.-Y., Shen, F., Yang, W., Yue, D.: Learning robust and discriminative low-rank representations for face recognition with occlusion. Pattern Recogn. **66**, 129–143 (2017)
2. Zhang, Z., et al.: Inductive structure consistent hashing via flexible semantic calibration. IEEE Trans. Neural Netw. Learn. Syst. **32**, 4514 (2020)
3. Zhang, Z., Lai, Z., Xu, Y., Shao, L., Wu, J., Xie, G.: Discriminative elastic-net regularized linear regression. IEEE Trans. Image Process. (TIP) **26**(3), 1466–1481 (2017)
4. Wu, Q., Shen, C., Wang, P., et al.: Image captioning and visual question answering based on attributes and external knowledge. IEEE Trans. Pattern Anal. Mach. Intell. 1 (2017)
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014). https://doi.org/10.1109/CVPR.2014.81
6. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection. Comput. Vis. Pattern Recogn., 779 (2016)
7. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: IEEE Conference on Computer Vision & Pattern Recognition, pp. 6517–6525 (2017)
8. Hartigan, J.A., Wong, M.A.: A K-means clustering algorithm. Appl. Statist. **28**(1), 100 (1979)
9. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv e-prints (2018)
10. Setio, A., Ciompi, F., Litjens, G., et al.: Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE Trans. Med. Imaging **35**(5), 1160–1169 (2016)
11. Cheng, J.Z., Ni, D., Chou, Y.H., et al.: Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in CT scans. Sci. Rep. **6**, 24454 (2016)
12. Liu, S., Qi, L., Qin, H., et al.: Path aggregation network for instance segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2018)
13. Lin, T.Y., Dollar, P., Girshick, R., et al.: Feature pyramid networks for object detection. IEEE Comput. Soc. (2017)
14. Bochkovskiy, A., Wang, C.Y., Liao, H.: YOLOv4: optimal speed and accuracy of object detection (2020)
15. He, K., Zhang, X., Ren, S., et al.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2014)
16. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks (2017)
17. Liu, W., Anguelov, D., Erhan, D., et al.: SSD: single shot MultiBox detector. In: European Conference on Computer Vision. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
18. Wang, C.Y., Liao, H., Wu, Y.H., et al.: CSPNet: a new backbone that can enhance learning capability of CNN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE (2020)
19. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. Comput. Sci. (2014)

Breast Ultrasound Images Clustering Analysis Using Deep Clustering Method

Cheng Huang and Jinrong Cui$^{(\boxtimes)}$

College of Mathematics and Informatics, South China Agricultural University,
Guangzhou 510642, China
`tweety1028@163.com`

Abstract. With the expanding volume of medical data, the supervised algorithms have the difficulty in obtaining labels of the medical data. To address this problem, many researchers have introduced unsupervised learning algorithms to analyze medical data and to mine the potential information among different samples. Simultaneously, with the brisk growing of deep learning, the association of unsupervised algorithms and deep learning is becoming further intimate. In this paper, we will use clustering algorithm which is commonly used in unsupervised learning to analyze breast ultrasound images. The Improved Deep Embedded Clustering (IDEC) algorithm will be introduced in this paper, which is broad used in the current deep clustering methods to do clustering analysis. Meanwhile we have added the graph embedding strategies in IDEC algorithm to enhance the clustering performance. The algorithm includes two phases: pre-training phase to capture hidden features and fine- tuning phase to enhance clustering result. We have exerted our algorithm on breast ultrasound images dataset to perform clustering analysis, and the experiment results indicate that the clustering performance of this algorithm is better than other traditional clustering algorithms.

Keywords: Unsupervised learning · Clustering analysis · Deep clustering · Medical data

1 Introduction

To derive the hidden information of image, supervised learning algorithms based on deep learning exert fundamental algorithms and large-scale labeled materials for a range of image analysis [1]. Nevertheless, In the area of medical image analysis, manual description of medical image requires a series of medical and computer knowledge. In recent years, the medical image depository continues to enlarge. But there has not been an equivalent development in the number of labeled samples. Furthermore, the sources of medical image samples are extremely broad such as organs or human tissues images on X-ray and a range of images on Magnetic Resonance (MR), Faced with the variety of complex medical materials, the learning cost is a primary problem too.

In several paper, a series of approaches have been used to address these challenges including method based on one-shot learning. This algorithm uses some relatively small-scale labeled materials to train and deep model. Introducing transfer learning exert images from others general dataset as a feature learning phase [2]. Some approaches exert unsupervised learning algorithm to capture the feature of the medical data, where the purpose is to acquire medical image information. However, only exerting unsupervised learning algorithm to capture the hidden information of the medical data is not enough, and more unsupervised learning methods should be introduced to obtain better performance. Alternatively, the clustering method has been introduced to discriminate dissimilar samples. Clustering method plays an essential role in unsupervised learning methods. In image segmentation [3–5] and unsupervised image classification [6–9] areas, many clustering methods and models has been used.

Normally, medical image materials have the characteristics of large quantity, multiple data sources and complex picture structure. Breast ultrasound image is one of the typical image types. For breast ultrasound images, many researches have done a series of analysis [10–12]. Our purpose is to introduce an unsupervised learning method to obtain the main features from breast ultrasound images. Then exerting a clustering algorithm distinguishes these images and maximize the separation of images between different categories. Therefore, we introduced an improved Deep Embedding Clustering (IDEC) [13] algorithm that combines a deep model and a clustering algorithm. In order to enhance clustering performance, we join graph embedding strategy to maintain the neighbor relationship between samples. Meanwhile, we experimented on a breast ultrasound images dataset to verify the accuracy of proposed method. Final we have compared the proposed algorithm with other traditional clustering algorithms. Experimental results show that proposed algorithm performs better than other traditional Clustering algorithms. This may provide new thoughts for the area of medical image analysis.

2   Relate Works

2.1   Deep Clustering

Recently, a series of clustering methods have been published, and which introduces the deep models to learn main features representations and uses these hidden representations to enhance clustering performance. Deep clustering network (DCN) [14] was proposed in a framework which include two part: dimensional reduction and K-means clustering. Deep embedded clustering (DEC) [15] was introduced in a deep stacked autoencoder (SAE) [16] to optimize the feature extraction model by pre-train dataset, and then a novel clustering objective function was proposed in [17] and which is based on KL divergence. DEC optimized through iteratively self-training. These Clustering methods have displayed superiority beyond the tradition algorithms. While, these approaches are rarely exerted in the area of medical data analysis. This article applies the deep clustering method to medical image analysis, which can achieve better clustering performance than traditional methods, and can be suitable adapted to large-scale datasets.

2.2 AutoEncoder

Autoencoder is a neural network model of self-supervised learning, which include encodes and decodes. Samples were inputted into encoder and were outputted from decoder, the training strategies was introduced by comparing the similarity between input samples and decoded samples. The principal part of the stacked autoencoder is fully connected neural networks, which has a hidden layer z to capture the key features of the input sample. the function of encoder and decoder is $\bar{x} = F_E(F_D(x))$.

Where $F_E()$ indicates encoder mapping to learn hidden features z, $F_D()$ indicates decoder mapping to reconstruct input sample x, $\bar{x}$ indicates reconstruct sample for x.

Following are two general applications of autoencoders.

Under-complete autoencoder. The purpose of undercomplete autoencoder is to capture the key information of input sample, thus, it designs networks that keep the dimensions of hidden layer less than input layer.

Denoising autoencoder. The purpose of denoising autoencoder is to recover the input sample that destroyed by noise. The denoising autoencoder has better anti-noise performance than autoencoder, and the recovered images look clearer than autoencoder.

In this paper, under-complete autoencoder is introduced and the reconstruction loss function is used so that the autoencoder can learn the essential features of the input sample in pre-train phase.

2.3 Improved Deep Embedding Clustering

Improved Deep Embedded Clustering (IDEC) can initialize the hidden features with pre-training phase which using encoder and decoder. The decoder is removed after pre-training phase. In fine-tuning phase the remaining encoder is used to minimize objective function:

$$L = L_r + L_c \tag{1}$$

where L_r and L_c are reconstruction loss and clustering loss defined by Mean Squared Error (MSE) and KL divergence:

$$L_r = \sum_{i=1}^{n} \|x_i - F_D(z_i)\|_2^2 \tag{2}$$

$$L_c = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3}$$

In formula (2), L_r is so-call reconstruction loss which means the similarity between input sample and decoded sample. In formula (3), q_{ij} indicate the probability that the sample x_i can be clustered as clustering group j calculated by Student's t-distribution:

$$q_{ij} = \frac{(1 + \|z_i - \mu_i\|^2)^{-1}}{\sum_j (1 + \|z_i - \mu_i\|^2)^{-1}} \tag{4}$$

And p_{ij} in (2) indicate the target distribution derived by q_{ij}:

$$p_{ij} = \frac{q_{ij}^2 / \sum\limits_{i=1}^{n} q_{ij}}{\sum\limits_{j=1}^{c} (q_{ij}^2 / \sum\limits_{i=1}^{n} q_{ij})} \tag{5}$$

3   The Proposed Method
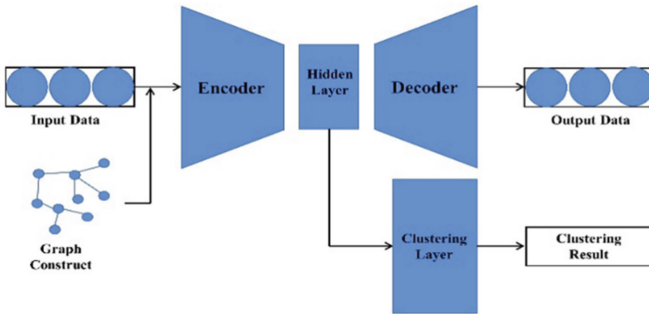
Fig. 1. The net construction of improved deep embedding clustering with graph embedding

Definition. Given a dataset $\{X\}$ include n unlabel samples. The aim of clustering is to cluster n unlabel samples into k categories. In Fig. 1 we have gave the net construction of the improved deep embedding clustering with graph embedding. Generally, the dimension of samples is high-dimension. Thus, IDEC algorithm introduces an under-complete autoencoder to reduce the dimension of high-dimensional samples. Meanwhile, the key information and features of high-dimensional samples are also learned during the phase of dimension reduction. The hidden features are defined z. Using hidden features z can reduce the overhead of the clustering phase.

3.1   Graph Embedding Strategy

Inspired by a hypothesis in the area of subspace learning that if a couple of samples in the high-dimension space are near, then the low-dimensional features extracted from them should be nearby in the low-dimension space [18]. To satisfy this intention, we institute neighbor relationships by following neighbor graph strategy:

$$\min \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| z_i - z_j \right\|_2^2 G_{i,j} \tag{6}$$

where $G \in R^{n \times n}$ indicate the matrix of nearest neighbor graph relationship from $\{ X \}$. The graph matrix G computed as follows:

$$G_{i,j} = \begin{cases} 1, & x_i \in \Delta(x_j) \ or \ x_i \in \Delta(x_j) \\ 0, & otherwise \end{cases} \tag{7}$$

where $\Delta(x_i)$ indicates the nearest neighbor of sample x_i under the Euclidean distance metric.

We impose this graph strategy on both pre-training and fine-tunning layers. The loss of pre-training layer is defined as follow:

$$L = L_r + \gamma L_g \tag{8}$$

where L_r and L_g defined as formulas (2) and (6) respectively, γ is a trade-off coefficient.

3.2   Fine-Tunning Layer and Initialization

After pre-training phase, encoder can capture hidden features of input samples. However, the hidden features that are only training trained by pre-training phase, and may not perform so well on the clustering task. Thus, we save the weight of model trained in the pre-training phase and remove the decoder part from the autoencoder. Then encoder part is connected with fine-tunning layer. Fine-tunning layer includes tunning and clustering part. The purpose of fine-tunning layer is to train an encoder that is able to which can capture features of clustering while obtaining enhanced clustering performance. Before training of fine-tunning layer, we initialize clustering centers by employing k-means. The loss of fine-tunning layer includes three parts: reconstruction loss, clustering loss and graph embedding loss as follow:

$$L = L_r + \gamma_1 L_c + \gamma_2 L_g \tag{9}$$

where L_r,L_c and L_g defined as same as formula (8) and (6).γ_1 and γ_2 is the trade-off coefficient of clustering loss and graph embedding loss respectively.

3.3   Parameter Updating

In this paper, there are three types of parameters to update: the weights of autoencoder, cluster centers and the parameter of target distribution. The clustering centers are updated as parameters, which is convenient to further extract the features suitable for clustering. Meanwhile, updating target distribution make the clustering group distribution more obvious. The whole method is displayed in Algorithm 1.

3.4   Method Summary

In general, inheriting the excellent properties of IDEC and graph strategy, Our proposed method has the potential to obtain the hidden representation of input samples while preserving the neighbor relationship among samples. By this way, the performance of clustering is further improved.

Algorithm 1: Improved Deep Embedding Clustering with Graph Embedding

---

Input: Data $\{X\}$; Parameter α; Maximum iterations:$Maxiter$
Update interval:T;Batch size:B_s;Stopping thresholdξ;
Output: Clustering result:$label$.
Initialization: Construct k-nearest neighbor graphs as formula (4); Initialize weights of autoencoder, Cluster centers μ according to Section 3.2
For $iter \in (0,1,2 \dots Maxiter)$ **do**
  if $iter \% T = 0$ **then**
    Compute hidden features points of input data $z_i = F_E(x_i)$
    Compute soft target P, Q according to (4), (5)
    Reserve last labels $s_{last} = s_i$;Update predicted labels
    $s_i = arg \max_j q_{ij}$
    if $sum(s_i \neq s)/n$ **then**
    Stop
End for

---

4   Experiment

4.1   Dataset

We impose the proposed method on breast ultrasound images which is published in MedMNIST [19]. The image dataset includes 780 ultrasound gray images of breast. The shape of origin images is 28×28 pixel. In this experiment, the shape of images will be converted into a 784-dimensional vector. The specific of this dataset is shown in Fig. 2 and Table 1.
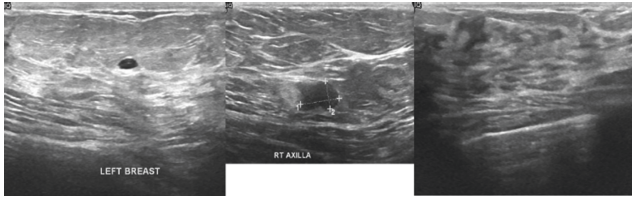
Fig. 2. Data set sample display

Table 1. The detail of BreastMNIST dataset

Name	Example number	Classes	Dimension
BreastMNIST	780	2	784

4.2 Experiment Work

Compared Methods. To demonstrate the effectiveness of our proposed algorithm, we compare our algorithm with a series of baseline. The mainly compared method is IDEC and DEC. When the reconstruction loss remains at zero, DEC can be regarded as a special situation of IDEC. The $AE + k$-means which means a two-stage processing includes pre-training autoencoder phase which like IDEC pre-training phase and performing k-means after pre-training phase.

Parameters Setting. The encoder and decoder network are combined into an under-complete autoencoder which is stacked by four layers fully connected neural network with neural numbers $[a, 500, 500, 1000, 10]$ and $[10, 1000, 500, 500, a]$, where a means the dimension of input samples. We introduce nonlinearity function 'ReLU' [20] as the activation function for all fully connected layers. In the pre-training phase and the fine-tuning phase, we use the optimizer 'Adam' [15] with learning rate $\lambda = 0.0003$ at pre-training phase and $\lambda = 0.0005$ at fine-tunning phase (We did experiments with learning rates ranging from 0.0001 to 0.0005 on fine-tunning phase) to optimization. We also use the coefficient of clustering loss in IDEC setting $\gamma_1 = 0.1$ and set the coefficient of graph embedding loss $\gamma = 0.001, \gamma_2 = 0.001$ in pre-training phase and fine-tunning phase. The threshold of stop training is assigned to $\delta = 0.1\%$.

Evaluation Metric. We introduce unsupervised clustering accuracy (ACC) as evaluation metric [13].

4.3 Experiment Result

Experimental results on the breast ultrasound dataset are demonstrated in Tables 2. In order to ensure the validity of the experimental results, we run pre-training 400 epochs with 256 batch size among compared methods which require pre-training phase. After 400 epochs pre-training phase, fine-tunning phase also keep 256 batch size. In order

Table 2. Comparison of experiment result in accuracy (%)

Methods	ACC
k-means	55.00%
AE+ *k*-means	60.90%
DEC	62.05%
IDEC	63.08%
IDEC with graph embedding (Ours)	**72.05%**

to better describe the influence of super parameters on the experiment, we adjusted the learning rate from 0.001 to 0.005 and the graph loss coefficient from 0.001 to 0.005, respectively.

Table 3. Adjusting different learning rate test accuracy (%). Fixing learning rate at 0.0005 test accuracy (%) under different graph embedding loss coefficients.

Learning rate	ACC	Graph loss	ACC
0.0001	56.54%	0.001	70.90%
0.0002	55.26%	0.002	71.28%
0.0003	69.36%	0.003	71.95%
0.0004	68.72%	0.004	72.05%
0.0005	70.90%	0.005	71.54%
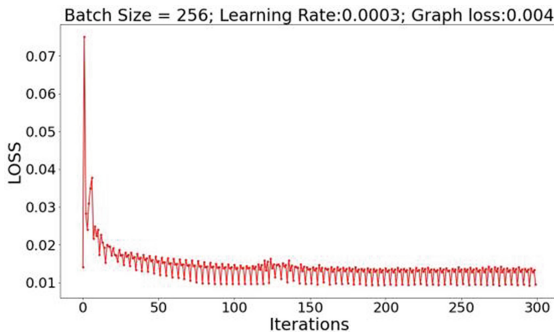
Fig. 3. Losses during fine-tunning phase on BreastMNIST. Note: Learning rate is 0.0005 and graph embedding loss is 0.004

We can notice the following results from Table 2, 3 and Fig. 3, 4: 1) IDEC with graph embedding notably outperforms the other compared methods on the breast ultra-sound databases. 2) By modifying the learning rate, ACC has increased significantly. By modifying graph embedding loss coefficient, ACC kept a relatively stable level. 3) The
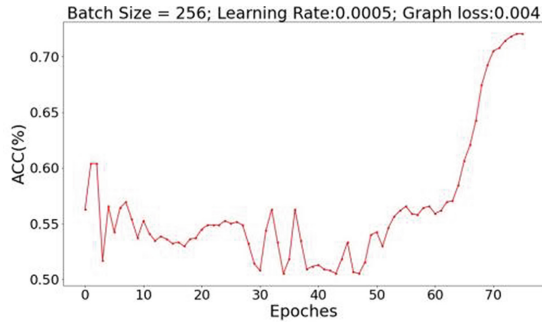
Fig. 4. Accuracies during fine-tunning phase on BreastMNIST. Note: Learning rate is 0.0005 and graph embedding loss is 0.004

loss of IDEC with graph embedding shows an oscillating trend which constantly oscillate between 0.02 and 0.01. 4) The clustering accuracy of IDEC with graph embedding fluctuates at first, and gradually rises after training for multiple epochs, and reaching the threshold at 72.05%.

5   Conclusion

In this paper, we proposed a novel method called IDEC with graph embedding. This method can be used for cluster of breast ultrasound images. Based on antoencoder and KL divergence, method introduces the graph embedding strategy to reserve the neighbor relationship of input samples. The excellent performance of IDEC with graph embedding are validated on BreastMNIST dataset. The IDEC with graph embedding may contribute extra insight for the unsupervised learning of medical image analysis.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE, Piscataway, NJ (2016)
2. Ahn, E., Kim, J., Kumar, A., Fulham, M., Feng, D.: Sparsity-based convolutional kernel network for unsupervised medical image analysis. arXiv preprint arXiv:1807.05648 (2018)
3. Vishnuvarthanan, G., Rajasekaran, M.P., Subbaraj, P., Vishnuvarthanan, A.: An unsupervised learning method with a clustering approach for tumor identification and tissue segmentation in magnetic resonance brain images. Appl. Soft Comput. **38**, 190–212 (2016)
4. Moriya, T., et al.: Unsupervised segmentation of 3D medical images based on clustering and deep representation learning. In: Gimi, B., Krol, A. Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging, p. 1057820. SPIE, Bellingham (2018)

5. Ahn, E., et al.: Saliency-based lesion segmentation via background detection in Dermoscopic images. IEEE J. Biomed. Health Inform. **21**(6), 1685–1693 (2017)

6. Moriya, T., et al.: Unsupervised pathology image segmentation using representation learning with spherical k-means. In: Gurcan,M N., Tomaszewski J E. Medical Imaging 2018: Digital Pathology, p. 1058111. SPIE, Bellingham (2018)

7. Zhang, Z., Liu, L., Shen, F., Shen, H.T., Shao, L.: Binary multi-view clustering. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1774–1782 (2018)

8. Zhang, Z., et al.: Highly- economized multi-view binary compression for scalable image clustering. In: 15th European Conference on Computer Vision, pp. 717–732. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01258-8_44

9. Gao, G., Yang, J., Jing, X.Y., Shen, F., Yang, W., Yue, D.: Learning robust and discriminative low-rank representations for face recognition with occlusion. Pattern Recogn. **66**, 129–143 (2017)

10. Rahmawaty, M., Nugroho, H.A., Triyani, Y., Ardiyanto, I., Soesanti, I.: Classification of breast ultrasound images based on texture analysis. In: 1st International Conference on Biomedical Engineering , pp. 1–6. IEEE, Piscataway, NJ (2016)

11. Xu, Y., Wang, Y., Yuan, J., Cheng, Q., Wang, X., Carson, P.L.: Medical breast ultrasound image segmentation by machine learning. Ultrasonics **91**, 1–9 (2019)

12. Moon, W.K., Lee, Y.W., Ke, H.H., Lee, S.H., Huang, C.S., Chang, R.F.: Computer- aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Comput. Methods Programs Biomed. **190**, 105361 (2020)

13. Guo, X., Gao, L., Liu, X., Yin, J.: Improved deep embedded clustering with local structure preservation. In: International Joint Conference on Artificial Intelligence, pp. 1753–1759. Morgan Kaufmann, Burlington (2017)

14. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: European Conference on Computer Vision (ECCV), pp. 139–156. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_9

15. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning (ICML), pp. 740–749. ACM, New York (2016)

16. Suk, H.-I., Lee, S.-W., Shen, D.: Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct. Funct. **220**(2), 841–859 (2013). https://doi.org/10.1007/s00429-013-0687-3

17. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(86), 2579−2605 (2008)

18. Kang, Z., Peng, C., Cheng, Q.: Clustering with adaptive manifold structure learning. In: International Conference on Data Engineering (ICDE), pp.79–82 (2017)

19. Yang, J., Shi R., Ni B.: MedMNIST classification decathlon: a lightweight automl benchmark for medical image analysis. arXiv preprint arXiv:2010.14925 (2020)

20. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. J. Mach. Learn. Res. **15**, 315–323 (2011)

Information Fusion for the Devices of Internet of Things

Comparing Methods of Imputation for Time Series Missing Values

Renkang Geng, Mingran Li[✉], Mingxu Sun[✉], and Yujie Wang

School of Electrical Engineering, University of Jinan, Jinan 250022, Shandong, China
834331749@qq.com, cse_sunmx@ujn.edu.cn

Abstract. Due to the rapid development of modern information engineering, a lot of data are used in machine learning and data cleaning and data mining of the hot research fields, such as a large portion of the data algorithm and related data model are built for complete data set, But in our real life and work, the absence of data exists in a large number of data collection, collation, transmission, storage and other links, it causes many obstacles and difficulties to build a model for complete data. The general way of dealing with missing values for simple delete, that deal with missing value method is a simple convenient but can cause: two aspects of the problem and the inconvenience caused by the original data set to reduce, reduce the reliability of the data, especially in the case of data loss is bigger, can cause a large number of data sets to reduce and missing, This has caused a lot of trouble to our work and research, so we need to find a more efficient and better method than direct deletion. In order to better solve the above problems, we mainly fill in the missing values of time series data, which has become an urgent problem to be solved. In this paper, mean filling, median filling, mode filling, PCA-EM filling and other methods are used to fill traffic data. By comparing these methods, the filling effect of each method is evaluated.

Keywords: Missing value · Filling method · Traffic data

1 Introduction

As early as the early 20th century, foreign scholars began to study data quality issues, such as data missing and data cleaning. In the late 1940s, the study of data missing problem set off a boom, and experts and scholars proposed various solutions to the missing values. Yates. F, a famous statistician, proposed a method to fill in the missing values because there were too much missing experimental data to complete the data analysis. This method showed a good effect in the different analysis [1]. Then the filling method became a research craze, followed by the mean filling, regression filling, clustering filling, hot card filling, multiple filling and many other classical methods. Entering the 21st century, the processing methods of data missing have become mature, and few new filling ideas have been put forward, most of which are based on the improvement and

application of the current field [2, 3]. For example, in 2003, Gebatista et al. analyzed and compared the four missing data processing methods of supervised learning, and proved the performance superiority of the k-nearest neighbour filling algorithm in terms of filling means [4]. In 2018, Zakaria et al. used ambient temperature and humidity monitoring data to evaluate four filling methods (mean filling, regression filling, multiple filling, and nearest neighbour filling). In 2019, Little et al. conducted the foregoing analysis on the latest statistical processing methods of missing data, providing practical application information [5].

The complexity of traffic data, including road network data, road name data, road surface conditions, road name emergencies, weather influence and other factors, under the action of these complex factors, whether the characteristics of traffic data can be accurately extracted and the data analysis and summary, has always been the most important problem in the study of traffic data [6].

The study of traffic data is of great significance to the improvement of urban traffic. In terms of information integration advantages and combination efficiency, it can help us build a better traffic model, reduce traffic congestion and improve traffic safety. However, the lack of data may lead to the deviation of our research results, which is not conducive to the improvement of traffic conditions [7].

For data mining, the existence of missing values causes the following effects: the system loses a lot of useful information; The uncertainty in the system is more significant, the deterministic component contained in the system is more difficult to grasp, and the data containing missing values will make the mining process into chaos, leading to unreliable output. Data mining algorithms themselves are more committed to avoiding the model built by over-fitting data, which makes it difficult for them to handle incomplete data well through their own algorithms. Therefore, missing values need to be derived and filled in by special methods to reduce the gap between data mining algorithms and practical applications.

2   Related Work

At present, the processing of missing values is basically divided into three categories: delete, fill, and do not process. According to the absence of data sets and different research contents, different processing methods also show different effects [8]. Although the simple deletion method is easy to operate and fast, it is very easy to cause the loss of valuable information in the data set, and the data distribution is chaotic. Therefore, the application field of this method is very limited. Therefore, we mainly introduce several missing value filling schemes we adopted in this experiment.

Filling methods of missing values can be roughly divided into two categories, namely, statistical methods and machine learning methods. Most statistical methods make assumptions based on the data set itself, and then use the original data set to fill the missing data accordingly [9, 10]. Such methods do not consider the category of the data object itself, and the filling value is often affected by objects of other categories, resulting in poor accuracy of filling results. Common methods include EM filling algorithm, regression analysis, multiple interpolations and so on. Machine learning methods are generally used to classify or cluster missing data sets and then fill them. This kind

of method is rising with the boom of machine learning in recent years, and the representative methods include K nearest neighbor filling, Bayesian network and so on [11, 12].

Mean filling: this method will fill the average value of the data of the existing attribute to the data of the same attribute, this method is simple and convenient, easy to operate and save time and effort, can save cost and time.

Median filling: Similar to mean filling, this method puts together data with observed values for the same attribute and then takes the median to fill in the missing data for the same attribute.

Mode filling. In this experiment, mode filling takes the observed values that appear most in a fixed column to fill in the missing values of other pairs in this column [13].

Probabilistic PCA-EM filling: The first method is a dimension reduction-maximum expectation algorithm. The original data is processed by linear projection technology, and the variance of the data after dimensionality reduction is retained to the maximum extent of the characteristics of the original data [14]. Then, the variance after dimensional reduction is combined with the EM algorithm to calculate the maximum likelihood estimate of the parameter through iteration. Finally, the missing value is filled through the estimated value. This algorithm is particularly efficient in the case of a large amount of data [15, 16].

3   Methods

The specific steps of this experiment are as follows: first, read the traffic data and preprocess the data; then, judge the missing values and select the completion method; finally, complete the traffic data and evaluate the complete results. Thereinto, after we read the traffic flow data, we branch the data according to each day, and each column of data corresponds to the traffic flow data of different days in the same period of time. Therefore, each column of data is regarded as the same attribute, and the observed value under the same attribute is used to complete the missing value under the same attribute.

As for the evaluation index of the complete result of this experiment, I adopted the visual analysis method Q-Q diagram. Q-Q diagram refers to a scatter diagram, where Q represents quantile and a probability diagram. The comparison of two data is made by the visualization method. In the figure, the corresponding x coordinate is the real data under the fixed quantile, and the *y* coordinate is the predicted value (completion value) under the fixed quantile. These two quantiles are the same. If the scatter plot is roughly distributed on both sides of the line $y = x$, then we can determine that the two sets of data are related.

If there is a Q-Q chart, we can judge whether our prediction data are correlated with the complete data. However, we cannot quantitatively describe the relationship between the prediction data and the missing data through the Q-Q chart. Therefore, we adopt several other methods, such as MSE, RMSE, MAE, SMAPE, etc., which can intuitively compare the effect of the complete data. The formula for these indicators is as follows:

$$MSE = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |(y_i - \hat{y}_i)|$$

$$SMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{(|\hat{y}_i| + |y_i|)/2}$$

4 Results
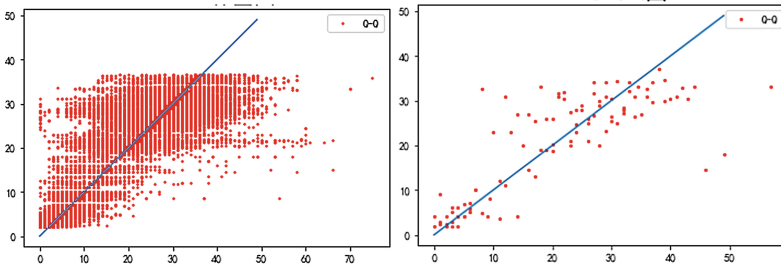
Fig. 1. The left figure is the original q-q chart after the mean filling of the data, and the right figure is the q-q chart drawn after the equal spacing sampling of the completed data. The abscissa is the real data, and the ordinate is the completed data. From the figure, we can find that the distribution of the complete data is similar to that of the real data.

Table 1. Values of each index under the mean filling

Index	Value
MSE	58.8
RMSE	7.7
MAE	5.4
SMAPE (%)	33.1

After filling with the mean value, we can intuitively see that most points are distributed on both sides of Y = X, so it can reflect that the two sets of data are roughly correlated. When the real traffic flow data is low, most of the points are distributed below the straight line formed by Y = X, and the predicted data of this point is less than the real data. When the real data is 30, our complete data is the most accurate; when the real data exceeds 30, the real data exceeds the predicted data (Figs. 1, 2 and Tables 1, 2).
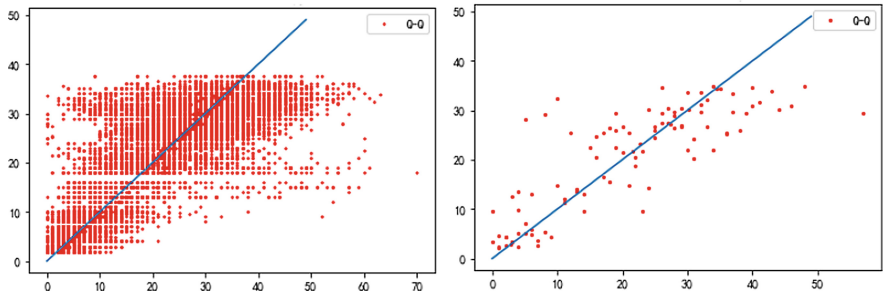
Fig. 2. The figure above is the Q-Q diagram after median filling of data, the abscissa is the real data, and the ordinate is the complete data. As can be seen from the figure, the effect of median filling is similar to that of mean filling. When the real data is small, the complete data is larger. When the real data is large, the complete data is small.

Table 2. Values of each index under median filling

Index	Value
MSE	59.8
RMSE	7.7
MAE	5.4
SMAPE (%)	33.3
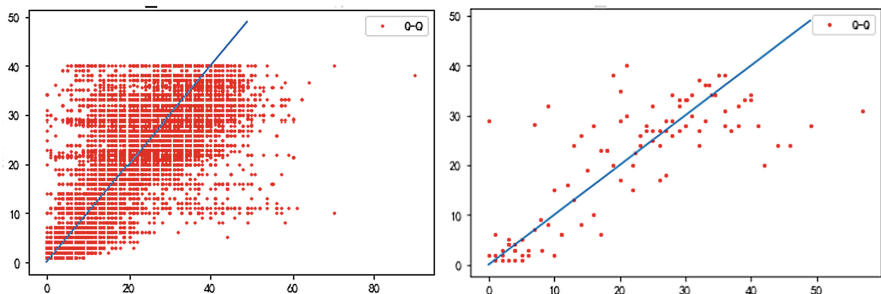
Fig. 3. The figure above is the Q-Q graph of data after mode filling. The abscissa is the real data, and the ordinate is the complete data. It can be found in the figure that, compared with mean filling and median filling, the distribution of some points in this case is farther from the base line, indicating that the effect of mode filling is general.

According to the figure and table, we can find that the filling effect of the median is similar to that of mean filling, and the distribution of scattered points is roughly the same as that of mean filling. The filling effect is best when the real data is between 0–10 and about 30. For the traffic flow data used in this study, the accuracy of median filling is slightly lower than that of mean filling. The value of MSE was 59.8, while SMAPE reached 33.3% (Fig. 3).

Table 3. Values of each index under mode filling

Index	Value
MSE	69.9
RMSE	8.4
MAE	5.8
SMAPE (%)	38.2

From the data in the Table 3, we can see that the effect of mode filling is significantly worse than that of mean filling and median filling. SMAPE is up to 38.2%, about 5 percentage points higher than other filling methods. However, the MSE value reaches 69.9, which is about 10 more than the MSE value of other filling methods, proving that mode filling is not suitable for traffic flow data (Fig. 4).
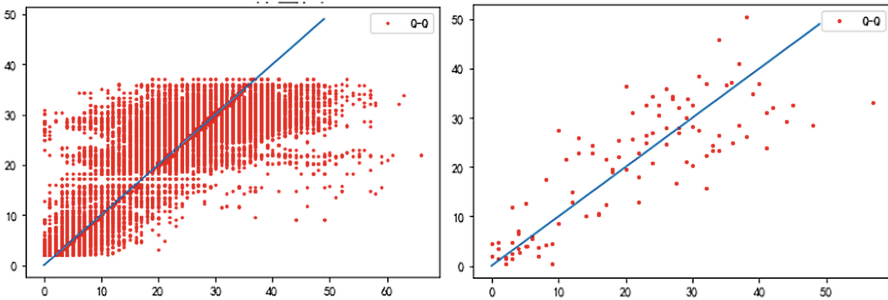
Fig. 4. The figure above is the Q-Q diagram after PCA-EM filling of the data. The abscissa is the real data, and the ordinate is the completed data. As can be seen from the figure, no matter what the value of the real data is, the scattered points are always distributed on both sides of the base line, indicating that the filling effect of PCA-EM is good.

Table 4. Values of each index under PCA-EM filling

Index	Value
MSE	56.7
RMSE	7.5
MAE	5.4
SMAPE (%)	33.0

As can be seen from the figures in the Table 4, PCA-EM filling is the best filling method among the four methods, with a SMAPE value of 33.0% and an MSE value of 56.7, both of which are the least among the four methods. Therefore, according to the comprehensive analysis of this experiment, for the filling of traffic flow data, the

PCA-EM filling has the best effect, the mean filling and the median filling have similar effects, and the mode filling has the worst effect.

5 Conclusion

At present, the research on the problem of missing data has gradually matured, and the processing of missing data is related to various research fields, showing a diversified development. This paper summarizes the research background, causes and types of missing data at home and abroad. All kinds of processing methods are introduced, and filling methods are described in detail. The classical filling methods are compared and summarized, and the latest improved methods of each kind of methods are summarized and compared. At the same time, the commonly used evaluation indexes of data filling effect are introduced from the perspective of parameters and fitting.

References

1. Fisher, R.A., Yates, F.: Statistical Tables: For Biological, Agricultural and Medical Research. Oliver and Boyd (1938)
2. Ma, L., Sun, B., Li, Z.: Bagging likelihood-based belief decision trees. In: 20th International Conference on Information Fusion (FUSION), Xi'an, China, 1–6 (2017). http://ieeexplore.ieee.org/abstract/document/8009664/
3. Geng, R., Sun, B., Ma, L., Zhao, Q., Shen, T.: Anomaly-aware in sequence data based on MSM-H with EXPoSE. In: 40th Chinese Control Conference (CCC 2021), Shanghai, China (2021)
4. Batista, G.E., Monard, M.C.: An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. **17**(5–6), 519–533 (2003)
5. Sun, B., Cheng, W., Ma, L., Goswami, P.: Anomaly-aware traffic prediction based on automated conditional information fusion. In: International Conference on Information Fusion (FUSION), Cambridge, United Kingdom, pp. 2283–2289. IEEE (2018)
6. Leduc, G.: Road traffic data: collection methods and applications. In: Working Papers on Energy, Transport and Climate Change, vol. 1, no. 55, pp. 1–55 (2008)
7. Sun, B., Cheng, W., Bai, G., Goswami, P.: Correcting and complementing freeway traffic accident data using mahalanobis distance based outlier detection. Tehnicki Vjesnik Tech. Gazette **24**(5), 1597–1607 (2017)
8. Scheffer, J.: Dealing with missing data (2002)
9. Lv, Y., Duan, Y., Kang, W., et al.: Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intell. Transp. Syst. **16**(2), 865–873 (2014)
10. Ma, L., Sun, B., Han, C.: Learning decision forest from evidential data: the random training set sampling approach. In: 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, China (2017)
11. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, Hoboken (2019)
12. Sun, B., Cheng, W., Goswami, P., Bai, G.: An overview of parameter and data strategies for K-nearest neighbours based short-term traffic prediction. In: ACM International Conference Proceeding Series 2017, pp. 68–74. ACM (2017)
13. Marlin, B.: Missing Data Problems in Machine Learning (2008)
14. Sun, B., Ma, L., Shen, T., et al.: A robust data-driven method for muti-seasonal and heteroscedastic IoT time series preprocessing. In: Wireless Communications and Mobile Computing (WCMC), p. 6692390 (2021)

15. Yu, L., Snapp, R.R., Ruiz, T., et al.: Probabilistic principal component analysis with expectation maximization (PPCA-EM) facilitates volume classification and estimates the missing data. J. Struct. Biol. **171**(1), 18–30 (2010)
16. Sun, B., Cheng, W., Goswami, P., et al.: Short-term traffic forecasting using self-adjusting k-nearest neighbours. IET Intell. Transp. Syst. **12**(1), 41–48 (2018)

Clustering-XGB Based Dynamic Time Series Prediction

Haoxuan Sun[1], Kun Zhang[2], Tingting Wang[1(✉)], Wanfeng Ma[1(✉)], and Qinjun Zhao[1]

[1] School of EE, University of Jinan, Jinan 250022, China
202021100395@mail.ujn.edu.cn, elephantfff@163.com
[2] Shandong Non-metallic Materials Institute, Jinan 250031, China

Abstract. This work analyzes time series and find the rules and statistical characteristics from the numerous data. According to the purpose of the time series analysis, we find the rules and conduct the future time forecast. This paper is mainly based on the similarity of time series. Based on clustering results, XGB is used to reflect the relationship between similarity and clusters' weights and to predict the value. Overall, it is a time series prediction model based on clustering and XGB regulated weights. The process of model prediction is realized by using instances in dataset, and the relationship between similarity and weights is obtained by using XGB.

Keywords: Time series · KMEANS clustering · XGBoost

1 Introduction

As a highly integrated and integrated application of the new generation of information technology, the Internet has strong penetrating power, strong driving effect and good comprehensive benefits [1–3]. It is the computer, Internet and mobile communication networks after the development of the information industry is another driving force. Logistics industry, as an important part of the internet of things industry chain, with its characteristics of high market maturity, wide market prospects and big investment opportunities, will become a key area of logistics networking industry development in the next few years [4,5].

With the development of big data of artificial intelligence, various fields have been developed to different extent, artificial intelligent products appear more and more in more industries [6]. There have also been rapid developments in the area of transport. Quality time series obtained through bus cards, detectors, cameras, communication equipment, Internet, etc. However, due to the periodicity of time and the influence of noise, how to utilize time series is still a problem to be solved [7,8].

2   Related Conception

2.1   Euclidean Distance

Much of the time of Euclidean equations is also known as euclid distance, which is a familiar distance [9,10]. It is a calculation of the distance between multidimensional vectors, usually defined in terms of distances in m-dimensional vector spaces, and you can think of it as a point in a vector space with a higher dimension. The distance perception between them can be seen as one from this point to the origin, or as the actual distance from one to another [11].

$$distance(X, Y) = \sqrt{\sum (x_{ti} - y_{ti})^2} \quad t = 1, 2, ..., n \tag{1}$$
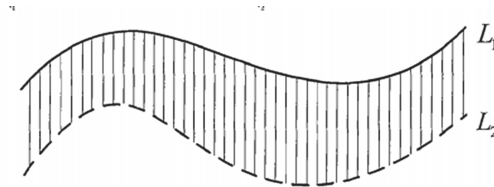
Fig. 1. Time series Euclidean distance schematic diagram

2.2   Clustering

In unsupervised learning, clustering is to train data values without standard classification, reveal the internal rules of the data, and automatically divide the data into similar clusters [12].

KMEANS. In 1967, MacQueene proposed the KMEANS algorithm, one of the simplest and most common clustering methods [13,14]. The similarity of KMEANS is reflected in the distance between samples. The closer you are, the more likeness you have. The degree of similarity directly affects the classification criteria. However, most people use the countdown of distance to express similarities, making the two positively related. Most of the distance is from Europe or Manhattan.

2.3   Evaluation

Evaluation methods can be understood in engineering theory and definitions, just as learning achievement is used to represent a student' s learning performance.

Algorithm 1. Kmeans algorithm

Input: $D = \{x_1, x_2, ..., x_m\}$;

        Cluster number k.

1: Randomly select k samples from D as the initial mean vector $\{\mu_1, \mu_2, ..., \mu_k\}$
2: **repeat**
3:    $C_i = \phi(1 \leq i \leq k)$
4:    **for** $j = 1, 2, .., m$ **do**
5:        Calculate the distance between sample x_j and each mean vector $\mu_i(1 \leq j \leq k)$:
          $d_{ji} = \|x_j - \mu_i\|_2$;
6:        The cluster marker of x_j is determined according to the nearest mean vector
          :$\lambda_j = arg\ min_{i \in \{1,2,...,k\}} d_{ji}$;
7:        Assign the sample x_j to the corresponding cluster: $C_{\lambda_j} = C_{\lambda_j} \bigcup \{x_j\}$;
8:    **end for**
9:    **for** i=1,2,...,k **do**
10:       Calculate the new mean vector$\mu_i' = \frac{1}{|C_i|} \sum_{x \in C_i} \boldsymbol{x}$;
11:       **if** $\mu_i' \neq \mu_i$ **then**
12:          Update the current mean vector μ_i to μ_i'
13:       **else**
14:          Leave the current vector unchanged
15:       **end if**
16:    **end for**
17: **until** None of the current mean vectors have been updated
Output: Cluster partition$C = \{C_1, C_2, ..., C_k\}$

MAE. Mean Absolute Error is referred to as MAE for the purpose of finding the difference between the predicted value and the real value and the Absolute Error [15].

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - y_i'| \tag{2}$$

$\boldsymbol{R^2}$. R-squaredis a relative measure that measures the sum of residuals of existing models far from the standard of the sum of residuals of the benchmark model [16].

$$SSR = \sum_{i=1}^{N} w_i(\overline{y_i} - \widehat{y_i})^2 \tag{3}$$

$$SST = \sum_{i=1}^{N} w_i(\overline{y_i} - y_i)^2 \tag{4}$$

so,

$$R^2 = 1 - \frac{SSR}{SST} \tag{5}$$

MAPE. Mean Absolute Percentage Error (MAPE) is one of the most popular indicators for evaluating predictive performance [17].

$$M = \frac{1}{n} \sum_{t=1}^{n} \frac{A_t - F_t}{A_t} \tag{6}$$

where A_t represents the actual value and F_t is the predicted value.

2.4   Grid Search

Grid Search, a common method, is mainly used for tuning parameters. The model is tested by using different parameters in turn in order to find the parameters corresponding to the best effect. Is to use exhaustive search for hyperparameters, for each case to solve. The super parameters are all the cases of the pre-specified parameters, select from the inside to find the best performance of the set of parameters [18].

2.5   XGBOOST

It can be referred to as XGB, which was developed by Dr. Tianqi Chen from the University of Washington in the United States [19]. The idea of the algorithm is to grow a tree by continuously adding trees and continuously splitting features. Each addition of a tree is actually learning a new function to fit the residual error predicted last time. When we complete the training and get k trees, we need to predict the score of a sample. In fact, according to the characteristics of the sample, it will fall to a corresponding leaf node in each tree, and each leaf node corresponds to a score. Finally, we only need to add up the corresponding scores of each tree to get the predicted value of the sample.

$$\widehat{y} = \Phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \tag{7}$$

$$where\ F = \{f(x) = w_{q(x)}\}(q : R^m \to T, w \in R^T)$$

$w_{q(x)}$ is the fraction of leaf node q, and f(x) is one of the regression trees. Objective function is

$$Obj = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{8}$$

$$where\ \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

The objective function above has two parts, the former is the difference between the real and the predicted, the latter is the regularization term, T is the number of leaf nodes, w is the fraction of leaf nodes. γ controls the number of leaf nodes, λ controls the fraction of leaf nodes will not be too large, can prevent overfitting.

As shown in the following example, two decision trees are trained
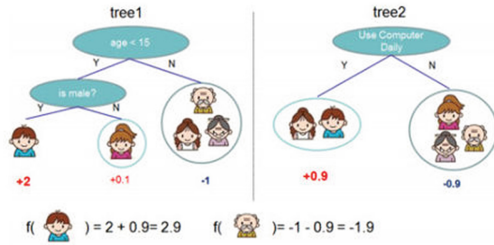
Fig. 2. A simple example of XGB

3 Clustering Prediction and Comparison Based on Multiple Similarity Degrees

3.1 Data Preprocessing

The dodge dataset used in this example has a large number of "−1" values, which means there are many defaults. If missing values are not taken into account, −1 will have a significant impact on the predicted values of the data, making it difficult to process subsequent data.
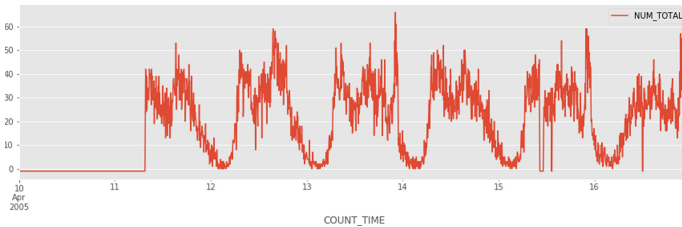
Fig. 3. An overview of several days of raw data

As shown in Fig. 3, in the absence of data for several consecutive days, a large number of consecutive missing values are found in data and a significant number continue to be missing. For example, one day on 10 May 2005. In order to eliminate the disturbance of missing values, combined with the fact that traffic data at traffic junctions is on a one-day cycle, all day data are used to fill in the missing value. In order to minimize changes in the sequence, the time points of missing data are added to the average of the current points and all valid data is grouped. The corresponding values of 288 points of time in the daily HT data are calculated every five minutes to fill the data with the average values in each point in time over a 24-h period.

The average data are rounded to integers to facilitate operation and reduce the data space occupied.

3.2   Generating Time Series

Because time series represent different time periods, a similar time period between different dates is selected as data for the data clustering. Experiments show that the length of half a day is suitable for the size of the search step. 1248 Because it is a 40-min phased step over eight steps, which is usually considered a medium-term forecast rather than a short-term one. Half-day traffic data in a five-minute group selects 144 data as search steps in the time series, which is expressed as

$$x_i = \{t_i, t_{i+1}, \ldots, t_{i+n}\} \quad n \text{ is the search step size} \tag{9}$$

In addition, the next period of time series is selected for prediction, and the data with the step length of 1, 2, 4, and 8 are selected for prediction, which is expressed as

$$y_i = \{t_{i+n+1}, t_{i+n+2}, t_{i+n+3}, t_{i+n+4}\} \quad n \text{ is the search step size} \tag{10}$$

The sequence of datasets for the production experiment is shown in Fig. 4.

var1(t-143)	var1(t-142)	var1(t-141)	var1(t-140)	var1(t-139)	var1(t-138)	var1(t-137)	var1(t-136)	var1(t-135)	var1(t-134)	...	var1(t-5)	var1(t-4)	var1(t-3)	var1(t-2)	var1(t-1)	var1(t)	var1(t+1)	var1(t+2)	var1(t+4)	var1(t+8)
8	8	7	7	7	6	6	6	6	5	...	24	25	26	26	26	25	24	24	25	25
8	7	7	7	6	6	6	6	6	5	...	25	26	26	26	25	24	24	25	25	26
7	7	7	6	6	6	6	5	5	5	...	26	26	26	25	24	24	25	25	25	25
7	7	6	6	6	6	5	5	5	5	...	26	26	25	24	24	25	25	25	25	25
7	6	6	6	6	5	5	5	5	5	...	26	25	24	24	25	25	25	25	25	24
6	6	6	6	5	5	5	5	5	4	...	25	24	24	25	25	25	25	25	26	24
6	6	6	5	5	5	5	5	4	4	...	24	24	25	25	25	25	25	25	25	24
6	6	5	5	5	5	5	4	4	4	...	24	25	25	25	25	25	25	26	25	26
6	5	5	5	5	5	4	4	4	5	...	25	25	25	25	25	25	26	25	24	25
5	5	5	5	5	4	4	4	5	4	...	25	25	25	25	25	26	25	25	24	25

Fig. 4. A sequence of total data sets

In order to have a better generation effect, the data were divided into test set and training set, which were different from each other. The data was broken up to simulate the real data prediction results. The train_test_split method was used to take 25% of the whole data as the predicted data.

3.3   Clustering Similarity

Clustering of Euclidean Distance. The Kmeas algorithm used in this section is data clustering. Under the condition of unsupervised learning, the algorithm pays more attention to the similarity of sequences, and obtains the distribution of time series roughly. In the Kmeas algorithm, most people use Euclidean distance and Manhattan distance to represent the similarity measure of time series.

```
array([0.12571429, 0.10857143, 0.10857143, 0.18285714, 0.17142857,
       0.21142857, 0.09142857])
```

Fig. 5. Cluster number distribution of distance

Classification. As can be seen, the following figure shows the distribution of the predicted data according to the distance based clustering method is shown in the figure below

As can be seen from the Fig. 5, the main purpose of using the test set is to simulate the effect of the algorithm in the actual scene, showing whether there is a good generalization effect and then better application of improvement and optimization in the real situation.

The result of clustering analysis shows that the clustering data has changed obviously. Taking into account only the morphological characteristics of traffic flow time series and the original purpose of clustering them for seven days, the values of each class should tend to be averaged.

Similarity. After clustering all traffic flow time series, we can learn about grouping and find the average of the characteristics and tags of each cluster to represent the total value in the current cluster. The similarity between the predicted sequence and the seven clusters can be calculated. A distance matrix can be generated from the formula 1. The following is the distance of the sequence

```
array([[198.38870561, 312.30204145, 147.5663145 , 221.7634799 ,
        132.98371663, 179.37599818, 275.94923531]])
```

Fig. 6. Distance between sequence and each cluster

Knowing the similarity, we can use the relationship between the similarity to calculate the weight W, which is used to predict the data value. That is to say, there is a mapping relationship between similarity and weight.In order to better show the mapping relationship between similarity and distance, the method of XGB will be used in this paper.

3.4   XGB's Method to Calculate the Weight

In fact, since we do not know the exact relationship between similarity and weight, the relationship between similarity and weight of each category may not be the positive correlation that we think the greater the similarity is, the higher the weight is.

Therefore, the method of XGB model was chosen in order to better fit the relationship between the two. As long as the distance between each category is required, the weight of each category can be obtained, and the predicted value can be obtained directly.

Construction of New Data Set. Calculate the distance of 7 categories of data, and sort them from small to large. These 7 clusters are taken as features of the training data. Label is the weight of each of the seven clusters.

Table 1. New data set

dist1	dist2	...	dist7	w_1	w_2	...	w_7
	...		... ...		...	...	...

In the above experiment, after the similarity of a sample data is obtained, the distance between the predicted sequence and 7 clusters is sorted, and the set parameters are determined by Grid Search, and the sum of the 7 weight parameters is 1. The parameter with the least difference between the predicted value and the actual value, that is, the one with the best effect, is selected as the label of the current data. The distance of seven classes and the parameter of weight W are combined as a sample data.

After all the original test set data processing is completed, the data set of Table 1 is constructed. It is used to calculate the weight directly. 80% of the data are divided into training model XGB, and the rest are tested to see whether the model is good or bad.

Prediction of XGB. The training set of the data set is put into the training model, and the test set is used to evaluate the model and check the effect of the model.

Because the predicted weight has no constraint that the sum is 1, it is necessary to normalize the predicted samples, set one well, and express the relationship with each category.

Analysis of the Model. Since the size of the XGB data set will be much smaller than the previous data set, it is possible that the prediction of the data will be less effective.

However, the XGB method has its own advantages, which can more accurately get the appropriate weight, and get the proportion relationship between each category and the predicted value. Forecasts are much more accurate. The XGB model preprocessing and data storage after each step will not consume too much time and space, which is about the same as the Euclidian distance. But if the use of DTW, enough time, there will be a better effect. Due to time constraints, only XGB is used for comparative analysis of Euclidean distance here.

The following figure shows the weights of the predicted results.

```
array([[0.09259243, 0.503311  , 0.09004612, ..., 0.1951583 , 0.02927206,
        0.1178335 ],
       [0.16224124, 0.2085919 , 0.15434733, ..., 0.10209591, 0.08707408,
        0.08930299],
       [0.16820015, 0.17057562, 0.14777775, ..., 0.08558471, 0.11823782,
        0.17418884],
       ...,
       [0.2544834 , 0.23196718, 0.14363924, ..., 0.08511011, 0.07576442,
        0.1274816 ],
       [0.11999956, 0.11634909, 0.14853086, ..., 0.5316235 , 0.101652  ,
        0.02974296],
       [0.08982415, 0.1501841 , 0.20426954, ..., 0.16044119, 0.03047957,
        0.0235962 ]], dtype=float32)
```

Fig. 7. Weight of prediction

Due to the advantages of XGB itself, the accuracy of the prediction data is much better. So, if you have a larger amount of data, it will work better for the model store, and XGB has a loss function to prevent overfitting.

| 9.97287936201832 | −2.6165895439404885 | 0.9350044508417967 |

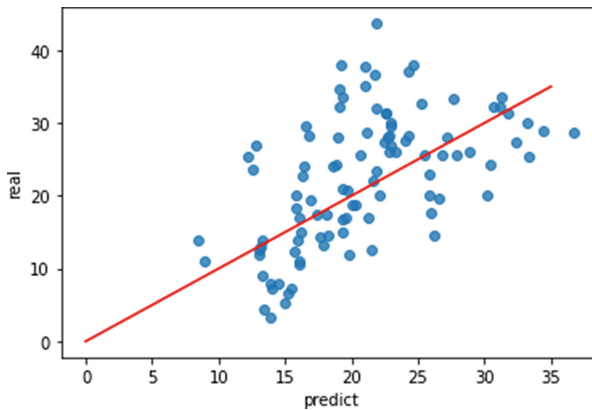**Fig. 8.** MAE,R2,MAPE

The Q-Q diagram of XGB is as follows

Fig. 9. XGB's Q-Q figure

Even if the data set is relatively small, the effect is still good, indicating the feasibility of this method is relatively strong.

4   Conclusion and Prospect

Euclidean distance will still be widely used in practice as a similarity, and XGB can be used if necessary, because the grid search will make the parameters more reasonable, and it will have better prediction and generalization effect.

In 2019, the total investment scale of smart transportation in China exceeded 227.8 billion yuan. Even under the influence of the epidemic, the transportation investment of 10 million yuan project in the first quarter of 2020 increased by 15% year on year.

As a highly intensive integration and comprehensive application of a new generation of technology, the Internet of Things has the characteristics of high knowledge intensity, wide application range, great growth potential, strong driving force and good comprehensive benefits. China's 12th Five-Year Plan has accumulated nine fields of projects: smart industry, smart agriculture, smart logistics, smart transportation, smart power grid, smart environmental protection, smart security, smart medical care, and smart home. These areas of concentrated development almost cover all aspects of our social production and life.

In particular, with the improvement of machine learning and deep learning algorithm accuracy, the improvement of computing power and the decrease of cost, the cost of intelligent transportation is constantly reduced and the advantages of scale are constantly emerging, which greatly promote the implementation and application of the algorithms listed in this paper.

References

1. Weiwei, W.A.N.G., Xinghua, S.H.A.N.: Study on regular pattern of railway passener flow in three-daw holiday based on clustering method of time series. Railw. Comput. Appl. **04**, 23–27 (2015)
2. Geng, R., Sun, B., Ma, L., Zhao, Q., Shen, T.: Anomaly-aware in sequence data based on MSM-H with EXPoSE. In: 40th Chinese Control Conference (CCC 2021), Shanghai, China (2021)
3. Sun, B., Cheng, W., Goswami, P., Bai, G.: Short-term traffic forecasting using self-adjusting k-nearest neighbours. IET Intell. Transp. Syst. **12**(1), 41–48 (2018)
4. Ji, M., Xiao, L.: A dynamic k-means clustering algorithm for time series data. Comput. Digit. Eng. **48**(8), 1852–1857 (2020). https://doi.org/10.3969/j.issn.1672-9722.2020.08.007
5. Ma, L., Sun, B., Ziyi, L.: Bagging likelihood-based belief decision trees. In: 20th International Conference on Information Fusion (FUSION). Xi-An, China, pp. 1–6 (2017). http://ieeexplore.ieee.org/abstract/document/8009664/ http://ieeexplore.ieee.org/abstract/document/8009664/
6. Sun, B., Wei, C., Liyao, M., Prashant, G.: Anomaly-aware traffic prediction based on automated conditional information fusion. In: International Conference on Information Fusion (FUSION), Cambridge, UK, pp. 2283–2289. IEEE (2018)
7. Lin, Q.: Research on Feature Screening and Clustering Analysis of Time Series Data - A Case Study of the CSI 300 Index. Southwestern University of Finance and Economics (2017)

8. Sun, B., Cheng, W., Goswami, P., Bai, G.: An overview of parameter and data strategies for K-nearest neighbours based short-term traffic prediction. In: ACM International Conference Proceeding Series, pp. 68–74. ACM (2017)
9. Zhang, G.: Research and Application on Interval Time Series Clustering Based on DTW. Northwest Normal University (2020)
10. Ma, L., Sun, B., Han, C.: Learning decision forest from evidential data: the random training set sampling approach. In: 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, China (2017)
11. Chen, H., Liu, C., Sun, B.: Survey on similarity measurement of time series data mining. Control Decision **32**(001), 1–11 (2017)
12. Sun, B., Ma, L., Shen, T., et al.: A robust data-driven method for multi-seasonal and heteroscedastic IoT time series preprocessing. Wirel. Commun. Mobile Comput. 6692390 (2021)
13. Lai, Y.: Study on Real-Time Prediction of Arrival Time for Floating Transit Vehicle. Chongqing University (2011)
14. Sun, B., Cheng, W., Bai, G., Goswami, P.: Correcting and complementing freeway traffic accident data using mahalanobis distance based outlier detection. Tehn. Vjesn. Techn. Gazette **24**(5), 1597–1607 (2017)
15. Lyu, Z.: Price Forecast and Comparative Study of Stock Index Futures Based on Machine Learning Algorithms. Zhejiang University (2020)
16. Sun, B., Cheng, W., Goswami, P., Bai, G.: Short-term traffic forecasting using self-adjusting k-nearest neighbours. IET Intell. Transp. Syst. **12**(1), 41–48 (2018). https://doi.org/10.1049/iet-its.2016.0263
17. Jiang, D., Pei, J., Zhang, A.: DHC: a density-based hierarchical clustering method for time series gene expression data. BIBE 393–400 (2003)
18. Ashish, S., Dale, E.: Clustering for multivariate time series data. In: Proceeding of the American Control Conference Anehorage, May, 2002, pp. 586–591 (2002)
19. Zheng, C.Z.L.: Shape clustering on time series data. In: Proceedings of Information Technology and Environmental System Sciences (ITESS), vol. 3, pp. 1249–1253 (2008)

Time Series Data Imputation Using Expectation-Maximization with Principal Component Analysis

Renkang Geng, Jing Cao[(⊠)], Qinjun Zhao[(⊠)], and Yujie Wang

School of Electrical Engineering, University of Jinan, Jinan 250022, Shandong, China
caojingjn@163.com, cse_zhaoqj@ujn.edu.cn

Abstract. Data quality is the basis of data analysis and determines the effect and depth of data analysis. Missing values are a very important factor affecting data quality. Since machine learning algorithms have been used to process data, the processing of missing values has become an important field of machine learning. For the vast majority of data, the first step of data analysis is often to complete the missing values of data. With the increasing complexity of current social traffic conditions and the increasingly serious urban road congestion, traffic data, as the most intuitive reflection of urban road conditions, has great application value and application potential. The quality of traffic data is directly related to whether we can accurately predict the traffic flow and judge the road traffic conditions so as to effectively govern the urban traffic problems. Therefore, it is very important to choose appropriate algorithms and methods to fill in the missing values in traffic data quickly and accurately. In this paper, taking traffic flow data as an example, we create different missing value ratios for the data and use the PCA-EM algorithm to fill in the missing value. Through the experimental results, we have a preliminary evaluation of the comprehensive performance of the PCA-EM algorithm when it is used to fill in the missing value of traffic data.

Keywords: Traffic flow data · Missing value · PCA-EM

1 Introduction

It is of great significance to study traffic data for improving urban congestion and reducing traffic anomalies [1, 2]. If the traffic data can be accurate processing and analysed, can effectively predict the future traffic flow, and carries on the anomaly detection of a traffic incident, so that the relevant traffic department, the traffic in an orderly, the guidance of the timely handling of traffic accidents, channel distribution vehicles, improve urban congestion and facilitate public travel, enhance people's quality of life, To ensure the safety of citizens [3, 4]. For the analysis of traffic data, the key step lies in whether the missing values can be accurately filled, which seriously affects the accuracy of our analysis [5].

Since the beginning of the last century, people have begun to pay attention to the problem of data quality, so the lack of data has deeply affected many ongoing studies. Lack of data will affect task analysis and make it more difficult, resulting in deviation of results, which will greatly reduce the work efficiency of statistical work [6]. Especially when there is a systematic difference between complete observation and incomplete observation, there will be great limitations and deviations in the conclusions made by using mathematical methods and conventional statistical methods on incomplete data sets [7, 8]. Such structure cannot replace the original data. The progress of time series missing value preprocessing methods enables us to use better methods for research. Missing data filling is one of the important research contents in data clarity and data mining [9].

The generation of data missing value will cause problems in many ways, for data development and in the field of data mining aspects of the work, the existence of data missing value, will produce a lot of influence, data sets to produce missing value can cause a lot of useful information is missing [10, 11], and some programs only in view of the complete data sets, which reduces the work efficiency, also, for instance, For KNN algorithm, it is very excellent in terms of classification accuracy, and the accuracy of decision tree classification is much higher than that of the largest category classification. However, missing values have a very important impact on the classification ability of KNN [12, 13]. In the KNN algorithm, data sets containing missing values are classified, and the classification accuracy is completely dependent on the size of the missing proportion.

The larger the missing proportion, the smaller the classification accuracy. If the data is almost completely missing [14], the classification accuracy can reach 0. The uncertainty in the data set may cause confusion of the data distribution and the reduction of the uncertainty in the data set. This situation will be more obvious, and the certainty in the data set will be more difficult to grasp. The inclusion of missing values in a data set can cause confusion in the data mining process and add a lot of unreliability to the data output [15, 16].

2   Related Work

Due to the different reasons for data missing, there are different types of data missing. In order to deal with all kinds of data missing problems more conveniently, it is necessary to classify data missing. In this way, we can find out the solution to the problem more efficiently and specifically, so that our data filling is closer to the real value and the result is more accurate [17, 18]. We can classify the missing data according to the relationship between the complete data and the missing data. This classification method is called missing mechanism classification, which is generally divided into three categories: 1) Completely random missing [19]: Completely random loss refers to the distribution of the missing value in a data set is random, no rules to follow, this has to do with his own properties, and there is no relationship between adjacent data, for example, when we study student's result, the lack of a course grade has nothing to do with other results, so we can't find a rule to fill the missing value from other grades. 2) Random missing: Random deletion index data deletion is only related to the value of the complete attribute.

3) Non-random missing: this kind of missing data is related to its own attributes and other data in the data set. Generally speaking, such missing data cannot be deleted directly [20].

Probabilistic PCA-EM filling [21]: Probabilistic PCA-EM is an algorithm that uses a variational EM algorithm to process missing data. Essentially, it is a method that considers the probability distribution of each variable. After determining the probability function of the principal element and error, the model is established by the EM algorithm. Firstly, the implicit variable X and the probability distribution of the observed data were obtained by standard centralized processing of the original training sample data. Then, the expectation was calculated and the maximum likelihood estimate of the hidden variable was calculated by using the existing estimate of the hidden variable. Then maximize the maximum likelihood value to calculate the value of the parameter. The objective function of the Probabilistic PCA-EM algorithm is the lower bound of the logarithmic likelihood of data. In high dimensional space, the computation required for each iteration of the PCA-EM algorithm is much smaller than that of traditional algorithms.

3   Methods

The specific steps of this experiment are as follows: First reads raw traffic data and will be treated as raw data branch according to each day, each column data corresponding to the same period under different days of traffic flow data, that is to say, each column data into a set of common attributes of the data, then we can make different missing value in each column proportion, finally, use PCA - EM algorithm to fill each column of missing value.

In this experiment, we used the control variable method to create missing values of different proportions for the same data set and then used the PCA-EM algorithm to fill in the missing values. Finally, the filling effect was compared and evaluated. In this way, we can effectively compare the filling effect of the PCA-EM algorithm when the proportion of missing values is different, and get the best filling effect of the PCA-EM algorithm when facing the traffic flow data under what condition the missing values are in. In this experiment, the missing values artificially created accounted for 10%, 30%, 50%, 70% and 90%, respectively.

In order to compare the distribution of real data and filled data, we draw a Q-Q graph and draw the baseline of Y = X. The essence of the Q-Q graph is that, firstly, real data and filled data are arranged according to the quantile and corresponding real data and a filled data under the same quantile are taken to draw a coordinate point as the coordinate (x, y), and then the whole Q-Q graph is drawn through iteration. In addition to the Q-Q diagram, in order to accurately quantify the relationship between real data and filled data, we also used MSE, RMSE, MAE, SMAPE and other indicators to intuitively analyze the effect of completion. The indices are formulated as follows:

$$MSE = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \hat{y}_i \right)^2$$

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(y_i - \hat{y}_i \right)^2}$$

$$MAE = \frac{1}{m} \sum_{i=1}^{m} \left| (y_i - \hat{y}_i) \right|$$

$$SMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{\left| \hat{y}_i - y_i \right|}{\left(\left| \hat{y}_i \right| + |y_i| \right)/2}$$

4   Results
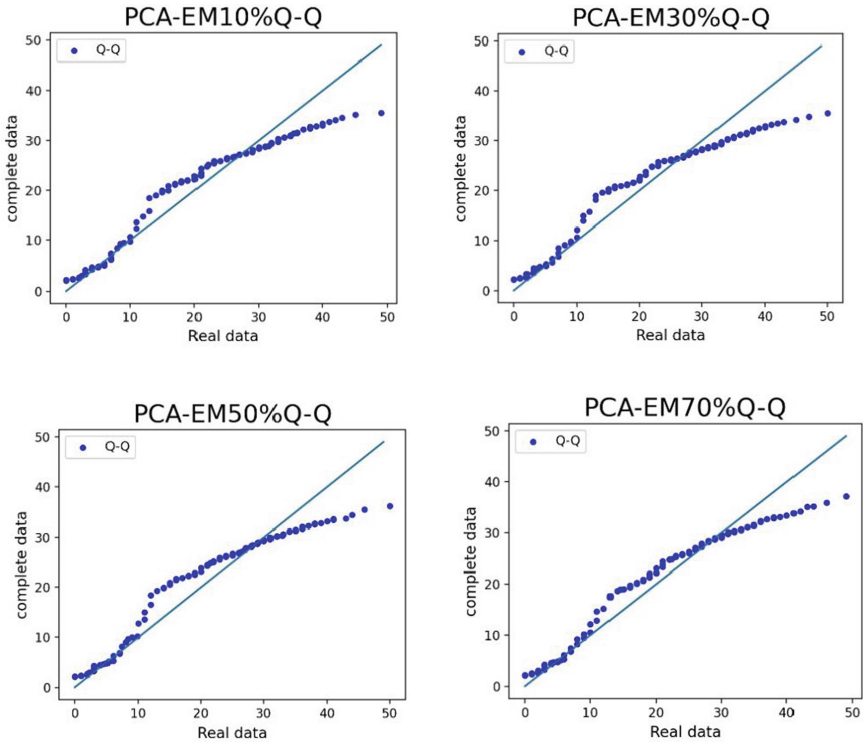
Fig. 1. The figure shows the Q-Q diagram completed by the PCA-EM algorithm after we have created different missing value ratios for the data. The abscissa is the real data, and the ordinate is the completed data. It can be seen that the complete effects in the four cases are generally similar. When the real data is less than 25, the complete data is larger than the real data, while when the real data is greater than 25, the complete data is smaller.

Table 1. Under different missing value ratio, the value of each indicator

Index	10%	30%	50%	70%
MSE	56.6	56.7	56.0	57.0
RMSE	7.5	7.5	7.5	7.5
MAE	5.4	5.4	5.3	5.4
SMAPE (%)	33.4	33.0	33.2	33.7

From all the data in the table, we can find that when the missing value accounts for 30%, the PCA-EM algorithm has the best filling effect, and SMAPE is 33.0%. In the numerical analysis of MSE, when the missing value accounted for 50%, the value of MSE was the lowest, only 56.0. In general, the overall filling effect of the PCA-EM algorithm is good. In the case of different missing values, the filling accuracy of the PCA-EM algorithm is slightly different, and the difference is small, indicating that the overall optimization of the PCA-EM algorithm is good when it is used to fill the traffic flow data. We finally proved that the PCA-EM algorithm is most suitable for filling traffic flow data with missing values accounting for about 30% (Table 1 and Fig. 1).

5 Conclusion

In this paper, we first made a general introduction to data missing and summarized several categories of data missing. Then it describes the influence of missing values in traffic data on data analysis, and the importance of studying the missing values of traffic data for improving urban congestion and reducing traffic accidents. During the experiment, we selected the traffic flow data and adjusted the proportion of missing values. PCA-EM method was used to fill the traffic flow data with different proportion of missing values successively, and the corresponding filling effect was obtained. Finally, we have a preliminary evaluation of the comprehensive performance of the PCA-EM algorithm when it is used to fill in the missing values of traffic data. In the future, I will continue to make further research in this field.

References

1. Jie, L., Van Zuylen, H.J.: Road traffic in China. Procedia Soc. Behav. Sci. **111**, 107–116 (2014)
2. Liu, Z., Yue, X., Zhao, R.: The cause of urban traffic congestion and countermeasures in China. Urban Stud. **11**, 90–96 (2011)
3. Sun, B., Cheng, W., Goswami, P., Bai, G.: An overview of parameter and data strategies for k-nearest neighbours based short-term traffic prediction. In: ACM International Conference Proceeding Series 2017, pp. 68–74. ACM (2017)
4. Sun, B., Cheng, W., Bai, G., Goswami, P.: Correcting and complementing freeway traffic accident data using Mahalanobis distance based outlier detection. Tehnicki Vjesnik-Technical Gazette **24**(5), 1597–1607 (2017)

5. Wen, H., Sun, J., Zhang, X.: Study on traffic congestion patterns of large city in China taking Beijing as an example. Procedia Soc. Behav. Sci. **138**, 482–491 (2014)
6. Li, J., Walker, J.L., Srinivasan, S., et al.: Modeling private car ownership in China: investigation of urban form impact across megacities. Transp. Res. Rec. **2193**(1), 76–84 (2010)
7. Sun, B., Ma, L., Shen, T., et al.: A robust data-driven method for muti-seasonal and heteroscedastic IoT time series preprocessing. In: Wireless Communications and Mobile Computing (WCMC), p. 6692390 (2021)
8. Ma, L., Sun, B., Han, C.: Learning decision forest from evidential data: the random training set sampling approach. In: 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, China (2017)
9. Bramer, M.: Principles of Data Mining. Springer, London (2007). https://doi.org/10.1007/978-1-84628-766-4
10. Scheffer, J.: Dealing with missing data (2002)
11. Ahmed, M.S., Cook, A.R.: Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques (1979)
12. Ma, L., Sun, B., Li, Z.: Bagging likelihood-based belief decision trees. In: 20th International Conference on Information Fusion (FUSION), Xi-An, China, pp. 1–6 (2017). http://ieeexplore.ieee.org/abstract/document/8009664/
13. Geng, R., Sun, B., Ma, L., Zhao, Q., Shen, T.: Anomaly-aware in sequence data based on MSM-H with EXPoSE. In: 40th Chinese Control Conference (CCC 2021), Shanghai, China (2021)
14. Zhang, S., Li, X., Zong, M., et al.: Learning k for kNN classification. ACM Trans. Intell. Syst. Technol. (TIST) **8**(3), 1–19 (2017)
15. Liu, P., Lei, L., Wu, N.: A quantitative study of the effect of missing data in classifiers. In: The Fifth International Conference on Computer and Information Technology (CIT 2005), pp. 28–33. IEEE (2005)
16. Sun, B., Cheng, W., Goswami, P., et al.: Short-term traffic forecasting using self-adjusting k-nearest neighbours. IET Intel. Transp. Syst. **12**(1), 41–48 (2018)
17. Sun, B., Cheng, W., Ma, L., Goswami, P.: Anomaly-aware traffic prediction based on automated conditional information fusion. In: International Conference on Information Fusion (FUSION), Cambridge, United Kingdom, pp. 2283–2289. IEEE (2018)
18. Zhang, S., Zhang, C., Yang, Q.: Data preparation for data mining. Appl. Artif. Intell. **17**(5–6), 375–381 (2003)
19. Bhaskaran, K., Smeeth, L.: What is the difference between missing completely at random and missing at random? Int. J. Epidemiol. **43**(4), 1336–1339 (2014)
20. Marlin, B.M., Zemel, R.S.: Collaborative prediction and ranking with non-random missing data. In: Proceedings of the Third ACM Conference on Recommender Systems, pp. 5–12 (2009)
21. Yu, L., Snapp, R.R., Ruiz, T., et al.: Probabilistic principal component analysis with expectation maximization (PPCA-EM) facilitates volume classification and estimates the missing data. J. Struct. Biol. **171**(1), 18–30 (2010)

Time Series Prediction with Preprocessing and Clustering

Haoxuan Sun[1], Shuai Lin[2], Lin Han[1(✉)], Jidong Feng[1(✉)], and Mingxu Sun[1]

[1] School of EE, University of Jinan, Jinan 250022, China
201921200627@mail.ujn.edu.cn, fjd_844575264@163.com
[2] Shandong Non-metallic Materials Institute, Jinan 250031, China

Abstract. This paper studies the similarity of time series, and studies the influence of weight on prediction results on the basis of clustering. We first introduce the practical significance and research purpose of the selected topic, summarizes the current research situation at home and abroad, and summarizes the research content of this paper. Second, we describe related concepts. Later, based on Dodger data set, we study the flow of total prediction data of time series. First of all, feature extraction of the data, pre-processing work, the original data generation time series. Then the data are processed and divided into training data and test data for the convenience of subsequent processing. Then the clustering algorithm was used to divide the time series into categories, and seven categories were divided according to the characteristics of one week time cycle. The average value of each category is calculated to replace the characteristics of the current category, and then the similarity is compared. Finally, the weight of each category is calculated by similarity degree, and then the data is predicted. MAE, R-squared, MAPE and other indicators were used to analyze and evaluate the forecast data.

Keywords: Time series · Kmeans clustering · Similarity

1 Introduction

With the development of artificial intelligence with big data, various fields have developed to varying degrees [1], and more and more artificial intelligence products appear in more industries [2]. In the field of transportation, there is also rapid development. Mass time series obtained through bus cards, detectors, cameras, communication equipment, the Internet, etc. [3–5]. However, due to periodicity and high noise, how to make use of time series is still an unresolved problem [6].

As an important part of the industry chain of the Internet of Things, the transportation Internet of Things has the characteristics of high market maturity, mature industry sensing technology, and strong government support [7]. In the slogan of building "digital city" and "smart city", intelligent transportation

system has been applied on a large scale in many cities, with broad market prospects and huge investment opportunities [8,9]. It will become a key area for the development of the Internet of Things industry in the next few years.

2 Related Conception

2.1 Euclidean Distance

The Euclidean metric (most of the time also called the Euclidean distance) is a very familiar distance [10]. It's a calculation of distances between multidimensional vectors usually defined in terms of distances in m-dimensional vector space, you can think of it as a point in a higher dimensional vector space, and the distance between them can be thought of as the distance from this point to the origin, or it can be the actual distance from this point to another point [11] (Fig. 1).

$$distance(X, Y) = \sqrt{\sum (x_{ti} - y_{ti})^2} \quad t = 1, 2, ..., n \tag{1}$$
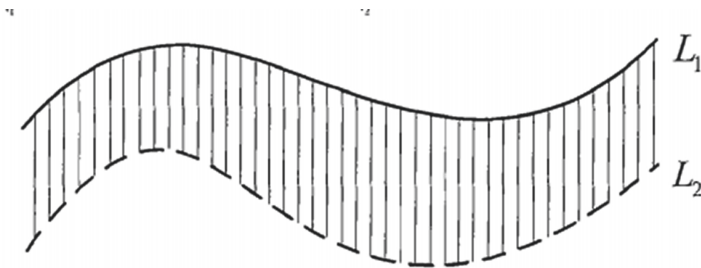
Fig. 1. Time series Euclidean distance schematic diagram

2.2 Softmax

Softmax function in artificial neural network, the activation function used mostly in this, is the sigmoid multivariate versions. Softmax can be thought of as the operation of Arg Max, mainly used for activation functions, which are also approximations of smoothing [12].

Given a one-dimensional vector, Softmax function maps it to a probability distribution. The softmax function $R^n \rightarrow R^n$ is defined by the following formula

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}, for \ 1, ..., n \ and \ x = [x_1, ..., x_n]^T \in R^n \tag{2}$$

2.3   Clustering

In unsupervised learning, clustering is to train data values without standard classification, reveal the internal rules of the data, and automatically divide the data into similar categories [13,14].

   The purpose of clustering is to classify many unrelated subsets in the data set, that is, to divide the data samples into different clusters. Clustering can be a separate process to find the rules of the data set. Clustering is also an essential part of the learning task, providing data for subsequent work.

KMEANS. In 1967, MacQueenE proposed the KMEANS algorithm, which is the most common clustering method and also one of the simplest algorithms. The similarity of KMEANS is represented by the distance between samples. The closer the distance, the higher the similarity will be [15]. Most people will use the reciprocal of distance to express the similarity, so that the two become a positive correlation. And the distance mostly choose European distance or Manhattan distance.

2.4   Evaluation

The evaluation method, just like using academic performance to represent the good or bad learning conditions of students, can be understood in the engineering theory and definition, in order to better optimize the algorithm.

MAE. Mean Absolute Error is referred to as MAE for the purpose of finding the difference between the predicted value and the real value and the Absolute Error [16]. MAE can be obtained by averaging its values.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - y_i'| \tag{3}$$

MAPE. Mean Absolute Percentage Error (MAPE) is one of the most popular indicators for evaluating predictive performance [17].

$$M = \frac{1}{n} \sum_{t=1}^{n} \frac{A_t - F_t}{A_t} \tag{4}$$

where A_t represents the actual value and F_t is the predicted value.

3   Clustering with Multiple Similarity Measurements

3.1   Data Preprocessing

The Dodger data set used in this example has a large number of '−1' values, which means that there are many missing values. If missing values are not taken
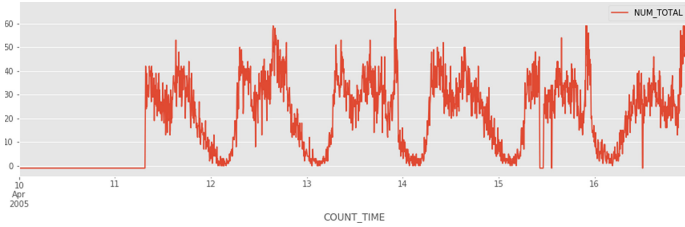
Fig. 2. An overview of several days of raw data

into account, '−1' will have a great impact on the predicted value of the data, making it very difficult to process the subsequent data.

Figure 2 shows a large number of consecutive missing values, with no data for several consecutive days. For example, one day on May 10, 2005. In order to restore the trend of time series as much as possible, all valid data are grouped and the missing data is filled into the average value of the current time point. The corresponding values of 288 time points a day were calculated (data were obtained every 5 min), and the average value of each time point within 24 h was taken.

	time	NUM_TOTAL
0	0.00	8
1	0.05	8
2	0.10	7
3	0.15	7
4	0.20	7
5	0.25	6
6	0.30	6
7	0.35	6
8	0.40	6
9	0.45	5
10	0.50	5

Fig. 3. Filled data

Rounding the average value of the data to integers to facilitate calculation and reduce the data footprint. As shown in Fig. 3. This shows the trend of the day, as well as the change of the vehicle over the day. In this paper, the characteristics of traffic flow time, using data analysis of the relevant content of data processing. Due to the large amount of data used, Pandas library is used to assist the calculation.

3.2   Reconstruct Time Series

Since the resulting data is not a time series but a timestamp, permutation and combination are required to generate the time series. Since the time period of each day represented by each time series is different, the similar time period between different days is selected as the data clustering. Through the experiment, it is found that the length of half a day as the search step is appropriate. 1248 is because more than 8 steps (40 min) is considered to be a medium term forecast rather than a short term forecast. Half a day's traffic flow data (144 pieces of data in 5 min for each group) were selected as the search step of the time series, and expressed as

$$x_i = \{t_i, t_{i+1}, \ldots, t_{i+n}\} \quad n = 143 \tag{5}$$

In addition, the next period of time series is selected for prediction, and the data with the step length of 1, 2, 4, and 8 are selected for prediction, which is expressed as

$$y_i = \{t_{i+n+1}, t_{i+n+2}, t_{i+n+3}, t_{i+n+4}\} \quad n = 143 \tag{6}$$

With the help of Python's Numpy and Pandas library, the sequence of datasets for the production experiment is shown in Fig. 4 below

var1(t-143)	var1(t-142)	var1(t-141)	var1(t-140)	var1(t-139)	var1(t-138)	var1(t-137)	var1(t-136)	var1(t-135)	var1(t-134)	...	var1(t-5)	var1(t-4)	var1(t-3)	var1(t-2)	var1(t-1)	var1(t)	var1(t+1)	var1(t+2)	var1(t+4)	var1(t+8)
8	8	7	7	7	6	6	6	6	5	...	24	25	26	26	26	25	24	24	25	25
8	7	7	7	6	6	6	6	5	5	...	25	26	26	26	25	24	24	25	25	26
7	7	7	6	6	6	6	5	5	5	...	26	26	26	25	24	24	25	25	25	25
7	7	6	6	6	6	5	5	5	5	...	26	26	25	24	24	25	25	25	25	25
7	6	6	6	6	5	5	5	5	5	...	26	25	24	24	25	25	25	25	25	24
6	6	6	6	5	5	5	5	5	4	...	25	24	24	25	25	25	25	25	26	24
6	6	6	5	5	5	5	5	4	4	...	24	24	25	25	25	25	25	25	25	24
6	6	5	5	5	5	5	4	4	4	...	24	25	25	25	25	25	25	26	25	26
6	5	5	5	5	5	4	4	4	5	...	25	25	25	25	25	25	26	25	24	25
5	5	5	5	5	4	4	4	5	4	...	25	25	25	25	25	26	25	25	24	25

Fig. 4. A sequence of total data sets

In order to have a better generation effect, the data were divided into test set and training set, which were different from each other, and the order was random. The data was broken up to simulate the real data prediction results. The train_test_split method was used to take 25% of the whole data as the predicted data.

3.3   Clustering Similarity

Clustering with Euclidean Distance. The Kmeans algorithm used in this section is data clustering, which is a partition-based clustering method. Since you are using a time series as your data, you need to find similar sequences. Clustering analysis can be used to cluster time series to a certain extent. In the

case of unsupervised learning, the algorithm will pay more attention to sequence similarity for sequence classification, and then roughly obtain the distribution of time series. Therefore, the method to express the similarity measure between samples is very important. In the Kmeans algorithm, the Euclidean distance and Manhattan distance used by most people represent the similarity measure of time series. But the Euclidean distance is more common. But the limitation of Euclidean distance is that one dimension is much bigger than the others, and the distance is excessively affected by this dimension.

Python's Sklearn library can use the Kmeans algorithm, but the default Euclidean distance is not easy to modify.

Classification. It can be seen that the clustering methods based on these distances have some differences in the distribution of classification of prediction data, as shown in the figure below

```
array([0.12571429, 0.10857143, 0.10857143, 0.18285714, 0.17142857,
       0.21142857, 0.09142857])
```

Fig. 5. Cluster number distribution of two distances

As can be seen from the Fig. 5, the main purpose of using the test set is to simulate the effect of the algorithm in the real scenario and show whether the algorithm has a good generalization effect, so as to better apply the improvement and optimization to the real scenario.

It can be clearly seen from the clustering results that the clustering data situation. Considering the morphological characteristics of traffic flow time series and the purpose of clustering for 7 days, all kinds of values should tend to average. The result of clustering is even.

Similarity Weights. After clustering all sequences, the grouping situation is obtained, and the average value of feature and label of each cluster is calculated to represent the current cluster. The similarity between the predicted sequence and the 7 clusters can be calculated. The distance matrix of the formula 1 can be generated. In the process of calculating the distance, we use the above mentioned distance metrics to calculate the distance. And through the reciprocal, can be directly converted to generate A Update the weight matrix W, dist related to A, and obtain the current final weight matrix after the modification of Formula 2.

```
array([[198.38870561, 312.30204145, 147.5663145 , 221.7634799 ,
        132.98371663, 179.37599818, 275.94923531]])
```

Fig. 6. Distance between sequence and each cluster

```
array([[0.00504061, 0.00320203, 0.00677661, 0.00450931, 0.00751972,
        0.00557488, 0.00362385]])
```

Fig. 7. Similarity between sequence and each cluster

Meanwhile, set the weights w1, w2, ..., w7, W = sofetmax (A), and finally the data can be predicted (Figs. 6 and 7)

$$predict = W \times Y_m = [24.35835927, 24.12020417, 21.55549108, 21.44721639] \quad (7)$$

Y_m is the label of each cluster.

3.4   Calculation of Total Data Value

A set of data obtained at this point. Iterate the remaining test set data. The data of all test sets are completed and the algorithm is finished. So now I have my prediction for the next time, and I do the same thing and I get my prediction after time 1, 2, 4, 8 (Fig. 8).

	tn+1	tn+2	tn+4	tn+8
0	12.107919	12.167536	11.333130	12.605538
1	30.857143	31.500000	32.214286	31.571429
2	18.315708	19.278737	19.036072	21.510601
3	23.068570	22.870936	24.329254	21.738762
4	17.198980	18.747107	16.647945	17.276340
...	...	...	...	...
12558	14.408070	17.267959	17.929496	22.149490
12559	13.604976	14.614037	14.665070	16.120974
12560	28.903030	29.597521	29.757800	27.339606
12561	19.573403	20.247104	22.313147	21.755453
12562	32.181818	34.909091	34.727273	36.363636

Fig. 8. Total data of prediction

3.5   Analysis

In the Euclidean distance prediction data, compare the predicted data at four times with the real data in the test set, and draw the following figure
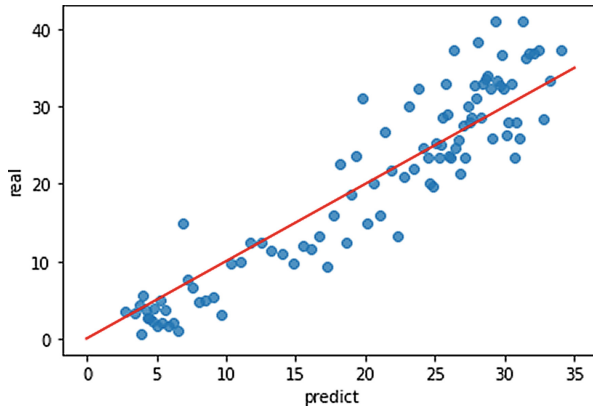
Fig. 9. The Q-Q graph of the method

As can be seen from the picture, the weight vector can reflect the data characteristics to some extent, but we can see that there is a part of the data close to the position of horizontal axis 0. Since the data set was recorded 15 years ago, the conditions at that time could not be compared with now, so the record was caused by many missing data, power outages and other problems, which could not be avoided. The domestic traffic environment is improved to ensure people's personal safety, and protective measures are gradually in place. However, it can be seen that on the oblique line of 45°, a large number of data sets indicate the rules of most fitting sequences of data.

The clustering prediction of Euclidean distance takes about 21 s to calculate 100 data. Therefore, the Kmeans clustering algorithm based on the Sklearn library is feasible and effective, and it is easier to use in actual scenes.

As can be seen from Fig. 9, it is obvious to see the gap between the predicted data of each category and the real data.

Looking at the values of MAE, R2 and MAPE after the completion of the prediction, we can clearly see that the actual difference between the evaluation values of the data at the four moments is not very big.

As can be seen from Fig. 10, the further back the predicted data is, the worse the accuracy of the prediction will be, and the higher the prediction needs to be after a long period of time. Therefore, the closer the moment of the predicted data is to the time series, the higher the accuracy of the prediction will be.

第t n+ 1 时刻的预测值
5.541900932250141   0.3957427310493161   0.49763984017579377
第t n+ 2 时刻的预测值
5.587645373660939   0.377720614522142   0.4969103345177101
第t n+ 4 时刻的预测值
5.647849973231797   0.3556265976099703   0.5171025866886609
第t n+ 8 时刻的预测值
5.7866137881857425   0.2963791757598844   0.5277307158591259

Fig. 10. Estimates of the three distances

4   Conclusion and Prospect

Through the analysis of the characteristics of the time series of vehicle flow, combined with the actual scene application, this paper studies a method of combining the various similarity distances to forecast data, and analyzes the different results of the application of several similarity measures. The core idea is as follows: the improved DTW algorithm is used to measure the similarity distance, the Kmeans clustering algorithm is used to obtain the centroid of multiple clustering results of the total data, the classification of clustering results is obtained, and the prediction of the final data is completed by combining with the similarity, and the evaluation of the validity index of clustering results is completed.

In 2019, the total investment scale of smart transportation in China exceeded 227.8 billion yuan. Even under the influence of the epidemic, the transportation investment of 10 million yuan project in the first quarter of 2020 increased by 15% year on year. In February 2021, Shandong Provincial Department of Public Security issued the Guiding Opinions on Strengthening the Application of Urban Road Traffic Signal Control, which required the integration of big data, intelligent new thinking and new technology to further improve the application ability of urban road traffic signal control. The algorithm used in this paper is efficient, accurate and economical. In particular, with the improvement of machine learning and deep learning algorithm accuracy, the improvement of computing power and the decrease of cost, the cost of intelligent transportation is constantly reduced and the advantages of scale are constantly emerging, which greatly promote the implementation and application of the algorithms listed in this paper.

References

1. Wang, W., Shan, X.: Study on regular pattern of railway passenger flow in three-day holiday based on clustering method of time series. Railw. Comput. Appl. **04**, 23–27 (2015)

2. Geng, R., Sun, B., Ma, L., Zhao, Q., Shen, T.: Anomaly-aware in sequence data based on MSM-H with EXPoSE. In: 40th Chinese Control Conference, CCC 2021, Shanghai, China (2021)
3. Sun, B., Cheng, W., Goswami, P., Bai, G.: Short-term traffic forecasting using self-adjusting k-nearest neighbours. IET Intel. Transp. Syst. **12**(1), 41–48 (2018)
4. Ji, M., Xiao, L.: A dynamic k-means clustering algorithm for time series data. Comput. Digit. Eng. **48**(8), 1852–1857 (2020). https://doi.org/10.3969/j.issn.1672-9722.2020.08.007
5. Lin, Q.: Research on Feature Screening and Clustering Analysis of Time Series Data - A Case Study of the CSI 300 Index. Southwestern University of Finance and Economics (2017)
6. Ma, L., Sun, B., Ziyi, L.: Bagging likelihood-based belief decision trees. In: 20th International Conference on Information Fusion (FUSION), Xi-An, China, pp. 1–6 (2017). http://ieeexplore.ieee.org/abstract/document/8009664/
7. Sun, B., Wei, C., Liyao, M., Prashant, G.: Anomaly-aware traffic prediction based on automated conditional information fusion. In: International Conference on Information Fusion (FUSION), pp. 2283–2289. IEEE, Cambridge, United Kingdom (2018)
8. Zheng, C.Z.L.: Shape clustering on time series data. In: Proceedings of Information Technology and Environmental System Sciences, ITESS 2008, vol. 3, pp. 1249–1253 (2008)
9. Plant, C., Wohlschhiger, A.M., Zherdin, A.: Interaction-based clustering of multivariate time series. In: The 9th IEEE International Conference on Data Mining, ICDM 2009, Miami, Florida, USA, 6–9 December 2009, pp. 914–919 (209)
10. Sun, B., Cheng, W., Goswami, P., Bai, G.: An overview of parameter and data strategies for k-nearest neighbours based short-term traffic prediction. In: 2017 ACM International Conference Proceeding Series, pp. 68–74. ACM (2017)
11. Ma, L., Sun, B., Han, C.: Learning decision forest from evidential data: the random training set sampling approach. In: 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, China (2017)
12. Li, F., Tan, L., et al.: On the data-mining oriented methods for clustering time series. Comput. Sci. **027**(012), 76–80 (2000)
13. Zijian, T.: Time Series Forecast via Similar Fluctuate Pattern. Hefei University of Technology (2016)
14. Sun, B., Cheng, W., Bai, G., Goswami, P.: Correcting and complementing freeway traffic accident data using Mahalanobis distance based outlier detection. Tehnicki Vjesnik (Tech. Gaz.) **24**(5), 1597–1607 (2017)
15. Liu, C.: Research on Interactive Prediction of Airport Noise Monitoring Points Based on Time Series Similarity Measure. Nanjing University of Aeronautics and Astronautics
16. Sun, B., Ma, L., Shen, T., et al.: A robust data-driven method for muti-seasonal and heteroscedastic IoT time series preprocessing. Wirel. Commun. Mob. Comput. (WCMC) **2021**, 1–11 (2021). Article ID 6692390
17. Lai, Y.: Study on Real-Time Prediction of Arrival Time for Floating Transit Vehicle. Chongqing University (2011)

Using an Ensembled Boosted Model for IoT Time Series Regression

Shuai Lin[1(✉)], Kun Zhang[1], Renkang Geng[2], and Liyao Ma[2]

[1] Shandong Non-Metallic Materials Institute, Jinan 250031, Shandong, China
i531inshuai@126.com

[2] School of Electrical Engineering, University of Jinan, Jinan 250022, Shandong, China
202021200778@mail.ujn.edu.cn, cse_sunb@ujn.edu.cn

Abstract. As a typical regression model, time series prediction is a very important part of machine learning. With the development of urban roads and the increase of car ownership, traffic data is more closely related to machine learning. As a common time-series data in our life, traffic flow data has great research value and a wide range of application fields. Compared with the general time series, traffic flow data has larger data volume, stronger volatility, and higher requirements for accuracy and speed of prediction, but traditional algorithms often fail to achieve these goals. With the development of ensemble algorithm, it has outstanding performance in the field of classification and regression. Therefore, we choose XGB algorithm as the core algorithm of this experiment. In this paper, we introduce the working principle of the XGBoost algorithm, the acquisition of the traffic flow data used in the experiment, and the feature extraction of the traffic flow data in detail. Finally, we use the XGBoost algorithm to model and output the prediction results. In addition, we modified some very important parameters in XGBoost, such as iteration model, iteration number, etc., to explore the influence of each parameter on prediction accuracy when the XGBoost algorithm is used to predict traffic flow data.

Keywords: Traffic flow data · IoT Time Series · Ensemble learning · XGBoost

1 Introduction

Since entering the 21st century, with the development of science and technology and civilization, the quickening pace of life, we follow the travel demand is becoming bigger and bigger [1], strong ability of infrastructure in China has brought the rapid development of urban road, greatly convenient for people to travel, travel choice, as the most convenient and flexible car became the most used means of travel, The number of motor vehicles in China has been increasing year after year, and electric vehicles have been developing rapidly [2]. However, the explosion of the number of cars has also brought

great challenges to urban road traffic, leading to traffic jams, accidents and other abnormal situations, which has brought great inconvenience and risks to people's lives [3]. Therefore, the management of traffic congestion has become one of the major social problems in the 21st century. If the traffic flow data can be accurately predicted, the traffic congestion can be effectively controlled, and the traffic pressure can be alleviated [5].

The traffic flow data itself has a certain trend and regularity, but due to the complexity of time and the heterogeneity of space, the data itself also has strong volatility in a short time, especially when some abnormal conditions occur on the road, the traffic flow data may change abruptly [8]. With the development of machine learning, more and more machine learning algorithms have been used in the prediction of time series data. Traffic flow data, as a typical representative of time series, has great research value and use-value [9].

There are many algorithms that are good for machine learning to predict traffic flow data, including statistical model, non-parametric regression model, neural network model and support vector machine model [10]. However, due to the high demand for traffic data on the prediction speed and accuracy of the algorithm, the traditional modelling methods have certain limitations [12]. Therefore, in recent years, more and more integration algorithms have been used in the field of time series prediction, such as Adaboost (Adaptive Boosting) [13], GBDT (Gradient Boosting Decision Tree) [14] etc. In 2016, Tianqi Chen has published a new algorithm, XGBoost (Extreme Gradient Boost) algorithm [15] was proposed, which was optimized on the basis of GBDT, and greatly improved the speed and efficiency of computing, which not only guaranteed.

In this paper, we extract the features of the traffic flow data and then forecast the traffic flow data based on the modelling of the XGBoost algorithm. Through the prediction results, we judge the prediction accuracy of XGBoost for such data and discuss the influence of different parameters in XGBoost on the prediction accuracy. Finding out the parameter configuration that XGBoost algorithm is suitable for predicting traffic flow data.

2   Related Work

Ensemble Learning [16] is an algorithm that builds multiple learners and then combines them together in a certain way, which is mainly divided into three categories, Bagging, Boosting and Stacking. Here we mainly introduce Bagging and Boosting.

Bagging uses a sampling method with putting back to generate training data. It randomly sampled the initial training set through multiple rounds of putting back, and multiple training sets were generated in parallel, which corresponded to the training of multiple base learners (there is no strong dependency relationship between them). Then, these base learners were combined to construct a strong learner. By increasing the randomness of samples, the variance can be reduced.

In essence, the XGBoost algorithm is a method based on Tree structure combined with ensemble learning, whose basic Tree structure is CART (Classification and Regression Tree) [17]. A CART regression tree is assumed to be a binary tree, and the sample space is divided by constantly splitting features. The core idea is that in the input space where the training data set is located, each region is recursively divided into two subregions and the output value of each sub-region is determined. The criteria for delimiting molecular regions depend on the type of tree, and the square error minimization criterion is usually used for regression trees. The objective function generated by the CART regression tree is:

$$\sum\nolimits_{x_i \in R_m} (y_i - f(x_i))^2$$

The XGBoost model is also a special GBDT. The idea is to grow a tree by repeated feature splitting, adding a tree at a time (in effect, learning a new function) to fit the residual of the previous prediction. XGBoost can customize a set of objective functions, with the help of Taylor expansion (you only need to know the first and second derivatives of the loss function to find the loss function) into a quadratic function of one variable, to get the extreme point and extreme value. The target function of XGBoost is as follows:

$$L(\phi) = \sum\nolimits_i l(y_i, \hat{y}_i) + \sum\nolimits_k \Omega(f_k)$$

GBDT algorithm only uses first-order Taylor expansion for the loss function, while XGBoost uses second-order Taylor expansion for the loss function. On the one hand, the second derivative expansion of the XGBoost loss function can increase the accuracy, and on the other hand, it can customize the loss function, because the second derivative Taylor expansion can approximate a large number of loss functions more accurately. On the basis of GBDT, XGBoost improves the computing speed, accuracy and optimizes some functions. For example, for samples with missing eigenvectors, the sparse perception algorithm adopted by XGBoost can automatically learn its splitting direction.

In order to effectively prevent the algorithm from over-fitting, there are two main solutions in the field of machine learning. One is to reduce the number of characteristic variables we select, and the other is regularization [18]. Regularization measures the smoothness of our fitting curve and is usually used as a penalty term in the loss function. In machine learning, the objective function of most algorithms is composed of the loss function and additional terms, such as the objective function of XGBoost. This extra term is the regularization term. Intuitive understanding from the polynomial model, the arbitrary function can use polynomial fitting, to help our fitting curve is smooth, is each different power coefficient of x before, a zero (or very small value) the situation is, the more that each x, in the final fitting polynomial to effectively distinguish between different weights.

$$f(x_i) = w_0 x_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots \cdots + w_n x_n$$

Regularization is to get rid of the items with low weight and keep the items with high weight, thus reducing the order of magnitude of the eigenvector. Regularization has the advantage of our retained all the characteristics of the variables, and not to deleting feature variables, but to distinguish the importance of the feature vector, because, in the actual problem, each variable can be more or less a little influence on the prediction of every variable has the value of it, we don't want to remove them [20], However, we also hope that some characteristic variables will not excessively affect the curve to prevent over-fitting, so we adopt the regularization method.

3 Methods

The traffic flow data used in this experiment is typical time-series data. Specific acquisition and processing methods are as follows: at a node on a certain road, read the traffic flow of this section within 5 min at a fixed time interval of 5 min. A total of 116,927 pieces of vehicle flow data with 5-min intervals in the past 406 days were taken as the data set for this XGBoost time series prediction. In order to enrich the feature vectors, we built some lags on the data set, and the number of lags is the number of feature vectors in this experiment. The number of feature vectors not only affects the degree of fitting but also affects the operation speed of the algorithm. Therefore, we should first process the data appropriately and select the appropriate number of lags as the feature of this prediction.

Usually, in the time series data, typically built of 8–128 lags to machine learning characteristic vector. In Bin Sun's 2018 paper [20], he tried to model different amounts of lags in the XGBoost algorithm, let XGBoost algorithm itself for the importance of these characteristic variables selection, shows the lag takes different values, the impact on the modelling accuracy. We found that when the number of lags was 144, it basically covered all the useful feature vectors of the traffic flow data of the time series for the XGBoost algorithm, and the calculation amount was not large. Therefore, 144 lags (144 eigenvectors) were used for this prediction. Each lag is subtracted 5 min from the previous lag, which constitutes a basic method to predict the traffic flow data at the next time with the characteristics of 144 data (data of half a day in the past). Finally, the feature vectors are given to the XGBoost algorithm for learning, and the prediction is given (Fig. 1).

In this experiment, we configured the way of cross-validation. For each prediction, three times of cross-validation were needed to output the final predicted value. In addition, we compared the influence of different parameters in XGBoost on the accuracy of our prediction in this experiment.
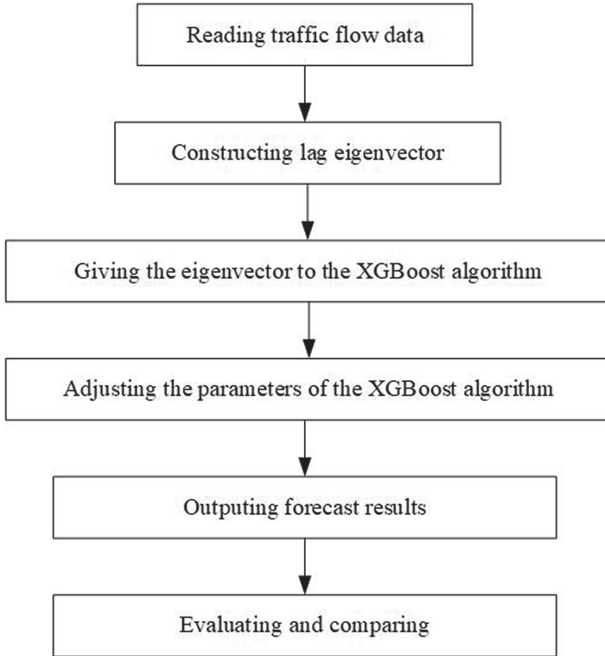
Fig. 1. The logical diagram shows the steps of our experiment. Extract the original data, construct the feature vector, and then send the feature vector to XGBoost, and adjust the XGBoost related parameters to output the prediction results, and finally evaluate.

4   Results

The prediction of time series is a typical regression algorithm in machine learning. For the evaluation of the regression algorithm, we chose MSE, RMSE, MAE and R-squared as four indexes. The formula is as follows:

$$MSE = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |(y_i - \hat{y}_i)|$$

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\hat{y}_i - y_i)^2}$$

In order to explore the influence of different parameters in XGBoost on the accuracy of traffic flow data prediction. We set several groups of different parameters for the XGBoost algorithm and made predictions under the same data set. The predicted results are as follows (Table 1):

Table 1. Impact of Booster parameters on predicted results

Booster	GBTree	Gbliner
MSE	64.8627	69.7946
MAE	5.7458	6.0019
RMSE	8.0537	8.3543
R-squared	0.9060	0.8988

Booster parameters are used to select the base model for each iteration. GBTree is a tree-based model and Gbliner is a linear model. We found that GBTree model was more accurate in predicting traffic flow data. R-squared is closer to 1 (Table 2).

Table 2. The effect of the Max_depth parameter on the predicted results.

Max_depth	3	5	7	9
MSE	62.7124	63.6834	66.4980	69.6052
MAE	5.6909	5.7103	5.8174	5.9586
RMSE	7.9191	7.9802	8.1546	8.3430
R-squared	0.9090	0.9077	0.9036	0.8991

Max_depth is the maximum depth of the tree. Used to control the overfitting of the model. The higher the value of Max_depth is, the more specific model learning is. From the table, we can see that when Max_depth is 3, the prediction effect is the best. With the increase of the value, the prediction accuracy decreases (Table 3).

Table 3. The effect of the Min_child_weight parameter on the predicted results.

Min_child_weight	1	2	3	5	6
MSE	64.8627	64.2205	65.0793	64.8114	64.7080
MAE	5.7458	5.7456	5.7620	5.7538	5.7567
RMSE	8.0537	8.0138	8.0671	8.0505	8.0441
R-squared	0.9060	0.9069	0.9057	0.9060	0.9062

Min_child_weight is the sum of the minimum sample weights, which is used to avoid overfitting. When its value is large, the model can avoid learning local special samples, but it may also lead to under-fitting. In this experiment, when the value of Min_child_weight is 3, the algorithm performs the best, and the value of R-squared is 0.9069. We find that the high or low value of Min_child_weight will slightly affect the accuracy of prediction (Table 4).

Table 4. The effect of the N_estimators parameter on the predicted results.

N_estimators	10	20	50	100	200	400
MSE	65.4463	62.8453	63.6407	64.8627	66.7749	68.9352
MAE	5.7727	5.6796	5.6986	5.7457	5.8298	5.9387
RMSE	8.0899	7.9275	7.9775	8.0537	8.1716	8.3027
R-squared	0.9051	0.9089	0.9077	0.9060	0.9032	0.9001

N_estimators is the number of the largest tree generated and also the maximum number of iterations. The experimental results show that when the value of N_estimators is 20, the prediction effect is the best, and R-squared reaches 0.9089. When the number of iterations is between 20 and 50, the best effect is achieved. However, when the number of iterations is too large or too small, the prediction accuracy of the XGBoost model for traffic flow data drops slightly.

5   Conclusion

In this experiment, we predict the traffic flow data based on XGBoost algorithm. We take the traffic flow of a fixed road in a 5-min interval in 406 days as the original data set and select the traffic flow data of the past half-day as the feature each time to predict the traffic flow of the next moment. Experimental results show that the XGBoost algorithm can effectively adapt to the traffic flow data, and the prediction effect is good. The value of R-squared is generally around 0.9. After that, we compared the influence of three very important basic parameters in the XGBoost algorithm on the experiment effect. In general, the adjustment of parameters has little influence on the accuracy of the prediction of this data set, with only some small changes. It turns out that XGBoost itself is pretty well optimized. In the future, we will conduct further exploration in two aspects. On the one hand, we will select more traffic flow data sets of different roads to train our model and improve the accuracy and application range of the model. On the other hand, continue to study the influence of different parameters in XGBoost on traffic flow data prediction.

References

1. Jie, L., Van Zuylen, H.J.: Road traffic in China. Procedia Soc. Behav. Sci. **111**, 107–116 (2014)
2. Liu, Z., Yue, X., Zhao, R.: The cause of urban traffic congestion and countermeasures in China. Urban Stud. **11**, 90–96 (2011)
3. Sun, B., Cheng, W., Goswami, P., Bai, G.: An overview of parameter and data strategies for k-nearest neighbours based short-term traffic prediction. In: ACM International Conference Proceeding Series 2017, pp. 68–74. ACM (2017)
4. Wen, H., Sun, J., Zhang, X.: Study on traffic congestion patterns of large city in China taking Beijing as an example. Procedia Soc. Behav. Sci. **138**, 482–491 (2014)

5. Sun, B., Cheng, W., Bai, G., Goswami, P.: Correcting and complementing freeway traffic accident data using Mahalanobis distance based outlier detection. Tehnicki Vjesnik-Technical Gazette **24**(5), 1597–1607 (2017)
6. Sun, B., Ma, L., Shen, T., et al.: A robust data-driven method for muti-seasonal and heteroscedastic IoT time series preprocessing. In: Wireless Communications and Mobile Computing (WCMC), p. 6692390 (2021)
7. Li, J., Walker, J.L., Srinivasan, S., et al.: Modeling private car ownership in China: investigation of urban form impact across megacities. Transp. Res. Rec. **2193**(1), 76–84 (2010)
8. Lv, Y., Duan, Y., Kang, W., et al.: Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intell. Transp. Syst. **16**(2), 865–873 (2014)
9. Chen, Y., Shu, L., Wang, L.: Traffic flow prediction with big data: a deep learning based time series model. In: 2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 1010–1011. IEEE (2017)
10. Ma, L., Sun, B., Han, C.: Learning decision forest from evidential data: the random training set sampling approach. In: 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, China (2017)
11. Ahmed, M.S., Cook, A.R.: Analysis of Freeway Traffic Time-Series Data By Using Box-Jenkins Techniques (1979)
12. Shi, D., Ding, T., Ding, B., et al.: Traffic speed forecasting method based on nonparametric regression. Comput. Sci. **43**(2), 224–229 (2016)
13. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
14. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat., 1189–1232 (2001)
15. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
16. Sagi, O., Rokach, L.: Ensemble learning: a survey. Wiley Interdiscip. Rev. Data Min. Knowl. Disc. **8**(4), e1249 (2018)
17. Lewis, R.J.: An introduction to classification and regression tree (CART) analysis. In: Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California, p. 14 (2000)
18. Ma, L., Sun, B., Li, Z.: Bagging likelihood-based belief decision trees. In: 20th International Conference on Information Fusion (FUSION), Xi-An, China, pp. 1–6 (2017). http://ieeexplore.ieee.org/abstract/document/8009664/
19. Bickel, P.J., Li, B., Tsybakov, A.B., et al.: Regularization in statistics. TEST **15**(2), 271–344 (2006)
20. Geng, R., Sun, B., Ma, L., Zhao, Q., Shen, T.: Anomaly-aware in sequence data based on MSM-H with EXPoSE. In: 40th Chinese Control Conference (CCC 2021), Shanghai, China (2021)
21. Sun, B., Cheng, W., Goswami, P., et al.: Short-term traffic forecasting using self-adjusting k-nearest neighbours. IET Intel. Transp. Syst. **12**(1), 41–48 (2018)

Dynamic Time Warping Based Clustering for Time Series Analysis

Kun Zhang[1], Shuai Lin[1], Haoxuan Sun[2], Liyao Ma[2(✉)], and Junpeng Xu[3]

[1] Shandong Non-metallic Materials Institute, Jinan 250031, China
[2] School of Electrical Engineering, University of Jinan, Jinan 250022, China
`cse_maly@ujn.edu.cn`
[3] Shandong Huasheng Pesticide Machinery Co. Ltd., Linyi 276017, China

Abstract. This paper proposes a prediction method based on time series similarity. Based on clustering, Dynamic Time Warping (DTW) algorithm is used to find the influence of similarity and weight on the prediction results. Time series is a structure that records data in time sequence. The characteristics of multiple data at each time point are the same and comparable. According to people's purpose to find the rule of time series, and to the future time forecast. The first chapter introduces the background of the topic. The second chapter mainly introduces the time series, clustering algorithm, similarity, DTW distance and other basic theories involved in this paper. In the third chapter, we study the method of the total forecast data of time series. DTW distance is used for clustering to obtain the similarity with each class and then predict the data.

Keywords: K-MEANS · DTW · Similarity · Dynamic weighting

1 Introduction

With the development of artificial intelligence (AI) and big data, various fields have developed to varying degrees, and more and more artificial intelligence products appear in more industries [1–3]. In the field of transportation, there is also rapid development. Mass time series obtained through bus cards, detectors, cameras, communication equipment, the Internet of things, etc. However, due to periodicity and high noise, how to make use of time series is still a developing problem [4,5].

Although the development and use of AI in China is very strong, the use in the field of intelligent transportation still needs to be improved [6,7]. The penetration rate of intelligent transportation systems in the United States has reached more than 85%, and even cities like New York and Los Angeles have reached 90% [8].

2   Background and Related Work

2.1   DTW Distance

In some complicated cases, the expression of the relationship between two time series (or between similar time series) can not effectively use the traditional Euclidean distance measure to express the relationship of similarity degree [9, 10]. In general, sequences, taken as a whole, should have very similar shapes, laws, and properties, but they do not align along the X-axis. Therefore, before we can compare their similarities, we need to regularize one or more sequences under the timeline to get a better balance. DTW is an effective method to achieve this regularization distortion. DTW calculates the similarity between two time series to extend the sequence through dynamic programming, find one-to-one or many-to-one points, and calculate the distance between the sequences. Therefore, in 1994, Bemdt and Clifford proposed a dynamic Time Warping algorithm (DTW for short) after research. The DTW distance is mainly to find the shortest distance between sequences after planning [11].

Compared with the Euclidean distance, the length of the two sequences for calculating the DTW distance can be arbitrary and not equal to 0. DTW algorithm can detect similar positions between two sequences and make correspondence, such as points A and C in Fig. 1, rather than points A and B in Euclidean distance. However, the calculation is much more complicated than the Euclidean distance. Two time series $L_1 = \{p_1, p_2, ..., p_i, ..., p_m\}$ and $L_2 = \{q_1, q_2, ..., q_j, ..., q_n\}$ (where i = 1, 2, ..., m, j = 1, 2, ..., n)(m and n may not be equal). To calculate the DTW distance, first construct an m * n matrix, as shown in Fig. 2. Calculate the distance matrix
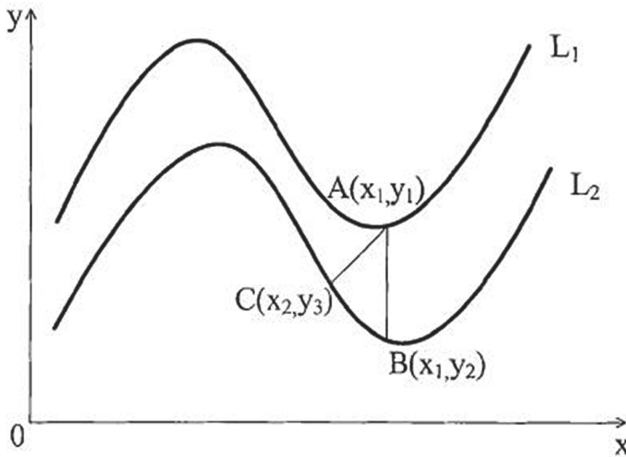

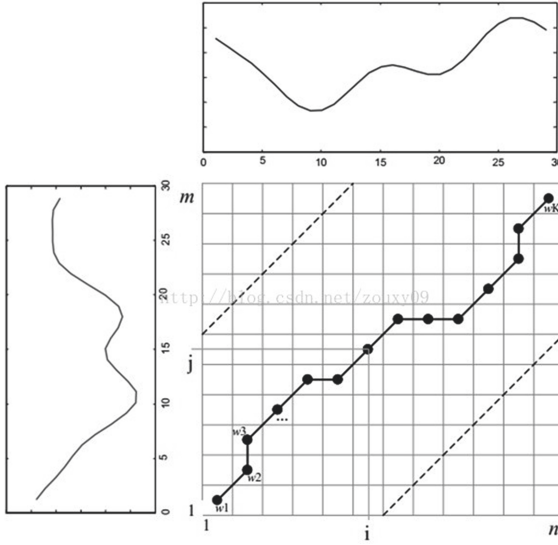**Fig. 1.** Dynamic time wraping distance principle

Fig. 2. DTW calculation steps

$$D = \begin{bmatrix} d_{11} \ d_{12} \ ... \ d_{1j} \ ... \ d_{1m} \\ d^{21} \quad ... \quad . \quad . \quad . \\ . \qquad . \quad . \quad ... \quad . \\ d_{i1} \quad ... \quad ... \quad d_{ij} \ ... \ d_{im} \\ . \qquad\qquad . \\ d_{n1} \ d_{n2} \ ... \ d_{nj} \ ... \ d_{nm} \end{bmatrix} \tag{1}$$

where, the matrix element represents the Euclidean distance from the i-th point in the matrix to the JTH point in the matrix, which can be calculated by using Formula 1, that is, the distance between each point in the matrix and each point in the matrix can be calculated to obtain the similarity. Then you need to find a path through several nodes of the matrix. This path can be called a dynamic time warped path, which is used to represent, and the distance of the path is called the DTW distance. The KTH element is defined as the mapping between reflection and, in fact, the grid point through which the path passes, and this is the process of calculating the nearest path between the two sequences. So that gives us $max(m,n) \leq l < m+n-1$. The path should meet the following three constraints:

(1) Boundary conditions: $w_1 = (1,1)$ and $w_k = (m,n)$. The regular path can only start from the lower left and go to the right, up and top right until it reaches the position of (m,n).
(2) Continuity condition: if $w_k = (a,b)$, then $w_{k+1} = (a',b')$ shall satisfy $(a'-a) \leq 1 \& (b'-b) \leq 1$, that is, the dimensions in the time series can only be calculated with adjacent points and proceed in the prescribed order.

This ensures that every point in the sum of the two time series L_1 and L_2 can be matched without missing out.

(3) Monotone condition: if $w_k = (a, b)$, then $w_{k+1} = (a', b')$ needs to meet the condition $(a' - a) \geq 0 \& (b' - b) \geq 0$. This ensures that the path goes down in order of direction.

Satisfying the above three constraints, there are many paths, but the DTW distance is the path with the smallest overall normalization loss, i.e.

$$DTW(L_1, L_2) = min(\frac{\sum_{k=1}^{l} w_k}{l}) \tag{2}$$

Here, add up the distances of all the points on the path to get the cumulative distance s(m, n). The cumulative distance has a linear relationship with the cost, and the smaller its value is, the smaller the wrapping cost is. Dynamic cumulative distance, its calculation formula is as follows:

$$s(i, j) = s(p_i, q_j) + min\{s(i - 1, j - 1), s(i - 1, j), s(i, j - 1)\} \tag{3}$$

The cumulative distance is the DTW distance we need.

It can be seen that the calculation of distance, after the operation of DTW algorithm, all the data in the sequence will be traversed at least once. However, it can better represent the similarity between the sequences. However, the disadvantage of DTW distance is that both the time complexity and the space complexity are much higher than the calculation of Euclidean distance, which is a problem that cannot be ignored [12].

2.2   Softmax

In artificial neural networks, the most common activation function is the multivariate version of sigmoid [13]. Soft Max can be thought of as an ArgMax operation, primarily to activate functions, which is also a smooth approximation.

Given a one-dimensional vector, Softmax function maps it to a probability distribution. The softmax function $R^n \to R^n$ is defined by the following formula

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}, for\ 1, ..., n\ and\ x = [x_1, ..., x_n]^T \in R^n \tag{4}$$

2.3   Clustering

In unsupervised learning, cluster analysis is the training of unstandardized classification of data, revealing the inherent laws of the data and automatically classifying data into similar categories.

The purpose of clustering is to classify many irrelevant subsets of a dataset, that is, to divide the data samples into different clusters. A cluster can be a separate procedure for finding data set properties. Internal indicators of cluster

results do not require external data. The distance between the sample point and the cluster center is used to express the quality of the clustering analysis results. Clustering is also an important part of learning tasks, providing data for future work.

And a good clustering, there are many characteristics, specific performance in practical applications, but also

(1) Good scalability.
(2) The ability to process different data.
(3) Noise processing.
(4) Insensitive to sequential samples.
(5) Constraints.
(6) Interpretation and ease of use.

KMEANS. In 1967, MacQueene proposed the KMEANS algorithm, one of the simplest and most common clustering methods. The similarity of KMEANS is reflected in the distance between samples. The closer you are, the more likeness you have. The degree of similarity directly affects the classification criteria. However, most people use the countdown of distance to express similarities, making the two positively related. Most of the distance is from Euclidean distance or Manhattan distance [14].

Algorithm 1. Kmeans algorithm

Input: $D = \{x_1, x_2, ..., x_m\}$;
     Cluster number k.
1: Randomly select k samples from D as the initial mean vector $\{\mu_1, \mu_2, ..., \mu_k\}$
2: **repeat**
3:     $C_i = \phi(1 \le i \le k)$
4:     **for** $j = 1, 2, .., m$ **do**
5:         Calculate the distance between sample x_j and each mean vector $\mu_i(1 \le j \le k)$:
         $d_{ji} = \|x_j - \mu_i\|_2$;
6:         The cluster marker of x_j is determined according to the nearest mean vector
         :$\lambda_j = arg\ min_{i \in \{1,2,...,k\}} d_{ji}$;
7:         Assign the sample x_j to the corresponding cluster: $C_{\lambda_j} = C_{\lambda_j} \bigcup \{x_j\}$;
8:     **end for**
9:     **for** i=1,2,...,k **do**
10:        Calculate the new mean vector $\mu_i' = \frac{1}{|C_i|} \sum_{x \in C_i} x$;
11:        **if** $\mu_i' \ne \mu_i$ **then**
12:            Update the current mean vector μ_i to μ_i'
13:        **else**
14:            Leave the current vector unchanged
15:        **end if**
16:     **end for**
17: **until** None of the current mean vectors have been updated
Output: Cluster partition$C = \{C_1, C_2, ..., C_k\}$

2.4   Evaluation

Evaluation methods can be understood in engineering theory and definitions, just as learning achievement is used to represent a student's learning performance [15]. Evaluate the rationality of existing models and algorithms and compare with better models.

MAE. Mean Absolute Error is referred to as MAE for the purpose of finding the difference between the predicted value and the real value and the Absolute Error. MAE can be obtained by averaging its values.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - y_i'| \tag{5}$$

3   Methods

3.1   Clustering of Time Series

After compiling time data, the scale of the time series used for comparison was selected. Experiments show that half a day's length is suitable for the search step. 1248 is considered to be a medium-term forecast rather than a short-term forecast as it goes beyond eight steps. The generated sample sequence is characterized by half-day data sequences and the size of the search step 1248 is tabulated. Generates a time series as a data set. The clustering algorithm compares the size of sequences by comparing the similarity between sequences, and classifies them precisely so as to prepare for the traffic flow prediction at the next traffic intersections.

$$x_i = \{t_i, t_{i+1}, \ldots, t_{i+n}\} \tag{6}$$
$$y_i = \{t_{i+n+1}, t_{i+n+2}, t_{i+n+3}, t_{i+n+4}\} \tag{7}$$

n is length of search step.

The most common K-mean algorithm is chosen for clustering time series. Taking into account the periodicity of the time series, the 7-day cycle per week will be selected for cluster prediction. Because each time series represents a different time period, similar time periods between different dates are selected as data for the data cluster. First, choose the value of the k class, k is 7 elements, K points are randomly selected. When calculating Kage, if the data skew of some samples in the dataset is too large, it will make operation difficult and not easy to converge. However, it is better suited to analyze large amounts of data than hierarchical clustering.

DTW distance represents the similarity between sequences, which is better than Euclidean distance, etc., to find sequences that are similar between sequences. In the process of Kmeans clustering, DTW distance can be used to calculate the distance between the two to classify the classes. But you can choose the Tslearn library to carry out the Kmeans algorithm to calculate the DTW

distance. However, due to the high time complexity, the calculation of DTW distance is a process of dynamic programming, so the termination condition cannot be the same as the Euclidean distance when each mean is iterated to no update. The clustering can only be ended when the number of iterations of the better clustering is better. In the case of sacrificing certain accuracy, the running time of the algorithm is reduced and the overall operation efficiency of the model is improved.

After clustering all time series, the following table is given,

Table 1. Number of categories

Clusters	Amount
1	Number1
2	Number2
3	Number3
4	Number4
5	Number5
6	Number6
7	Number7

Seeing 7 clusters from 1 to 7, and the number of sequences in each category. The number of categories should be relatively equal.

3.2 Mean Value of Each Cluster

All the sequences of each category make up a group so that it is easy to calculate the average value for each attribute of the entire group and represent a sequence of current clusters. Other clusters did the same and obtained a sequence corresponding to the average of the seven clusters. The data from the sequence can be used as predictive data for each galaxy cluster. The following

$$mean\_clusters^{(i)} = \{t_1^{(i)}, t_2^{(i)}, ..., t_{143}^{(i)}, t_{144}^{(i)}, ..., t_{147}^{(i)}\} \quad i = 1, 2, ..., 7 \qquad (8)$$

i is the cluster.

3.3 Similarity Metric

Since the data is a time series and there are no specific classification criteria, clustering analysis can be used to cluster such sequences to some extent [?]. Under the condition of unsupervised learning, the algorithm pays more attention to the similarity of sequences, and obtains the time series distribution roughly.

The predicted sequence is

$$sequence_of_feature = \{t_1, t_2, ..., t_{144}\} \tag{9}$$

There is no vehicle data in the last four moments of Formula 8, because the last four data are the data we need, so the predicted time series is only 144 features. What needs to be predicted is

$$prediction_of_label = \{t_{145}, t_{146}, t_{147}, t_{148}\} \tag{10}$$

And is used to calculate the similarity of 7 clusters, that is, the DTW distance is obtained by Formula 3. The distance matrix with seven clusters can be obtained

$$dist = [d_1, d_2, ..., d_7] \tag{11}$$

3.4 Similarity Weights and Data Prediction

By understanding the similarity, we can calculate the W-weighted value by using the relationship between similarity degrees and predict the value of the data. In other words, there is a mapping relationship between similarity and weight. The distance used here represents W's weight, and the greater the distance, the smaller the W_i value of the class.

From Formula 11, we can know which clusters of the data to be predicted belongs to, and the weight is needed to calculate the data to be predicted for Formula 9 and Formula 10. The predicted value is closely related to the similarity matrix A, as shown in the following formula

$$A = \frac{1}{dist} = \{\frac{1}{d_1}, \frac{1}{d_2}, ..., \frac{1}{d_7}\} \tag{12}$$

Initialize a weight matrix W,

$$W = |w_1, w_2, ..., w_7| \tag{13}$$

Since the predicted values are related to the values of the 7 clusters, only the degree of correlation is different, so the values of the row need to be added up to one, so as to show the relationship between the predicted data and the prediction values of the 7 clusters. So we need to sum each row of W to 1.

Using the Formula (4), we get

$$W = softmax(A) \tag{14}$$

So we know that the total number is

$$Forecast\_data = W \times mean\_clusters[:, 144 :] = \{pred_1, pred_2, pred_4, pred_8\} \tag{15}$$

This is the current data that needs to be predicted. Perform the same operation to complete the prediction after the future time 1, 2, 4 and 8.

Put the predicted value into Formula 4 to get the value of the evaluation coefficient.

4   Conclusion and Prospect

Based on the analysis of the characteristics of vehicle flow time series and the practical application in the field, the method of combining the similar distance with the forecast data is studied. The core idea is to measure the similarity distance by using the improved DTW algorithm, to obtain the center of multiple clustering results of the total data by K-mean clustering analysis, and to get the result classification.

Although in theory, DTW distance can better represent the similarity of sequences, due to the high time complexity, it can not complete the clustering. The effect is good, but the time is too long, not suitable for the real environment. The completion of clustering will also take a long time to calculate the data.

Although the theoretical complexity of DTW algorithm is $O(n^2)$, with the cheap hardware, and because a desktop with GPU can support more than 200 terminals, it can support more than 50 intersections according to the calculation of 4 terminals at each intersection, which fully meets the needs of the main roads in general cities.

By 2019, China's total investment in smart transport will exceed 227.8 billion yuan. Investment in transport for 10 million projects will increase by 15% in the first quarter of 2020, even under the impact of the epidemic. In February 2021, the Shandong provincial public security bureau issued guidelines for strengthening the application of urban road traffic signal control, calling for the integration of big data, intelligent new thinking and new technologies to further improve the ability to apply urban traffic signals control. The algorithm adopted in this paper is efficient, accurate, economical and reasonable, especially with the improvement of the precision of machine learning and depth learning algorithm, the enhancement of computational power and the reduction of cost, and intelligent transportation costs continue to drop, scale advantage continues to appear, which greatly promotes the landing and application of these algorithms.

References

1. Zhang, G.: Research and Application on Interval Time Series Clustering Based on DTW. Northwest Normal University (2020)
2. Geng, R., Sun, B., Ma, L., Zhao, Q., Shen, T.: Anomaly-aware in sequence data based on MSM-H with EXPoSE. In: 40th Chinese Control Conference, CCC 2021, Shanghai, China (2021)
3. Sun, B., Cheng, W., Goswami, P., Bai, G.: Short-term traffic forecasting using self-adjusting k-nearest neighbours. IET Intel. Transp. Syst. **12**(1), 41–48 (2018)
4. Li, S.: Clustering for Time Series Based on the Feature. Guangxi Normal University (2014)
5. Ma, L., Sun, B., Ziyi, L.: Bagging likelihood-based belief decision trees. In: 20th International Conference on Information Fusion (FUSION), Xi-An, China, pp. 1–6 (2017). http://ieeexplore.ieee.org/abstract/document/8009664/
6. Lafuente-Rego, B., Vllar, J.A.: Clustering of time series using quantite autocovariances. Adv. Data Anal. Classif. **10**(3), 391–415 (2016)

7. Sun, B., Cheng, W., Bai, G., Goswami, P.: Correcting and complementing freeway traffic accident data using Mahalanobis distance based outlier detection. Tehnicki Vjesnik (Tech. Gaz.) **24**(5), 1597–1607 (2017)
8. Plant, C., Wohlschhiger, A.M., Zherdin, A.: Interaction-based clustering of multivariate time series. In: The 9th IEEE International Conference on Data Mining, ICDM 2009, Miami, Florida, USA, 6–9 December 2009, pp. 914–919 (2009)
9. Sun, B., Wei, C., Liyao, M., Prashant, G.: Anomaly-aware traffic prediction based on automated conditional information fusion. In: International Conference on Information Fusion (FUSION), Cambridge, United Kingdom, pp. 2283–2289. IEEE (2018)
10. Ying, L.: Research on Clustering Methods for Time Series. Liaoning Normal University (2012)
11. Ma, L., Sun, B., Han, C.: Learning decision forest from evidential data: the random training set sampling approach. In: 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, China (2017)
12. Sun, B., Cheng, W., Goswami, P., Bai, G.: An overview of parameter and data strategies for k-nearest neighbours based short-term traffic prediction. In: 2017 ACM International Conference Proceeding Series, pp. 68–74. ACM (2017)
13. Jianle, S.: Stock Price Trend Prediction Research Based on Time Series Similarity. Chongqing Jiaotong University (2014)
14. Ashish, S., Dale, E.: Clustering for multivariate time series data. In: Proceedings of the American Control Conference, Anchorage, May 2002, pp. 586–591 (2002)
15. Sun, B., Ma, L., Shen, T., et al.: A robust data-driven method for multi-seasonal and heteroscedastic IoT time series preprocessing. Wirel. Commun. Mob. Comput. (WCMC) **2021**, 1–11 (2021). Article ID 6692390

Voltage Estimation of PH Meter Calibrator Using Integration of Kalman/FIR and R-T-S Smoothing Method

Wanjie Ren[1(✉)], Xia Li[2], Guoxing Hu[1], Rui Tuo[1], and Chen Cai[1]

[1] Shandong Institute of Non-metallic Materials, Jinan 250000, Shandong, China
renwanjie53@126.com
[2] Shandong Academy of Pharmaceutical Science, No. 989 Xinlu Street,
High-tech Zone, Jinan 250000, Shandong, China

Abstract. In order to further improve the accuracy of voltage estimation of PH meter calibrator, this paper introduce the scheme for the voltage estimation of PH meter calibrator using integration of Kalman filter (KF)/finite impulse response (FIR) filter and Rauch-Tung-Striebel (R-T-S) smoothing method. Firstly, the state-space for the voltage estimation is designed. Secondly, the integrated method of KF/FIR and R-T-S smoothing method is investigated. The proposed method employs the Kalman filter as the forward filter, as for the backward smoothing, the proposed R-T-S smoothing method is adopted. Then, the final estimation of the voltage is computed the average of the output of the KF/R-T-S smoothing-integrated method. By taking practical tests, the performance of the proposed integration of KF/FIR filter and R-T-S smoothing method is verified.

Keywords: KF/FIR · R-T-S smoothing · Voltage estimation

1 Introduction

The voltage estimation is essential for PH meter calibration, and its accuracy directly affects the accuracy of the PH meter calibrator. In terms of data fusion filters, Kalman filter is one of the most well-known examples [2], which has bee widely used in the field of the navigation [1]. However, it's supposed to point out that the Kalman filter (KF)'s performance depends on the accurate of the noise statistics [6], which is hard to obtain. Therefore, researchers proposed the finite impulse response (FIR) filter to overcome this shortcoming [5]. And this method has been used in the field of navigation [3,4].

In this work, the voltage estimation of PH meter calibrator using integration of Kalman filter (KF)/finite impulse response (FIR) filter and Rauch-Tung-Striebel (R-T-S) smoothing method will be proposed. In this work, contributions can be reflected in following aspects:

(1) For the voltage estimation, the state-space has been designed.
(2) The voltage estimation of PH meter calibrator using integration of KF/FIR filter and R-T-S smoothing method has been proposed.
(3) The performance of the proposed integration of KF/FIR filter and R-T-S smoothing method is verified by taking a practical test.

2   State Space for the Voltage Estimation

In this section, the state-space for the voltage estimation will be designed. The Eq. (1) is the state equation of the KF/FIR filter used for the voltage estimation.

$$\mathbf{Vol}_{k|k-1} = \mathbf{A}\mathbf{Vol}_{k-1} + \mathbf{w}_k, \tag{1}$$

where k is the time index, $\mathbf{Vol}_k$ represents the state vector when the time index is k, in this work, the $\mathbf{Vol}_k$ only includes the voltage value at the time index k, and we set $\mathbf{A} = 1$, which represents the state transition matrix, and the $\mathbf{w}_k \sim (0, \mathbf{Q}_k)$ represents the system noise.

The Eq. (2) is the measurement equation of the KF/FIR filter used for the voltage estimation.

$$\mathbf{z}_k = \mathbf{H}\mathbf{Vol}_{k|k-1} + \mathbf{v}_k, \tag{2}$$

where $\mathbf{z}_k$ represents the observation vector when the time index is k, in this work, the $\mathbf{z}_k$ is the voltage value measured by the PH meter calibrator at the time index k, the measurement conversion matrix is represented by $\mathbf{H}$ and the $\mathbf{w}_k \sim (0, \mathbf{R}_k)$ represents the measurement noise.

3   KF/FIR Filter for the Voltage Estimation

From the model (1) and (2), we can see that the data fusion model is linear, therefore, the KF/FIR filter is used as data fusion filter, which will be introduced in this section. With the model (1) and (2) of the KF/FIR filter used for the voltage estimation, in the Algorithm 1, the pseudo code has been shown. The state prediction and state update of the Kalman filter are shown in the algorithm, at lines 3–4 and lines 5–7 respectively.

4   The R-T-S Smoothing Method

The R-T-S smoothing method will be investigated in this section, which is based on the (1) and (2). The R-T-S smoothing method is one of the most famous smoothing method, it requires the outputs of the forward filter, and we adopt the KF as the forward filter in this work, which is listed in Algorithm 1, and

---

Algorithm 1: The KF/FIR filter based on the model (1) and (2)

Data: $\mathbf{z}_k$, $\mathbf{Vol}_0$, $\mathbf{P}_0$, $\mathbf{Q}$, $\mathbf{R}$, q^E

Result: $\hat{\mathbf{Vol}}_k$, $\hat{\mathbf{P}}_k$

1 **begin**
2   **for** $k = 1 : N$ **do**
3   **if** $k < q^E$ **then**
4             $\hat{\mathbf{Vol}}_{k|k-1} = \mathbf{A}\hat{\mathbf{Vol}}_{k-1}$;
5             $\hat{\mathbf{P}}_{k|k-1} = \mathbf{A}\hat{\mathbf{P}}_{k-1}\mathbf{A}^T + \mathbf{Q}$;
6             $\mathbf{K}_k = \hat{\mathbf{P}}_{k|k-1}\mathbf{H}_k^T \left(\mathbf{H}_k\hat{\mathbf{P}}_{k|k-1}\mathbf{H}_k^T + \mathbf{R}\right)^{-1}$;
7             $\hat{\mathbf{Vol}}_k = \hat{\mathbf{Vol}}_{k|k-1} + \mathbf{K}_k \left(\mathbf{z}_k - \mathbf{H}\hat{\mathbf{Vol}}_{k|k-1}\right)$;
8             $\hat{\mathbf{P}}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\hat{\mathbf{P}}_{k|k-1}$;
9   **else**
10             $m = k - q^E + 1$, $t = m + M^E - 1$;
11             $\tilde{\mathbf{Vol}}_t = \hat{\mathbf{Vol}}_t$;
12             $\tilde{\mathbf{P}}_t = \hat{\mathbf{P}}_t$;
13   **for** $b = m + M^E : k$ **do**
14                 $\tilde{\mathbf{Vol}}_{b|b-1} = \mathbf{A}\tilde{\mathbf{Vol}}_{b-1}$;
15                 $\tilde{\mathbf{P}}_{b|b-1} = \mathbf{A}\tilde{\mathbf{P}}_{b-1}\mathbf{A}^T + \mathbf{Q}$;
16                 $\mathbf{G}_b = \left[\mathbf{H}^T\mathbf{H} + \left(\mathbf{A}\mathbf{G}_{b-1}\mathbf{A}^T\right)^{-1}\right]^{-1}$;
17                 $\mathbf{K}_b = \mathbf{G}_b\mathbf{H}^T$;
18                 $\tilde{\mathbf{Vol}}_b = \tilde{\mathbf{Vol}}_{b|b-1} + \mathbf{K}_b \left(\mathbf{z}_b - \mathbf{H}\tilde{\mathbf{Vol}}_{b|b-1}\right)$;
19                 $\tilde{\mathbf{P}}_b = (\mathbf{I} - \mathbf{K}_b\mathbf{H})\tilde{\mathbf{P}}_b(\mathbf{I} - \mathbf{K}_b\mathbf{H})^T + \mathbf{K}_b\mathbf{R}\mathbf{K}_b{}^T$;
20   **end for**
21             $\hat{\mathbf{Vol}}_k = \tilde{\mathbf{Vol}}_b$;
22             $\hat{\mathbf{P}}_k = \tilde{\mathbf{P}}_b$;
23   **end if**
24   **end for**
25 **end**
26 † N is the last time index
27 † q^E is the dimension of $\hat{\mathbf{Vol}}_k$

---

the R-T-S smoothing method uses the KF's outputs $\hat{\mathbf{Vol}}_k$ and $\mathbf{P}_k, k \in [1, N]$. With the model (1) and (2)of the R-T-S smoothing method used for the voltage estimation, in Algorithm 2, the pseudo code has been shown. It can be seen from the algorithm that the R-T-S smoothing is a reverse smoothing.

Algorithm 2: R-T-S smoothing algorithm for the model (1) and (2)

Data: $\mathbf{z}_k$, $\mathbf{Q}$, $\mathbf{R}$, $\hat{\mathbf{Vol}}_k$, $\mathbf{P}_k$, $k \in [1, N]$
Result: $\hat{\mathbf{Vol}}_k^S$, $\hat{\mathbf{P}}_k^S$

1 **begin**
2 $\quad$ $\hat{\mathbf{Vol}}_N^S = \hat{\mathbf{Vol}}_N$;
3 $\quad$ $\hat{\mathbf{P}}_N^S = \hat{\mathbf{P}}_N$;
4 $\quad$ **for** $k = N : 2$ **do**
5 $\quad\quad$ $\hat{\mathbf{P}}_{k|k-1} = \mathbf{A}\hat{\mathbf{P}}_{k-1}\mathbf{A}^T + \mathbf{Q}$;
6 $\quad\quad$ $\mathbf{K}_{k-1}^S = \hat{\mathbf{P}}_{k-1}\mathbf{A}^T\hat{\mathbf{P}}_{k|k-1}^{-1}$;
7 $\quad\quad$ $\hat{\mathbf{Vol}}_{k-1}^S = \hat{\mathbf{Vol}}_{k-1} + \mathbf{K}_{k-1}^S \left(\hat{\mathbf{Vol}}_k^S - \mathbf{A}\hat{\mathbf{Vol}}_{k-1} \right)$;
8 $\quad\quad$ $\hat{\mathbf{P}}_{k-1}^S = \hat{\mathbf{P}}_{k-1} + \mathbf{K}_{k-1}^S \left(\hat{\mathbf{P}}_k^S - \hat{\mathbf{P}}_{k|k-1} \right) \left(\mathbf{K}_{k-1}^S \right)^T$;
9 $\quad$ **end for**
10 **end**
11 † Superior S denotes smoothness

---

5   The Integration of KF/FIR Filter and R-T-S Smoothing Method

In this section, the integration of Kalman and R-T-S smoothing algorithm for the voltage estimation will be investigated. With the Algorithm 1 and Algorithm 2, and in Algorithm 3, the integration of Kalman and R-T-S smoothing algorithm with the model (1) and (2) can be shown. The structure of the integration of Kalman and R-T-S smoothing method for the voltage estimation is shown in Fig. 1. From the algorithm and the figure, one can infer easily that the proposed integration of Kalman and R-T-S smoothing method includes three steps:

(1) The forward filter (lines 2–8).
(2) The backward smoothing (lines 9–16).
(3) The $\hat{\mathbf{Vol}}$ is the mean of the $\hat{\mathbf{Vol}}_k, k \in [1, N]$.
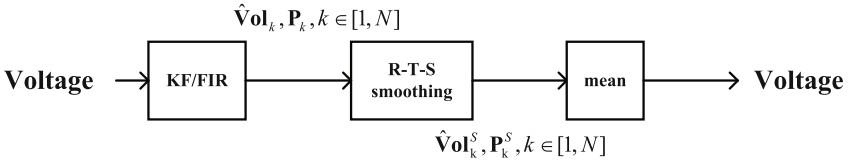

Fig. 1. The structure of the integration of Kalman and R-T-S smoothing method for the voltage estimation.

Algorithm 3: The integration of KF/FIR filter and R-T-S smoothing method for the model (1) and (2)

Data: $\mathbf{z}_k$, $\mathbf{Vol}_0$, $\mathbf{P}_0$, $\mathbf{Q}$, $\mathbf{R}$, q^E
Result: $\hat{\mathbf{Vol}}$

1 **begin**
2   **for** $k = 1 : N$ **do**
3   **if** $k < q^E$ **then**
4              $\hat{\mathbf{Vol}}_{k|k-1} = \mathbf{A}\hat{\mathbf{Vol}}_{k-1}$;
5              $\hat{\mathbf{P}}_{k|k-1} = \mathbf{A}\hat{\mathbf{P}}_{k-1}\mathbf{A}^T + \mathbf{Q}$;
6              $\mathbf{K}_k = \hat{\mathbf{P}}_{k|k-1}\mathbf{H}_k^T \left(\mathbf{H}_k\hat{\mathbf{P}}_{k|k-1}\mathbf{H}_k^T + \mathbf{R}\right)^{-1}$;
7              $\hat{\mathbf{Vol}}_k = \hat{\mathbf{Vol}}_{k|k-1} + \mathbf{K}_k \left(\mathbf{z}_k - \mathbf{H}\hat{\mathbf{Vol}}_{k|k-1}\right)$;
8              $\hat{\mathbf{P}}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\hat{\mathbf{P}}_{k|k-1}$;
9   **else**
10             $m = k - q^E + 1$, $t = m + M^E - 1$;
11             $\tilde{\mathbf{Vol}}_t = \hat{\mathbf{Vol}}_t$;
12             $\tilde{\mathbf{P}}_t = \hat{\mathbf{P}}_t$;
13   **for** $b = m + M^E : k$ **do**
14                 $\tilde{\mathbf{Vol}}_{b|b-1} = \mathbf{A}\tilde{\mathbf{Vol}}_{b-1}$;
15                 $\tilde{\mathbf{P}}_{b|b-1} = \mathbf{A}\tilde{\mathbf{P}}_{b-1}\mathbf{A}^T + \mathbf{Q}$;
16                 $\mathbf{G}_b = \left[\mathbf{H}^T\mathbf{H} + \left(\mathbf{A}\mathbf{G}_{b-1}\mathbf{A}^T\right)^{-1}\right]^{-1}$;
17                 $\mathbf{K}_b = \mathbf{G}_b\mathbf{H}^T$;
18                 $\tilde{\mathbf{Vol}}_b = \tilde{\mathbf{Vol}}_{b|b-1} + \mathbf{K}_b \left(\mathbf{z}_b - \mathbf{H}\tilde{\mathbf{Vol}}_{b|b-1}\right)$;
19                 $\tilde{\mathbf{P}}_b = (\mathbf{I} - \mathbf{K}_b\mathbf{H})\tilde{\mathbf{P}}_b (\mathbf{I} - \mathbf{K}_b\mathbf{H})^T + \mathbf{K}_b\mathbf{R}\mathbf{K}_b^T$;
20   **end for**
21             $\hat{\mathbf{Vol}}_k = \tilde{\mathbf{Vol}}_b$;
22             $\hat{\mathbf{P}}_k = \tilde{\mathbf{P}}_b$;
23   **end if**
24   **end for**
25     $\hat{\mathbf{Vol}}_N^S = \hat{\mathbf{Vol}}_N$;
26     $\hat{\mathbf{P}}_N^S = \hat{\mathbf{P}}_N$;
27   **for** $k = N : 2$ **do**
28         $\hat{\mathbf{P}}_{k|k-1} = \mathbf{A}\hat{\mathbf{P}}_{k-1}\mathbf{A}^T + \mathbf{Q}$;
29         $\mathbf{K}_{k-1}^S = \hat{\mathbf{P}}_{k-1}\mathbf{A}^T\hat{\mathbf{P}}_{k|k-1}^{-1}$;
30         $\hat{\mathbf{Vol}}_{k-1}^S = \hat{\mathbf{Vol}}_{k-1} + \mathbf{K}_{k-1}^S \left(\hat{\mathbf{Vol}}_k^S - \mathbf{A}\hat{\mathbf{Vol}}_{k-1}\right)$;
31         $\hat{\mathbf{P}}_{k-1}^S = \hat{\mathbf{P}}_{k-1} + \mathbf{K}_{k-1}^S \left(\hat{\mathbf{P}}_k^S - \hat{\mathbf{P}}_{k|k-1}\right)\left(\mathbf{K}_{k-1}^S\right)^T$;
32   **end for**
33     $\hat{\mathbf{Vol}} = \frac{1}{N}\sum\limits_{k=1}^{N}\hat{\mathbf{Vol}}_k$;
34 **end**

---

6   Experimental Results

In this test, we will conduct a real experiment to proving the effectiveness of the KF/FIR filter and R-T-S smoothing integrated method with the model (1) and (2), which is proposed in this work. Firstly, the details of the experiment will be introduced, secondly, the differences of test results between the KF/FIR filter and the integration of KF/FIR filter and R-T-S smoothing method will be investigated. In the test, the sample times are set as 20, and the reference voltage value is 100.05 mV. Figure 2 displays the errors estimated by the measurement, KF/FIR filter, and the integration of KF/FIR + R-T-S smoothing. Here, the error estimated by the measurement means the errors of the measurement. From the figure, one can infer that both the KF/FIR filter and KF/FIR + R-T-S smoothing are able to estimate the voltage value, and to further illustrate the effectiveness of the proposed algorithm, the cumulative distribution functions (CDFs) of the measurement, KF/FIR filter, and KF/FIR + R-T-S smoothing are shown in Fig. 3. The root mean square errors (RMSEs) of the measurement, KF/FIR filter, and KF/FIR + R-T-S smoothing are listed in Table 1. When the reference value is 100.0096, Table 2 shows the final outputs of KF/FIR filter and KF/FIR + R-T-S smoothing. It has been clearly shown in the figure that the KF/FIR + R-T-S smoothing proposed has the smallest error compared with the KF/FIR, which verify the effectiveness of the KF/FIR + R-T-S smoothing integrated method.

Table 1. The RMSEs of KF/FIR filter and KF/FIR + R-T-S smoothing.

Methods	RMSE (mV)
KF/FIR	0.20233
KF/FIR + R-T-S smoothing	0.18532

Table 2. The final outputs of KF/FIR filter and KF/FIR + R-T-S smoothing when the reference value is 100.0096.

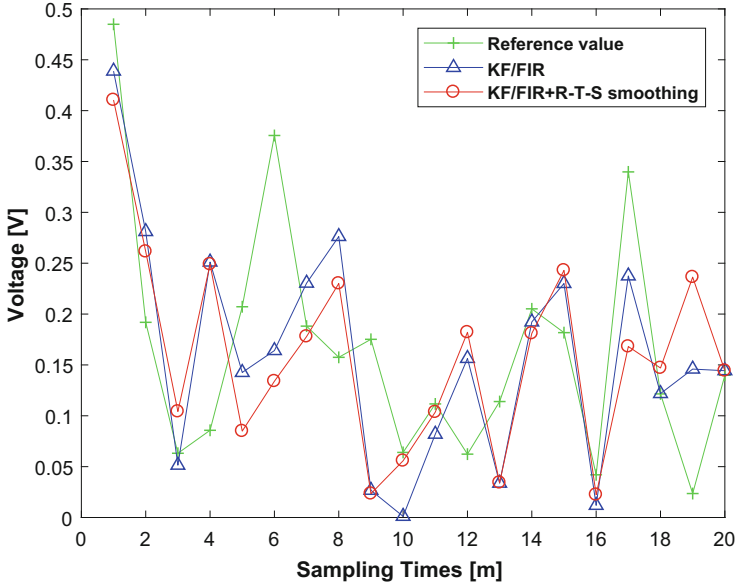| Methods	Value (mV)
KF/FIR	100.0356
KF/FIR + R-T-S smoothing	100.0302

Fig. 2. The errors estimated by the measurement, KF/FIR filter, and the integration of KF/FIR filter and R-T-S smoothing method (KF + R-T-S smoothing).
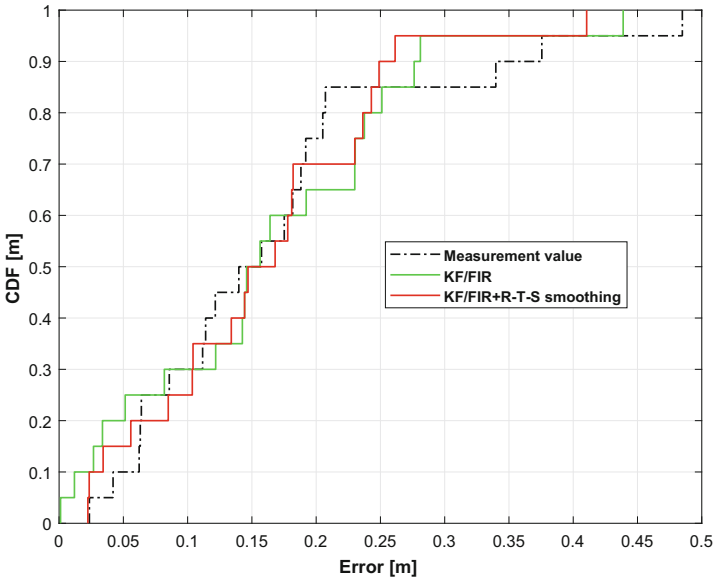

Fig. 3. The CDFs of the measurement, KF/FIR filter, and KF/FIR + R-T-S smoothing.

7   Conclusion

In this work, the voltage estimation of PH meter calibrator using integration of KF/FIR filter and R-T-S smoothing method has been proposed. The contribution of this work is reflected in the following aspects:

(1) The state-space for the voltage estimation has be designed.
(2) The voltage estimation of PH meter calibrator using integration of KF/FIR filter and R-T-S smoothing method has been proposed.
(3) The performance of the proposed integration of KF/FIR filter and R-T-S smoothing method is verified by taking a practical test.

References

1. Basso, M., Galanti, M., Innocenti, G., Miceli, D.: Triggered INS/GNSS data fusion algorithms for enhanced pedestrian navigation system. IEEE Sens. J. **20**(13), 7447–7459 (2020)
2. El-keyi, A., Kirubarajan, T., Gershman, A.B.: Robust adaptive beamforming based on the Kalman filter. IEEE Trans. Signal Process. **53**(8), 3032–3041 (2005)
3. Shmaliy, Y.S., Zhao, S., Ahn, C.K.: Optimal and unbiased filtering with colored process noise using state differencing. IEEE Signal Process. Lett. **26**(4), 548–551 (2019)
4. Xu, Y., Ahn, C.K., Shmaliy, Y.S., Chen, X., Li, Y.: Adaptive robust INS/UWB-integrated human tracking using UFIR filter bank. Measurement **123**, 1–7 (2018)
5. Zhao, S., Huang, B., Liu, F.: Localization of indoor mobile robot using minimum variance unbiased FIR filter. IEEE Trans. Autom. Sci. Eng. **15**(2), 410–419 (2016)
6. Zhao, S., Shmaliy, Y., Shi, P., Ahn, C.K.: Fusion Kalman/UFIR filter for state estimation with uncertain parameters and noise statistics. IEEE Trans. Industr. Electron. **64**(4), 3075–3083 (2017)

Human Tracking Using Distributed Dual-EKF Filter

Jing Cao[1], Jidong Feng[1], Wanjie Ren[2], Wanfeng Ma[1], Mingran Li[1], and Yuan Xu[1(✉)]

[1] School of Electrical Engineering, University of Jinan, Jinan 250022, Shandong, China
xy_abric@126.com
[2] Shandong Institute of Non-metallic Materials, Jinan 250000, Shandong, China

Abstract. Considering the accuracy of the indoor navigation system, an inertial navigation system (INS)/ultra wide band (UWB) technology integrated human tracking using distributed dual-EKF filter will be proposed. In this algorithm, the local data filter we use is composed of four dual-EKF filters to deal with noise more effectively. At the same time, we also take into account the uncertainty of system parameters. It employs the system parameters as the state vector. In the next step, the output of the local filter is input to the main filter for fusion and the best estimate is provided. The experimental results show that this scheme can effectively reduce the positioning error.

Keywords: Iterated dual-EKF · Indoor human localization · Ultra wide band · Inertial measurement unit

1 Introduction

Into the 21st century, with the rapid development of artificial intelligence and wireless Internet, people have put forward higher requirements for navigation performance. We are familiar with the Global Positioning System (GPS) to provide users with positioning services, which greatly promotes the progress of human production and life [4]. However, it is difficult for GPS signals to penetrate solid walls and obstacles, which causes indoor GPS to not work properly. In order to meet the demand for indoor high-precision positioning, how to ensure the accuracy of indoor navigation has become a hot topic [7].

Kalman filter (KF) is a highly efficient recursive filter [5]. KF can be used in any dynamic system with uncertain information to make a basis prediction for the next direction of the system. Once accompanied by various interferences, the Kalman transform can always point out what actually happened [6]. In recent years, KF have been implemented in many different ways. The standard KF is only suitable for linear systems. For nonlinear systems, we can use extended Kalman filter (EKF) for optimal estimationcite [2].

Taking into account the positioning error of indoor human tracking, an inertial navigation system (INS)/ultra wide band (UWB) integrated human tracking using distributed dual-EKF filter will be proposed. In this algorithm, the dual-EKF filters is used as the local data fusion filter, which considers the uncertainty of system parameters, it employs the system parameters as the state vector. Next, input the output of the local filter to the main filter for fusion and provide the best estimate. After a lot of experimental verification, it can be determined that this scheme can effectively reduce the positioning error.

The structure of this article is as follows. Robot localization scheme is discussed in Sect. 2. The distributed dual-EKF filtering algorithm is discussed in Sect. 3. In Sect. 4, the performance of distributed dual-EKF algorithm is verified by simulation results. Section 5 summarizes this article.

2   Fusion Model

The fusion model used in this work will be designed in this section. In this work, the time of arrival (TOA)-based UWB is employed, and the two-dimensional (2D) position is considered [3]. Thus, the UWB employed in this work requires at least three UWB reference nodes (RNs) to provide the ranges $r_l, l \in [1, m]$ from the UWB RNs to the UWB blind node (BN). Here, m means the numbers of the UWB RNs. The system chart of the distributed dual-EKF filter is shown in Fig. 1. It includes four local dual-EKF filters and a main filter. The local dual-EKF is employed to provide the $\mathbf{s}(t)^l$ and $\mathbf{P}(t)^l$, and the output of the distributed dual-EKF filter is fused by the main filter.
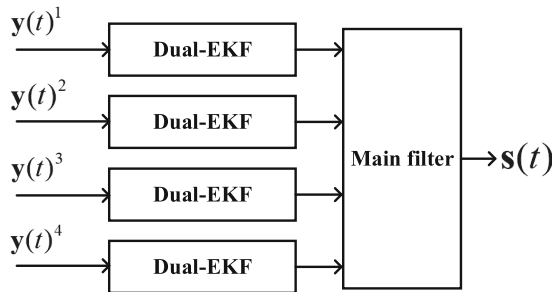

Fig. 1. Structure diagram of distributed dual-EKF filter.

Equation (1) shows the state equation for the local filter in the fusion model, expressing position error and velocity error as state variables.

$$\mathbf{s}(t)^l = \underbrace{\begin{bmatrix} 1 & 0 & \nabla t & 0 \\ 0 & 1 & 0 & \nabla t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{M}} \mathbf{s}(t-1)^l + \mathbf{w}(t)^l, \tag{1}$$

where $\mathbf{s}(t)^l = \begin{bmatrix} \nabla x_t^l & \nabla y_t^l & \nabla V x_t^l & \nabla V y_t^l \end{bmatrix}$ is the state vector of the lth local filter at the time index t. At time t, $\left(\nabla x_t^l, \nabla y_t^l\right)$ and $\left(\nabla V x_t^l, \nabla V y_t^l\right)$ represent the position error and velocity error in the x and y directions, respectively. ∇t represents the sample time, $\mathbf{w}(t)^l \sim \mathcal{N}\left(0, \mathbf{Q}^l\right)$ is the process noise.

The measurement equation used in this work can be written as follows according to [1].

$$
\begin{aligned}
\mathbf{y}(t)^l = \delta r_{l,t}^2 &= (r_{l,t}^I)^2 - (d_{l,t}^U)^2 \\
&= 2\left(x_t^{I,l} - x^l\right)\nabla x_t^l + 2\left(y_t^{I,l} - y^l\right)\nabla y_t^l - \left(\left(\nabla x_t^l\right)^2 + \left(\nabla y_t^l\right)^2\right), \\
&= \mathrm{h}(\mathbf{s}(t)^l) + \gamma_t^l
\end{aligned}
\tag{2}
$$

where $\left(x_t^I, y_t^I\right)$ is the INS position, $\left(x_t^l, y_t^l\right)$ is the lth UWB RN's position, $\gamma(t)^l \sim \mathcal{N}\left(0, \mathbf{R}^l\right)$ is the measurement noise.

3   Distributed Dual-EKF Filter

In this section, the dual-EKF filter, which is used as the local filter based on the model (1) (2) will be designed. Firstly, we rewrite the model (1) (2) as follows:

$$
\begin{cases}
\mathbf{s}(t)^l = \mathbf{M}(t-1)^l \mathbf{s}(t-1)^l + \mathbf{w}(t)^l \\
\mathbf{y}(t)^l = \mathrm{h}(\mathbf{s}(t)^l) + \gamma_t^l \\
\mathbf{M}(t)^l = \mathbf{M}(t-1)^l + \mathbf{w^M}(t)^l
\end{cases},
\tag{3}
$$

In this work, the system parameter $\mathbf{M}(t)^l$ is used as the state vector, which means that the uncertain of the system is considered in modified fusion model; the $\mathbf{w^M}(t)^l \sim N\left(0, \mathbf{Q^{M,l}}\right)$ is the noise of the $\mathbf{M}(t)^l$. Based on the model (3), the dual-EKF filter can be designed in the following section.

Firstly, with the initial value $\mathbf{s}(0)^l$ and $\mathbf{P_s}(0)^l$, the $\mathbf{s}(t)^l$ and $\mathbf{P_s}(t)^l$ can be predicted at the time index t via Eqs. (4) (5).

$$
\mathbf{s}(t)^l = \mathbf{M}(t-1)^l \mathbf{s}(t-1)^l + \mathbf{w}(t)^l,
\tag{4}
$$

$$
\mathbf{P_s}(t)^l = \mathbf{M}(t-1)^l \mathbf{P_s}(t-1)\left(\mathbf{M}(t-1)^l\right)^T + \mathbf{Q}^l,
\tag{5}
$$

where $\mathbf{P_s}(t)^l$ represents the estimation error of $\mathbf{s}(t)^l$. Then, the $\mathbf{s}(t)^l$ and $\mathbf{P_s}(t)^l$ can be updated at the time index t via Eqs. (6) (7) (8).

$$
\mathbf{K}(t)^l = \mathbf{P_s}(t)^l \left(\mathbf{H}(t)^l\right)^T [\mathbf{R}(t)^l + \mathbf{H}(t)^l \mathbf{P_s}(t)^l \left(\mathbf{H}(t)^l\right)^T]^{-1},
\tag{6}
$$

$$
\mathbf{s}(t)^l = \mathbf{s}(t)^l + \mathbf{K}(t)^l [\mathbf{y}(t)^l - h_t(\mathbf{s}(t)^l)],
\tag{7}
$$

$$
\mathbf{P_s}(t)^l = [\mathbf{I} - \mathbf{K}(t)^l \mathbf{H}(t)^l] \mathbf{P_s}(t)^l,
\tag{8}
$$

where $\mathbf{H}(t)^l = \frac{\partial h_t(\mathbf{s}(t)^l)}{\partial \mathbf{s}(t)^l}$.

Based on the $\mathbf{s}(t)^l$ and $\mathbf{P_s}(t)^l$, the $\mathbf{M}(t)^l$ and $\mathbf{P_M}(t)^l$ can be predicted at the time index t via Eqs. (9) (10).

$$\mathbf{M}(t)^l = \mathbf{M}(t-1)^l + \mathbf{w^M}(t)^l\,,\tag{9}$$

$$\mathbf{P_M}(t)^l = \mathbf{P_M}(t-1) + \mathbf{Q^{M,l}}\,,\tag{10}$$

And then, $\mathbf{M}(t)^l$ and $\mathbf{P_M}(t)^l$ can be updated at the time index t via Eqs. (11) (12) (13).

$$\mathbf{K}(t)^{\mathbf{M},l} = \mathbf{P_M}(t)^l \left(\mathbf{H}(t)^{\mathbf{M},l}\right)^T [\mathbf{R}(t)^{\mathbf{M},l} + \mathbf{H}(t)^{\mathbf{M},l}\mathbf{P_M}(t)^l \left(\mathbf{H}(t)^{\mathbf{M},l}\right)^T]^{-1}\,,\tag{11}$$

$$\mathbf{M}(t)^l = \mathbf{M}(t)^l + \mathbf{K}(t)^{\mathbf{M},l}[\mathbf{y}(t)^l - h_t(\mathbf{s}(t)^l)]\,,\tag{12}$$

$$\mathbf{P_M}(t)^l = [\mathbf{I} - \mathbf{K}(t)^{\mathbf{M},l}\mathbf{H}(t)^{\mathbf{M},l}]\mathbf{P_M}(t)^l\,,\tag{13}$$

where $\mathbf{H}(t)^{\mathbf{M},l} = \frac{\partial h_t(\mathbf{M}(t)^l)}{\partial \mathbf{M}(t)^l}$.

With the local EKF's output $\mathbf{s}(t)^l$ and $\mathbf{P_s}(t)^l$, the main filter works to provide the optimal output by fusing $\mathbf{s}(t)^l$ and $\mathbf{P_s}(t)^l$ via Eqs. (14) (15).

$$\mathbf{s}(t) = \mathbf{P_s}(t)\left(\left(\mathbf{P_s}(t)^1\right)^{-1}\mathbf{s}(t)^1 + \left(\mathbf{P_s}(t)^2\right)^{-1}\mathbf{s}(t)^2 + \left(\mathbf{P_s}(t)^3\right)^{-1}\mathbf{s}(t)^3\right.$$
$$\left. + \left(\mathbf{P_s}(t)^4\right)^{-1}\mathbf{s}(t)^4\right)\,,\tag{14}$$

$$\mathbf{P_s}(t) = \left(\left(\mathbf{P_s}(t)^1\right)^{-1} + \left(\mathbf{P_s}(t)^2\right)^{-1} + \left(\mathbf{P_s}(t)^3\right)^{-1} + \left(\mathbf{P_s}(t)^4\right)^{-1}\right)^{-1}.\tag{15}$$

4   Experimental Testing

This chapter will design experiments to prove the performance of the proposed algorithm. Firstly, clarify the experimental environment and experimental application equipment. Secondly, compare the proposed method with distributed EKF filters, UWB and INS.

4.1   Experimental Environment

In this experiment, one inertial measurement unit (IMU), one UWB BN and four UWB RNs were used. The author carried a fixed bracket, and the UWB BN was fixed on the bracket. The base station is UWB RN, and UWB RN is fixed at a known point in the test site. The positions of UWB BN and UWB RN should be on the same horizontal plane as possible. The IMU is placed on the feet of the target person in the correct direction. The test environment, experimental equipment and target personnel are shown in Figs. 2 and 3.

Fig. 2. Test environment.

Fig. 3. Experimental equipment and target personnel.

4.2 Performance Analysis of the Proposed Algorithm

In the following, we will discuss the performance of the above algorithm. For a more convenient observation, we compare the positioning error of the proposed method with the distributed EKF filter, UWB and INS. It can be seen that this method can effectively improve accuracy and reduce errors.

Figure 4 shows the estimated path provided by INS, distributed EKF filter and distributed dual-EKF filter, represented by green lines, black dashed lines, purple lines and red lines, respectively. Moreover, the UWB RNs are also shown in the figure. From the figure, one can infer easily that the INS's path has obvious error accumulation, and the UWB has the stable solution. The comparison shows that the distributed EKF filter and the distributed double EKF filter can provide accurate paths. Figure 5 shows the CDF of the positioning error of UWB, distributed EKF filter and distributed dual-EKF filter. Through experiments, it is known that the distributed dual-EKF can better reduce the error. The proposed method's error is about 0.27 m, and that value of the distributed EKF filter is 0.3 m. Moreover, the Table 1 shows the RMSE of the INS, UWB, distributed EKF filter, and distributed dual-EKF filter. It can be confirmed that the distributed dual-EKF proposed in this experiment is better than other algorithms.
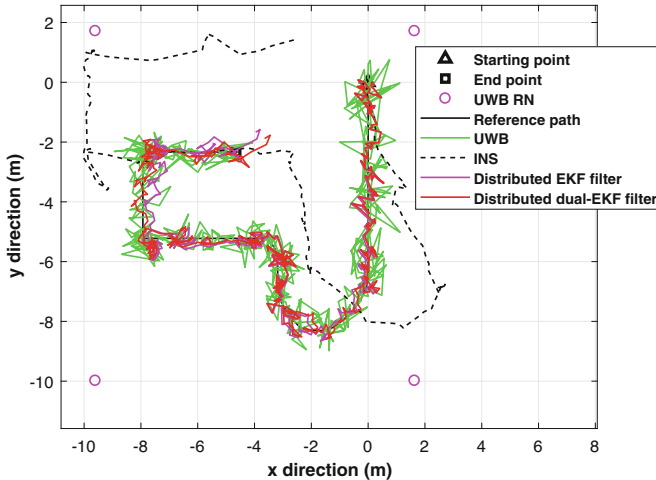

Fig. 4. The estimated path is provided by distributed dual EKF, distributed EKF, UWB and INS.
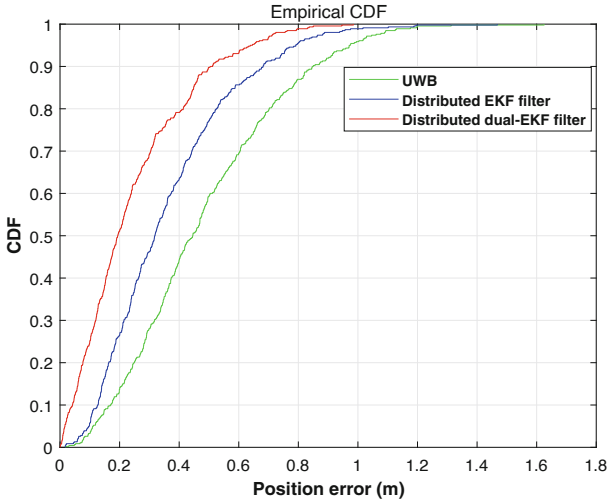
Fig. 5. The CDFs of the distributed dual-EKF, distributed EKF, and UWB.

Table 1. The RMSE(m) generated by INS, UWB, EKF and Dual-EKF

Filter	RMSE	
	East	North
INS	1.62	1.98
UWB	0.39	0.39
EKF	0.36	0.23
Dual-EKF	0.32	0.22

5   Conclusion

The positioning algorithm based on the distributed dual-EKF filter proposed in this paper can better improve the positioning error of human tracking. In this algorithm, it uses four dual-EKF filters as the local data fusion filter, the output of the local filter is used as the input of the main filter, and the fusion provides the best estimate. Through a large number of experiments in this article, it can be proved that the dual-EKF filter can better reduce the error.

References

1. Bu, L., Zhang, Y., Xu, Y.: Indoor pedestrian tracking by combining recent INS and UWB measurements. In: 2017 International Conference on Advanced Mechatronic Systems (ICAMechS), pp. 244–248. IEEE (2017)
2. Jiménez, A.R., Seco, F., Prieto, J.C., Guevara, J.: Indoor pedestrian navigation using an INS/EKF framework for yaw drift reduction and a foot-mounted IMU. In: 2010 7th Workshop on Positioning, Navigation and Communication, pp. 135–143. IEEE(2010)
3. Li, J., Ke, X.: Study on autonomous navigation based on pulsar timing model. Sci. China, Ser. G: Phys. Mech. Astron. **52**(2), 303–309 (2009)
4. Li, X., Zhang, X., Ren, X., Fritsche, M., Wickert, J., Schuh, H.: Precise positioning with current multi-constellation global navigation satellite systems: GPS, GLONASS Galileo and BeiDou. Sci. Rep. **5**(1), 1–14 (2015)
5. Wang, Y., Li, X.: An improved robust EKF algorithm based on sigma points for UWB and foot-mounted IMU fusion positioning. J. Spat. Sci. **66**(2), 329–350 (2021)
6. Xu, Y., Shmaliy, Y.S., Li, Y., Chen, X.: UWB-based indoor human localization with time-delayed data using EFIR filtering. IEEE Access **5**, 16676–16683 (2017)
7. Yang, J., et al.: A height constrained adaptive Kalman filtering based on climbing motion model for GNSS positioning. IEEE Sens. J. **17**(21), 7105–7113 (2017)

Evaluation and Analysis of Basic-Level Aircraft Maintenance and Support Capabilities

Yanli Gao[1(✉)], Shaokang Ji[1], Jianling Qu[1], and Mingran Li[2]

[1] Qingdao Branch of Naval Aviation University, No. 2, Siliu Middle Road,
Qingdao 266000, China
gaoyanli2003@163.com
[2] School of Electrical Engineering, University of Jinan, No. 336, Nanxinzhuang West Road,
Jinan 250022, China

Abstract. This paper provides a scientific evaluation of the basic-level aircraft maintenance and support capabilities. According to the principles of objectivity and representativeness, one evaluation index system is selected and constructed, and then the weight of each index is determined by the analytic hierarchy process combined with the Delphi method. Using the fuzzy comprehensive evaluation method, the basic-level aircraft maintenance and support capabilities evaluation model is constructed and applied. The analysis and verification results show that the model can effectively evaluate the basic-level aircraft maintenance and support capabilities.

Keywords: Aircraft maintenance · Support capabilities

1 Introduction

Maintenance and support capabilities are the abilities to ensure a good completion of maintenance tasks, which is of great significance to the flight safety of aircraft. In recent years. With the smooth and well progress of the reform process, our army's new aircraft have continued to enter the working list, and the demand for aviation equipment maintenance and support has increased. For this reason, it is necessary for us to carry out an evaluation of aircraft maintenance and support capabilities. The results of the evaluation can help the development plan of the army and are of great significance to improving the combat effectiveness of the army. This paper conducts an evaluation study on the aircraft maintenance and support capabilities at the basic-level.

Considering the actual situation of basic-level aircraft maintenance and support activities, taking basic-level aircraft maintenance and support capabilities as the research object, based on the construction of an index system, using analytic hierarchy process, Delphi method, and comprehensive evaluation techniques to study the basic-level aircraft comprehensive evaluation model maintenance and support. Later, we analyse the security maintenance capabilities barrier system maintenance and support tasks to complete, and then pointed out that according to the results reflected in the constraints of the basic-level and maintenance and support the existence of the phenomenon and to suggest improvements.

2   Constructing the Index System

To evaluate basic-level aircraft maintenance and support capabilities, we must first construct an evaluation index system, which is an important basis for carrying out evaluation activities. This paper follows objectivity, hierarchy, representation, independence of the principles to construct the evaluation index system. It uses Delphi method to screen evaluation indices, the expert group on screening results given by consulting the relevant experts to summarize the collective opinion, after several soliciting opinions, feedback and adjustments, analysis and sorting out relatively unified results (Table 1).

Table 1. Basic-level aircraft maintenance and support capabilities evaluation index system

Target layer	First level indices	Secondary level indices	Index type
Basic-level aircraft maintenance and support capabilities	Maintenance human resources A1	Professional allocation rate B11	Quantitative
		Locomotive maintenance style B12	Qualitative
		Professional technical level of personnel B13	Qualitative
		Personnel exchange status B14	Qualitative
	Maintenance facility equipment A2	Facilities and equipment types matching rate B21	Quantitative
		Allocation rate of equipment and facilities B22	Quantitative
		Facility equipment maintenance level B23	Qualitative
		Facility and equipment safety level B24	Qualitative
	Spare parts equipment guarantee A3	Spare parts equipment matching rate B31	Quantitative
		Spare parts equipment supply capacity B32	Qualitative
		Spare parts equipment storage level B33	Qualitative
		Integrity rate of spare parts and equipment B34	Quantitative
	Information resources A4	Operation card completeness B41	Qualitative

Table 1. (*continued*)

Target layer	First level indices	Secondary level indices	Index type
		Completeness of maintenance image data B42	Qualitative
		Completeness of equipment technical data B43	Qualitative
		Applicability of equipment technical data B44	Qualitative
	Organizational management level A5	Perfection of rules and regulations B51	Qualitative
		Division of powers B52	Qualitative
		Maintenance training situation B53	Qualitative
		Aircraft maintenance culture construction level B54	Qualitative

3 Determining the Weights of the Evaluation Indices

In the comprehensive evaluation, each index in the evaluation index system has a large or small effect on the evaluated object, and its importance is different. Therefore, it needs to be weighted. Weight is a numerical value that measures the degree of influence of each index relative to the evaluated object. In the process of comprehensive evaluation, determining the weight of each evaluation index is a very sensitive and critical link, because whether the weight is reasonable or not will be directly determined. Affect the reliability of the evaluation results.

This paper chooses the analytic hierarchy process combined with the Delphi method as the method to determine the index weights in the basic-level aircraft maintenance and support capabilities evaluation work, and the Delphi method is used to construct the decision matrix. The realization process is shown in Fig. 1.

According to the constructed decision matrix to solve the weight of each index, this paper adopts the sum method, and the specific solution is as follows:

Assume A is a decision matrix, each column is normalized to get $B = \left(b_{ij}\right)_{n \times n}$, where

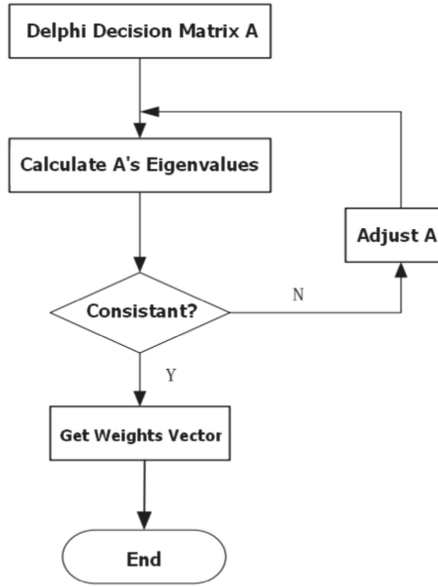$$\overline{b_{ij}} = \frac{b_{ij}}{\sum_{k=1}^{n} b_{kj}} (i, j = 1, 2, \ldots, n) \tag{1}$$

Fig. 1. AHP-Delphi empowerment flow chart

Then conduct addition by rows to get the vector $C = (c_1, c_2, \ldots, c_n)^T$, where

$$c_i = \sum_{j=1}^{n} b_{ij} (i = 1, 2, \ldots, n) \tag{2}$$

After normalization of $C = (c_1, c_2, \ldots, c_n)^T$, the weight of each index is obtained:

$$w_i = \frac{C_i}{\sum_{k=1}^{n} C_k} (i = 1, 2, \ldots, n) \tag{3}$$

The comparison results obtained by the Delphi method may be contrary to the original intention or self-contradictory, which will cause the consistency of the decision matrix to deviate. Therefore, to ensure the reliability of the determined weights, the matrix must be checked.

First, find the largest characteristic root of the decision matrix:

$$\lambda_{max} = \frac{1}{n} \sum_{i=1}^{n} \frac{(Aw)_i}{w_i} \tag{4}$$

The quantitative index used to measure the degree of consistency is called the consistency index, which is generally represented by CI:

$$CI = \frac{\lambda max - n}{n - 1} \tag{5}$$

In previous work, it has been introduced that there are five first-level indices under the basic-level aircraft maintenance and support capabilities T, which are maintenance human resources A1, maintenance equipment and facilities A2, spare equipment support A3, information resources A4, and organizational management level A5, which are now determined the five-level indices right weight, the expert scoring results in Table 2.

Table 2. T-Ai decision matrix

T	A1	A2	A3	A4	A5
A1	1	3	2	5	8
A2	1/3	1	1/2	3	5
A3	1/2	2	1	4	7
A4	1/5	1/3	1/4	1	3
A5	1/8	1/5	1/7	1/3	1

According to the formula (1), we first normalize each column of the matrix to obtain:

$$B = \begin{bmatrix} 0.4633 & 0.4592 & 0.5138 & 0.3750 & 0.3333 \\ 0.1544 & 0.1531 & 0.1284 & 0.2250 & 0.2083 \\ 0.2317 & 0.3061 & 0.2569 & 0.3000 & 0.2917 \\ 0.0927 & 0.0510 & 0.0642 & 0.0750 & 0.1250 \\ 0.0579 & 0.0306 & 0.0367 & 0.0250 & 0.0417 \end{bmatrix}$$
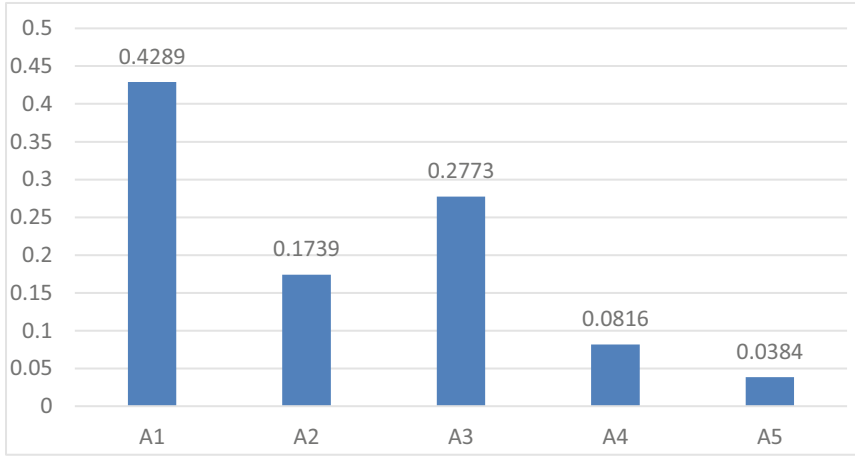
Later, sum the matrix B by rows to get C:

$$C = \begin{bmatrix} 2.1446 \\ 0.8693 \\ 1.3863 \\ 0.4079 \\ 0.1919 \end{bmatrix}$$

Normalize the matrix C to obtain the weight of each index at the first level

$$w_T = \begin{bmatrix} 0.4289 \\ 0.1739 \\ 0.2773 \\ 0.0816 \\ 0.0384 \end{bmatrix}$$

The following is the consistency test of the decision matrix of T-Ai

$$AW = \begin{bmatrix} 1 & 3 & 2 & 5 & 8 \\ 1/3 & 1 & 1/2 & 3 & 5 \\ 1/2 & 2 & 1 & 4 & 7 \\ 1/5 & 1/3 & 1/4 & 1 & 3 \\ 1/8 & 1/5 & 1/7 & 1/3 & 1 \end{bmatrix} \begin{bmatrix} 0.4289 \\ 0.1739 \\ 0.2773 \\ 0.0816 \\ 0.0384 \end{bmatrix} = \begin{bmatrix} 2.2200 \\ 0.8921 \\ 1.4344 \\ 0.4098 \\ 0.1936 \end{bmatrix}$$

Thus, maximum eigenvalues are:

$$\lambda_{max} = \frac{1}{n} \sum_{i=1}^{n} \frac{(Aw)_i}{w_i} = \frac{1}{5} \sum_{i=1}^{5} \frac{(Aw)_i}{w_i} = 5.1094$$

Consistency index value

$$CI = \frac{\lambda max - n}{n - 1} = \frac{5.1094 - 5}{5 - 1} = 0.0273$$

By looking up the table, we know that the correction value of the fifth-order matrix is $RI = 1.12$, then $CR = CI/RI = 0.0273/1.12 = 0.0244 < 0.1$, indicating that the decision matrix has passed the test and has good consistency.

4   Constructing a Comprehensive Evaluation Model

This work uses fuzzy method to construct the evaluation model.

4.1   Construction of Domain of Factors

Domain of factors $\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ within which factors the assessed object is composed of and u_n means the n-th evaluation index.

4.2   Construction of Domain of Comment Levels

The domain of comment levels $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ is a collection of various evaluation results of the evaluated object, which is used to describe the state after the evaluation of each factor, that is, to define each interval of the state of the evaluated object. Among them, it represents the total number of state intervals. People usually use appropriate vocabulary to describe the comment levels of different situations. For example, to evaluate the risk levels of epidemics in various regions, you can use $V = \{$high risk, medium risk, low risk$\}$ to describe. For this paper, to evaluate the basic level aircraft maintenance and support, we can use $V = \{$excellent, good, fair, bad$\}$ to describe, and then assign the corresponding score interval to this comment.

4.3   Construction of Fuzzy Relation Matrix R

Individually evaluate the underlying factors that affect the evaluated object to obtain the fuzzy relationship matrix $\mathbf{R}$:

$$R = \begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1m} \\ r_{21} & r_{22} & \ldots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \ldots & r_{nm} \end{bmatrix} \tag{6}$$

where $r_{ij} (i = 1, 2, \ldots, n; j = 1, 2, \ldots, m)$ represents degree of the i-th factor in the U domain belonging to the j-th element of the V domain. Generally, we have the rule that r_{ij} = the number of people whose u_i is v_j divided by total number of people who participated in the evaluation. After normalization like this, the matrix R has no dimensions, and there is no need to further process the matrix.

4.4   Multi-factor Fuzzy Comprehensive Evaluation

Next, the calculated fuzzy weight vector W and the corresponding fuzzy relationship matrix R are integrated. This integration requires a suitable synthesis operator and gives the result vector Z of each evaluated object. The result vector Z is used to represent the degree to which each evaluated object belongs to the domain of each comment level

from the overall point of view. The formula model of fuzzy comprehensive evaluation is as follows:

$$Z = W°R = (w_1, w_2, \ldots, w_n) \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} = (z_1, z_2, \ldots, z_n) \qquad (7)$$

To evaluate the results more intuitively, it is necessary for vector Z to be addressed. That is, to put it into a specific overall score by mathematical calculations. The method of processing vector Z generally uses the maximum membership rule and the weighted average rule. As the maximum membership rule only uses the maximum value of $z_i (i = 1, 2, \ldots, m)$ and cannot fully utilize other information brought by Z, the result is relatively unneutral, while the weighted average rule can evaluate the result vector Z is integrated with the parameters of the comment rating domain, so that the final evaluation result is more scientific **Error! Reference source not found**. Therefore, this paper uses the weighted average vector of the Z.

4.5 Selection of Composition Operator

When integrating the weight vector and fuzzy relationship matrix of each evaluation index of the basic-level aircraft maintenance and support capabilities, the choice of composition operator is very important. Different composition operators correspond to different evaluation models, and the resulting evaluation results are also different. Each is different. In this paper, the M ($\bullet$, $\oplus$) operator can be used to balance the weight of all factors and fully reflect the essential characteristics of the evaluated object.

$$z_j = \min\left\{1, \sum_{i=1}^{n} \left(w_i \cdot r_{ij}\right)\right\}, j = 1, 2, \ldots, m \qquad (8)$$

5 Application and Analysis

to be able to accurately evaluate the aircraft maintenance and support capabilities of the maintenance organization, based on the construction of an evaluation index system, this paper combines the analytic hierarchy process and the Delphi method to determine the weights of indices at all levels, and uses FCE for comprehensive evaluation to construct basic-level aircraft maintenance and support capabilities integrated evaluation model, evaluation and analysis of the maintenance organization, as shown in Fig. 2.
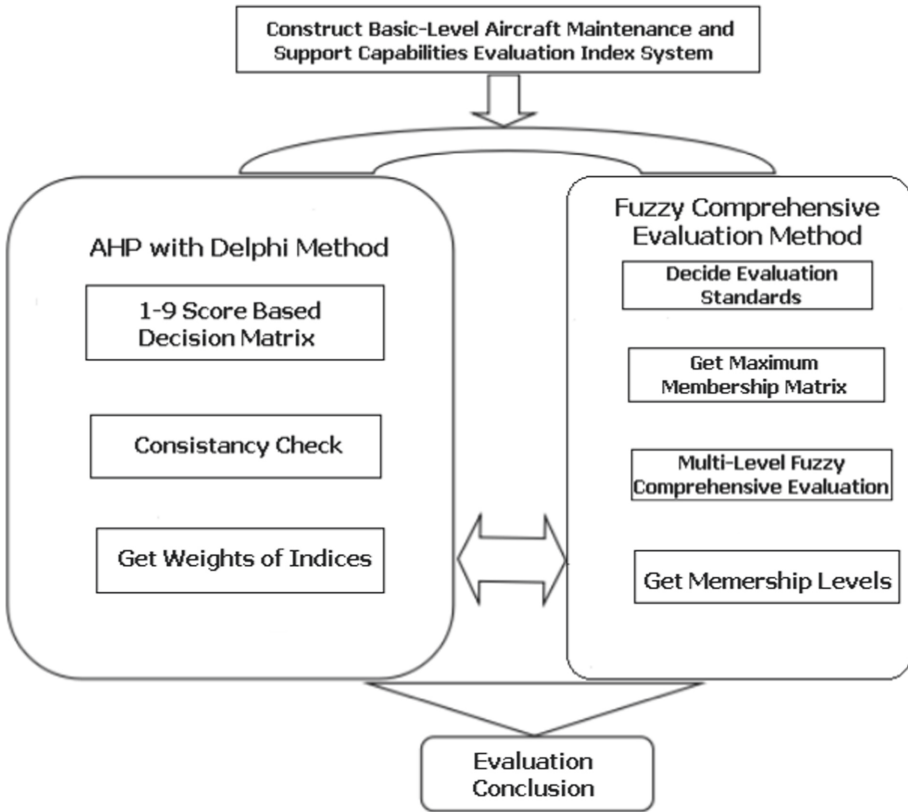
Fig. 2. Basic-level aircraft maintenance and support capabilities evaluation system model

The experiments include 15 officers of the unit to judge the actual situation of the evaluation indices, equipment maintenance and support personnel, including experts, ordinary employees and so on, the final evaluation of the results of statistical results of evaluation indices are shown in Table 3.

The fuzzy relationship matrix of maintenance human resources:

$$R_{A1} = \begin{bmatrix} 8/15 & 1/3 & 2/15 & 0 \\ 13/15 & 2/15 & 0 & 0 \\ 7/15 & 8/15 & 0 & 0 \\ 1/3 & 3/5 & 1/15 & 0 \end{bmatrix}$$

Fuzzy relationship matrix of maintenance equipment and facilities:

$$R_{A2} = \begin{bmatrix} 7/15 & 1/3 & 1/15 & 2/15 \\ 4/15 & 3/5 & 2/15 & 0 \\ 7/15 & 8/15 & 0 & 0 \\ 3/5 & 4/15 & 2/15 & 0 \end{bmatrix}$$

Table 3. Basic-level aircraft maintenance and support capabilities evaluation basic information collection form.

Evaluation items	Evaluation results	Comment Level			
		Excellent	Good	Normal	Bad
Maintenance human resources	Professional allocation rate	8	5	2	0
	Locomotive maintenance style	13	2	0	0
	Professional technical level of personnel	7	8	0	0
	Personnel exchange status	5	9	1	0
Maintenance facilities and equipment	Facilities and equipment types supporting rate	7	5	1	2
	Facility equipment quantity allocation rate	4	9	2	0
	Facility equipment maintenance level	7	8	0	0
	Facility and equipment safety	9	4	2	0
Spare parts equipment guarantees	Spare parts equipment matching rate	5	9	1	0
	Spare parts equipment supply capacity	6	8	1	0
	Spare parts equipment storage level	9	6	0	0
	Intact rate of spare parts and equipment	5	8	2	0
Information resources	Operation card completeness	12	3	0	0
	Completeness of repair image data	6	7	1	1
	Completeness of equipment technical data	13	2	0	0
	Applicability of equipment technical data	13	2	0	0
Organizational management level	Perfection of rules and regulations	13	2	0	0
	Division of powers	12	3	0	0
	Maintenance training situation	8	7	0	0

(*continued*)

Table 3. (*continued*)

Evaluation items	Evaluation results	Comment Level			
		Excellent	Good	Normal	Bad
	The level of maintenance culture	4	9	2	0

Fuzzy relationship matrix of spare parts equipment guarantee:

$$R_{A3} = \begin{bmatrix} 1/3 & 3/5 & 1/15 & 0 \\ 2/5 & 8/15 & 1/15 & 0 \\ 3/5 & 2/5 & 0 & 0 \\ 1/3 & 8/15 & 2/15 & 0 \end{bmatrix}$$

Fuzzy relationship matrix of information resources:

$$R_{A4} = \begin{bmatrix} 4/5 & 1/5 & 0 & 0 \\ 2/5 & 7/15 & 1/15 & 1/15 \\ 13/15 & 2/15 & 0 & 0 \\ 13/15 & 2/15 & 0 & 0 \end{bmatrix}$$

Fuzzy relationship matrix of organizational management level:

$$R_{A5} = \begin{bmatrix} 13/15 & 2/15 & 0 & 0 \\ 4/5 & 1/5 & 0 & 0 \\ 8/15 & 7/15 & 0 & 0 \\ 4/15 & 3/5 & 2/15 & 0 \end{bmatrix}$$

The calculated index weights are as follows (Table 4):

Table 4. Weights of each evaluation index

Target layer	First level indices	Weights	Secondary level indices	Weights
Basic-level aircraft maintenance and support capabilities	Maintenance human resources A1	0.4289	Professional allocation rate B11	0.5681
			Locomotive maintenance style B12	0.1333
			Professional technical level of personnel B13	0.2410
			Personnel exchange status B14	0.0576

(*continued*)

Table 4. (*continued*)

Target layer	First level indices	Weights	Secondary level indices	Weights
	Maintenance facility equipment A2	0.1739	Facilities and equipment types matching rate B21	0.2083
			Allocation rate of equipment and facilities B22	0.1010
			Facility equipment maintenance level B23	0.0480
			Facility and equipment safety level B24	0.6427
	Spare parts equipment guarantee A3	0.2773	Spare parts equipment matching rate B31	0.4209
			Spare parts equipment supply capacity B32	0.1064
			Spare parts equipment storage level B33	0.0517
			Integrity rate of spare parts and equipment B34	0.4209
	Information resources A4	0.0816	Operation card completeness B41	0.1171
			Completeness of maintenance image data B42	0.0509
			Completeness of equipment technical data B43	0.2695
			Applicability of equipment technical data B44	0.5625
	Organizational management Level A5	0.0384	Perfection of rules and regulations B51	0.3273
			Division of powers B52	0.5359
			Maintenance training situation B53	0.0916
			Aircraft maintenance culture construction level B54	0.0453

Based on the obtained weights, the composite operator is used to integrate the weight vector and the fuzzy relationship matrix, and then the integration result is processed according to the weighted average method to obtain the scores of the evaluation results at all levels.

Maintenance human resources A1, according to the fuzzy transformation formula (7), we can get:

$$Z_{A1} = w_{A1}^{\circ} R_{A1}$$

$$= \begin{bmatrix} 0.5681 & 0.1333 & 0.2410 & 0.0576 \end{bmatrix} \circ \begin{bmatrix} 8/15 & 1/3 & 2/15 & 0 \\ 13/15 & 2/15 & 0 & 0 \\ 7/15 & 8/15 & 0 & 0 \\ 1/3 & 3/5 & 1/15 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5502 & 0.3702 & 0.0796 & 0 \end{bmatrix}$$

The evaluation result vector of maintenance facilities and equipment A2 is:

$$Z_{A2} = w_{A2}^{\circ} R_{A2}$$

$$= \begin{bmatrix} 0.2083 & 0.1010 & 0.0480 & 0.6427 \end{bmatrix} \circ \begin{bmatrix} 7/15 & 1/3 & 1/15 & 2/15 \\ 4/15 & 3/5 & 2/15 & 0 \\ 7/15 & 8/15 & 0 & 0 \\ 3/5 & 4/15 & 2/15 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5322 & 0.3270 & 0.1130 & 0.0278 \end{bmatrix}$$

The evaluation result vector of spare parts equipment guarantee A3 is:

$$Z_{A3} = w_{A3}^{\circ} R_{A3}$$

$$= \begin{bmatrix} 0.4209 & 0.1064 & 0.0517 & 0.4209 \end{bmatrix} \circ \begin{bmatrix} 1/3 & 3/5 & 1/15 & 0 \\ 2/5 & 8/15 & 1/15 & 0 \\ 3/5 & 2/5 & 0 & 0 \\ 1/3 & 8/15 & 2/15 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.3542 & 0.5544 & 0.0913 & 0 \end{bmatrix}$$

The information resource A4 evaluation result vector is:

$$Z_{A4} = w_{A4}^{\circ} R_{A4}$$

$$= \begin{bmatrix} 0.1171 & 0.0509 & 0.2695 & 0.5625 \end{bmatrix} \circ \begin{bmatrix} 4/5 & 1/5 & 0 & 0 \\ 2/5 & 7/15 & 1/15 & 1/15 \\ 13/15 & 2/15 & 0 & 0 \\ 13/15 & 2/15 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.8351 & 0.1581 & 0.0034 & 0.0034 \end{bmatrix}$$

The organization management level A5 evaluation result vector is (Fig. 3):

$$Z_{A5} = w_{A5}^{\circ} R_{A5}$$

$$= \begin{bmatrix} 0.3273 & 0.5359 & 0.0916 & 0.0453 \end{bmatrix}^{\circ} \begin{bmatrix} 13/15 & 2/15 & 0 & 0 \\ 4/5 & 1/5 & 0 & 0 \\ 8/15 & 7/15 & 0 & 0 \\ 4/15 & 3/5 & 2/15 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.7733 & 0.2207 & 0.0060 & 0 \end{bmatrix}$$

According to a fuzzy calculation results, maintenance of human resources $Z_{A1} = \begin{bmatrix} 0.5502 & 0.3702 & 0.0796 & 0 \end{bmatrix}$, the table shows there are 55.02 percent of human resource evaluation of the maintenance personnel of the participating units for outstanding, 37.02% of eligible staff evaluated as good, as well as 7.96% of eligible staff evaluation in general, Draw a histogram of the evaluation result vector of each first-level index as shown in Fig. 4.
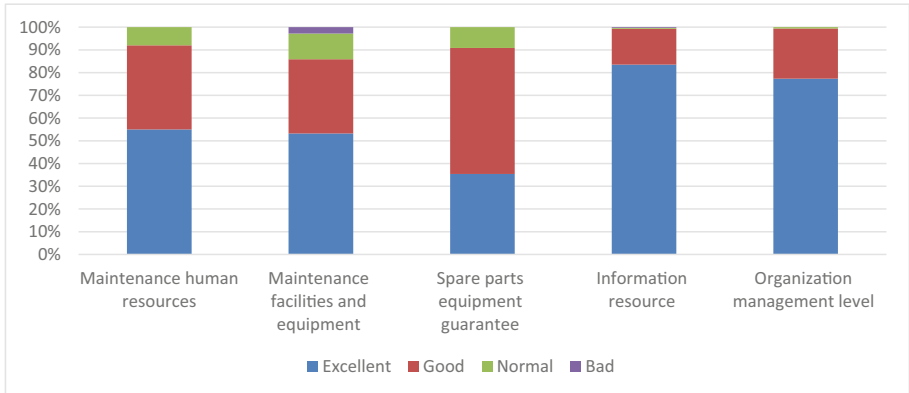
Fig. 3. First-level fuzzy comprehensive evaluation

According to the formula $S = Z \cdot V^T$, the total score of each first-level index can be obtained.

Maintenance human resources A1 total score:

$$S_{A1} = Z_{A1} \cdot V^T = \begin{bmatrix} 0.5502 & 0.3702 & 0.0796 & 0 \end{bmatrix} \cdot \begin{bmatrix} 95 \\ 80 \\ 60 \\ 25 \end{bmatrix} = 86.6610$$

Total score for maintenance facilities and equipment A2:

$$S_{A2} = Z_{A2} \cdot V^T = \begin{bmatrix} 0.5322 & 0.3270 & 0.1130 & 0.0278 \end{bmatrix} \cdot \begin{bmatrix} 95 \\ 80 \\ 60 \\ 25 \end{bmatrix} = 84.1939$$

Spare parts equipment guarantees A3 total score:

$$S_{A3} = Z_{A3} \cdot V^T = \begin{bmatrix} 0.3542\ 0.5544\ 0.0913\ 0 \end{bmatrix} \cdot \begin{bmatrix} 95 \\ 80 \\ 60 \\ 25 \end{bmatrix} = 83.4792$$

Information resources A4 total score:

$$S_{A4} = Z_{A4} \cdot V^T = \begin{bmatrix} 0.8351\ 0.1581\ 0.0034\ 0.0034 \end{bmatrix} \cdot \begin{bmatrix} 95 \\ 80 \\ 60 \\ 25 \end{bmatrix} = 92.2721$$

Organizational management level A5 total score:

$$S_{A5} = Z_{A5} \cdot V^T = \begin{bmatrix} 0.7733\ 0.2207\ 0.0060\ 0 \end{bmatrix} \cdot \begin{bmatrix} 95 \\ 80 \\ 60 \\ 25 \end{bmatrix} = 91.4869$$

Table 5. First-level evaluation result of each indices

Index	Result vector Z	Score S
A1	$\begin{bmatrix} 0.5502\ 0.3702\ 0.0796\ 0 \end{bmatrix}$	86.6610
A2	$\begin{bmatrix} 0.5322\ 0.3270\ 0.1130\ 0.0278 \end{bmatrix}$	84.1939
A3	$\begin{bmatrix} 0.3542\ 0.5544\ 0.0913\ 0 \end{bmatrix}$	83.4792
A4	$\begin{bmatrix} 0.8351\ 0.1581\ 0.0034\ 0.0034 \end{bmatrix}$	92.2721
A5	$\begin{bmatrix} 0.7733\ 0.2207\ 0.0060\ 0 \end{bmatrix}$	91.4869

According to the evaluation result table, the indices with excellent evaluation results include information resources and organizational management level, and the indices with good evaluation results include maintenance human resources, maintenance equipment facilities and spare parts equipment guarantee.

The comprehensive evaluation can be done. According to Table 5, the target T fuzzy relationship matrix is:

$$R_T = \begin{bmatrix} Z_{A1} \\ Z_{A2} \\ Z_{A3} \\ Z_{A4} \\ Z_{A5} \end{bmatrix} = \begin{bmatrix} 0.5502 & 0.3702 & 0.0796 & 0 \\ 0.5322 & 0.3270 & 0.1130 & 0.0278 \\ 0.3542 & 0.5544 & 0.0913 & 0 \\ 0.8351 & 0.1581 & 0.0034 & 0.0034 \\ 0.7733 & 0.2207 & 0.0060 & 0 \end{bmatrix}$$

Then the result vector of the basic-level aircraft maintenance and support capabilities is:

$$Z_T = w_T^{\circ} R_T$$

$$= \begin{bmatrix} 0.4289 & 0.1739 & 0.2773 & 0.0816 & 0.0384 \end{bmatrix}^{\circ} \begin{bmatrix} 0.5502 & 0.3702 & 0.0796 & 0 \\ 0.5322 & 0.3270 & 0.1130 & 0.0278 \\ 0.3542 & 0.5544 & 0.0913 & 0 \\ 0.8351 & 0.1581 & 0.0034 & 0.0034 \\ 0.7733 & 0.2207 & 0.0060 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5246 & 0.3908 & 0.0796 & 0.0051 \end{bmatrix}$$

The result vector representation from an overall evaluation of the analysis, of which 52.46% of aircraft maintenance and support personnel of the participating units of the grass-roots level evaluation is outstanding, 39.08% of eligible staff evaluated as good, as well as 7.96% of eligible staff evaluation in general, while in addition, the remaining 0.51% is evaluated as poor.

According to the scores of the four evaluation levels V = [95, 80, 60, 25], the final score of the fuzzy comprehensive evaluation of the basic-level aircraft maintenance and support capabilities can be obtained as:
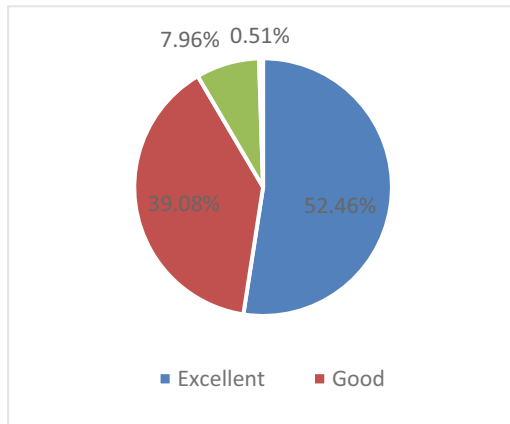
Fig. 4. Target layer evaluation results.

$$S_T = Z_T \cdot V^T = \begin{bmatrix} 0.5246 & 0.3908 & 0.0796 & 0.0051 \end{bmatrix} \cdot \begin{bmatrix} 95 \\ 80 \\ 60 \\ 25 \end{bmatrix}$$

$$= 86.0015$$

Finally, it can be concluded that the comprehensive score of basic-level aircraft maintenance and support capabilities is 86.0015 points, which is reflected in the evaluation as good and tends to be excellent. The evaluation result is basically consistent with the actual maintenance and support situation, which verifies the effectiveness of the evaluation model on the basic-level aircraft maintenance and support capabilities. According to the first-level index evaluation score, the unit performed relatively well in terms of information resources and organizational management level, while there is still room for improvement in the performance of maintenance human resources, maintenance equipment and facilities, and spare equipment support, which is mapped to the second-level The indices are mainly the lack of the allocation of maintenance professionals, the level of maintenance technology, the matching of facilities and equipment, and the completeness of spare parts and equipment.

6   Conclusion

This paper first studies the comprehensive evaluation method. According to the constructed index system and the actual evaluation problem, the fuzzy comprehensive evaluation method is selected for evaluation. Later, the evaluation method is used to construct the evaluation model system. The system not only solves the basic-level aircraft maintenance and support capabilities evaluation and analysis problem but also provides conclusions and suggestions for improvement. The verification of the proposed method of this paper proves the feasibility of using the model system to solve the basic-level aircraft maintenance and support capabilities evaluation problem. It also provides a basis for basic-level aircraft maintenance and support evaluation.

References

1. Jie, L., Van Zuylen, H.J.: Road traffic in China. Procedia Soc. Behav. Sci. **111**, 107–116 (2014)
2. Zhao, S., Huang, B.: Trial-and-error or avoiding a guess? Initialization of the Kalman filter. Automatica **121**, 109184 (2020)
3. Zhao, S., Shmaliy, Y.S., Liu, F.: Fast Kalman-like optimal unbiased FIR filtering with applications. IEEE Trans. Signal Process. **64**(9), 2284–2297 (2016)
4. Wang, J.: Evaluation modeling on command capability of command organization. Command Control Simul. **38**(2) (2016)
5. Zhang, W., Yu, M.: A comprehensive evaluation of war injury care and treatment capacity in field medical facilities based on the analytic hierarchy process (AHP) method. Mil. Med. Sci. **41**(12) (2017)
6. Cao, C., Liu, J.: Evaluation method of POL support capabilities of naval aviation station in war. J. Naval Univ. Eng. **22**(5) (2010)

Robust Unscented Kalman Filter for Target Tracking Based on Mahalanobis Distance

Bingbing Gao[1], Wenmin Li[1(✉)], Longqiang Ni[2], and Wei Wang[1]

[1] School of Automation, Northwestern Polytechnical University, Xi'an, China
cocogalaxy@163.com
[2] Northwest Institute of Mechanical and Engineering, Xianyang, China

Abstract. A robust unscented Kalman filtering (UKF) is presented on the basis of the theory of Mahalanobis distance to enhance radar target tracking's robustness against abnormal observation information. This method firstly uses the concept of Mahalanobis distance to identify the abnormal observation involved in the radar tracking; and then, a scaling factor with robust property is determined and used into the innovation covariance of classical UKF to weaken the negative effect of aberrant observations on system estimation. The designed robust UKF could effectively enhance the filter's robustness and improve the tracking accuracy of radar system. The simulation outcomes verify that the designed robust UKF has a better performance than the classical UKF and RUKF, leading to superior tracking performance for radar system.

Keywords: Radar target tracking · Unscented Kalman filter · Abnormal observations identification · Mahalanobis distance

1 Introduction

Target tracking technology has received an extensive application for military and civil fields like artillery radar, remote warning, civil aviation transportation, and so on [1–3]. It is constantly developing and updating in order to achieve the accurate tracking with the increasingly complex environment such as the human interference factors in the information war, Therefore, accurate target tracking problem has become the current research hotspot [3].

Filtering technology is the key for target tracking [2, 3]. At present, Kalman filtering (KF) is extensively applied in target tracking fields such as missile tracking, artillery radar system due to its advantages, for example strong real-time performance, small amount of calculation, high efficiency and stable prediction results. KF can obtain the optimal state estimation under the principle of the minimum mean square error for linear Gaussian system [4]. However, for the radar target tracking system, the state model and measurement model of a target cannot both be linear in the same coordinate system [3]. Under this condition, KF shows its shortcomings in dealing with nonlinear problems [4, 5]. In the field of target tracking, there are two commonly used methods to solve

the above problem: one is extended Kalman filtering (EKF) [5, 6]. EKF transforms the nonlinear filtering problem into a linear KF one by using the first-order linearization of nonlinear function. However, the accuracy of state estimation for the EKF is poor when the system is highly nonlinear [7]; in addition, it greatly limits the application scope because EKF needs to compute complex Jacobian matrix in the process of implementation. The other is unscented Kalman filtering (UKF) [8]. Compared with the EKF, there is no need for UKF to calculate Jacobian matrix and its calculation amount is equivalent to EKF [9]. Furthermore, UKF can achieve higher approximation accuracy and effectively overcomes the shortcomings of EKF in estimation accuracy and filtering stability [10, 11]. However, UKF has poor robustness against abnormal observations. The estimation accuracy of the filter will be seriously deteriorated when the radar is subject to abnormal interference or equipment failure, which will has an impact on the system tracking performance [12, 13].

The existing researches mostly use H-infinity, noise statistics estimation and adjustment factor to enhance the robust performance of UKF for aberrant observation. Zhao et al. proposed a robust UKF by introducing H-infinity strategy into classical UKF. However, it is not suitable for the condition of random aberrant observation [14]. By using the maximum a posteriori estimation, Shi and Han established an adaptive UKF, which is actually integrated the classical UKF with the traditional Sage-Husa estimator [12]. However, in this method the determined forgetting factor is suboptimal since it is chosen by experience. Meng et al. proposed an adaptive UKF based on covariance matching method to estimate the observation noise covariance [15]. However, the steady-state estimation error limits the further improvement of UKF filtering accuracy. Soken and Hajiyev designed a robust UKF by embedding a robust factor to regulate the filter gain, thus the robustness of UKF against abnormal observation can be improved [16]. However, the robust factor in this method is also determined by experience, which leads to suboptimal filtering results.

In this paper, a robust UKF is proposed on the basis of Mahalanobis distance for enhancing the robust performance of radar single target tracking system against abnormal observation information. A novel method of abnormal observation identification is established based on the Mahalanobis distance. Next, a scaling factor with robust property is further determined and embedded into the innovation covariance matrix of classical UKF to adjust the filter gain for decreasing the negative impact of aberrant observation on state estimation. The stimulation on target tracking for a radar system has been conducted to comprehensively evaluate the proposed methodology.

2   System Model of Radar Target Tracking

Based on the two-dimensional Cartesian coordinate system, the system model for radar target tracking is established [17]. It is assumed that the target's motion obeys the CV model, its azimuth and slant range measured by a radar [18].

2.1 State Model

Supposed that a target P is moving in two-dimensional space x-y, and its position and velocity at time k can be expressed by

$$X_k = \left[x_k, \dot{x}_k, y_k \dot{y}_k\right]^{\mathrm{T}} \tag{1}$$

It is assumed that P moves at approximately constant velocity in horizontal (x-axis) and vertical (y-axis) directions, and there is additive system noise W_k in both directions [19]. Then, the state model of the target motion in Cartesian coordinate system is described by

$$X_k = F_{k/k-1} X_{k-1} + W_k \tag{2}$$

where W_k is additive system noise; $F_{k/k-1}$ is system transformation matrix, and its specific form is:

$$F_{k/k-1} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3}$$

2.2 Observation Model

The slant range and azimuth of a target is measured by a radar, thus the observation model of radar target tracking system can be established as [20]

$$Z_k = h(X_k) + V_k = \begin{bmatrix} \sqrt{x_k^2 + y_k^2} \\ ar\tan\left(\frac{y_k}{x_k}\right) \end{bmatrix} + V_k \tag{4}$$

where V_k is the observation noise.

3 Robust Unscented Kalman Filter Based on Mahalanobis Distance

By utilizing the theory of Mahalanobis distance, we firstly develop a novel way of aberrant observation identification in this section. Then, we further propose a new robust UKF to improve the radar target tracking's robust performance against aberrant observations.

3.1 Classical UKF

By combining (2) and (4), the nonlinear Gaussian discrete model is considered as follows

$$X_k = F_{k/k-1} X_{k-1} + W_k \tag{5}$$

$$Z_k = h(X_k) + V_k \tag{6}$$

where $X_k \in \mathbf{R}^n$ is the system state variable, $F_{k/k-1}$ is discrete state transform matrix, $Z_k \in \mathbf{R}^m$ is the observation variable, $h(\cdot)$ is nonlinear observation function, W_k and V_k are the unrelated Gaussian noise processes with zero-mean, whose variances satisfy

$$E\left[W_k W_k^\mathrm{T}\right] = Q, \quad E\left[V_k V_k^\mathrm{T}\right] = R, \quad E\left[W_k V_k^\mathrm{T}\right] = 0 \tag{7}$$

The procedures of classical UKF is described as:

Step 1: *Initialization.*
State estimate and its covariance are initialized as

$$\begin{cases} \hat{X}_0 = E[X_0] \\ P_0 = E[(X_0 - \hat{X}_0)(X_0 - \hat{X}_0)^\mathrm{T}] \end{cases} \tag{8}$$

Step 2: *Prediction.*
The calculation of the state prediction and its error covariance matrix follow as the KF because of the system's linear state model

$$\hat{X}_{k/k-1} = F_{k/k-1}\hat{X}_{k-1} \tag{9}$$

$$P_{k/k-1} = F_{k/k-1}P_{k-1}F_{k/k-1}^\mathrm{T} + Q \tag{10}$$

Step 3: *Sigma point selection.*
The Sigma points are selected according to the state prediction $\hat{X}_{k/k-1}$ and associated error covariance matrix $P_{k/k-1}$

$$\begin{cases} \xi_{i,k/k-1} = \hat{X}_{k/k-1} & i = 0 \\ \xi_{i,k/k-1} = \hat{X}_{k/k-1} + a\left(\sqrt{nP_{k/k-1}}\right)_i & i = 1, 2, \cdots, n \\ \xi_{i,k/k-1} = \hat{X}_{k/k-1} - a\left(\sqrt{nP_{k/k-1}}\right)_{i-n} & i = n+1, n+2, \cdots, 2n \end{cases} \tag{11}$$

where $\left(\sqrt{nP_{k/k-1}}\right)_i$ represents the ith column of $\sqrt{nP_{k/k-1}}$; $a \in \mathbf{R}$ is adjustment parameter, which is determined by the trial and error method, and it controls Sigma points' distribution around $\hat{X}_{k/k-1}$.

Step 4: *Update.*
The selected Sigma points in (11) are transformed by observation function

$$\gamma_{i,k/k-1} = h\left(\xi_{i,k/k-1}\right)(i = 0, 1, \cdots, 2n) \tag{12}$$

The observation prediction and its error covariance matrix were calculated by

$$\hat{Z}_{k/k-1} = \sum_{i=0}^{2n} \omega_i \gamma_{i,k/k-1} \tag{13}$$

$$P_{\hat{Z}_{k/k-1}} = \sum_{i=0}^{2n} \omega_i \left(\gamma_{i,k/k-1} - \hat{Z}_{k/k-1}\right)\left(\gamma_{i,k/k-1} - \hat{Z}_{k/k-1}\right)^\mathrm{T} + R \tag{14}$$

The cross-covariance matrix is calculated as

$$P_{\hat{X}_{k/k-1}\hat{Z}_{k/k-1}} = \sum_{i=0}^{2n} \omega_i \left(\xi_{i,k/k-1} - \hat{X}_{k/k-1} \right) \left(\gamma_{i,k/k-1} - \hat{Z}_{k/k-1} \right)^{\mathrm{T}} \qquad (15)$$

where
$$\begin{cases} \omega_i = 1 - \dfrac{1}{a^2} & i = 0 \\ \omega_i = \dfrac{1}{2na^2} & i = 1, 2, \cdots, n \end{cases}$$
Determine the gain matrix

$$K_k = P_{\hat{X}_{k/k-1}\hat{Z}_{k/k-1}} P_{\hat{Z}_{k/k-1}}^{-1} \qquad (16)$$

Update the state estimation and its error covariance

$$\hat{X}_k = \hat{X}_{k/k-1} + K_k \left(Z_k - \hat{Z}_{k/k-1} \right) \qquad (17)$$

$$P_k = P_{k/k-1} - K_k P_{\hat{Z}_{k/k-1}} K_k^{\mathrm{T}} \qquad (18)$$

Step 5: Return to Step 2, and perform the following time's filtering solution until all data is conducted.

3.2   Mahalanobis Distance Based Robust Unscented Kalman Filter

3.2.1   Identification of Aberrant Observation

In statistics, we usually use a criterion named Mahalanobis distance to detect outliers of multivariate data [21, 22]. Assume that a multi-dimensional vector $x = (x_1, x_2, \cdots, x_p)^{\mathrm{T}}$, whose mean is $\mu = (\mu_1, \mu_2, \cdots, \mu_p)^{\mathrm{T}}$ and the covariance matrix is Σ, we can define the Mahalanobis distance as follow [22].

$$D(x) = \sqrt{(x - \mu)^{\mathrm{T}} \Sigma^{-1} (x - \mu)} \qquad (19)$$

We can identify aberrant observations for radar target tracking system via the concept of Mahalanobis distance, thus the innovation of filter is defined as

$$\tilde{Z}_k = Z_k - \hat{Z}_{k/k-1} \qquad (20)$$

When a Gaussian system does not have any aberrant observations, the innovation $\tilde{Z}_k$ will be a multivariate Gaussian distribution $N\left(0, P_{\tilde{Z}_{k/k-1}}\right)$, where $P_{\tilde{Z}_{k/k-1}}$ is innovation covariance and its specific form is [4, 15].

$$P_{\tilde{Z}_{k/k-1}} = P_{\hat{Z}_{k/k-1}} = \sum_{i=0}^{2n} \omega_i \left(\gamma_{i,k/k-1} - \hat{Z}_{k/k-1} \right) \left(\gamma_{i,k/k-1} - \hat{Z}_{k/k-1} \right)^{\mathrm{T}} + R \qquad (21)$$

We denote following formula through the definition of Mahalanobis distance's in (19)

$$\beta_k = D^2(\tilde{\mathbf{Z}}_k) = \tilde{\mathbf{Z}}_k^{\mathrm{T}} \mathbf{P}_{\tilde{\mathbf{Z}}_{k/k-1}}^{-1} \tilde{\mathbf{Z}}_k \tag{22}$$

It is known that $D^2(\tilde{\mathbf{Z}}_k)$ obeys the χ^2 distribution according to the statistical knowledge [4].

$$\beta_k = D^2(\tilde{\mathbf{Z}}_k) \sim \chi_m^2 \tag{23}$$

When a significance level $\alpha(0 < \alpha < 1)$ is given, there is a threshold $\chi_{m,\alpha}^2$ to generate the relationship as below in the light of χ^2 test theory

$$P\{\beta_k > \chi_{m,\alpha}^2\} = \alpha \tag{24}$$

Therefore, the following criteria can be constructed to identify the aberrant observation of the system

$$\begin{cases} H_0 : \beta_k \leq \chi_{m,\alpha}^2 \text{ without aberrant observation} \\ H_1 : \beta_k > \chi_{m,\alpha}^2 \text{ with aberrant observation} \end{cases} \tag{25}$$

where $\chi_{m,\alpha}^2$ is the pre-set test threshold.

After completing aberrant observation identification of system model, a robust UKF is proposed on the basis of Mahalanobis distance in the following to resist the aberrant observations for improving the robustness of tracking system.

3.2.2   Calculation of the Robust Factor

Robust UKF is to modify observation noise covariance matrix by a robust factor, which can affect the $\mathbf{P}_{\tilde{\mathbf{Z}}_{k/k-1}}$ and reduce gain matrix, leading to weakened influence of abnormal observation on system solution. Therefore, its key is how to determine the robust factor.

The core of the proposed method is to embed the robust factor κ_k into the classical UKF's innovation covariance, so the proposed method's innovation covariance can be formulated as

$$\mathbf{P}_{\tilde{\mathbf{Z}}_{k/k-1}}^* = \sum_{i=0}^{2n} \omega_i \left(\gamma_{i,k/k-1} - \hat{\mathbf{Z}}_{k/k-1} \right) \left(\gamma_{i,k/k-1} - \hat{\mathbf{Z}}_{k/k-1} \right)^{\mathrm{T}} + \kappa_k \mathbf{R} \tag{26}$$

Substituting (26) into (22), the following nonlinear function can be achieved:

$$g(\kappa_k) = \beta_k^* - \chi_{m,\alpha}^2 = \tilde{\mathbf{Z}}_k^{\mathrm{T}} \left(\mathbf{P}_{\tilde{\mathbf{Z}}_{k/k-1}}^* \right)^{-1} \tilde{\mathbf{Z}}_k - \chi_{m,\alpha}^2 \tag{27}$$

The robust factor can be calculated by solving the equation when we let $g(\kappa_k) = 0$. The Newton iterative method [9, 21] with good convergence is adopted in this paper to solve the above nonlinear equation, and the following formula is obtained

$$\kappa_k(i+1) = \kappa_k(i) - \frac{g[\kappa_k(i)]}{g'[\kappa_k(i)]} \tag{28}$$

where i is the iteration number, and $g'[\cdot]$ represents the derivative of $g[\cdot]$.

Taking (27) into (28), it will be

$$
\kappa_k(i+1) = \kappa_k(i) - \frac{\tilde{\boldsymbol{Z}}_k^{\mathrm{T}} \left(\boldsymbol{P}_{\tilde{\boldsymbol{Z}}_{k/k-1}}^{*}\right)^{-1} \tilde{\boldsymbol{Z}}_k^{\mathrm{T}} - \chi_{m,\alpha}^2}{\tilde{\boldsymbol{Z}}_k^{\mathrm{T}} \left[\left(\boldsymbol{P}_{\tilde{\boldsymbol{Z}}_{k/k-1}}^{*}(i)\right)^{-1}\right] \tilde{\boldsymbol{Z}}_k^{\mathrm{T}}}
\tag{29}
$$

By using the matrix formula [9]

$$
\frac{d}{dt}(\boldsymbol{M}^{-1}) = -\boldsymbol{M}^{-1}\frac{d\boldsymbol{M}}{dt}\boldsymbol{M}^{-1}
\tag{30}
$$

it is easy to get

$$
\kappa_k(i+1) = \kappa_k(i) + \frac{\tilde{\boldsymbol{Z}}_k^{\mathrm{T}} \left(\boldsymbol{P}_{\tilde{\boldsymbol{Z}}_{k/k-1}}^{*}(i)\right)^{-1} \tilde{\boldsymbol{Z}}_k - \chi_{m,\alpha}^2}{\tilde{\boldsymbol{Z}}_k^{\mathrm{T}} \left[\left(\boldsymbol{P}_{\tilde{\boldsymbol{Z}}_{k/k-1}}^{*}(i)\right)^{-1} \boldsymbol{R} \left(\boldsymbol{P}_{\tilde{\boldsymbol{Z}}_{k/k-1}}^{*}(i)\right)^{-1}\right] \tilde{\boldsymbol{Z}}_k} \quad (i = 0, 1, 2, \cdots)
\tag{31}
$$

where $\boldsymbol{M}$ represents a matrix, which is invertible about t.

The above iteration's initial value is set as $\kappa_k(0) = 1$ when using (31) for iterative calculation, and the iterative result of each step is brought into (22) for calculating β_k^{*}. The iteration ends at the time of $\beta_k^{*} \leq \chi_{m,\alpha}^2$, and the last iterative result is the determined robust factor.

3.2.3 Proposed Methodology

The proposed methodology is summarized as the Fig. 1, which has the following procedures:

Step 1: *Initialization.*
The filter is initiated as (8).

Step 2: *Prediction.*
Trough (9) and (10), the state prediction $\hat{X}_{k/k-1}$ and error covariance $\boldsymbol{P}_{k/k-1}$ is obtained.

Step 3: *Update.*

i)   The innovation $\tilde{\boldsymbol{Z}}_k$ and error covariance $\boldsymbol{P}_{\hat{z}_{k/k-1}}$ of the filter are calculated by (20) and (21).
ii) Set $\kappa_k(0) = 1$ and calculate β_k^{*};
iii) If $\beta_k^{*} \leq \chi_{m,\alpha}^2$,
•   Perform the classical UKF steps (15)–(18) to achieve the $\hat{X}_k$.

Else,

- κ_k is solved by (31) until it meets $\beta_k^*(i) \leq \chi_{m,\alpha}^2$.
- Update innovation covariance $\boldsymbol{P}_{\tilde{\boldsymbol{Z}}_{k/k-1}}^*$
- Execute the steps (15)–(18) in classical UKF to renewal the $\hat{\boldsymbol{X}}_k$.

Step 4: Return to Step 2, and perform the following time's filtering solution until all data is conducted.
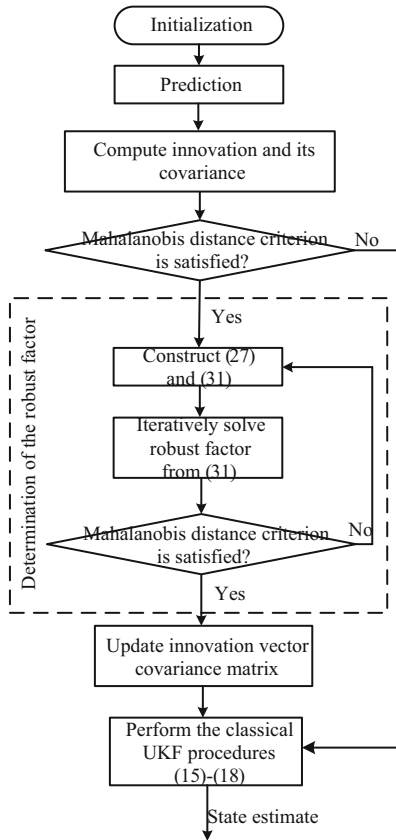

Fig. 1. The calculation process of the MRUKF

4   Simulation Validation and Analysis

The Mahalanobis distance based robust UKF (MRUKF) presented in this paper is applied to radar target tracking system. Compare the results of Monte Carlo simulations of the

MRUKF with the classical UKF and the robust UKF (RUKF) in [16]. Monte Carlo runs $M = 10$.

It is assumed that the covariance of V_k is set to

$$R = diag\left(5^2, 0.005^2\right) \tag{32}$$

We try to evaluate the MRUKF when there exists abnormal observations, and supposed that the observation noise covariance suddenly increases to 50 times of its real value in [400s, 600s]. Therefore, the real covariance matrix of observation noise can be described as

$$R_k = \begin{cases} R & k < 400 \\ 50 \cdot R & 400 \leq k \leq 600 \\ R & 600 < k < 1000 \end{cases} \tag{33}$$

Figures 2, 3, 4, and 5 show the target's position and velocity errors for the *x*-axis and *y*-axis calculated from the classical UKF, RUKF and MRUKF. We can conclude from Figs. 2, 3, 4, and 5 that:

(i) For (0s, 400s) and (600s, 1000s), since the abnormal observation is not introduced into the target tracking system, the classical UKF, RUKF and the proposed MRUKF algorithms all can quickly and accurately estimate the target's position and velocity, and the three algorithms' tracking precision is very close.

(ii) During the time of [400s, 600s], the classical UKF's estimation precision is obviously decreased because of the aberrant observation for the target tracking system. RUKF can reduce the negative impact of aberrant observation on state estimation and improve the estimation precision of classical UKF. But, there is always a violate fluctuation in the estimation error curve of RUKF. Compared with the above two methods, the proposed MRUKF significantly have a more accurate estimation for the radar target tracking system by embedding a robust factor to regulate the filter gain. Its estimation error is significantly smaller than both the mentioned classical UKF and RUKF. Different from RUKF, the robust factor in MRUKF is directly derived from the concept of Mahalanobis distance, which is not determined by artificial experience, thus it has better regulation effect.

Figures 6, 7, 8, and 9 depicts the intuitive comparison of the position and velocity's average absolute errors for the target in *x*-axis and *y*-axis directions achieved from the classical UKF, RUKF and MRUKF during [400s, 600s] and other time interval respectively. Further, the maximum velocity and position errors achieved from the classical UKF, RUKF and MRUKF during [400s, 600s] are also calculated and intuitively compared in Figs. 10 and 11. The calculated data in Figs. 6, 7, 8, and 9 as well as Figs. 10 and 11 also fully confirm the above conclusion.

The simulation results show that when the radar target tracking system involves abnormal observation, the estimation accuracy of proposed MRUKF is better than the mentioned classical UKF and RUKF. At the same time, it significantly enhances the robustness of filter, thus improving the target tracking performance of the radar.
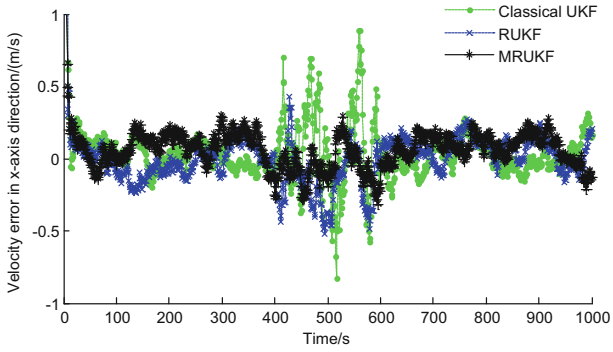
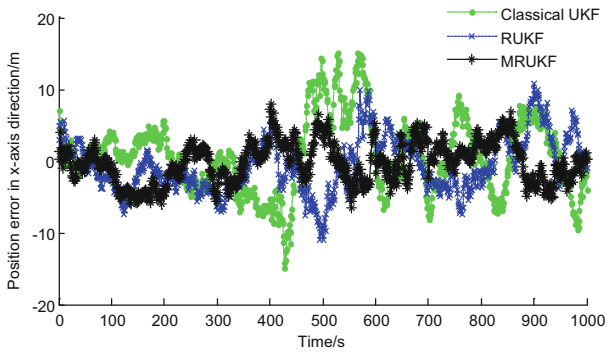**Fig. 2.** Velocity errors in *x*-axis by the classical UKF, RUKF and MRUKF

Fig. 3. Position errors in *x*-axis by the classical UKF, RUKF and MRUKF
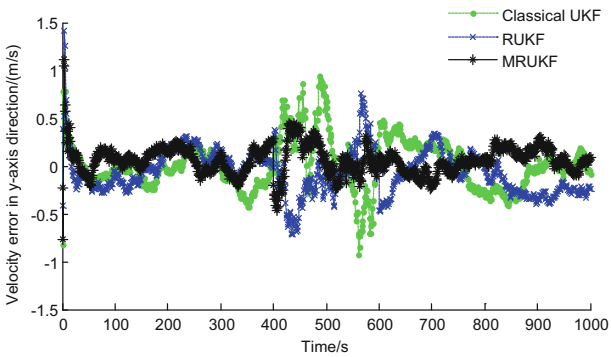

Fig. 4. Velocity errors in *y*-axis by the classical UKF, RUKF and MRUKF
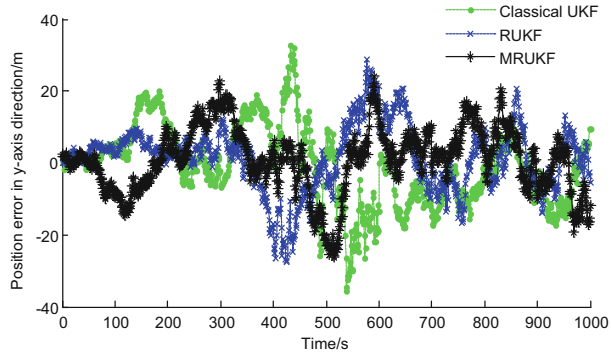
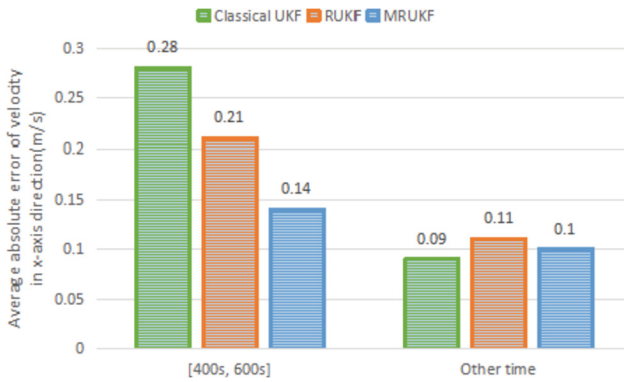**Fig. 5.** Position errors in *y*-axis by the classical UKF, RUKF and MRUKF

Fig. 6. Average absolute error of velocity in *x*-axis by the classical UKF, RUKF and MRUKF
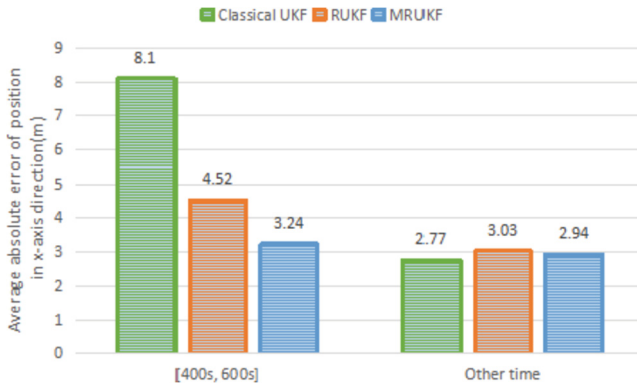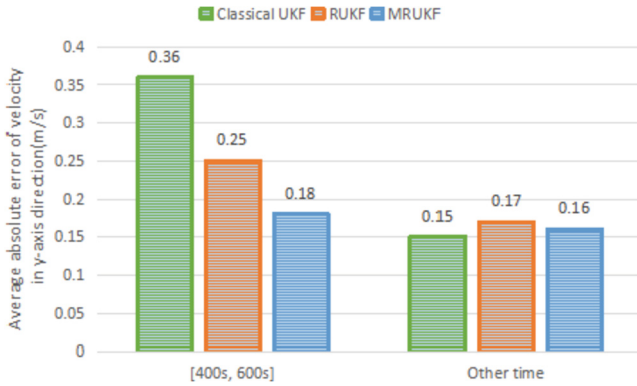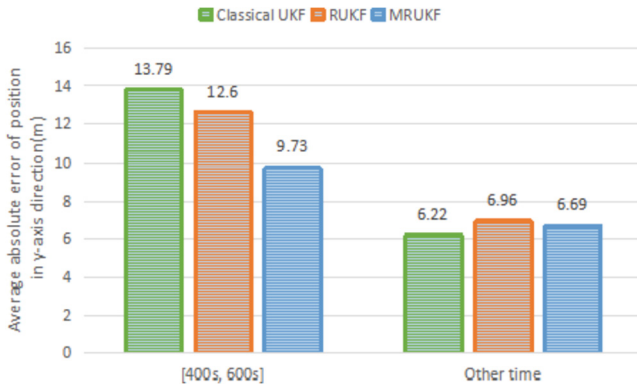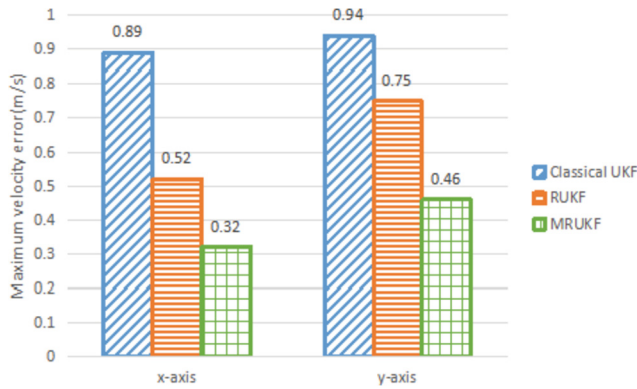

Fig. 7. Average absolute error of position in *x*-axis by the classical UKF, RUKF and MRUKF

Fig. 8. Average absolute error of velocity in *y*-axis by the classical UKF, RUKF and MRUKF

Fig. 9. Average absolute error of position in *y*-axis by the classical UKF, RUKF and MRUKF

Fig. 10. Maximum velocity error by the classical UKF, RUKF and MRUKF during the time interval [400s, 600s]
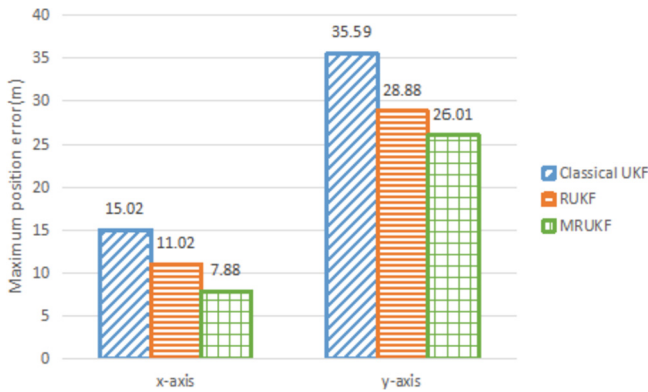
Fig. 11. Maximum position error by the classical UKF, RUKF and MRUKF during the time interval [400s, 600s]

5 Conclusions

This paper describes a MRUKF algorithm to enhance the robustness of radar target tracking system against aberrant observation information. A method of the identification of abnormal observation information for target tracking system is established based on the method of Mahalanobis distance; According to this, a scaling factor with robust property is further determined and embed into the classical UKF's innovation covariance matrix to adjust the filter gain and restrain the negative effect of abnormal observation on system estimation. The simulation outcomes verify that the designed MRUKF has better estimation performance compared to the classical UKF and RUKF when the system involves abnormal observations, which effectively improves the robustness of the filter and significantly enhances the target tracking performance for the radar system.

References

1. Chen, J., Li, J., Yang, S., et al.: Weighted optimization-based distributed Kalman filter for nonlinear target tracking in collaborative sensor networks. IEEE Trans. Cybern. **47**(11), 3892–3905 (2017)
2. Han, M., Tang, X., Meng, J., et al.: Target tracking algorithm based on improved Kalman filter and mean-shift. Tact. Missile Technol. **193**(1), 121–129 (2019)
3. Zheng, Z., Ouyang, D., Li, L.: Study on typical nonlinear filtering algorithm during radar tracking target. Fire Control Radar Technol. **4**, 51–55 (2017)
4. Chang, G.: Kalman filter with both adaptivity and robustness. J. Process Control **24**(3), 81–87 (2014)

5. Chang, L., Hu, B., Li, A., et al.: Transformed unscented Kalman filter. IEEE Trans. Autom. Control **58**(1), 252–257 (2012)
6. Wang, L., Li, G., Qiao, X., et al.: An adaptive UKF algorithm based on maximum likehood principle and expectation maximization algorithm. Acta Autom. Sin. **38**(7), 1200–1210 (2012)
7. Gao, B., Gaoge, H., et al.: Cubature Kalman filter with both adaptability and robustness for tightly-coupled GNSS/INS integration. IEEE Sens. J. **21**(13), 14997–15011 (2021)
8. Gao, B., Gaoge, H., et al.: Cubature rule-based distributed optimal fusion with identification and prediction of kinematic model error for integrated UAV navigation. Aerosp. Sci. Technol. **109**, 106447 (2021)
9. Wang, Y., Sun, S.: Adaptively robust unscented Kalman filter for tracking a maneuvering vehicle. J. Guid. Control. Dynam. J. Aerosp. Eng. **37**(5), 1696–1701 (2014)
10. Xiong, K., Zhang, H.Y., Chan, C.W.: Performance evaluation of UKF-based nonlinear filtering. Automatica **42**(2), 261–270 (2016)
11. Hajiyev, C., Soken, H.E.: Robust adaptive unscented Kalman filter for attitude estimation of pico satellites. Int. J. Adapt. Control Signal Process. **28**(2), 107–120 (2014)
12. Shi, Y., Han, C.: Adaptive UKF method with application to target tracking. Acta Autom. Sin. **37**(6), 755–759 (2011)
13. Cho, S.Y., Choi, W.S.: Robust positioning technique in low cost DR/GPS for land navigation. IEEE Trans. Instrument. Measur. **55**(4), 1132–1142 (2006)
14. Zhao, B., Guo, K., et al.: Speed observer design for permanent magnet synchronous motor based on H_∞ robust SUKF algorithm. J. Jilin Univ. (Eng. Technol. Edn). **185**(4), 1017–1022 (2016)
15. Meng, Y., Gao, S., Zhong, Y., et al.: Covariance matching based adaptive unscented Kalman filter for direct filtering in INS/GNSS integration. Acta Astronaut. **120**, 171–181 (2016)
16. Soken, H.E., Hajiyev, C.: Pico satellite attitude estimation via robust unscented Kalman filter in the presence of measurement faults. ISA Trans. **49**(3), 249–256 (2010)
17. Zhan, R., Wan, J.: Iterated unscented Kalman filter for passive target tracking. IEEE Trans. Aerosp. Electron. Syst. **43**(3), 1155–2263 (2007)
18. Zhang, H., Dai, G., et al.: Unscented Kalman filter and its nonlinear application for tracking a moving target. Optik **124**(20), 4468–4471 (2013)
19. Zhou, W., Hou, J.: A new adaptive robust unscented Kalman filter for improving the accuracy of target tracking. IEEE Access. **7**, 2169–43536 (2019)
20. Duan, Z., Li, X.R., et al.: Sequential unscented Kalman filter for radar target tracking with range rate measurements. In: 2005 7th International Conference on Information Fusion, 25–28 July (2005)
21. Chang, G., Liu, M.: An adaptive fading Kalman filter based on Mahalanobis distance. Proc. Inst. Mech. Eng. Part G: J. Aerosp. Eng. **229**(6), 1114–1123 (2015)
22. Chang, G.: Robust Kalman filtering based on Mahalanobis distance as outlier judging citerion. J. Geodesy **88**, 391–401 (2014)

Analysis of Chaotic Behavior in Single Mode NH3 Molecular Laser

Hongyan Zang[1(✉)], Shourong Zhang[2], and Tengfei Lei[1]

[1] Collaborative Innovation Center of Memristive Computing Application (CICMCA),
Qilu Institute of Technology, Jinan 250200, China
`379177780@qq.com`
[2] Shandong Vanform High Energy Physics Technology, Jinan 250200, China

Abstract. In this paper, a single-mode NH3 molecular laser system is introduced. Firstly, the basic dynamic behaviors of the laser system such as bifurcation diagram, Lyapunov exponent and complexity C0 are analyzed by numerical simulation using MATLAB, and the chaotic characteristics of the laser system and its sensitivity to parameters are described. At the same time, in order to further study the influence of its parameters, the relationship between the system period and the existence of chaos is studied by using the complexity C0 under the change of two parameters.

Keywords: NH3 molecular laser · Chaotic · Complexity

1 Introduction

Chaos is the internal randomness of deterministic systems, and Chaos is a universal phenomenon in nature. With people's understanding of nature, chaos phenomenon was discovered in meteorology in 1963. The first chaos model was established [1], which opened people's research on chaotic system and chaotic dynamics [2]. More than ten years ago, most researchers focused on the construction of chaotic systems. However, in recent years, it has been gradually discovered that chaotic phenomena exist widely in many engineering systems, such as motor systems [3], power system, DC/DC converter system [4] and laser system [5].

In the 1980s, many scientists have found chaos in a series of lasers such as CO2 laser, XE laser, He-Ne laser, NH3 laser and semiconductor laser. NH3 laser system is a class of strong nonlinear and multi-coupling complex system. The chaotic behavior of NH3 laser system has been paid much attention and research by engineers and researchers. In paper [5], a parameter mismatch scheme is adopted for a kind of complex laser system synchronized and the security of two laser chaotic systems are analysed and tested. Ref [6] reviewed the random number generators in laser chaos physics, pointed out the advantages and disadvantages of various random number generators, and described the new research results in this direction. The team of Southwest University studied the laser chaotic system [7–9] by using finite-time master-slave control, delay feedback control,

parallel control and series control methods respectively, and finally achieved stable control. Based on this, more attention should be paid to the study of laser chaos mechanism. The study of laser chaos mechanism is particularly important for the synchronous control of laser chaotic system [10–12]. Laser chaos is also studied in secure communication and optical fiber communication [13]. Image encryption based on traditional chaotic system is easy to be implemented in the circuit. However, it is also limited by the inherent problems of electronic circuits such as low speed, low complexity, high attenuation, high cost and so on [14]. Applying optical chaos generated by semiconductor lasers to image information security processing is expected to effectively solve the problems of low bandwidth and low transmission efficiency in traditional image information encryption and transmission.

The mathematical model of a class of single-mode lasers is studied in this paper. Firstly, the phase diagram is used to show the attractor of the whole system. Secondly, using bifurcation diagram, Lyapunov exponent and complexity C0 analysis methods, the influence of parameters on the system is studied. Finally, in order to further expatiate the strength of the interaction between the parameters and the system, the system complexity under the change of two parameters is used to study it. The simulation results show that the parameter a has a great influence on the system. The study of laser chaos in this paper provides important value for its application in encryption and optical fiber communication.

2   Chaotic System Model of Single-Mode NH3 Molecular Laser

In the 1970s, the Maxwell-Bloch equation describing the uniformly broadened single-mode laser is derived by Haken. After mathematical derivation and simplification for the single-mode NH3 molecular laser. It is called the Lorenz-Haken laser system. Through scale change, it is concluded that this equation is consistent with the classical Lorenz equation in form. And its simplified dynamic equation is:

$$\begin{cases} \dot{x} = a(y - x) \\ \dot{y} = cx - y - xz \\ \dot{z} = xy - bz \end{cases} \tag{1}$$

In the formula, x represent the electric field intensity of the laser; y is macroscopic planned strength of laser working medium, z represent the reverse ion number density of atoms in the laser working material. Parameter a and c represent the pump power and b represents the pump parameter. The equilibrium points of system (1) can be easily calculated by setting the left-hand side to zero. The system (1) has three special equilibria E1(0, 0, 0), $(\sqrt{b(c-1)}, \sqrt{b(c-1)}, c-1)$, E2$(-\sqrt{b(c-1)}, -\sqrt{b(c-1)}, c-1)$. The linearization analysis is carried out on the system at the equilibrium point E $= (x_0, y_0, z_0)$, and the Jacobian matrix is obtained as:

$$J = \begin{bmatrix} -a & a & 0 \\ c - z_0 & -1 & -x_0 \\ x_0 & y_0 & -b \end{bmatrix} \tag{2}$$

when $a = 1.4253$, $c = 50$, $b = 0.2778$, the eigenvalues at E1 are: $\lambda_{11} = -9.6572$, $\lambda_{12} = 7.2319 > 0$, $\lambda_{13} = -0.27778$; The eigenvalues at E2 and E3 are $\lambda_{21} = -2.7117$, $\lambda_{22} = 0.0043 + 3.7828i > 0$, $\lambda_{23} = 0.0043 - 3.7828i > 0$, so system (1) that the Lorenz-Haken laser system enters a chaotic state. The three Lyapunov exponent of the system are LE1 $= 0.33 > 0$, LE2 $= 0$, LE3 $= -2.6 < 0$, which accord with the law of Lyapunov exponents of three-dimensional chaotic system (+ 0 −).And it is found through simulation that there is a typical chaotic attractor in system (1), as shown in Fig. 1(a). The Lyapunov exponent of the system is shown in Fig. 1(b).
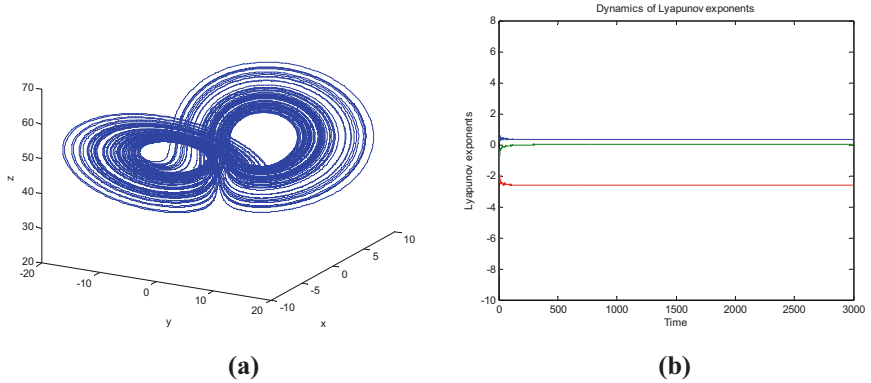

Fig. 1. Chaotic attractor and Lyapunov exponent of single-mode NH3 molecular laser system. (a) Chaotic attractor, (b) Lyapunov exponent.

3   Influence of Parameters on Laser Chaotic System

For laser chaotic system, the influence of parameters a, b and c on the system is analyzed by bifurcation diagram, Lyapunov exponential spectrum (LE spectrum) and complexity C0.When a varies from 0 to 3 and other parameters keep unchanged, the bifurcation diagram of the state variable x and the corresponding Lyapunov exponent are shown in Fig. 2. The system is periodic when $a \in [0, 1.2)$, and largest Lyapunov exponent is negative. The system is chaotic when $a \in [1.2, 3)$, and largest Lyapunov exponent is positive. Thus it can be seen that the bifurcation diagram is consistent with the Lyapunov exponent. As can be seen from Fig. 3, when the system is in periodic state, the complexity C0 is small; when the system is in chaotic state, the complexity C0 of the system is large.
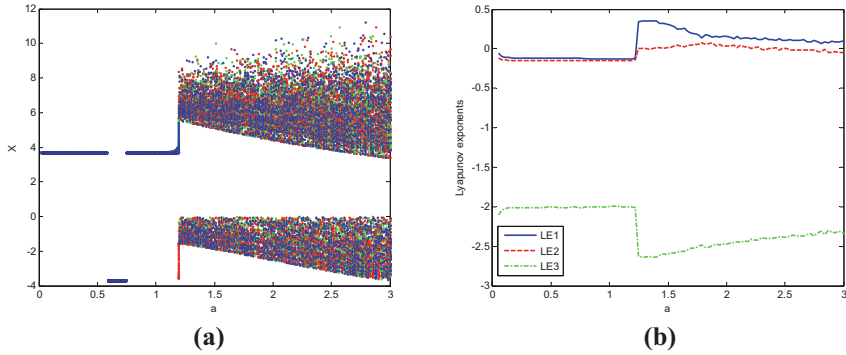
Fig. 2. Bifurcation diagram and Lyapunov exponent of the system changing with *a*. (a) Bifurcation diagram. (b) Lyapunov exponents.
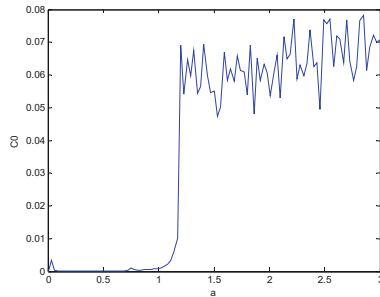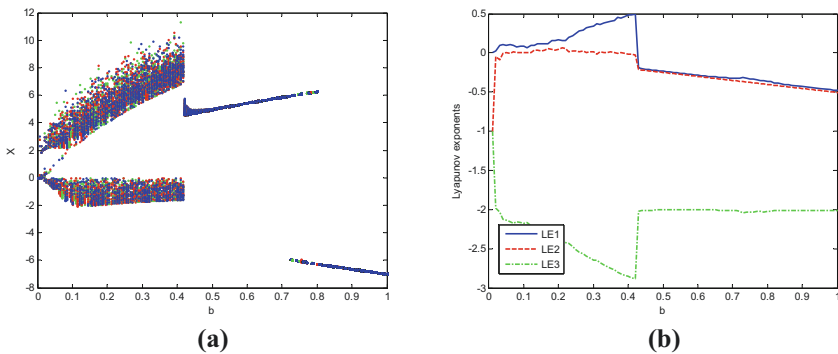

Fig. 3. System complexity C0 of *a*.

Fig. 4. Bifurcation diagram and Lyapunov exponent of the system changing with *b*. (a) Bifurcation diagram. (b) Lyapunov exponents.

When *b* varies from 0 to 1 and other parameters keep unchanged, the bifurcation diagram of the state variable *x* and the corresponding Lyapunov exponent are shown in Fig. 4.The system is fixed point when $b \in [0.45, 1]$, and largest Lyapuno exponent is negative. The system is chaotic when $b \in [0, 0.45)$, and largest Lyapunov exponent is

positive. The system is in periodic state, the complexity C0 is small; when the system is in chaotic state, the complexity C0 of the system is large, as shown in Fig. 5.
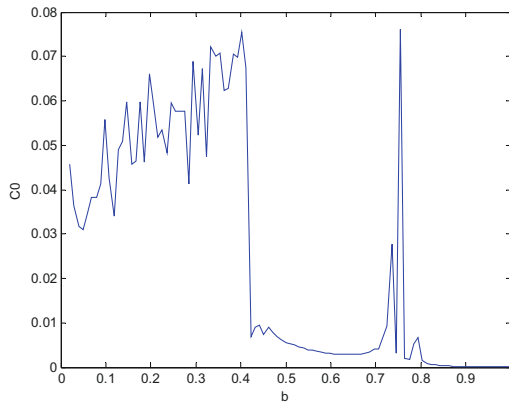

Fig. 5. System complexity C0 of b.

When c varies from 20 to 60, and other parameters keep unchanged, the bifurcation diagram of the state variable x and the Lyapunov exponent are shown in Fig. 6.The system is fixed point when $c \in [20, 35]$, and largest Lyapunov exponent is negative. The system is chaotic when $c \in [35.5, 60)$, and largest Lyapunov exponent is positive. The system is in periodic state, the complexity C0 is small; when the system is in chaotic state, the complexity C0 of the system is large, as shown in Fig. 7.
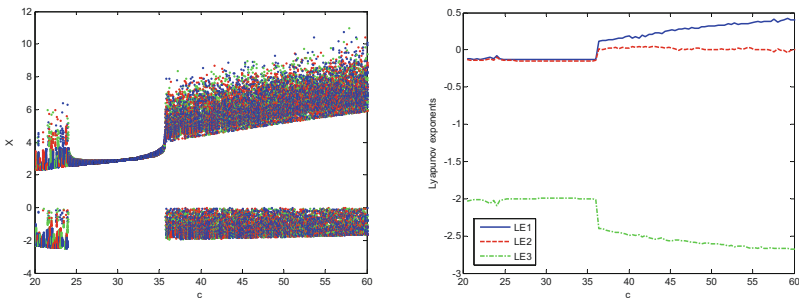

Fig. 6. Bifurcation diagram and Lyapunov exponent of the system changing with c. (a) Bifurcation diagram. (b) Lyapunov exponents.
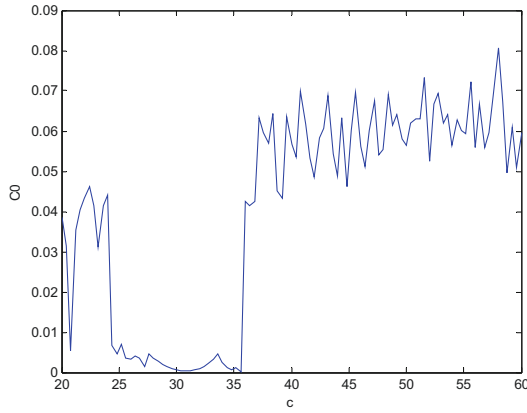
Fig. 7. System complexity C0 of *c*.

4   The Influence of Two Parameters on Chaotic System

This paper describes the bifurcation space of the system, taking the complexity C0 and the maximum Lyapunov exponent as standard quantities. As can be seen from Fig. 8, the parameters of the system are mutually restricted. But the influence of the parameters of the system shown in above research, is a case of only one parameter changing. In order to further study the influence of parameter changes on the system, we give the bifurcation graph with two parameter changing together.
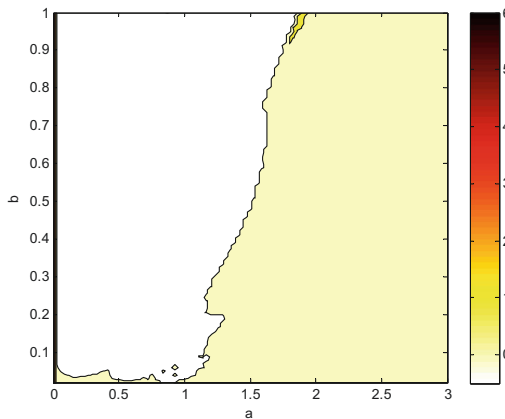

Fig. 8. The situation when the system changes with *a* and *b* simultaneously

5   Conclusions

In this paper, the chaotic dynamic characteristics of the system are analyzed by bifurcation diagram, Lyapunov exponent, and complexity analysis. The bifurcation space of the system is studied by using the complexity C0 under the change of two parameters. The simulation results show that the system has chaotic motion parameter range. The study of laser chaos dynamics in this paper provides a theoretical basis and reference value for the study of laser chaos synchronization control and optical fiber communication.

Conflicts of Interest. The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Lorenz, E.N.: Deterministic nonperiodic flow. J. Atmos. Sci. **20**(2), 130–141 (1963)
2. Chen, G., Ueta, T.: Yet another chaotic attractor. Int. J. Bifurcat. Chaos **9**(7), 1465–1466 (1999)
3. Lei, T., Kang, J., Liu, Y., et al.: Chaotic characteristics and feedback control of synchronous reluctance motor system. J. Shaoyang. Univ. (Nat. Sci. Edn.) **14**(03), 34–39 (2017)
4. Fakhfakh, L., Ammous, A.: New simplified model for predicting conducted EMI in DC/DC converters. Electr. Eng. **99**(3), 1087–1097 (2016)
5. Mu, P., Pan, W., Li, N., et al.: Performance of chaos synchronization and security in dual-chaotic optical multiplexing system. Acta Phys. Sin. **64**(12), 210–219 (2015)
6. Li, P., Wang, Y.: Research progress in physical random number generator based on laser chaos for high speed secure communication. Laser Optoelect. Prog. **51**(06), 13–24 (2014)
7. Zhao, J., Wang, J., Wang, H.: The study of finite-time stability active control method for Lorenz-Haken laser chaotic system. Acta Phys. Sin. **11**, 79–87 (2012)
8. Ding, L., Wu, J., Xia, G.: Suppression of time delay feedback signatures in a semiconductor laser with double optical feedback. Acta Phys. Sin. **60**(1), 014210 (2011)
9. Fan, L., Xia, G., Wu, Z.: Chaotic parallel synchronization of optoelectronic feedback semiconductor lasers. Acta Phys. Sinca. **58**(2), 989 (2009)
10. Pan, M., Meng, Y., Chen, Z.: Dynamic characteristics of optoelectronic oscillation system with double delay feedback for generating laser chaos. Chinese J. Lasers **45**(6), 0606002 (2018)
11. Li, H., Li, J.,Fang, D.: Synchronization control of laser chaotic systems based on neural network finite-time observer. J. Air Fouce Eng. Univ. (Nat. Sci. Edn.) **32**(4) (2020)
12. Pusuluri, K., Meijer, H.G.E., Shilnikov, A.L.: (INVITED) Homoclinic puzzles and chaos in a nonlinear laser model. Commun. Nonlin. Sci. Numer. Simulat. **93**, 105503 (2021)
13. Yan, S.-L.: Chaotic laser parallel series synchronization and its repeater applications in secure communication. Acta Phys. Sin. **68**(17), 170502 (2019)
14. Meucci, R., Stefano Euzzor, F., Arecchi, T., Ginoux, J.-M.: Minimal universal model for chaos in laser with feedback. Int. J. Bifurcat. Chaos **31**(04), 2130013 (2021)

Collaborative Path Optimization Method for Flood Control Material Storage

Zhihao Li[1], Xinyao Wang[2,3], Shuhui Bi[1,3(✉)], and Qinjun Zhao[1]

[1] School of Electrical Engineering, University of Jinan, Jinan 250022, China
[2] HRG (Shandong) Intelligent Equipment Research Institute,
Jinan 250000, Shandong, China
[3] Shandong Guige Intelligent Technology Co., Ltd,
Jinan 250000, Shandong, China
`cse_bish@ujn.edu.cn`

Abstract. In order to reach its efficient storage and wide application, this paper analyzes and improves the outgoing efficiency of flood control material warehouse. Firstly, according to the characteristics of flood control material warehouse, the grid model of the warehouse is established, and the storage shelves are classified to store the goods with different use frequencies. Secondly, in view of the possible conflicts of robots in the operation of the warehouse, this paper proposes multiple strategies to improve the efficiency of goods delivery under the premise of avoiding conflicts. The reservation table is established to solve the problem of frontal collision of robots. Further, using the improved A* algorithm, each robot can transport goods along the shortest path. The dynamic weighting table is added to solve the multi-robot driving strategy of intersection conflicts, and the most urgent goods are guaranteed priority out of storage. Finally, the proposed program was verified by the simulation.

Keywords: A* algorithm · Appointment form · Dynamic weighting table

1 Introduction

With the development of logistics industry, intelligent storage has attracted more and more attention. The order allocation is one of the factors that affect the efficiency of warehouse [1]. In addition, how to distribute goods is also one of the research directions [2]. At the same time, the cooperation of multiple robots should also be considered to avoid collisions [3]. Given this premise, robots find the shortest path [4].

The characteristics of storehouses for flood control materials are obvious. Compared with the common warehouse, the sudden flood situation and the emergency of disaster relief make the warehouse of flood control materials have high requirements on the efficiency of outbound, but not excessive requirements

on the efficiency of inbound [5]. The materials of the same type or frequency of use are placed close to each other, which is convenient for management and can be transferred together when they are out of the warehouse [6].

In the process of warehouse delivery, multiple robots need to work at the same time to give full play to the advantages of warehouse space. Each robot searches for the shortest path and fetches its cargo. However, with the increase of robots, each robot cannot predict the route of other robots, so local optimal solutions and even conflicts often appear [7]. Therefore, when multiple robots are running in the warehouse, the central controller needs to coordinate the conflict uniformly, and then each robot calculates the shortest path according to the instructions of the central controller, so as to improve the overall efficiency [8].

In this paper, according to the characteristics of flood control materials and their intelligent storage, the warehouse needs to give priority to the transportation of materials in high demand according to the flood control work. So a storage model of flood control materials is established around the efficiency of warehouse delivery and the collaborative optimization of multi-robots.

This article will carry on the research through algorithm design and simulation. In section two, the reservation table and the improved A* algorithm are used to solve the path conflict problem of multi-robots, and the dynamic weight table is added to address the issue of prioritizing the transport of items in high demand in the event of a route conflict. In section three, the improved algorithm is compared with the simple A* algorithm, and the high efficiency of the algorithm is verified.

2   Algorithm Design

2.1   The Establishment of the Environment

Raster Map Building. This chapter proposes an efficient modular storage model to meet the storage requirements of different regions and materials. The small-scale storage model is adopted here, with a planned standard warehouse of 25 m long and 26 m wide. The most left row of the warehouse is the position of the picking table. The robot starts from the picking table, moves to the corresponding shelf to take down the goods, and sends the goods to the picking table, which is the end of the warehousing.

The warehouse is divided into 650 parts with each 1-m long and 1-m wide grid to form a standardized grid map, as shown in Fig. 1.

In the raster map, each shelf measuring 2 m long and 4 m wide occupies 8 grids. To ignore the problem of different pallet sizes caused by different materials, each pallet is set as a standard module with a length of 1 m and a width of 1 m, and the robot can only transport one pallet for each task. Each shelf is separated by a path with a width of 1 m, which serves as a transportation channel for warehouse robots and is distributed across the whole warehouse.
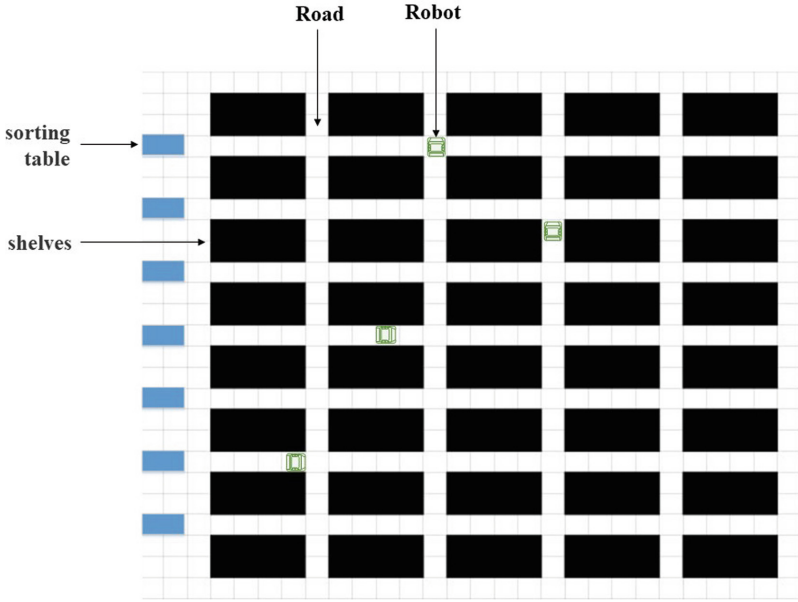

Fig. 1. Raster map and warehouse structure.

In order to minimize the carrying distance of shelves in the picking process, all the shelves in the warehouse are divided into three parts according to the frequency of use, as shown in Fig. 2. The two rows of shelves near the picking platform are high-frequency storage areas for storing goods with high frequency of use such as life jackets and tents. The third and fourth rows of shelves are medium frequency storage areas, which store oil, inspection lights and other goods. The last row of shelves for low-frequency storage areas, storage generators and other low-frequency and large goods.

Fig. 2. The division of shelf area.
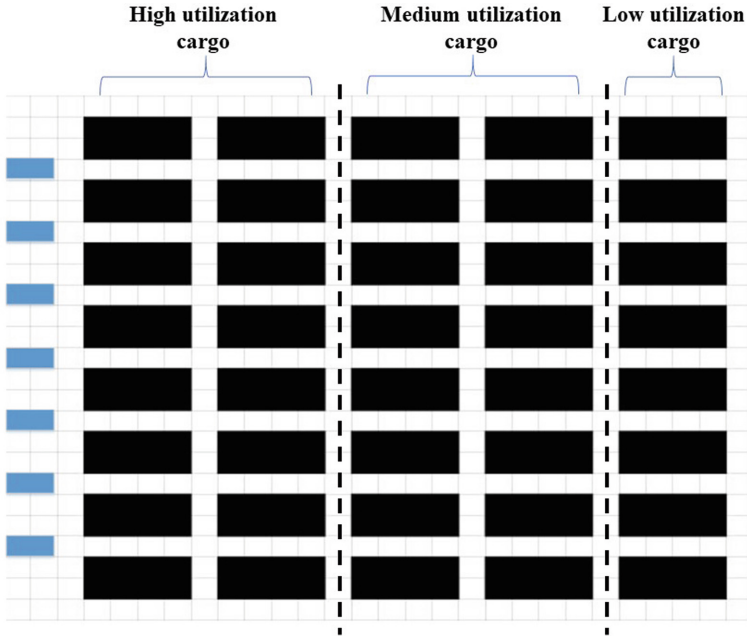
Appointment Form. In the intelligent storage system, multiple robots participate in the transportation of goods in and out of the warehouse at the same time. The environment is dynamic and changeable, and there will be congestion, collision and even deadlock. Typical situations can be divided into the following four categories, as shown in Fig. 3.
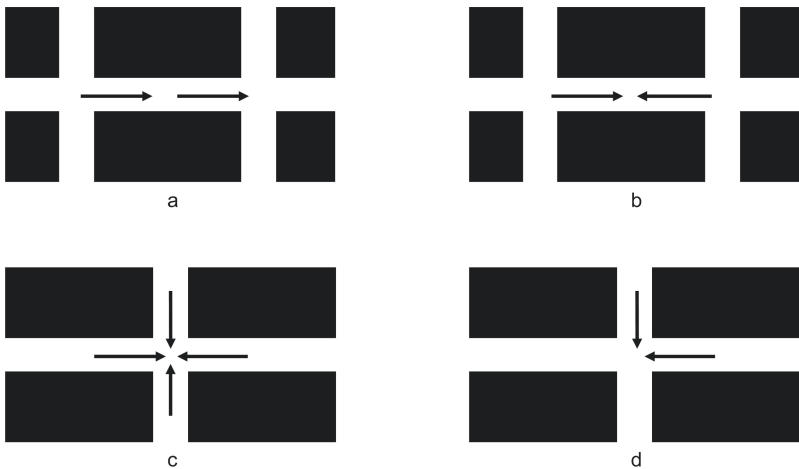

Fig. 3. Conflicting situations.

In case A, the risk of rear-end collision can be avoided by installing a distance sensor on the head of the robot and slowing down in advance after receiving the information.

However, case B, C and D still cannot be avoided by using distance sensor, so this paper uses reservation table to solve this problem. When two robots meet at an intersection, it must be decided that one robot first occupies the intersection and passes through it, while the other robot waits in place and continues to move after this robot passes. In other words, each grid can only be occupied by one robot at a certain time. The raster occupied by the robot in the raster map at the current time is recorded in the same table to form the reservation table, as shown in Fig. 4. The reservation table takes the warehouse grid model as the horizontal and vertical coordinates, and uses a two-dimensional array to record the state of each cell. The idle table is 0, otherwise it is 1. The grid occupied by the robot at each moment records the corresponding robot number, and other robots can decide to move forward or wait for another robot to pass by querying the reservation table at the current moment.

Fig. 4. Appointment Form.

2.2   Improved A* Algorithm

A* algorithm is A heuristic algorithm, which is the most effective direct search method to solve the shortest path in the static map. It sort the neighboring nodes

according to the evaluation function, and then select the optimal neighboring nodes for the next search until the target point. The warehouse map used in this paper is simple and neat, which is suitable for using A* algorithm to calculate the shortest path.

A starting point and a target point are first set. The robot starts from the starting point and expands the grid around the current grid point. Extended search mechanism usually has two kinds, four neighborhood search and eight neighborhood search. In this paper, the robot in the warehouse can move up, down, left and right, so the neighborhood search method is chosen. The robot calculates the cost of raster paths in the four directions of the current location extension, which is called the parent node. Select the direction with the least cost to move, and the position after moving becomes the new parent node, and continue to expand the grid with the new parent node as the center. Repeat this step until the robot reaches the target point. In this way, we can get an optimal path from the starting point to the target point with the least cost. The cost estimation function of A* algorithm is expressed as:

$$f(n) = g(n) + h(n) \tag{1}$$

In the formula, $g(n)$ represents the actual cost from the starting point to the current node n, which is generally expressed by distance or time. In this paper, time is used as a unified scalar to compare the cost of function $f(n)$. The expression $g(n)$ is

$$g(n) = \frac{d}{v} \tag{2}$$

Variable d is the actual moving distance of the robot from the starting point to the current node n, and v is the speed at which the robot travels at constant speed.

Variable $h(n)$ represents the heuristic estimated cost from the current node n to the target point, expressed as

$$h(n) = \frac{d_n}{v} \tag{3}$$

Variable d_n is the estimated shortest distance of the robot from the current node n to the target point, which is usually represented by Euclidean distance, Chebyshev distance and Hamandon distance. Here, the Hamilton distance is used.

$$d_n = abs(n.x - goal.x) + abs(n.y - goal.y) \tag{4}$$

Hamilton distance is the sum of the transverse and longitudinal distances between the current node n and the target point.

However, in the actual operation process, robots, especially the four-way shuttle robots commonly used in logistics warehouses, need a certain amount of time to complete the reversing action in the turning process, as shown in Fig. 5(a). Compared with the linear movement, the reversing time of the robot needs to be taken into account.
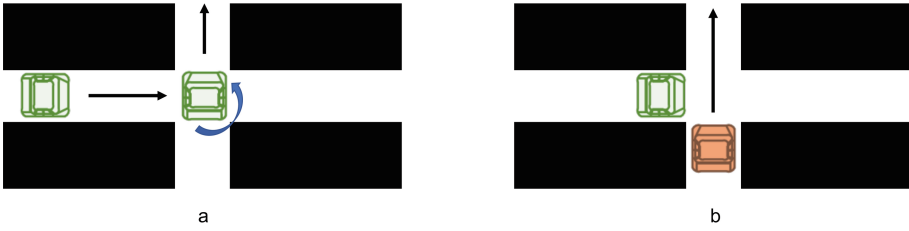
Fig. 5. Robot reversing and waiting actions.

In addition, when the robot encounters path conflict, if its weight is small, it needs to pause and wait for other robots to pass before continuing to move, as shown in Fig. 5(b). The waiting time should also be taken into account. Therefore, this paper improves the A* algorithm and updates the cost formula as follows:

$$f(n) = g(n) + h(n) + \sum_{j=1}^{m} t_{j(turn)} + \sum_{j=1}^{m} t_{j(wait)} \tag{5}$$

where, $\sum_{j=1}^{m} t_{j(turn)}$ is the sum of the extra time taken by the robot to turn from the starting point to the target point, and $\sum_{j=1}^{m} t_{j(wait)}$ is the extra time taken by the robot to wait in place due to path conflict. The improved A* algorithm reduces the turning times and path conflict times of the robot's moving path, and improves the overall efficiency of all robots in the storage model.

2.3   Dynamic Weighting Table

Under the joint action of the reservation table and A* algorithm, the frontal collision of robots is avoided. However, when multiple robots are about to pass the same intersection at the same time, the sequence of robots passing cannot be effectively determined by the control of the reservation table alone.

The general warehouse is determined according to the remaining time for the robot to reach the target point from the current conflict point, but in the flood control warehouse, this scheme may not be able to give priority to the most needed materials and affect the progress and efficiency of the flood control work. As shown in Fig. 6, at this conflict point, robots transporting low-demand materials are preferred to pass through, while robots transporting high-demand materials can only wait in place and then continue to move, resulting in lower efficiency.
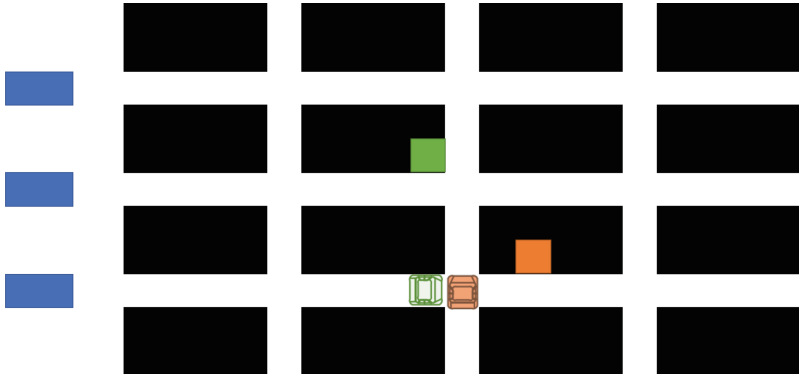
Fig. 6. Intersection conflict.

Combined with the timeliness and priority of flood control warehousing, this paper innovatively puts forward a dynamic weighting table to solve the problem of crossing order and ensure that the materials with the highest demand will be given priority for delivery.

The dynamic weighting table is a matrix table composed of robot label and corresponding dynamic weights, as shown in Table 1.

Table 1. Dynamic weighting table

Robot ID (r_i)	Weight (W_i)	Resource (R_i)	Journey (J_i)	$h_{(n)_i}$
r_1	$W_1 = [R_1, J_1, h_{(n)_1}]$	R_1	J_1	$h_{(n)_1}$
r_2	$W_2 = [R_2, J_2, h_{(n)_2}]$	R_2	J_2	$h_{(n)_2}$
...	...	...	...	...
r_i	$W_1 = [R_i, J_i, h_{(n)_i}]$	R_i	J_i	$h_{(n)_i}$

The dynamic weight is represented by W_i, which refers to the weight of robot r_i at the current moment. The matrix composition of weight W_i is as follows:

$$W_1 = [R_i, J_i, h_{(n)_i}] \tag{6}$$

where, R_i represents the priority level of materials in the transportation task currently performed by robot r_i. After receiving the task, the robot will get the weight R_i according to the type of goods until the end of the task. According to the classification of shelves, goods are also divided into three types here. R_i of goods used in high frequency, medium frequency and low frequency are assigned as 1, 2 and 3, respectively.

Variable J_i means whether robot r_i is carrying goods at the current moment, yes means robot r_i is in the shipment stage at this time, assign a value of 1, and no means robot r_i is in the delivery stage at this time, Assign a value of 2. Variable $h_{(n)_i}$ represents the heuristic time of the current position of robot r_i to reach the target point.

In the warehouse model in this paper, the priority of selecting weights is as follows: Firstly, the more frequent the material demand is, the more priority the corresponding robot will pass, which is the dominant factor. Secondly, if the robots in conflict are transported of the same kind of goods, the robot in the shipping state has higher priority than the robot in the receiving state. Thirdly, in the case of conflict between robots of the same material class and the same receiving/shipping state, the shorter the remaining heuristic time, the higher the priority to pass.

When multiple robots encounter conflict at the intersection, the central controller calls the dynamic weighting table. By comparing the dynamic weight W_i of the robots in the conflict situation, the sequence of passing is determined and the reservation table of corresponding robots is updated. The reservation table of the robot that passes first keeps 0 in the intersection grid, that is, the idle state remains unchanged. The reservation table of the waiting robot is updated to 1 in the intersection grid, which closes the intersection until the robot is allowed to pass through.

3   The Simulation Results

In order to verify the effectiveness of the algorithm, A* algorithm based on reservation table is compared with the optimization algorithm using dynamic weighted table (the optimization algorithm for short) designed in this paper in the warehouse raster map model established in this paper. The simulation experiment was carried out by MATLAB and the simulation experiment is as follows:

In the raster map established in this paper, 8 robots are arranged. The operation efficiency of A* algorithm and the optimization algorithm is compared by simulating 40 material transportation tasks in each group. In order to simulate the transportation characteristics of flood control material warehouse, this paper judged the efficiency of the algorithm by comparing the efficiency of high-frequency material shipment, high-frequency material pickup efficiency, medium-frequency material pickup efficiency and the time of the algorithm.
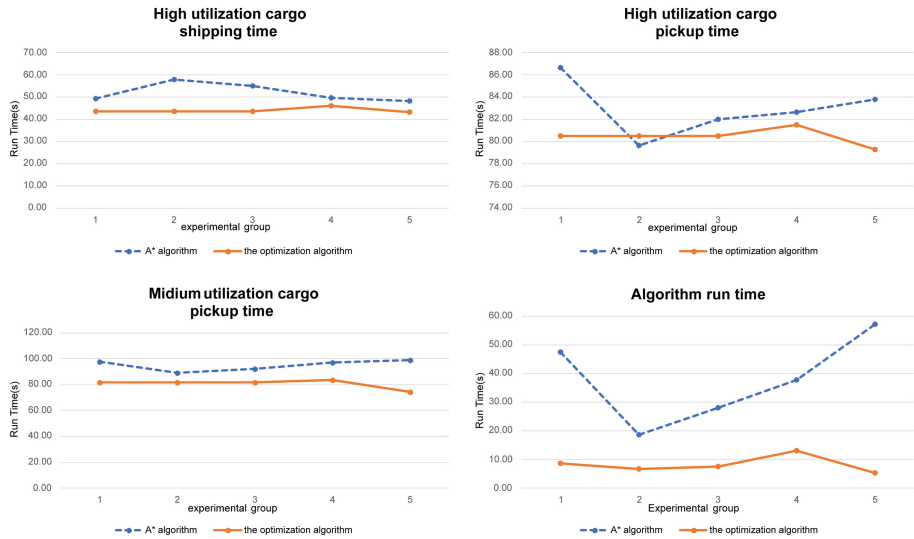
Fig. 7. Comparison of simulation experimental results between A* algorithm and optimization algorithm.

The simulation results are shown in Fig. 7. Figure 7 shows the comparison of the completion time of each index of the two algorithms in the five groups of transport tasks. Compared with A* algorithm, the four experimental indicators of The optimization algorithm have been improved to different degrees. Compared with A* algorithm, the optimization algorithm improves the average shipment efficiency by 14.89%, the average high-frequency pickup efficiency by 2.89%, and the average medium-frequency pickup efficiency by 15.1%. The addition of dynamic weighting table in the optimization algorithm does not have negative effect on intermediate frequency goods because it improves the priority of high frequency goods. In addition, due to the addition of the dynamic weighting table, the robot can make quick decisions when it encounters conflicts and avoid unnecessary path planning. As a result, the running efficiency of the optimization algorithm is improved by 75.05% compared with A* algorithm.

4   Conclusion

In this paper, based on the characteristics of flood control material warehouse, an improved A* algorithm under the reservation table and dynamic weighted table is proposed. The following experiments are completed: building the grid map of the warehouse and dividing the storage areas of goods with different frequency of use; joining the reservation table to prevent the robot from path conflict. This paper innovatively adds dynamic weighting table, which gives priority to high priority goods, and improves the operation efficiency of the warehouse under

flood control. The feasibility and efficiency of the proposed method are verified by simulation experiments.

References

1. Xia, D., Wu, Y., Wang, Y., Zou, X.: Order ordering optimization of "delivery to person" system based on intelligent robot. J. Shenzhen Univ. (Sci. Eng.) **36**(06), 696–701 (2019)
2. Zhou, Y., Xiang, Q., Yu, C., Chen, D., Guan, S.: Warehouse location optimization method considering material demand correlation and turnover rate. J. Donghua Univ. (Nat. Sci.) **46**(03), 414–420 (2020)
3. Gu, Y., Duan, J., Su, Y., Yuan, Y.: Path planning of warehouse logistics robot based on improved ant colony algorithm. J. Wuhan Univ. Technol. (Transp. Sci. Eng.) **44**(04), 688–693+697 (2020)
4. Pan, C., Guo, M.: Batch picking path planning simulation of warehouse logistics robots. Comput. Modernization **02**, 12–16 (2017)
5. Wang, W., Zheng, G., Wei, J., Huang, L., Han, Y., Hua, Q.: Multi-objective optimization of space allocation in intelligent storage for flood control and drought relief materials. Yellow River **42**(02), 27–33 (2020)
6. Lin, Y., Li, Q., Lu, P., Sun, Y., Wang, L., Wang, Y.: Collaborative optimization algorithm of intelligent warehouse location planning and AGV path planning. J. Softw. **31**(09), 2770–2784 (2020)
7. Cao, X., Zhu, M.: Decision optimization of collision avoidance for multi-autonomous guided vehicles based on conflict prediction. Comput. Integr. Manuf. Syst. **26**(08), 2092–2098 (2020)
8. Zhang, D., Sun, X., Fu, S., Zheng, B.: Multi-robot cooperative path planning in intelligent warehouse. Comput. Integr. Manuf. Syst. **24**(02), 410–418 (2018)

Detection of Rail Bottom Damage Defects Based on Recurrent Neural Network

Fengguang Zhou, Qinjun Zhao, Yuan Xu, Qinhua Xu, and Tao Shen[(✉)]

School of Electrical Engineering, University of Jinan, Jinan 250022, China
cse_st@ujn.edu.cn

Abstract. Today, the development of railway network has covered most of the country. This not only brings convenience to people, but also undoubtedly brings a lot of dangerous rail damages which will threaten the running safety of trains. One of the best ways to ensure the safety of railway transportation is to detect the possible damaged position in the rail before accidents. Under the circumstance, this paper studies the detection and identification of rail damages, establishes a set of rail bottom damage detection system based on Recurrent Neural Network, the detection accuracy is about 90%. Then, an adaptive moment estimation optimizer is added to the network model to realize the dynamic adjustment of the learning rate, and the training accuracy is also improved to over 95%. This system can quickly and accurately find out the damage at the bottom of the rail, which not only reduces the labor intensity of workers, but also improves the safety factor of railway transportation.

Keywords: Rail damage detection · Ultrasonic detect · Recurrent Neural Network · Adaptive moment estimation optimizer

1 Introduction

For a long time, railway transportation is the main lifeline of national economy. With the development of railway, the operation of a series of trains, high-speed trains and bullet trains has brought new vitality to modern transportation. However, with the increase of transportation process and service time, various kinds of damages will occur on the surface and inside of rail. Under the influence of these damages, coupled with the temperature stress caused by various extreme weather, the rail will suddenly break [1]. There is no doubt that this serious rail break will bring irreversible harm to the economy and personal safety of passengers. Therefore, it's of great significance to find a rapid detection and identification method for rail damage. Fortunately, with the development of electronic technology and computer technology, as well as the upgrading of nondestructive testing technology, it provides the possibility for rail internal inspection [2].

According to the investigation and analysis of NASA, there are about 70 kinds of nondestructive methods, which can be divided into six categories. However, in practical application, there are five common methods: ultrasonic testing, radiographic testing,

magnetic particle testing, penetrant testing and eddy current testing. In today's railway rail flaw detection field, ultrasonic flaw detection is more popular than other methods [3]. Compared with other conventional flaw detection methods, this method is suitable for internal defect detection, and has the advantages of wide detection range, high sensitivity, high efficiency, simple operation and low cost.

Railway is one of the earliest non-destructive testing departments in China. Non-destructive testing is a comprehensive applied science and technology, which detects macroscopic and microscopic defects by physical means without changing or affecting the performance of the tested object. Railway is one of the earliest non-destructive testing departments in China. Since the introduction of resonance ultrasonic flaw detector made in Switzerland in 1950s, the electronic components of rail flaw detector have experienced three development stages: tube-transistor-integrated circuit. The signal processing mode of rail flaw detector changes from analog to digital, and the display mode of rail flaw detection develops from A-type display to B-type display.

In China, the work of flaw detection workers in relevant departments is mainly to manually find the damaged parts from the flaw detection maps generated by ultrasonic flaw detection vehicles. However, with the continuous construction of new railways in China, the workload of flaw detection personnel will continue to increase. They need to look for damage from B-type display images with a long mileage on the playback software, which also includes a large number of repetitive B-type display images without value. Workers doing this kind of mechanical and tedious work for a long time will greatly reduce the correct rate of playback detection. Therefore, how to quickly identify and classify the rail flaw detection map is very important.

Fortunately, with the development of artificial intelligence and other related technologies, more and more scholars have applied this technology in all walks of life. In article [4], the author introduces the feature parameter extraction of vehicle audio signal and the recognition algorithm of recurrent neural network. In order to optimize the traditional recurrent neural network model, the feature layer is added to the input layer to improve the model structure. The results of processing the audio data of four vehicle types show that the model can effectively identify different vehicle types, and the recognition accuracy rate exceeds 80%, which can meet the basic recognition requirements. In the article [5], Wang K studied the transmission characteristics of rail damage acoustic emission signals in wheel-rail coupling, and proposed a vehicle-mounted rail damage detection method based on convolutional neural network. Moreover, based on the improved convolutional neural network algorithm of multiple acoustic emission events probability, the classification rules of rail damage stage were established, and the constructed damage feature library was classified and trained, and the recognition rate reached the expected effect. Sun C applied the deep convolution neural network to the damage identification of B-type display images, designed the damage identification network architecture, and trained and fine-tuned the architecture, which made the accuracy of the model superior to the existing system of rail inspection vehicles, reached the index requirements of manual analysis, and improved the detection accuracy [6]. Wang Y combined adaptive moment estimation (Adam) algorithm optimizer with convolutional neural network, obtained better classification effect by adding learning rate correction factor, and at the same time reduced iterative oscillation and insufficient classification

stability of image samples [7]. Paper [8] proposes the first use of Adam algorithm to fast and stably converge large-scale tap coefficients of polynomial nonlinear equalizer. Different from serial least-mean square adaptive algorithm, Adam algorithm is a parallel processing algorithm, which can obtain globally optimal tap coefficients without being trapped in locally optimal tap coefficients. Owing to parallel processing and global optimization, Adam algorithm has much better performance on resisting the timing error. In conclusion, Adam algorithm shows great potential for converging the tap coefficients of PNLE in PAM8-based optical interconnects.

In this paper, the generation principle of B-type display image is analyzed first, and then a set of intelligent detection system for rail bottom damage of flaw detection data based on Recurrent Neural Network is designed, which can automatically eliminate undamaged data and identify damaged parts. Finally, on this basis, adaptive moment estimation (Adam) optimizer is added, which can get better training effect detection accuracy. After a series of training, the accuracy of the system can easily reach and exceed the manual accuracy, and the time required is greatly reduced.

2 Generation Process of Flaw Detection Data

2.1 Rail Flaw Detection Equipment

B-type display image data used in this paper are all from GCT-8C flaw detector produced by Xingtai ultrasonic testing equipment co., ltd., Hebei province, and the detection speed are about 10 km/h. This flaw detector is a small hand-push flaw detection tool with nine ultrasonic probes, which can cover all parts of a rail. As shown in Fig. 1, among the nine probes, six 70° probes are responsible for the data of the rail head, which is the most complicated. Two 37° probes are responsible for collecting the data of rail waist and rail bottom, and the last 0° probe is mainly used to calibrate the horizontal position of the flaw detector, so as to ensure that the other probes correspond to the relevant parts of the rail.
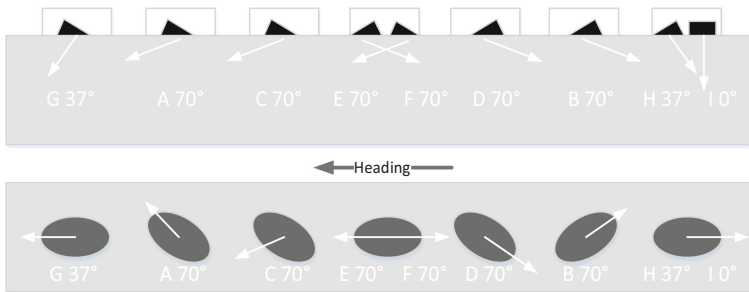

Fig. 1. Probe distribution diagram of flaw detector

2.2   Generation Principle of B-type Display Image

People call the propagation process of sound source vibration in medium (such as air) sound wave, and the mechanical wave with frequency higher than 20 kHz is called ultrasonic wave. In the field of flaw detection, ultrasonic has many advantages, such as good directivity, strong penetrating power, high sensitivity, small energy loss, etc. Moreover, ultrasonic will reflect and refract at heterogeneous interfaces, especially at gas-solid interfaces. Therefore, when the ultrasonic wave propagates to the interface between metal and defect, it will be partially or totally reflected. Each probe of the ultrasonic flaw detector will periodically transmit ultrasonic pulses and receive various reflected echoes at the same time. After the probe receives the reflected echo, it will be amplified and compared with the preset voltage threshold. If the sound pressure of the echo exceeds the current threshold, it means that the echo signal contains damage information, so it is necessary to draw a B-type display image in the output map. In the output result map, there are nine signal channels corresponding to nine ultrasonic probes of monorail flaw detector. The height of the drawing area from channel 1 to channel 6 is 48 pixels, corresponding to 48 sampling points. Starting from the first bit of the first data, when bit is 0, a point is drawn at the corresponding height. The drawing principle of the other three channels is the same.

2.3   Pretreatment of Flaw Detection Output Data

Railway transportation can be said to be an important part of the national strategic planning, so the rail flaw detection data obtained by flaw detection workers are not the original appearance of the data, but hexadecimal numbers after a series of encryption means. Before learning and training these data, it is necessary to convert the hexadecimal number to format, so as to obtain the original appearance of the data.

For B-type display images stored in hexadecimal numbers, each piece of data has special significance. Because this paper studies the detection of rail bottom damage, it takes the data of random section of rail bottom flaw detection as an example. Suppose the data obtained is 00 00 00 03 08 40 00 E3 FF, in which the first four bytes of data 00 00 00 03 are the current number of pulses, 08 is the current number of image channels, the next 40 00 is a mask for encrypting information, E3 is real data and the whole data ends with FF. To sum up, the flaw detection data of the rail bottom can be converted into 0xff, 0xff, 0xff, 0xff, 0xff, 0xff, 0xff, 0xff and 0xe3.

In order to indicate the position of data in sampling points, the drawing rules will use a mask of 1–2 bytes, that is, 8–16 bits, to indicate its position according to the different sampling points of different channels. For the bottom part of 7–8 channels, the number of sampling points is between 8 and 16 bytes, so a 2-byte (16-bit) mask is needed to indicate the source position of the data.

3   Construction of Recurrent Neural Network

3.1   Introduction of Recurrent Neural Network

Recurrent Neural Network (RNN) is an important branch of deep learning, because it can cyclically and recursively process historical data and model historical memory, and is suitable for processing information that is closely related in time and space series. It can learn complex mapping relationship from vector to vector, and it has self-connection inside, which can be used to process sequence data. Besides its biggest feature is that the results of each layer can be used as the input values of the next layer, so that its calculation results have the ability to remember the previous results and keep the data dependency.

At every moment, the output of RNN will be combined with the current input and the model state. As we can see from the following Fig. 2, on the left is a single recurrent neural network, which is composed of input vector x, hidden layer state s and output vector h, and other W, U and V are all weight parameters. Among them, the output of hidden layer has two functions, one is to feed back to itself circularly, and the other is to propagate to neurons in the next moment in sequence. Therefore, the main characteristic of RNN is that the hidden layer state s_t can memorize the sequence data, and the relationship between sequence information can be captured through the hidden layer, so that the sequence with any length can be processed with limited parameters.
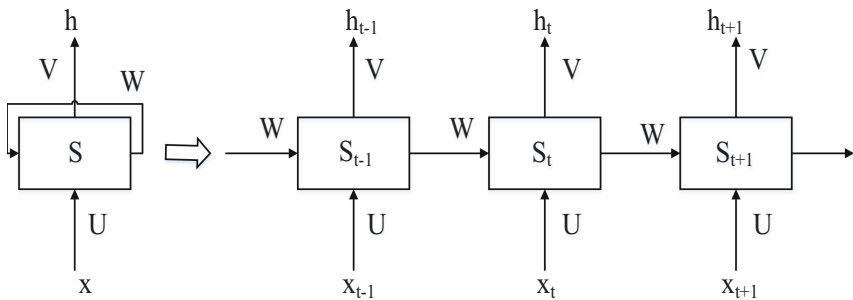

Fig. 2.   Structure model of Recurrent Neural Network

3.2   Input Data Processing

The purpose of this paper is to detect the existence of rail bottom damage defects, so it is necessary to extract the features of this part of data before it can be used as the input of the system. In the preparation work at the early stage of the experiment, the flaw detection data at the bottom of the rail are found out accurately by manual marking, which includes both damage data and normal data. In this experiment, 4014 pieces of data are shared, and each piece of data was 15, so the system input was a 4014*15-dimensional matrix. In the data collection stage, different flaw detectors are used by different workers, and their flaw detection speed may be different, which leads to slightly different characteristics

of the same injury data. Therefore, in the experiment, the activation function Sigmoid is used to normalize the original input matrix, that is, to normalize all the data to [0,1], which can reduce the system operation and the system error. The specific expression of this function is as shown in Eq. 1.

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

3.3 Specific Introduction of Training Model

There are three layers in this model, including input layer, hidden layer and output layer. The characteristic of RNN is that the propagation between layers includes forward propagation and backward propagation. RNN can be regarded as the result that the same neural network structure is replicated many times in time series. This replicated structure is called loop body, and how to design loop body is also the key to solve practical problems of this network model. The classification of this paper is mainly binary, so Sigmoid function is used in forward propagation, and cross entropy cost function is defined to represent the average error of all samples.

$$C = -\frac{1}{m} \sum_{x} \left[y l n a + (1 - y) l n (1 - a) \right] \tag{2}$$

In which C represents cost function, x represents samples, y represents actual value,a represents output value of input matrix calculated by activation function, and n represents total number of samples. When the error is bigger, the gradient is bigger, and the adjustment speed of parameters is wider, so the final training speed is faster. In the back propagation of RNN, it is assumed that the loss function is cross entropy. In all the forward propagation, the inverse calculation is carried out, and the gradient is calculated. The gradient descent algorithm is used to update the weight.

3.4 Adaptive Moment Estimation

Adaptive moment estimation (Adam) is a popular optimization method to estimate large-scale parameters in neural networks. As we all know, when calculating on the whole data set, as long as the learning rate is low enough, we can always get non-negative progress on the loss function. When using this method, it is necessary to set an initial learning rate first, and then introduce a second-order momentum based on this learning rate. This variable can effectively reduce the convergence speed of the system, which can effectively prevent the system from exceeding the optimal solution in the calculation process.

4   Training and Testing

As explained above, the flaw detection data of rail bottom are distributed in channel 7 and channel 8, and the example of damage is shown in Fig. 3, in which the marked with black box represents the bottom defect, and the unmarked one represents the normal echo or the weld echo signal under another flaw detection condition.
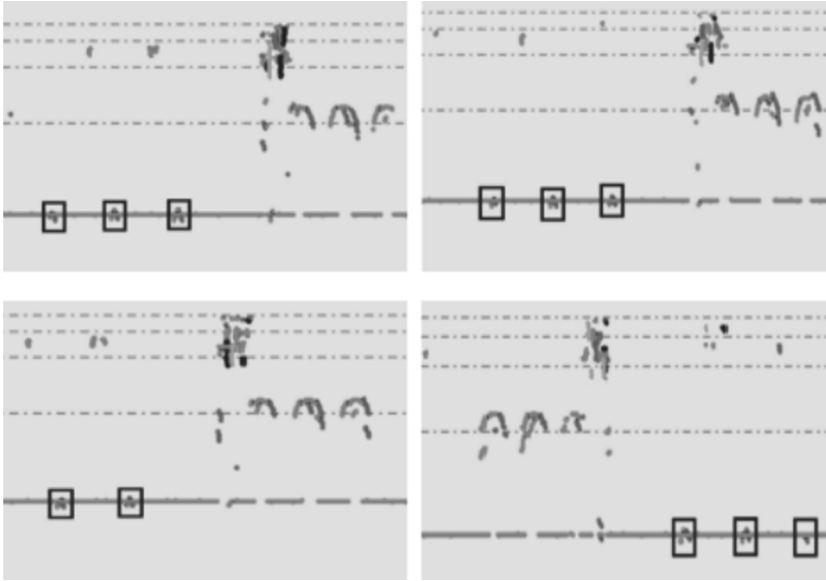

Fig. 3. Case of rail bottom damage

All the data used in this paper are from each line within the jurisdiction of Jinan Railway Bureau in Shandong Province in recent years, and the amount of data is sufficient. In the early stage of the experiment, the flaw detection data on the whole line were sliced, and the images containing the echoes at the bottom of the rail were extracted, and then marked and integrated. About 90% of the data set is selected as the training set, and the rest is used as the test set. The experimental results are shown in Fig. 4 (Table 1).

It can be seen from the waveform of the cost function of the experimental results that when the training times are the same, different learning rates will lead to great differences in training results. When the learning rate is set to 0.0001, the initial value of the cost function is relatively small, but the convergence speed in the later period is slow, that is, the average variance of the sample training results converges slowly and remains a large value at the end of training, which also leads to the low accuracy. However, when the learning rate is increased to 0.00018, the convergence speed is increased compared with that before, and the average variance of the sample training results converges to a smaller value at the end of training, which corresponds to an increase in the training accuracy.
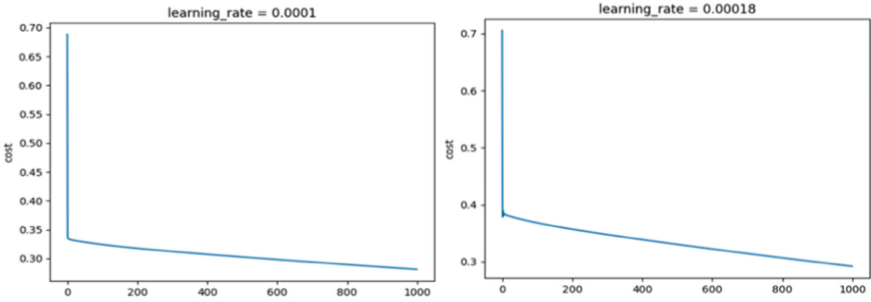
Fig. 4. Cost function image under different learning rate

Table 1. Training accuracy record

Training times	Learning rate	Training accuracy
100000	0.0001	85.85%
500000	0.0001	95.79%
100000	0.00018	91.56%
500000	0.00018	98.11%

When analyzing the training results of a single picture, we can find that the value of cost function decreases with the increase of training times, which means that the average variance of the whole sample training is getting smaller and smaller, that is, the accuracy of the network model is getting higher and higher, and finally reaches 98.11% after 500,000 training.
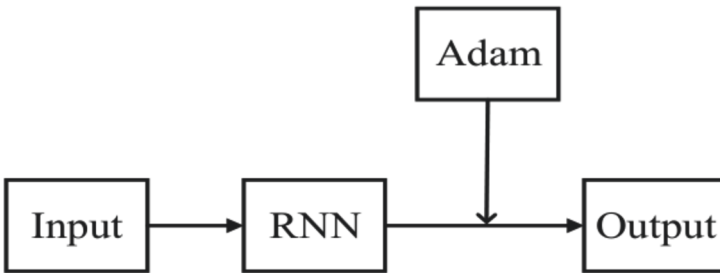

Fig. 5. System with Adam optimizer

However, when combined with adaptive moment estimation optimizer shown in Fig. 5, the system can automatically adjust the learning rate according to the training situation, so that better training results can be obtained. Adam is a variant of gradient descent algorithm, and it dynamically adjusts the learning rate of each parameter by using the first-order moment estimation and the second-order moment estimation of

gradient. However, the learning rate of each iteration parameter has a certain range, so the learning rate (step size) will not become large because of the large gradient, and the value of the parameter is relatively stable. When the system function is close to the optimal solution, the step size of system convergence is reduced, which can prevent the system from crossing the optimal solution in the calculation process and causing local divergence. Figure 6 shows the accuracy of the system after adding Adam optimizer. It can be seen from the figure that the system can get better training effect with less training times.
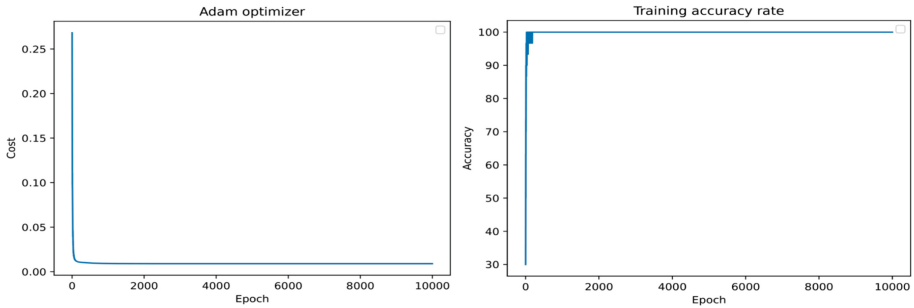

Fig. 6. Training results with Adam optimizer

Table 2. Test sample of rail bottom damage

Sample number	Whether it is defect	Expected output	Actual output
1	Yes	1	1
2	Yes	1	1
3	Yes	1	1
4	Yes	1	1
5	Yes	1	1
6	Yes	1	1
7	Yes	1	1
8	No	0	1
9	No	0	0
10	No	0	0

After the whole training, to ensure the reliability of the system designed in this paper, some brand-new data fragments were intercepted from the flaw detection data of previous years, which included both injury data and normal echo data. After a round of testing, 10 test results were randomly selected and counted in Table 2. From the table,

it can be seen that the accuracy of the test basically accords with the results obtained in the training process.

5 Conclusions

Firstly, this paper analyzes the present situation of railway development in China, and gives the damage caused by rail damage and the causes of these damages. Then, combined with some current research status, an automatic detection system of rail bottom damage based on Recurrent Neural Network is designed. What's more, combines with Adam optimizer, a better detection effect is obtained. After training more than 4,000 samples, the accuracy of the system can reach the accuracy in the actual working process. At the same time, the detection speed far exceeds the manual detection speed.

In the future research, if you want a better and more accurate classification effect, you can use SoftMax classifier to achieve multi-level classification of injuries. You can also use the same method to make a larger data set, so that after the system is fully trained, check the detection effect of the system.

References

1. Tian, G., Gao, B., Gao, Y., Wang, P., et al.: Overview of inspection and monitoring technology for railway rail defects. Chin. J. Sci. Instrum. **37**(08), 1763–1780 (2016)
2. Zhang, H., Song, Y., et al.: Summary of nondestructive testing and evaluation technology for rail defects. Chin. J. Sci. Instrum. **40**(02), 11–25 (2019)
3. Li, W.: Study on ultrasonic detection of rail damage and location. North University of China (2013)
4. He, J.: Vehicle recognition method based on recurrent neural network. W. Commun. Technol. **10**(44), 160–162 (2020)
5. Wang, K.: Research on Vehicle Identification Method of Rail Damage Based on Convolutional Neural Network. Harbin Institute of Technology (2017)
6. Sun, C., Liu, J.: Intelligent identification method of rail damage based on deep learning. China Rail. Sci. **39**(05), 51–57 (2018)
7. Wang, Y., Xu, P.: Classification of CNN electron microscope medical images based on improved Adam optimizer. J. Xi'an Univ. Posts Telecommun. **24**(05), 26–33 (2019)
8. Zhou, J., et al.: Adaptive moment estimation for polynomial nonlinear equalizer in PAM8-based optical interconnects. Opt. Express **27**(22), 32210–32216 (2019)

Short Term Load Forecasting Method Based on Full Convolution Deep Learning

Hai-hong Bian[1($\boxtimes$)], Xing-jian Shi[1], Qian Wang[1], and Li-kuan Gong[2]

[1] Nanjing Institute of Technology, Nanjing 211167, China
llmn0002@sina.com
[2] Power Dispatching Control Center of Shenzhen Power Supply Bureau,
Shenzhen 440304, China

Abstract. The traditional load forecasting methods can not take into account the time and space characteristics of load data at the same time, which leads to the low application efficiency of load forecasting methods. In order to solve this problem, a short-term load forecasting method based on full convolution deep learning is proposed. Preprocess the power load data, delete the abnormal samples, unify the load data format through normalization processing, design the relevant network parameters, determine the loss function, complete the design of the prediction model, use the sample data to train the prediction model, and predict the short-term power load after the model meets the prediction requirements. The experimental results show that: in the same experimental environment, the short-term power load forecasting method based on full convolution deep learning has high prediction accuracy, wide prediction range, and its application efficiency has been improved.

Keywords: Full convolution deep learning · Short term electricity · Load forecasting · Neural network

1   Introduction

Power system load forecasting is based on the historical data of power sector, economy, policy, meteorology, etc., to explore the influence of the historical data change characteristics of power load on the future power load, and to find the possible relationship between power load and various related factors, so as to correctly predict the future power load. According to the length of forecasting time, power sector load forecasting can be divided into annual forecasting, monthly forecasting and daily forecasting, as well as long-term, medium-term, short-term and ultra short-term load forecasting [1]. The coverage time of long-term forecast varies from the next few years to more than ten years, which is mainly used for power development planning and power grid development planning; Medium term forecast refers to the power load forecast in the next year (12 months), which is used to arrange the overhaul plan and the economic operation of hydropower plants; Short term load forecasting usually refers to the daily load forecasting in the next 24 h and the weekly load forecasting in the next 168 h.

The goal is to arrange daily and weekly power generation tasks for each power plant, including the determination of unit startup and shutdown, coordination of thermal power and hydropower, fuel procurement and supply, tie line exchange power, economic distribution of power load, reservoir dispatching and equipment repair, etc.; Ultra short term forecast refers to the forecast of 1 h, 0.5 h or even 10 min in the future, which is mainly used for safety monitoring, preventive control and emergency handling [2–4].

With the development of the power market in the world, people pay more and more attention to the field of load forecasting. Some scholars have made statistics on the works published in IEEE Trans. And other world-famous journals. The results show that people's interest in load forecasting has been increasing year by year since the early 1970s. In the mid-1980s, due to the shortage of energy, there is an urgent need for scientific load management and the expectation of accurate and adaptive load model. It makes the research investment of load forecasting more and more high [5]. Since the 1990s, with the vigorous development of power markets in various countries, power load forecasting has been paid more and more attention by scholars.

Power system short-term load forecasting has been studied for a long time. It is developed with the gradual development of EMS system in power system. With the introduction of data analysis and artificial intelligence technology, people have proposed prediction methods based on different principles. Each of these methods has its own advantages and disadvantages. No method is absolutely superior to the prediction method in any region or at any stage [6]. For short-term load forecasting, it is necessary to fully study the characteristics of power grid load change and analyze the factors related to load change, especially the weather factors, date types and other factors that have a greater relationship with short-term load change [7]. Many literatures have done a lot of research on load forecasting, mostly based on more advanced theory to improve the accuracy of load forecasting, which provides a strong guarantee for the economy and security of power system operation. At present, the research of load forecasting mainly focuses on the improvement of forecasting methods and the proposal of new methods [8]. According to the applied mathematical model and the development process of load forecasting, it can be divided into traditional classical methods and new methods of artificial intelligence technology [9–11]. Classical methods mainly refer to time series models of various statistical theories, while artificial intelligence technology is a modern forecasting method represented by soft computing. After application research, it is found that some traditional prediction methods have poor application efficiency [12–15].

Therefore, a short-term power load forecasting method based on full convolution deep learning is proposed. The full convolution neural network in deep learning is used to obtain the load data in real time, so as to achieve accurate forecasting and solve the problems in the traditional forecasting methods mentioned above.

2   Design of Short-Term Load Forecasting Method Based on Full Convolution Deep Learning

2.1   Power Load Data Preprocessing

Among the various attributes of historical load data and historical meteorological information, the data may be abnormal due to the unknown records or abnormal events.

These data not only have no reference value for forecasting itself, but also affect the load forecasting results to a large extent [16]. Therefore, it is necessary to determine which abnormal samples are and exclude them from the data input. The following abnormal data detection algorithm is based on the combination of distance and clustering. Firstly, the samples are divided by clustering algorithm; Secondly, for each sample, calculate its clustering distance, and then calculate its anomaly index. Then all samples are arranged in descending order according to the size of the anomaly index. Finally, the sample with the largest outlier index is determined as the outlier sample.

If the value of some continuous attributes of a sample is too large (greater than 10150), delete the sample. If the value of each attribute of a sample is missing, delete the sample. If some attributes in all sample data are the same or missing, delete this attribute. For each attribute U_{ok} and $k = 1, 2, \ldots, K$ in the dataset, if U_{ok} is a continuous attribute, the total average value Q_k and the total standard deviation SD_k of the attribute are calculated by using all the valid values of the attribute. Replace all missing values with the calculated overall average. If U_{ok} is a discrete attribute, a new missing value is used instead of all missing values, and it is regarded as a valid attribute value.

After the above processing, the attribute variables are used as input variables, and a clustering model is constructed by clustering algorithm, so that the samples are allocated to a cluster [17–19].

For each continuous attribute U_k, calculate and store its overall mean Q_k and overall standard deviation SD_k. For each cluster $p = 1, 2, \ldots, P$, calculate its sample size n_p. If U_k is a continuous variable, the average value Q_{pk} and standard deviation SD_{pk} of U_k in cluster p are calculated; If U_k is a discrete attribute with m different values, calculate the frequency η_{nkm} of each value in cluster p, and save the mode Q_{pk} of the attribute in cluster p. These statistics will be used to calculate the log likelihood distance $d(p, s)$ between a cluster p and a given sample s.

Given the group deviation index GDI of a sample s, it is actually the log likelihood distance $d(p, s)$ between s and its nearest sample p, which is used to measure the similarity between the sample and cluster p. The larger the distance is, the smaller the similarity is; otherwise, the greater the similarity is. $d(p, s)$ is the sum of the distance component $d(p, s)$ between each attribute value U_k and the corresponding attribute value in the cluster. $d_k(p, s)$ is called the variable deviation index VDI_k of the variable U_k.

When the attribute variable U_k is a continuous attribute, the deviation index VDI_k of attribute U_k relative to cluster p is calculated as follows:

$$d_k(p, s) = \frac{1}{2}\left[-N_p log\left(\Delta_k + \alpha_{pk}^2\right) - N_s log\left(\Delta_k + \alpha_{sk}^2\right) + N_{(p,s)} log\left(\Delta_k + \alpha_{\langle p,s\rangle k}^2\right)\right]$$
(1)

In the formula, N_p is the number of samples in cluster p, $\Delta_k = \alpha_k^2/6$ is a regulating factor to avoid the situation that the logarithm is meaningless when there is only one sample in the cluster. In the formula, α_k^2 is the sample variance of U_k in the whole data set, α_{pk}^2 is the sample variance of U_k in cluster p, N_s is the number of samples in the cluster, and α_{sk}^2 is the sample variance of U_k in cluster s, $N_{(p,s)}$ is the number of samples after cluster p contains sample s. obviously, $N_{(p,s)} = N_p + 1$.

$\alpha_{\langle p,s\rangle k}^2$ means that cluster p contains the sample variance of U_k after sample s.

After calculating the deviation index VDI_k of all attribute variables, the group deviation index of the sample can be calculated

$$GDI = d(p, s) = \sum_{K=1}^{K^A + K^B} d_k(p, s) \tag{2}$$

In the formula, K^A represents the number of continuous attributes of the sample, and K^B represents the number of discrete attributes of the sample. After obtaining the above calculation results, the anomaly index is calculated to measure the anomaly of the sample. It is the ratio of GDI of sample s divided by the average GDI of all samples in cluster p.

$$AI = \frac{GDI_s}{mean(GDI_p)} \tag{3}$$

The larger the calculated anomaly index AI is, the more abnormal the data sample is. Generally, observations with an anomaly index less than 1.5 will not be regarded as outliers, but observations with an index value greater than 2 may be outliers. After clearing the outliers in the short-term load data, the data samples are normalized [20].

Generally speaking, in the training and learning of historical load data, neural network will not directly use the original data, otherwise it will appear the phenomenon of neuron saturation, so it is necessary to normalize the load data. Before inputting the historical data into the neural network, it is normalized by formula (4). In order to convert it into the value of $[-1, 1]$ interval; At the end of the forecast, the result is changed back to the actual load value through formula 5.

$$y_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{4}$$

$$x_i = y_i(x_{max} - x_{min}) + x_{min} \tag{5}$$

In the formula, x_{max} and x_{min} represent the maximum and minimum values of the load in the load data sample, and x_i and y_i represent the values of the sample before and after normalization. After the processing, the features of power load data are extracted, and the forecasting model is designed. The extracted features are used as the input of the forecasting model to realize the short-term load forecasting.

2.2 Prediction Model Design Based on Deep Learning

The prediction model design combines the full convolutional neural network, and uses several parallel u-net neural networks to process short-term power load data. The structure of the full convolutional neural network is shown in Fig. 1.
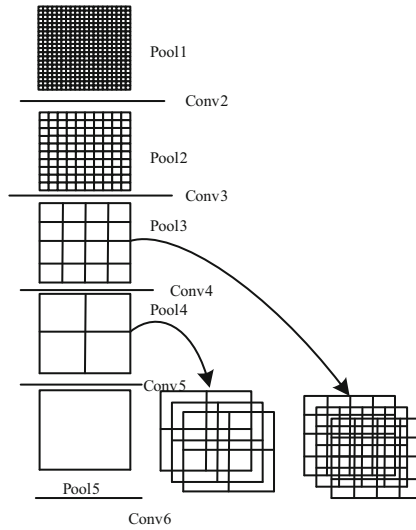
Fig. 1. Full Convolutional neural network structure diagram

The related parameter settings are shown in Table 1.

Table 1. Parameters of fully convolutional neural network

Modular	Level	Layer name	Convolution kernel/window size	Output feature map size
Search path	Level 1	Convolution layer 1	3 * 3	80 * 8 * 240/32
		Convolution layer 2	3 * 3	80 * 8 * 240/32
		Pooling layer	2 * 2	40 * 8 * 120/32
	Level2	Convolution layer 3	3 * 3	40 * 8 * 120/064
		Convolution layer 4	3 * 3	40 * 8 * 120/064
		Pooling layer 2	2 * 2	20 * 8 * 60/64
	Level3	Convolution layer 5	3 * 3	20 * 8 * 60/128
		Convolution layer 6	3 * 3	20 * 8 * 60/128
		Pooling layer 3	2 * * 2	20 * 8 * 30/128

(*continued*)

Table 1. (*continued*)

Modular	Level	Layer name	Convolution kernel/window size	Output feature map size
		Bidirectional convolution	3 * 3	10*8 * 30/128
Convolution	Level3	LSTM layer 1		10 * 8 * 30/256
		Convolution layer 7		10 * 8 * 30/256
		Bidirectional convolution		10 * 8 * 30/256
		LSTM layer 2		10 * 8 * 30/256
		Upper sampling layer 1	2 * 2	20 * 8 * 60/372
		Convolution layer 8	3 * 3	20 * 8*60/128
		Convolution layer 9	3 * 3	20 * 8 * 60/128
		Upper sampling layer 1	2 * 2	40 * 8 * 120/192
Expansion path	Level2	Convolution layer 10	3 * 3	40 * 8 * 120/64
		Convolution layer 11	3 * 3	40 * 8 * 120/64
		Upper sampling layer 3	2 * 2	80 * 8*240/96
	Level1	Convolution layer 12	3 * 3	80 * 8 * 240/32
		Convolution layer 13	3 * 3	80 * 8 * 240/32
Output		Convolution layer 14	2 * 2	80 * 8 * 240/32

The whole structure of the neural network includes three parts: downward contraction path, convolution LSTM module and upward expansion path. The main function of contraction path is to learn and extract important features from the input data. Convolution LSTM layer receives these features and fuses the adjacent features of Y axis to enhance and improve its own features. After expanding the path, the predicted data block is obtained. Because the y-axis dimension of the input data block is 8, the proposed neural network is equivalent to 8 parallel u-net neural networks [21, 22].

It is proposed that the shrinking path and expanding path of neural network consist of three different levels of shrinking module and expanding module respectively, and the characteristic dimension and channel number of different levels of modules are different. Each shrinking module contains a residual module and a pooling layer. Each residual module contains two convolution layers with the size of 3×3 convolution kernel [23, 24]. The convolution layer uses the strategy of filling 0 value to avoid dimension reduction. Each convolution layer is followed by a batch normalization layer and a relu activation function. The template size of pooling layer is 2×2, and the maximum pooling strategy is adopted. Every time the pooling layer passes through, the number of channels of the next feature is doubled. More feature channels can get more multi-scale features. The number of channels in the three levels of contraction path is 32, 64 and 1280 respectively.

The convolution LSTM layer is located between the contraction path and the expansion path. This algorithm uses two bidirectional convolution LSTM layers with 256 channels. This structure effectively connects eight parallel u-net neural networks in series. The features extracted from different u-net after shrinking path interact and fuse in convolution LSTM layer to improve the segmentation effect of network. There is a convolution layer with the size of 3×3 convolution kernel between two bidirectional convolution LSTM layers. The features enhanced by convolution LS TM layer will be sent to the expansion path.

The expansion path and contraction path are just opposite. It includes three expansion modules. Each expansion module contains an upper sampling layer and a residual module. The structure of each residual module is the same as that of the shrinkage path. The upsampling layer is equivalent to the reverse process of the pooling layer, and the template size is also 2×2. The maximum strategy is adopted. Every time it passes through the up sampling layer, the number of the next channels is halved. The number of channels in the three levels of expansion path is 128, 64 and 32 respectively. Like the original u-net structure, this network also adds three jump connections to connect the same level of contraction module and expansion module. It is helpful to learn segmentation details. Finally, a 1 * 1 convolution layer is used to reduce the dimension, and a three channel characteristic graph matrix is obtained. Finally, softmax layer is used for classification. Softmax layer can classify the known data and map the probability to the range of (0, 1). The calculation formula is as follows:

$$c_i = \frac{exp(b_i)}{\sum_{k=1}^{3} exp(b_k)} \tag{6}$$

In the formula, exp represents the exponential constant with natural number as the base, b_i and b_k represent the output of 1 * 1 convolution layer classified as class i and class k respectively, and c_i represents the probability of class i after classification. In the training

step, the loss function uses the daisy loss function to minimize the difference between the predicted result and the labeled result.

$$E_{dice} = 1 - \frac{2 \times \sum_{i=1}^{3} \sum v_i(x)g_i(x)}{\sum_{i=1}^{3} \sum \left(v_i^2(x)g_i^2(x)\right)} \tag{7}$$

In the formula, E_{dice} refers to the value of the day's loss function, $v_i(x)$ represents the probability that the prediction data x belongs to class i, and $g_i(x)$ represents the value that the labeled data x belongs to class i. Before forecasting, the forecasting model is trained, and the parameters of the forecasting model are adjusted continuously through the loss function until the forecasting requirements are met.

2.3   Short-Term Power Load Forecasting

Before the prediction, the bottom-up unsupervised pre training is used to initialize the parameters of the full convolution neural network, and then the top-down supervised training is used to fine tune the parameters of the whole network. The connection weight β and bias ε of the visible layer and the hidden layer in the model are continuously adjusted to fit the functional relationship between the input and output data, Then the overall cost function of the prediction model is as follows:

$$H(\beta, \varepsilon) = \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - y_i'\right)^2 + \frac{\lambda}{2} \sum_i \sum_j \sum_l \left(\beta_{i,j}^l\right)^2 \tag{8}$$

In the formula, y_i is the target output value of input x_i, y_i' is the corresponding prediction value, and λ is the weight attenuation parameter.

BP back propagation algorithm based on gradient descent is used to update the parameters of a single self encoder. However, when it is directly used in prediction model training, gradient dispersion is easy to occur and fall into local optimum. The core idea of the algorithm is to train only the AE with one hidden layer in the network each time. When the AE is optimal, the reconstructed visual layer is removed, and the output of the hidden layer is used as the input to train the next AE, until the last AE is optimal, the unsupervised pre training of SAE is completed. In the process of layer by layer training, the reconstructed data gets an output through a single AE, and then the cost function is minimized to update $\{\beta, \varepsilon\}$. the weight update rules are as follows:

$$\beta^l(k+1) = \beta^l(k) - \sigma \nabla_{\beta^l} C(\beta, \varepsilon) \tag{9}$$

$$\varepsilon^l(k+1) = \varepsilon^l(k) - \sigma \nabla_{\varepsilon^l} C(\beta, \varepsilon) \tag{10}$$

In the formula, $l = 1, 2, \ldots, N$, σ is the learning step of each iteration. If l is the l-compiler, the error of each node in the hidden layer is τ^l, then the partial derivative of the cost function C to the weight parameter is calculated as follows:

$$\nabla_{\beta^l} C(\beta, \varepsilon) = \tau^l + \lambda \beta^l \tag{11}$$

$$\nabla_{\varepsilon^l} C(\beta, \varepsilon) = \tau^l \tag{12}$$

$$\tau^l = \left[\beta^l \left(x^l - y\left(x^l\right) \right) g'(z) \right] f'(z) \tag{13}$$

In the formula, $y(x^l)$ represents the reconstructed output of x^l, $f'(z)$ and $g'(z)$ are the derivative values of the activation functions $g(z)$ and $f(z)$ respectively. After the unsupervised pre training, the prediction model is given initial weights, and the supervised parameters of the whole network are fine tuned by combining the connection weights β and bias ε of the visible layer and the hidden layer. The activation function value and the overall cost function of each layer of neuron node are calculated by using the above formula, and then the BP algorithm is used to calculate backward layer by layer. At this time, the cost function becomes $H(\beta, \varepsilon)$, The error of each node in the output layer and hidden layer of the prediction model is expressed as:

$$\begin{cases} \tau^l = -(y - y')f'(z) \\ \tau^l = \beta^l \tau^{l+1} f'(z) \end{cases} \tag{14}$$

The gradient descent method is used to update parameters repeatedly to minimize $H(\beta, \varepsilon)$. After the training, the training samples are used as the input of the forecasting model for feature learning, and the high-order feature representation of the reconstructed load, weather and day type data is obtained. After unsupervised network training and supervised parameter fine-tuning, the forecasting model is optimized, and the test samples are input into the forecasting model for short-term load forecasting. So far, the design of short-term load forecasting method based on full convolution deep learning is completed.

3 Experimental Research on Short-Term Load Forecasting Method Based on Full Convolution Deep Learning

3.1 Setting up the Experimental Environment

The Hadoop cluster built in the experiment consists of six PCs with the same configuration. Each PC has 2 G memory, soogb hard disk, dual core inter E7500 CPU, 2.93 GHz main frequency, and runs several CentOS Linux operating systems. One machine is the master node of master, which is responsible for the resource allocation and job scheduling of the whole cluster. The other five machines are slave nodes, which are mainly used to store data and run tasks. If the number of HDFS copies is 3, the slave node can run 4 map/reduce tasks at the same time. The Hadoop cluster topology is shown in Fig. 2.
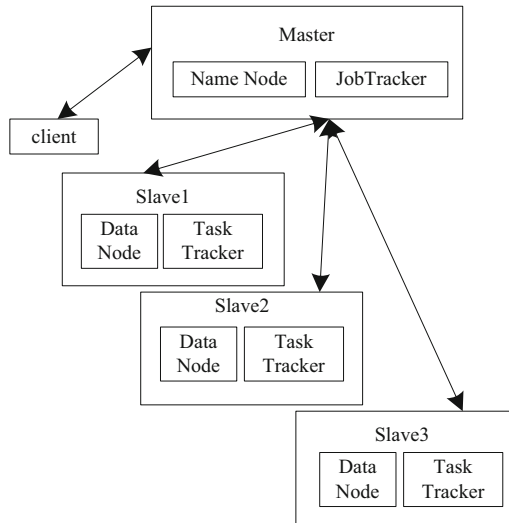
Fig. 2. Hadoop Cluster topology diagram

In the above experimental environment, the traditional three prediction methods are introduced and compared with the designed prediction method in the same experimental environment. According to the experimental results, each prediction method is compared and analyzed.

3.2   Preparation of Experimental Data Sets

The experimental data comes from the load data and influencing factor data collected by a power grid in a certain area. The amount of data is TB level, and the latitude is high. It is mainly structured and semi-structured data, which is in line with the characteristics of power big data. The influencing factors include daily maximum temperature, average temperature, minimum temperature, average wind speed, average humidity, average precipitation, sunshine type and season. The calculation of the correlation degree between each influencing factor and load is shown in Table 2.

Table 2. Results of the calculation of the correlation of each factor

Project	Parameter	Project	Parameter
Daily maximum temperature	0.87	Average humidity	0.66
Average temperature	0.71	Average precipitation	0.32
Minimum temperature	0.57	Sunshine type	0.60
Average wind speed	0.28	season	0.53

According to the calculation results, the daily maximum temperature, average temperature, sunshine type, average humidity and minimum temperature are selected as the five key factors affecting the load. The training sample is the power consumption data of a whole year, and the sampling interval is 1H. The composition of the training set is shown in Table 3.

Table 3. Training sample set

Parameter setting		Meaning
Input	x_1	Predicted daily maximum temperature
	x_2	Forecast daily average temperature
	x_3	Forecast daily average humidity
	x_4	Forecast daily sunshine type
	x_5	Forecast daily minimum temperature
	x_6–x_8	Forecast the load value of $t-1$, t and $t+1$ on the day before the day
	x_9–x_{11}	Load values of $t-1$, t and $t+1$ in the same week before the forecast day
Output	y	Load forecast value of forecast day t

Taking 24-h load data as test samples, relative error, average error, root mean square error and average absolute percentage error are used to measure load forecasting effect. The calculation formula of each index is as follows:

$$\chi_{ME} = \frac{1}{n+1} \sum_{i=0}^{n} (u(i) - v(i)), n = 23 \tag{15}$$

$$\chi_{RMSE} = \sqrt{\frac{\sum_{i=0}^{n} (u(i) - v(i))^2}{n+1}}, n = 23 \tag{16}$$

$$\chi_{MAPE} = \frac{1}{n+1} \sum_{i=0}^{n} \left| \frac{u(i) - v(i)}{v(i)} \right| \times 100\%, n = 23 \tag{17}$$

In the formula, $u(i)$ represents the actual load value of time i and $v(i)$ is the predicted value of time i. Taking the above indexes as the standard to measure the prediction accuracy, only one group of experiments is not enough to predict the practical application level of the method. Therefore, the second group of experimental plans is designed, and the second group of experiments takes the size of the prediction range as the measurement index, the larger the prediction coverage, the better the prediction ability.

3.3 Experimental Results and Analysis of Load Prediction Accuracy

In the load forecasting accuracy experiment, four different forecasting methods are used to forecast the daily load. The historical load value, weather attribute and day type data

of similar days are used at each time point. The forecasting method is used to forecast the power load change for five consecutive days, and the load forecasting error index is counted. The specific experimental results are shown in Table 4.

Table 4. Experimental results of prediction accuracy of different load forecasting methods

	Date	Relative error of daily maximum load/%	Average error/MW	Root mean square error/%	Percentage error of mean absolute value/%
Traditional prediction method 1	1 day	3.15	0.0850	4.11	21.05
	2 days	4.05	0.0904	3.97	23.53
	3 days	3.67	−0.0793	5.02	22.58
	4 days	4.26	−0.0993	4.95	19.75
	5 days	4.35	0.876	4.55	18.22
Traditional prediction method 2	1 day	3.66	0.0941	3.94	20.63
	2 days	4.05	0.0826	4.11	22.64
	3 days	4.91	0.0791	6.23	23.74
	4 days	4.52	−0.823	5.24	24.06
	5 days	3.86	0.0741	4.81	19.54
Traditional prediction method 3	1 day	4.62	0.0873	6.26	23.22
	2 days	3.91	0.0562	5.12	21.45
	3 days	3.87	0.0611	4.29	19.87
	4 days	4.03	−0.0438	3.94	26.52
	5 days	4.59	0.0506	4.01	20.47
The prediction method is proposed	1 day	1.72	−0.0141	2.09	16.21
	2 days	1.35	0.0106	2.62	17.32
	3 days	1.78	0.0151	2.50	17.26
	4 days	1.06	−0.0137	2.68	18.51
	5 days	1.14	0.0171	2.35	16.82

Comparing and observing the data in the table, it can be seen from the data in the table that the relative error, average error, root mean square error and absolute error of the three traditional prediction methods in the short-term prediction are higher than those of the proposed prediction methods, which are quite different from the actual load. The maximum relative error, average error, root mean square error and absolute error of the proposed prediction methods are 1.78%, 2.68 MW, 18.51% respectively 0.0171%. Therefore, the load forecasting method based on full convolution deep learning has high forecasting accuracy. The reason is that the input data is convoluted and pooled by full convolution deep learning, which improves the prediction performance.

3.4 Experimental Results and Analysis of Load Forecasting Coverage

In the experiment, in the experimental environment, different load forecasting methods are used to forecast short-term power load, the coverage area of load data processed by each forecasting method is calculated, and the experimental results are output through the third-party software. The specific content of load forecasting coverage area experiment results is shown in Fig. 3.
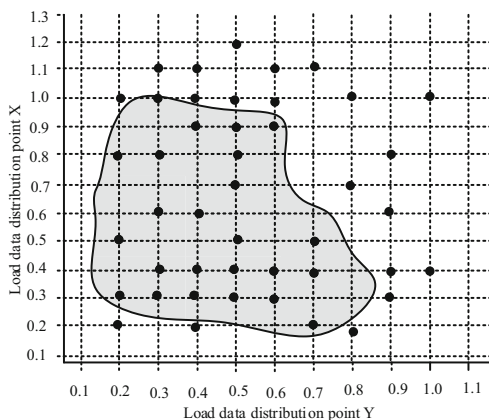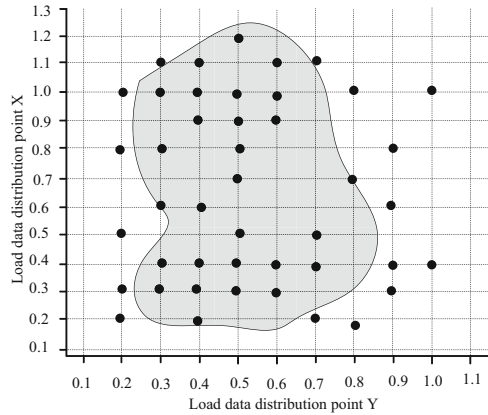

(a) Experimental results of traditional load forecasting method 1
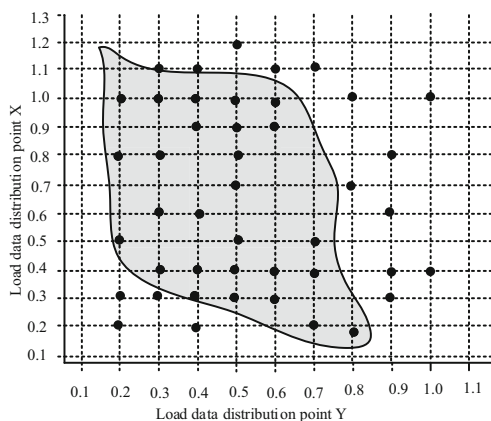
(b) Experimental results of traditional load forecasting method 2

Fig. 3. Experimental results of different load forecasting methods

(c) Experimental results of traditional load forecasting method 3
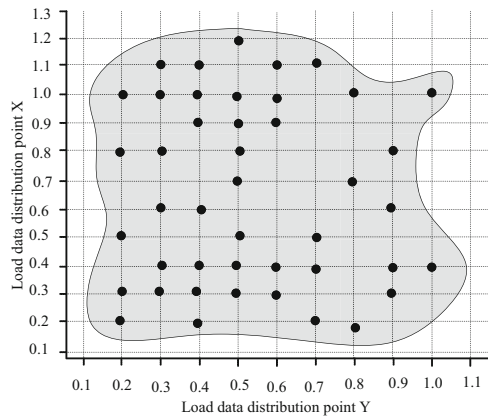
(d) Experimental results of the proposed load forecasting method

Fig. 3. continued

Through the observation of Fig. 3, it can be seen that the three traditional load forecasting methods can not cover all load data when processing load data. There are some missing data points outside the distribution point. Compared with the above, the proposed load forecasting method covers a larger area and covers all load data. According to the experimental results of prediction accuracy, the load forecasting method based on full convolution depth learning has high prediction accuracy and high application efficiency, which is better than the traditional load forecasting method.

4   Conclusion

With the rapid development of the economy and the advent of the electrical age, the demand for power is also increasing. The research on power load forecasting model has

become a hot topic. The accuracy of power load forecasting is directly related to the supply and demand balance of power grid, and affects the operation cost and security and stability of power grid. Therefore, a new model of power load forecasting is explored, it is very important to improve the accuracy of power load forecasting. In order to improve the accuracy of power load forecasting, this paper proposes a short-term power load forecasting method based on full convolution deep learning.

Load data has dual characteristics of time and space. In terms of time, load data has the characteristics of long-term trend, periodicity and close association with its own historical data. Because the load data has these characteristics, the algorithm model can fit it well. In space, the operation of power system is affected by external factors, and the load data produced also has the characteristics of randomness. According to the characteristics of power load data, after the design of the short-term power load forecasting method based on full convolution deep learning is completed, the experimental environment is built, the load data is collected, and the prediction accuracy and prediction area are compared by using different prediction methods. According to the two groups of experimental results, it is proved that the prediction method based on full convolution deep learning has higher accuracy and wider application range. It is suitable for practical projects. However, there are still some deficiencies in the design process. In the follow-up study, some shortcomings of this method will be studied and analyzed one by one to further improve the load forecasting method.

Reference

1. Xu, Y., Wu, Z., Zhu, H., et al.: Short-term load forecasting based on multi-scale convolutional neural network. J. Shenyang Univ. Technol. **42**(06), 618–623 (2020)
2. Zhao, B., Wang, Z., Ji, W., et al.: A short-term power load forecasting method based on attention mechanism of CNN-GRU. Power Syst. Technol. **43**(12), 4370–4376 (2019)
3. Chen, G., Teng, H.: Short-term load forecasting based on deep learning of hybrid neural networks. Water Resour. Power **38**(04), 193–196 (2020)
4. Chen, G., Teng, H.: Short-term load forecasting based on deep learning of hybrid neural networks. Water Resour. Power **56**(03), 91–96+102 (2019)
5. Zhao, H., Zhao, Y., Guo S.: Short-term load forecasting based on complementary ensemble empirical mode decomposition and long short-term memory. Electr. Power **53**(06), 48–55 (2020)
6. Zhao, W., Lin, R., Tang, W., et al.: Forecasting model of short-term PM2.5 concentration based on deep learning. J. Nanjing Norm. Univ. (Nat. Sci. Edn.) **42**(03), 32–41 (2019)
7. Li, H., Lin, J., Li, G., et al.: Short-term power load forecasting based on weighted cosine similarity and extreme learning machine. Sci. Technol. Eng. **20**(11), 4370–4374 (2020)
8. Xu, Y., Lu, Y., Zhu, B., et al.: Short-term load forecasting method based on FFT optimized ResNet model. Control Eng. China **26**(06), 1085–1090 (2019)
9. Luo, Y., Cai, Y., Qi, Y., et al.: Short-term power load forecasting algorithm based on maximum deviation similarity criterion BP neural network. Appl. Res. Comput. **36**(11), 3269–3273 (2019)

10. Li, Y.: Establishment and application of short-term power load forecasting. Comput. Simul. **28**(10), 316–319 (2011)
11. Liu, S., Liu, X., Wang, S., Muhammad, K.: Fuzzy-aided solution for out-of-view challenge in visual tracking under IoT assisted complex environment. Neural Comput. Appl. **33**(4), 1055–1065 (2021)
12. Gao, P., Li, J., Liu, S.: An introduction to key technology in artificial intelligence and big data driven e-learning and e-education. Mob. Netw. Appl. **26**, 2123–2126 (2021). https://doi.org/10.1007/s11036-021-01777-7
13. Liu, S., Liu, D., Srivastava, G., et al.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. **3**, 1–23 (2020)
14. Kalakova, A., Kumar, S., Jamwal, P.K., et al.: A novel genetic algorithm based dynamic economic dispatch with short-term load forecasting. IEEE Trans. Ind. Appl. **57**, 2972–2982 (2021)
15. Wang, Y., Chen, J., Chen, X., et al.: Short-term load forecasting for industrial customers based on TCN-LightGBM. IEEE Trans. Power Syst. **36**, 1984–1997 (2021)
16. Tavassoli-Hojati, Z., Ghaderi, S.F., Iranmanesh, H., et al.: A self-partitioning local neuro fuzzy model for short-term load forecasting in smart grids. Energy **199**, 117514 (2020)
17. Gilanifar, M., Wang, H., Sriram, L., et al.: Multitask Bayesian spatiotemporal gaussian processes for short-term load forecasting. IEEE Trans. Industr. Electron. **67**(6), 5132–5143 (2020)
18. Sheng, Z., Wang, H., Chen, G., et al.: Convolutional residual network to short-term load forecasting. Appl. Intell. **51**(4), 1–15 (2020)
19. Sharma, S., Majumdar, A., Arregui, V.E., et al.: Blind Kalman filtering for short-term load forecasting. IEEE Trans. Power Syst. **35**, 4916–4919 (2020)
20. Barman, M., Choudhury, N.: Season specific approach for short-term load forecasting based on hybrid FA-SVM and similarity concept. Energy **174**, 886–896 (2019)
21. Kassa, Y., Zhang, J., Zheng, D.: EMD-PSO-ANFIS based hybrid approach for short-term load forecasting in microgrids. IET Gener. Transm. Distrib. **14**(3), 470–475 (2019)
22. Yin, L., Xie, J.: Multi-temporal-spatial-scale temporal convolution network for short-term load forecasting of power systems. Appl. Energy **283**(6), 116328 (2020)
23. Li, C.: Designing a short-term load forecasting model in the urban smart grid system. Appl. Energy **266**, 114850 (2020)
24. Sadaei, H.J., Candid, D., Guimaraes, F.G., et al.: Short-term load forecasting by using a combined method of convolutional neural networks and fuzzy time series. Energy **175**, 365–377 (2019)

AI-Based Internet of Medical Things

Covid-19 Detection by Wavelet Entropy and Cat Swarm Optimization

Wei Wang[(⊠)]

Informatics Building, School of Informatics, University of Leicester, University Road,
Leicester LE1 7RH, UK
ww152@leicester.ac.uk

Abstract. The rapid global spread of COVID-19 poses a huge threat to human security. Accurate and rapid diagnosis is essential to contain COVID-19, and an artificial intelligence-based classification model is an ideal solution to this problem. In this paper, we propose a method based on wavelet entropy and Cat Swarm Optimization to classify chest CT images for the diagnosis of COVID-19 and achieve the best performance among similar methods. The mean and standard deviation of sensitivity is 74.93 ± 2.12, specificity is 77.57 ± 2.25, precision is 76.99 ± 1.79, accuracy is 76.25 ± 1.49, F1-score is 75.93 ± 1.53, Matthews correlation coefficient is 52.54 ± 2.97, Feature Mutual Information is 75.94 ± 1.53.

Keywords: COVID-19 · Wavelet entropy · Cat Swarm Optimization · Feedforward Neural Network · K-fold cross-validation

1 Introduction

Since the novel new coronavirus was first discovered in Wuhan, China in 2019, the virus has rapidly spread to various countries worldwide, causing massive deaths. The World Health Organization (WHO) has named the virus Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) [1] and defined the global pandemic it caused (COVID-19) as a Public Health Emergencies of International Concern (PHEIC) [2].

In addition to the extremely high transmission rate of SARS-CoV-2, the more complex diagnostic identification is a significant reason for the rapid spread. The most recognised method for the detection of COVID-19 is the reverse-transcription polymerase chain reaction (RT-PCR). Although RT-PCR offers high diagnostic accuracy, it has the disadvantage of being time-consuming (up to two days) and has a high false-negative rate [3]. The asymptomatic patients (who are not obviously symptomatic but are still infectious) significantly hampered rapid identification and monitoring of COVID-19 [4]. The search for a rapid and accurate COVID-19 test is an urgent task in the fight against the epidemic.

Medical studies have confirmed that the features in chest CT images can effectively reflect the focal status of COVID-19. The chest CT images combined with epidemiological and laboratory tests effectively assess COVID-19 [5]. However, manual-based

examinations are labour-intensive and time-consuming, and artificial intelligence is ideal for this process.

Lu, Z. (2016) [6] used a radial basis function (RBF) based method to detect brain disease, which can also be generalised to COVID-19 detection. Yao, X. et al. (2020) [7] proposed an experiment using wavelet entropy (WE) and biogeography-based optimisation (BBO) based method to classify chest CT images to diagnose COVID-19 and achieved good performance. Chen, Y. (2020) [8] proposed a method based on support vector machine (SVM) and grey-level co-occurrence matrix (GLCM) to classify chest CT images of heath people and COVID-19 patients.

This paper proposes a WE and Cat Swarm Optimization (CSO) based method for COVID-19 identification and obtains the highest performance among similar methods. The main contributions of the paper are: (i) The first classification model to combine WE and CSO. (ii) The best performance in WE based COVID-19 diagnostic approaches.

2   Dataset

The experiment used a dataset of COVID-19 CT images from the Fourth People's Hospital in Huai'an, China, taken from 66 as COVID-19-infected and 66 healthy individuals, for a total of 132 subjects. Each data sample consisted of CT images of the lungs and the corresponding accounting test results. These subjects had a maximum age of 24 years and a minimum age of 91 years and comprised 77 males and 55 females. The selection and labelling of the dataset performed by two specialist radiologists who simultaneously identified and labelled the CT images uploaded to the Picture Archiving and Communication System (PASC), selecting 1–4 CT images with suitable lesion identification for each subject infected with COVID-19. The portion of the dataset for healthy subjects was obtained from a random sample of 159 healthy subjects with physical examination results. The Fig. 1 show two image samples of the dataset.
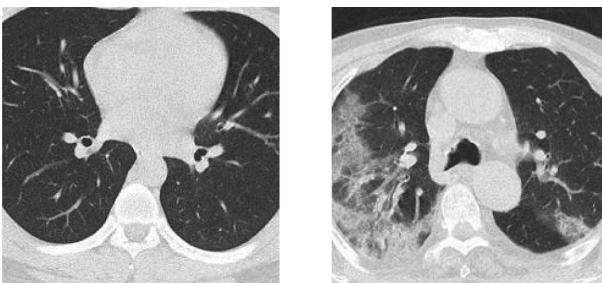

Fig. 1. CT Image sample of Health Control (*left*) and COVID Patient (*right*).

3 Methodology

3.1 Wavelet Entropy

The experiment uses the Discrete Wavelet Transform (DWT) as a feature extraction method, a popular method for signal analysis [9]. DWT inherits the main ideas of the Fourier Transform and addressed the disadvantages of the Fourier Transform for processing unstable signals by replacing the wireless length trigonometric basis with a finite length decaying wavelet basis while preserving the time and frequency information of the signal [10]. The Fourier Transform is defined as Eq. (1) and the DWT is defined as Eq. (2).

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt \tag{1}$$

where ω refers to frequency, t refers to time.

$$W(a, \tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\psi * \left(\frac{t-\tau}{a}\right)dt \tag{2}$$

where a represents scale, τ represents translation, ψ represents parent wavelet function, and t represents time.

DWT can use the multi-scale nature of orthogonal wavelets to effectively decompose images at different scales (resolutions) [11] and retain as much image information as possible. Image contours are preserved in large-scale spaces, and image details are preserved in the small-scale space. This property can effectively improve the recognition of small lesions in medical images. However, too much image information requires a large amount of space and computational cost [12]. Therefore, we introduced entropy to reduce the dimensionality of the dataset while retaining as much valid information as possible. Entropy is a stochastic statistic that can be approximated from the grey-scale probability histogram of an image. The Eq. (3) defines the entropy.

$$S(\alpha) = -\sum_i P(\alpha_i)log_b P(\alpha_i) \tag{3}$$

where α refers to grey levels and P is the probabilities of grey levels [13].

The original image signal is used as input, and after the first level wavelet decomposition, we can obtain four matrices, one low-frequency component (LL1) and three high-frequency components (HL1, LH1, HH1). The second level wavelet decomposition of the resulting low-frequency components yields a low-frequency component (LL2) and three new high-frequency components (HL1, LH1, HH1). Repeat the above procedure after a total of n matrix decompositions, one low-frequency component (LLn), and n high-frequency components (HL1, LH1, HH1,…, HLn, LHn, HHn) are finally obtained. The entropy is then extracted from these components represented in matrix form [14], allowing for a vector of entropy with a feature reduction to n+1 values [15]. These values can be used as input to the neural network [16]. The Fig. 2 shows a brief demonstration of the whole process.
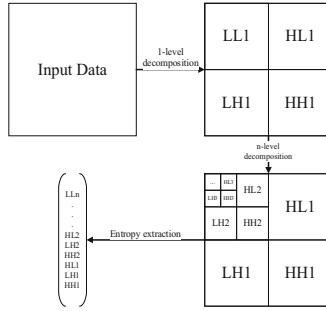
Fig. 2. DWT multi-level decomposition

3.2 Feedforward Neural Network

A feed-forward neural network (FNN) is a simple type of artificial neural network that is often used as a classifier for pattern classification [17]. FNN mainly consist of an input layer, one (superficial network) or more (deep network) hidden layers and an output layer. Inside the FNN, each layer (except the output layer) connected to the next [18], and parameters are propagated singularly from the input layer through the hidden layer to the output layer. The Fig. 3 illustrates the basic structure of an FNN. Deep FNN [19, 20] is our future research direction.
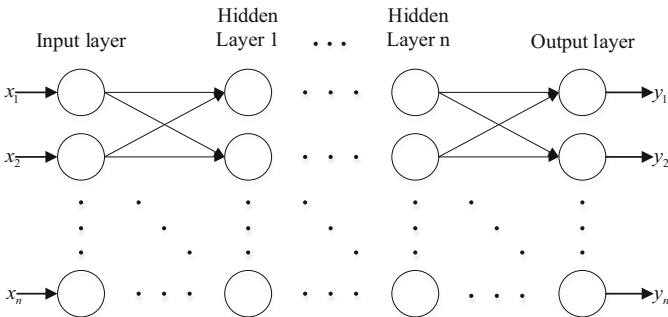
Fig. 3. The basic structure of Feedforward Neural Network

3.3 Cat Swarm Optimization

CSO is an algorithm that optimises values by mimicking the behaviour of cats in nature [21], using a predetermined ratio (MR) to determine the number of agents (cats) in both seeking and tracking modes [22]. There are four elements defined for the seeking mode:

- Seeking Memory Pool (SMP) defines the size of seeking memory for each cat.
- Seeking Range of the selected Dimension (SRD) provides the boundary conditions for change.

- Counts of Dimensions to Change (CDC) indicate the number of dimensions to change [23].
- Self-Position Considering (SPC) determines whether the current position of the cat is used as a seeking memory candidate coordinate.

Unlike the slowness of the seeking mode, the tracking mode simulates a cat's fast pursuit of a target, for a D-dimension space, the velocities of cats are represented as $V_i = (V_{i1}, V_{i2}, \cdots, V_{iD})$. The main idea of CSO is to solve the optimisation problem by uniting two sets of cats [24], as shown in Fig. 4.
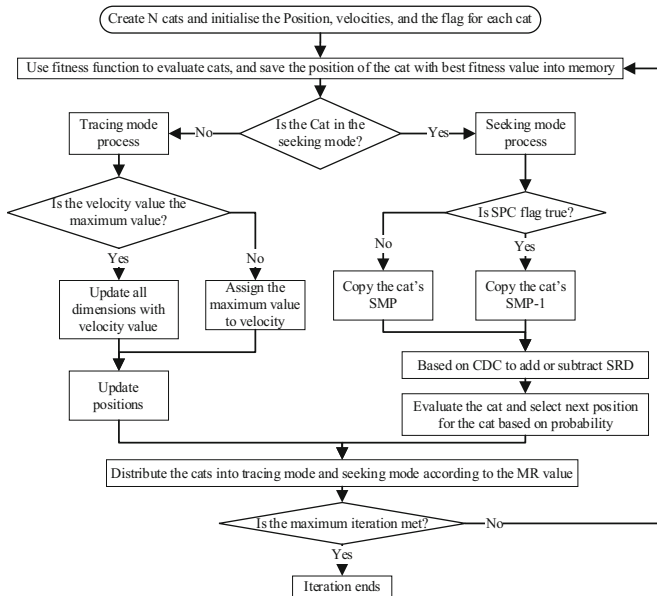
Fig. 4. Flowchart of the Cat Swarm Optimization algorithm

3.4 K-fold Cross Validation
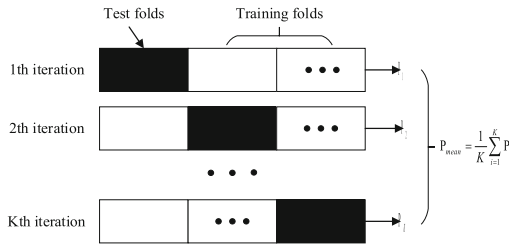
Fig. 5. Illustration of K-fold cross validation

K-fold cross validation is a data set partitioning method that effectively addresses the adverse effects of too small a data set [25]. The value of K is predetermined based on the size of the data set [26]. The K value affects the performance of the final model, with the larger value of K, the smaller the size of the dataset, but the more computational resources required [27, 28]. The Fig. 5 illustrates the K-fold cross validation.

4 Experiment Results and Discussions

4.1 WE Results

The Fig. 6 illustrates a four-level discrete wavelet transform sample image. The three most giant blocks in the image are the three high-frequency components produced by the first-level wavelet transform. Similarly, the high-frequency components produced by the second, third and fourth level wavelet transform could be found in the image. The low-frequency components of the first three wavelet transform are used for the next level wavelet transform. The only low-frequency component left is produced by the fourth level wavelet transform, which is the block at the top left corner in the image.
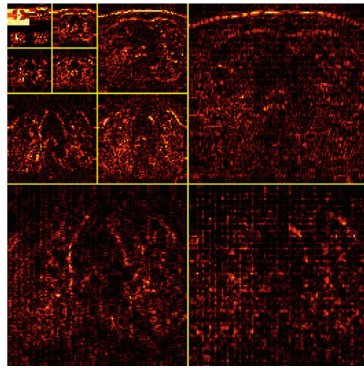

Fig. 6. Four-level discrete wavelet transform sample image.

4.2 Statistical Results

The experiments used 10-fold cross-validation [29] to ensure efficient dataset using and the results are unbiased [30]. The detail results of 10 times running are listed in the Table 1, in which the mean and standard deviation (Mean + SD) of sensitivity (Sen) is 74.93 ± 2.12, specificity (Spc) is 77.57 ± 2.25, precision (Prc) is 76.99 ± 1.79, accuracy (Acc) is 76.25 ± 1.49, F1-score (F1) is 75.93 ± 1.53, Matthews correlation coefficient (MCC) is 52.54 ± 2.97, Feature Mutual Information (FMI) is 75.94 ± 1.53.

Table 1. 10 runs of 10-fold cross-validation

Run	Sen	Spc	Prc	Acc	F1	MCC	FMI
1	72.97	78.38	77.14	75.68	75.00	51.43	75.03
2	77.70	76.35	76.67	77.03	77.18	54.06	77.18
3	74.32	79.73	78.57	77.03	76.39	54.13	76.42
4	73.65	79.05	77.86	76.35	75.69	52.78	75.72
5	78.38	81.08	80.56	79.73	79.45	59.48	79.46
6	74.32	76.35	75.86	75.34	75.09	50.69	75.09
7	72.97	76.35	75.52	74.66	74.23	49.35	74.24
8	77.70	73.65	74.68	75.68	76.16	51.39	76.17
9	73.65	79.05	77.86	76.35	75.69	52.78	75.72
10	73.65	75.68	75.17	74.66	74.40	49.33	74.41
Mean + SD	74.93 ± 2.12	77.57 ± 2.25	76.99 ± 1.79	76.25 ± 1.49	75.93 ± 1.53	52.54 ± 2.97	75.94 ± 1.53

4.3 Comparison to State-of-the-Art Approaches

The result of the research has achieved a better performance than other state-of-art approaches in multiple metrices. The detailed comparison in all metrices is summarised in the Table 2, which validates our approach as a viable and promising approach in the field of medical image classification.

Table 2. Performance comparison to other methods

Method	Sen	Spc	Prc	Acc	F1	MCC	FMI
RBFNN [6]	66.89 ± 2.43	75.47 ± 2.53	73.23 ± 1.48	71.18 ± 0.80	69.88 ± 1.08	42.56 ± 1.61	69.97 ± 1.04
WE-BBO [7]	72.97 ± 2.96	74.93 ± 2.39	74.48 ± 1.34	73.95 ± 0.98	73.66 ± 0.98	47.99 ± 2.00	73.66 ± 1.33
GLCM-SVM [8]	72.03 ± 2.94	78.04 ± 1.72	76.66 ± 1.07	75.03 ± 1.12	74.24 ± 1.57	50.20 ± 2.17	74.29 ± 1.53
Ours	74.93 ± 2.12	77.57 ± 2.25	76.99 ± 1.79	76.25 ± 1.49	75.93 ± 1.53	52.54 ± 2.97	75.94 ± 1.53

5 Conclusions

The potential of artificial intelligence to classify features in CT images cannot be ignored, and the rapid and accurate diagnosis it can bring is of considerable importance to the

medical field. The performance of our proposed method based on wavelet entropy and cat swarm optimisation also confirms the feasibility of a time-frequency based classification method for the COVID-19 CT image classification task.

References

1. Hotez, P.J., Fenwick, A., Molyneux, D.: The new COVID-19 poor and the neglected tropical diseases resurgence. Infect. Dis. Poverty **10**(1), 10 (2021)
2. Yuki, K., Fujiogi, M., Koutsogiannaki, S.: COVID-19 pathophysiology: a review. Clin. Immunol. **215**, 108427 (2020)
3. Fang, Y., et al.: Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology **296**(2), E115–E117 (2020)
4. Meng, H., et al.: CT imaging and clinical course of asymptomatic cases with COVID-19 pneumonia at admission in Wuhan, China. J. Infect. **81**(1), e33–e39 (2020)
5. Zhao, X., et al.: The characteristics and clinical value of chest CT images of novel coronavirus pneumonia. Clin. Radiol. **75**(5), 335–340 (2020)
6. Lu, Z.: A pathological brain detection system based on radial basis function neural network. J. Med. Imaging Health Inf. **6**(5), 1218–1222 (2016)
7. Yao, X., Han, J.: COVID-19 detection via wavelet entropy and biogeography-based optimization. In: Santosh, K.C., Joshi, A. (eds.) COVID-19: Prediction, Decision-Making, and its Impacts. LNDECT, vol. 60, pp. 69–76. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-9682-7_8
8. Chen, Y.: Covid-19 classification based on gray-level co-occurrence matrix and support vector machine. In: Santosh, K.C., Joshi, A. (eds.) COVID-19: Prediction, Decision-Making, and its Impacts. LNDECT, vol. 60, pp. 47–55. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-9682-7_6
9. Guido, R.C.: Nearly symmetric orthogonal wavelets for time-frequency-shape joint analysis: introducing the discrete shapelet transform's third generation (DST-III) for nonlinear signal analysis. Commun. Nonlinear Sci. Numer. Simul. **97**, 105685 (2021). Article ID 105685
10. Saritha, M., et al.: Classification of MRI brain images using combined wavelet entropy based spider web plots and probabilistic neural network. Pattern Recogn. Lett. **34**(16), 2151–2156 (2013)
11. Chui, C.K., et al.: Wavelet thresholding for recovery of active sub-signals of a composite signal from its discrete samples. Appl. Comput. Harmon. Anal. **52**, 1–24 (2021)
12. El-Dahshan, E.-S.A., et al.: Hybrid intelligent techniques for MRI brain images classification. Digit. Sig. Process. **20**(2), 433–441 (2010)
13. Yildiz, A., et al.: Application of adaptive neuro-fuzzy inference system for vigilance level estimation by using wavelet-entropy feature extraction. Exp. Syst. Appl. **36**(4), 7390–7399 (2009)
14. Phillips, P.: Pathological brain detection in magnetic resonance imaging scanning by wavelet entropy and hybridization of biogeography-based optimization and particle swarm optimization. Prog. Electromagnet. Res. **152**, 41–58 (2015)
15. Yang, M., Automatic brain tumor detection in MRI Scanning by wavelet entropy. J. Am. Geriatr. Soc. **64**(S2), S349–S349 (2016). Article ID 349
16. Phillips, P.: Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm. Neurocomputing **272**, 668–676 (2018)
17. Nakamura-Zimmerer, T., et al.: QRnet: optimal regulator design with LQR-augmented neural networks. IEEE Control Syst. Lett. **5**(4), 1303–1308 (2021)

18. Wang, S.-H., Muhammad, K., Hong, J., Sangaiah, A.K., Zhang, Y.-D.: Alcoholism identification via convolutional neural network based on parametric ReLU, dropout, and batch normalization. Neural Comput. Appl. **32**(3), 665–680 (2018)

19. Wang, S.-H., Zhang, Y.-D.: DenseNet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification. ACM Trans. Multimedia Comput. Commun. Appl. **16**(2s), 1–19 (2020). Article 60

20. Alcaraz, J.C., Moghaddamnia, S., Peissig, J.: Efficiency of deep neural networks for joint angle modeling in digital gait assessment. EURASIP J. Adv. Sig. Process. **2021**(1), 1–20 (2021). Article ID 10

21. Suresh, M., Sam, I.S.: Exponential fractional cat swarm optimization for video steganography. Multimedia Tools Appl. **80**(9), 13253–13270 (2021). https://doi.org/10.1007/s11042-020-10395-6

22. Wang, S.-H., Yang, W., Dong, Z., Phillips, P., Zhang, Y.-D.: Facial emotion recognition via discrete wavelet transform, principal component analysis, and cat swarm optimization. In: Sun, Yi., Lu, H., Zhang, L., Yang, J., Huang, H. (eds.) IScIDE 2017. LNCS, vol. 10559, pp. 203–214. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67777-4_18

23. Zhang, Y.-D., Sui, Y., Sun, J., Zhao, G., Qian, P.: Cat Swarm Optimization applied to alcohol use disorder identification. Multimedia Tools Appl. **77**(17), 22875–22896 (2018)

24. Saha, S.K., et al.: Cat swarm optimization algorithm for optimal linear phase FIR filter design. ISA Trans. **52**(6), 781–794 (2013)

25. Akbari, H., Sadiq, M.T., Rehman, A.U.: Classification of normal and depressed EEG signals based on centered correntropy of rhythms in empirical wavelet transform domain. Health Inf. Sci. Syst. **9**(1), 1–15 (2021). Article ID 9

26. Wang, S.-H.: COVID-19 classification by CCSHNet with deep fusion using transfer learning and discriminant correlation analysis. Inf. Fus. **68**, 131–148 (2021)

27. Wang, S.-H., Zhang, Y., Cheng, X., Zhang, X., Zhang, Y.-D.: PSSPNN: PatchShuffle stochastic pooling neural network for an explainable diagnosis of COVID-19 with multiple-way data augmentation. Comput. Math. Meth. Med. **2021**, 1–18 (2021). Article ID 6633755

28. Rajapandy, M., Anbarasu, A.: An improved unsupervised learning approach for potential human microRNA–disease association inference using cluster knowledge. Netw. Model. Anal. Health Inf. Bioinf. **10**(1), 1–16 (2021). Article ID 21

29. Bodaghi, S., et al.: Regularization of a nonlinear inverse problem by discrete mollification method. Comput. Meth. Differ. Eqn. **9**(1), 313–326 (2021)

30. Govindaraj, V.: Explainable diagnosis of secondary pulmonary tuberculosis by graph rank-based average pooling neural network. J. Ambient Intell. Humanized Comput. (2021). https://doi.org/10.1007/s12652-021-02998-0

A Short Survey on Deep Learning Models for Covid-19 Detection Based on Chest CT and X-ray Images

Wei Wang[✉]

Informatics Building, School of Informatics, University of Leicester, University Road,
Leicester LE1 7RH, UK
ww152@leicester.ac.uk

Abstract. The continued and rapid global spread of COVID-19 is taking a heavy toll on the global economy and human health, which has attracted the attention of professionals in various fields. Controlling the spread of this disease and reducing the threat to human life is of paramount importance. There are no clinically effective drugs for this disease. However, research on deep learning-based diagnostic systems for COVID-19 has yielded significant results and is expected to be an essential weapon in the fight against COVID-19 in the future. This paper provides a brief summary and evaluation of 15 studies on deep learning-based COVID-19 diagnostics, covering a total of 13 common pre-trained models and nine custom deep learning models in the COVID-19 dataset, and discusses the current challenges and future trends in this category of research. This paper aims to help healthcare professionals and researchers understand the advances in deep learning techniques for COVID-19 diagnosis to assist them in conducting relevant research to stop the further spread of COVID-19.

Keywords: COVID-19 · Coronavirus · Deep learning · Machine learning · Epidemic

1 Introduction

COVID-19 is an acute atypical respiratory disease caused by a novel coronavirus (SARS-CoV-2) infection. The main symptoms of COVID-19 are persistent cough, fever and breathlessness, which can be life-threatening in severe cases [1]. Furthermore, COVID-19 is highly infectious and mutable and can be transmitted by droplets from human respiration, making it extremely difficult to control [2]. Therefore, rapid diagnosis and isolation are essential to interrupt the spread of COVID-19.

The currently accepted standard for confirming the diagnosis of COVID-19 is the reverse transcription-polymerase chain reaction (RT-PCR) [3]. This test is cumbersome and time-consuming, and the yield of reagents easily influences the speed of detection. Most importantly, RT-PCR has a high rate of false negatives and often requires multiple tests to obtain accurate results [4]. Fortunately, with the rapid development of modern

technology, numerous solutions based on new medical facilities and intelligent tools are available to assist in diagnosing this type of disease, such as CT and X-ray, to obtain accurate images of human organs. By analysing the effects of these, specialist doctors can diagnose the subject accurately. However, due to the large number of COVID-19 infections, it is not enough to rely on manual diagnosis by specialist doctors.

Deep learning is a statistically based artificial intelligence technique widely used in healthcare and industry as a reliable image recognition technology [5]. Diagnostic systems based on deep learning techniques can automate the recognition and identification of CT and X-ray images. These systems could be fall into two main categories, pre-trained models based on migration learning and customised deep learning models for specific domains.

This paper provides a brief review of research on COVID-19 diagnostic systems based on deep learning techniques and provides statistics and evaluation of some of the currently available systems and frameworks. The second part of the paper presents statistics on COVID-19 diagnostic systems based on pre-trained models. The third section presents statistics on custom deep learning models for COVID-19 diagnosis. The fourth and fifth sections briefly discussed all models covered in the paper and the limitations of this work.

2 Pre-trained Models

A pre-trained model is one in which the parameters of an already trained model are migrated to a new model to help train the new model. Typically, training a new model from scratch requires a large number of computational resources and time costs based on a large-scale dataset. Since most of the data or tasks are correlated, transfer learning allows the model parameters already learned to be shared with the new model, thus speeding up the convergence of the new model and optimising the learning efficiency of the new model. Most of the standard pre-training models available for transfer learning are based on convolutional neural networks, such as ResNet, SqueezeNet, DenseNet, VGG, Inception, AlexNet, GoogleNet.

Minaee, Kafieh [6] trained ResNet18, ResNet50, SqueezeNet and DenseNet-121 based on transfer learning. They used 2000 chest X-ray images of normal people and 84 chest X-ray images of COVID-19 patients from publicly available datasets and performed data augmentation on COVID-19 patient data samples (420 samples after augmentation). Their experiments sampled the input images as 224*224 and were trained using a cross-entropy loss function and an ADAM optimiser with a learning rate of 0.0001. The pre-trained model with the highest accuracy obtained was SqueezeNet, which achieved 99.2%. Narin, Kaya [7] trained more pre-trained models based on transfer learning and tested models including ResNet50, ResNet101, ResNet152, InceptionV3, and InceptionResNetV2. They used a more balanced dataset of 100 chest X-ray images, including 50 COVID-19 patients and 50 healthy people's chest X-ray images. Due to the small size of the dataset, a 5-fold cross-validation method was used to reduce the negative impact of the dataset size limitation. In this experiment, the pre-trained model with the highest accuracy was ResNet101 with 96.1 ± 2.6.

The experimental results of Shamsi, Asgharnezhad [8] showed that CT images have richer features than X-ray images for better prediction results in AI-based COVID-19

diagnosis. Their experiments used 100 chest X-ray images (25 COVID-19 patient images and 75 healthy human images) and 746 chest CT images (349 COVID-19 patient images and 397 healthy human images) to predict the results of ResNet50, DenseNet-121, InceptionResNetV2, and VGG16 pre-trained models were trained based on transfer learning. It is worth mentioning that eight classifiers were tested separately for each model to ensure the fairness of the results. In the end, the highest accuracy was 98.6% (ResNet50 + Linear SVM) based on X-rays and 87.9% (ResNet50 + Linear SVM) based on CT images. Loey, Manogaran [9]'s research used 345 CT images from COVID-19 patients and 397 from healthy individuals, in a total of 742 chest CT images, to train five pre-trained models (ResNet50, AlexNet, VGG16, VGG19 and GoogleNet) based on transfer learning. They used data augmentation techniques to prevent possible overfitting problem and the Conditional Generative Adversarial Network (CGAN) to improve the classification performance. Also using CT images as a training dataset was a study by Ahuja, Panigrahi [10], who used 746 chest CT images (349 from COVID-19 patients and 397 from healthy individuals) for ResNet18, ResNet50, ResNet51 and Squeeze. ResNet101 and SqueezeNet, four pre-trained models were trained based on transfer learning. The experiments also used data augmentation techniques and ultimately achieved excellent performance, with 99.4% accuracy from the ResNet18 model. Table 1 and Table 2 shows the detail performance comparison of pre-trained models in above studies.

Table 1. Summary of pre-trained models on COVID-19 diagnosis based on chest X-ray datasets

Authors	Data size	Models	Acc (%)	Sen (%)	Spc (%)
Minaee, Kafieh [6]	Total = 2420 (Covid-19 = 420 Non-Covid = 2000)	SqueezeNet	99.2	98 ± 2.7	92.9 ± 0.9
		ResNet50	99	98 ± 2.7	89.6 ± 1.1
		ResNet18	98.9	98 ± 2.7	90.7 ± 1.1
		DenseNet-121	97.6	98 ± 2.7	75.1 ± 1.5
Shamsi, Asgharnezhad [8]	Total = 100 (Covid-19 = 25 Non-Covid = 75)	ResNet50	98.6 ± 2.1	99.9 ± 1.2	98.2 ± 2.8
		InceptionResNetV2	98 ± 3.2	96.3 ± 7.8	98.5 ± 3.5
		VGG16	96.6 ± 3.4	98.8 ± 9.9	98.3 ± 3.1
		DenseNet-121	96.4 ± 3.1	93.9 ± 9.3	97.2 ± 3.7
Narin, Kaya [7]	Total = 100 (Covid-19 = 50 Non-Covid = 50)	ResNet101	96.1 ± 2.6	74.3 ± 25.7	98.2 ± 1.6
		InceptionV3	95.9 ± 3.7	90.5 ± 6.7	96.4 ± 3.6
		ResNet50	95.1 ± 3.4	92.7 ± 5.9	95.3 ± 4
		ResNet152	93.4 ± 3.9	57.45 ± 35.4	96.8 ± 2.5
		InceptionResNetV2	93.5 ± 3.7	81.7 ± 11.1	95.5 ± 3.4

(Acc = Accuracy; Sen = Sensitivity; Spc = Specificity).

Table 2. Summary of pre-trained models on COVID-19 diagnosis based on chest CT datasets

Authors	Data size	Models	Acc (%)	Sen (%)	Spc (%)
Ahuja, Panigrahi [10]	Total = 746 Covid-19 = 349 Non-Covid = 397	ResNet18	99.4	100	98.6
		ResNet50	98.8	100	97.2
		ResNet101	97	97.9	95.8
		SqueezeNet	95.2	95.8	94.4
Shamsi, Asgharnezhad [8]	Total = 746 (Covid-19 = 349 Non-Covid = 397)	ResNet50	87.9 ± 5.8	86.5 ± 7.1	89.1 ± 5.4
		VGG16	86.5 ± 5.8	84.8 ± 8.2	88.1 ± 5.9
		DenseNet-121	85.9 ± 5.9	84.9 ± 8.4	86.8 ± 6.3
		InceptionResNetV2	84.3 ± 7.3	83.2 ± 9	91.9 ± 7.4
Loey, Manogaran [9]	Total = 742 (Covid-19 = 345 Non-Covid = 397)	ResNet50	82.64	80.85	91.4
		VGG16	78.05	75.53	93.3
		GoogleNet	77.07	75.53	82.9
		VGG19	76.54	88.3	91.4
		AlexNet	75.73	88.3	87.6

(Acc = Accuracy; Sen = Sensitivity; Spc = Specificity).

3   Custom Deep Learning Models

A custom deep learning model is a deep learning model specifically tailored to a particular application domain, often based on the evolution or adaptation of a deep learning model, the fusion of multiple deep learning models, or the combination of deep learning techniques with other artificial intelligence techniques. Because they are specifically designed for a particular domain, custom deep learning models can often achieve more specialised functionality and more accurate performance. However, without parameters inherited from pre-trained models, custom deep learning models must be trained from scratch, which implies higher computational and time costs than migration learning. This paper summarised and evaluated several custom deep learning models for Covid-19, which is as below.

Islam, Islam [11] proposed a CNN and Long Short-Term Memory (LSTM) 20-layer hybrid network (CNN-LSTM) for COVID-19 diagnosis. The network consists of 12 two-dimensional convolutional layers with a maximum pooling layer after every two or three convolutional layers, a fully connected layer connecting an LSTM network layer to the output layer at the end of the network to sort and classify the X-ray images after the analysed temporal features. The CNN-LSTM model was trained using a balanced dataset (1525 chest X-ray images from COVID-19 patients, 1525 chest X-ray images from pneumonia patients and 1525 chest X-ray images of the healthy population) and achieved reasonably good results (accuracy of 99.2%). The model proposed by Li, Zhang [12], COVID-GATNet, combined DenseNet and Graph Attention Networks (GAT). The experiment used a large dataset (containing 399 chest X-ray images from COVID-19 patients, 7399 chest X-ray images from pneumonia patients, and 10,192 chest X-ray images from the healthy population) to train and test the model. Due to the small number

of chest X-ray images from COVID-19 patients in the dataset, the COVID-19 class of the dataset was expanded using geometric transformation operations, and the expanded COVID-19 class consisted of 1197 X-ray images. The experiment achieved relatively good results, with an accuracy of 94.3%. Toraman, Alakus [13] presents a Convolutional CapsNet (Conv-CapsNet) for binary classification tasks and multiclassification tasks of COVID-19 chest X-ray images. The experiment also used data augmentation techniques to expand the dataset with 231 images from COVID-19 patients, resulting in 1050 images. The model achieved an accuracy of 97.23% $\pm$ 0.97%. Hussain, Hasan [14] used 2100 chest X-ray images (500 from COVID-19 patients, 800 from pneumonia patients and 800 from the healthy population) for a 22-layer CNN consisting of convolutional, pooling, dense, flatten and Leaky ReLU layers. A final accuracy of 99.1% $\pm$ 0.5% was achieved in the binary classification task of COVID-19 (COVID-19, Normal). Mukherjee, Ghosh [15] proposed a shallow CNN for COVID-19 diagnosis, which is only a 6-layer network consisting of the input layer, convolution layer, max-pooling layer, two dense layers, and output layer. After training with 321 chest X-ray images from COVID-19 patients and 321 chest X-ray images from people not infected with COVID-19, an accuracy of 99.69 was achieved in the binary classification task.

Many researchers have also proposed custom deep learning models for COVID-19 chest CT images classification. The method proposed by ELGhamrawy [16] uses a whale optimisation algorithm on top of CNN to select the most relevant features and tests three different classifiers, SVM, Naïve Bayes and Discriminant Analysis. After training with a dataset containing 583 chest CT images (432 from COVID-19 patients and 151 from viral pneumonia patients), an accuracy of 97.14% (SVM) was achieved. Amyar, Modzelewski [17] proposed an automatic classification segmentation method for classifying COVID-19 chest CT images. The framework consists of an encoder, two decoders and a multilayer perceptron (MLP), where the encoder is for reconstruction, the decoder for segmentation and the MLP for classification. The model was trained and tested using a dataset of 449 chest CT images from COVID-19 and 920 chest CT images from people without COVID-19 and achieved an accuracy of 94.67%. Yao and Han [18] presents a model using wavelet entropy for feature extraction, Feedforward Neural Network as the classification model and Biogeography-Based Optimisation as the optimisation algorithm. A balanced dataset consisting of 66 data from COVID-19 patients and 66 data from an uninfected COVID-19 population was used for training and testing. A final accuracy of 73.95 $\pm$ 0.98 was achieved. Chen [19] proposed a method with histogram equalisation (HE), grey scale co-occurrence matrix (GLCM), and SVM algorithm to identify COVID-19 chest CT images, and achieved 75.03% $\pm$ 1.12% accuracy on dataset that includes 296 chest CT images (148 from healthy people, 148 from COVID-19 patients). The Table 3 shows the detail performance comparison of all above models.

Table 3. Summary of custom deep learning methods for COVID-19 diagnosis

Model	Data type	Data size	Acc	Sen	Spc	Pre
Shallow CNN [15]	Chest X-ray Images	Total = 642 (Covid-19 = 321 Non-Covid = 321)	99.69	1.00	99.38	99.38
CNN_LSTM [11]	Chest X-ray Images	Total = 4575 (Covid-19 = 1525 Pneumonia = 1525 Non-Covid = 1525)	99.20	99.30	99.20	–
CoroDet (22-layer CNN) [14]	Chest X-ray Images	Total = 1300 (Covid-19 = 500 Normal = 800)	99.10 ± 0.5	94.20 ± 3.00	95.90 ± 3.70	97.70 ± 1.50
Conv-Capsnet [13]	Chest X-ray Images	Total = 2331 (Covid-19 = 231 Pneumonia = 1050 Normal = 1050)	97.23 ± 0.97	97.42 ± 1.81	97.04 ± 1.50	97.08 ± 1.42
WOA-CNN [16]	Chest CT Images	Total = 583 (Covid-19 = 432 Pneumonias = 151)	97.14	96.27	97.60	95.47
Encoder-Decoder with MLP [17]	Chest CT Images	Total = 1369 (Covid-19 = 449 Non-Covid = 920)	94.67	96.00	92.00	–
COVID-GATNet [12]	Chest X-ray Images	Total = 17990 (Covid-19 = 399 Pneumonia = 7399 Normal = 10192)	94.30	91.90	–	98.90
GLCM-SVM [19]	Chest CT Images	Total = 296 (Covid-19 = 148 Normal = 148)	75.03 ± 1.12	72.03 ± 2.94	78.04 ± 1.72	76.66 ± 1.07
WE-BBO [18]	Chest CT Images	Total = 132 (Covid-19 = 66 Non-Covid = 66)	73.95 ± 0.98	72.97 ± 2.96	74.93 ± 2.39	74.48 ± 1.34

(Acc = Accuracy; Sen = Sensitivity; Spc = Specificity; Pre = Precision).

4   Discussion

4.1   Open Discussion

This paper summarises 15 studies on building and testing artificial intelligence-based COVID-19 diagnostic systems. Five of these studies tested the performance of pre-trained models in COVID-19 diagnostics, and these studies covered 13 commonly used pre-trained models. The other nine studies discussed in this paper constructed and tested different custom deep learning models. The performance of the different models included in all studies is presented in Table 1, Table 2, and Table 3 with the selected performance

metrics being accuracy, sensitivity, specificity, and precision (the section of pre-trained models does not include precision.). Due to the differences in the datasets used in the different studies, it is impossible to make a valid comparison between the different models. However, in the pre-trained model section, ResNet50 was tested in all five studies about pre-trained models so that it is possible to compare its performance with other pre-trained models. ResNet50 has achieved the highest accuracy in [8, 9], second accuracy in [6, 10] also achieved the third accuracy of all five models tested in [7]. It shows that ResNet50 has stable performance in the diagnostic task of COVID-19. Due to the different number of datasets used in the different studies, it is not possible to provide a valid comparison regarding the performance of the chest CT image dataset and the chest X-ray image dataset in the COVID-19 task. However, in [8], both datasets were used to test different pre-trained models, and as seen in the results of that study, despite the fact that the amount of chest X-ray image dataset used in the study was much smaller than the chest CT image dataset, the results based on the chest X-ray image dataset still generally outperformed the chest CT image dataset.

4.2   Challenges and Future Trends

Although deep learning techniques based on chest CT images and chest X-ray images have achieved promising performance in detecting COVID-19, there are still many challenges before the technology can be widely used in the medical field. (i) Data shortage. Deep learning techniques often rely on a large number of data to achieve robust performance. However, as COVID-19 is an emerging disease, there is still a shortage of standard data. Almost all studies of COVID-19 diagnostic systems based on deep learning techniques have used multi-source patchwork datasets from the Internet, which lack a uniform standard for quality assurance and pose a significant challenge for different performance comparisons. Researchers need to design and develop more optimised algorithms to address the negative impact of performance from smaller datasets to overcome this challenge. Also, establishing a unified high-quality dataset is extremely important for the research of deep learning-based COVID-19 diagnostic systems. (ii) Insufficient computational resources. In addition, medical imaging technologies often result in images with high resolution, and analysis of such images requires extremely high computational costs. Most researchers resize image data before training the model, which causes some data loss and image model performance. As technology has evolved, corresponding solutions have emerged, such as cloud computing technologies that can provide higher computational power. Also, more new technologies are emerging, and hardware devices are upgrading quickly. The negative images caused by the lack of computing resources will slowly decrease. Researchers can experiment with newer technologies and hardware to analyse high-resolution medical images to achieve higher performance. (ii) Lower social awareness. RT-PCR as a PCR-based method is recognised and accepted worldwide as a well-established technique. In contrast, most people in the world are not aware of deep learning-based automated detection systems and find it difficult to build sufficient trust in them. Educational outreach about deep learning technology is critical before popularising the COVID-19 automated diagnostic system based on deep learning technology.

5   Limitations of the Study

The purpose of this paper is to provide a brief summary of the currently available AI-based diagnostic systems for COVID-19. Although the different techniques used and the resulting performance of several studies are presented, several shortcomings need to be addressed in future work. Firstly, the number of studies covered in this paper is too small to provide a simple comparison and does not allow valid conclusions to be drawn. Secondly, this work has not been replicated for all models, and it is not possible to compare the performance of individual models on the same dataset. Therefore, the model performance discussed in the paper is only informative and does not give an accurate picture of the performance strengths and weaknesses of the different models. Third, to unify the performance metrics of different studies, only accuracy, sensitivity, specificity, and precision are used to compare model performance, but this cannot provide a comprehensive assessment of model performance. Fourthly, only a brief description of the models is summarised, but not an in-depth and detailed description of the techniques used in the models. For more details on the different models, the reader should consult the relevant references.

6   Conclusions

COVID-19 continues to spread rapidly worldwide, setting new records for the cumulative number of diagnoses and deaths, posing a vital threat to the global economy and human lives. Deep learning-based COVID-19 diagnostic systems can automate the analysis of chest CT images and chest X-ray images and have achieved excellent performance, promising to be an essential weapon in the fight against COVID-19. However, no current studies indicate that COVID-19 diagnostic systems based on deep learning techniques can be used as authoritative COVID-19 diagnostic tools. This paper provides a brief summary and evaluation of COVID-19 diagnostic systems using different deep learning techniques and discusses current challenges and future trends. It is hoped that soon, with the concerted efforts of the community, medical professionals and researchers, accurate and efficient COVID-19 diagnostic systems based on deep learning techniques that are recognised in various fields will be available to control the spread of COVID-19 and reduce its threat.

References

1. Yuki, K., Fujiogi, M., Koutsogiannaki, S.: COVID-19 pathophysiology: a review. Clin. Immunol. 108427 (2020)
2. Chaudhuri, S., et al.: Modeling the role of respiratory droplets in Covid-19 type pandemics. Phys. Fluids **32**(6), 063309 (2020)
3. Fang, Y., et al.: Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology **296**(2), E115–E117 (2020)
4. Long, C., et al.: Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT? Eur. J. Radiol. **126**, 108961 (2020)
5. Wang, X., et al.: A comprehensive survey on convolutional neural network in medical image analysis. Multimed. Tools Appl. 1–45.https://doi.org/10.1007/s11042-020-09634-7

6. Minaee, S., et al.: Deep-COVID: predicting COVID-19 from chest X-ray images using deep transfer learning. Med. Image Anal. **65**, 101794 (2020)

7. Narin, A., Kaya, C., Pamuk, Z.: Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. arXiv preprint arXiv:2003. 10849 (2020)

8. Shamsi, A., et al.: An uncertainty-aware transfer learning-based framework for COVID-19 diagnosis. IEEE Trans. Neural Netw. Learn. Syst. (2021)

9. Loey, M., Manogaran, G., Khalifa, N.E.M.: A deep transfer learning model with classical data augmentation and CGAN to detect COVID-19 from chest CT radiography digital images. Neural Comput. Appl. 1–13 (2020)

10. Ahuja, S., Panigrahi, B.K., Dey, N., Rajinikanth, V., Gandhi, T.K.: Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices. Appl. Intell. **51**(1), 571–585 (2020). https://doi.org/10.1007/s10489-020-01826-w

11. Islam, M.Z., Islam, M.M., Asraf, A.: A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. Inform. Med. Unlocked **20**, 100412 (2020)

12. Li, J., et al. COVID-GATNet: A Deep Learning Framework for Screening of COVID-19 from Chest X-Ray Images. in 2020 IEEE 6th International Conference on Computer and Communications (ICCC). 2020. IEEE

13. Toraman, S., Alakus, T.B., Turkoglu, I.: Convolutional CapsNet: a novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks. Chaos Solitons Fractals **140**, 110122 (2020)

14. Hussain, E., et al.: CoroDet: a deep learning based classification for COVID-19 detection using chest X-ray images. Chaos Solitons Fractals **142**, 110495 (2021)

15. Mukherjee, H., et al.: Shallow convolutional neural network for COVID-19 outbreak screening using chest X-rays. Cogn. Comput. 1–14 (2021)

16. ELGhamrawy, S.M.: Diagnosis and prediction model for COVID19 Patients response to treatment based on convolutional neural networks and whale optimization algorithm using CT images. MedRxiv (2020)

17. Amyar, A., et al.: Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: classification and segmentation. Comput. Biol. Med. **126**, 104037 (2020)

18. Yao, X., Han, J.: COVID-19 detection via wavelet entropy and biogeography-based optimization. In: Santosh, K.C., Joshi, A. (eds.) COVID-19: Prediction, Decision-Making, and its Impacts. LNDECT, vol. 60, pp. 69–76. Springer, Singapore (2021). https://doi.org/10. 1007/978-981-15-9682-7_8

19. Chen, Y.: Covid-19 classification based on gray-level co-occurrence matrix and support vector machine. In: Santosh, K.C., Joshi, A. (eds.) COVID-19: Prediction, Decision-Making, and its Impacts. LNDECT, vol. 60, pp. 47–55. Springer, Singapore (2021). https://doi.org/10.1007/ 978-981-15-9682-7_6

IoT and AI Technology Used for COVID-19 Pandemic Control

Shu-Wen Chen[1,2]([✉]) and Xiao-Wei Gu[1]

[1] School of Math and Information Technology, Jiangsu Second Normal University, Nanjing 211200, China
chenshuwen@126.com
[2] State Key Laboratory of Millimeter Waves, Southeast University, Nanjing 210096, China

Abstract. The pandemic of COVID-19 is going on spreading in 2021, which has infected at least 170 million of people around the world. The healthcare systems are overwhelmed due to the virus infection. Luckily, Internet of Things (IoT) is one of the most effective paradigms in the smart world, in which artificial intelligence technology, like cloud computing and big data analysis, is playing a vital role in epidemic prevention and blocking COVID-19 spreading. For example, in terms of remote screening and diagnosis of the COVID-19 patients, AI technology based on machine learning and deep learning has significantly upgraded recently medical equipment and reshapes the workflow with minimal contact to patients, therefore medical specialists can make clinical decisions more efficiently, providing the best protection not only to patients but also specialists themselves. This paper hereby reviews the latest progress of IoT systems combined AI against COVID-19, and it also provide comprehensive detail on how to overcome the epidemic challenges along with directions towards the possible technology trends for future work.

Keywords: Internet of Things · COVID-19 · Epidemic prevention

1 Introduction

Coronavirus disease 2019 was officially named 'COVID-19' by the World Health Organization in February 2021. Since the first confirmed case of COVID-19, researchers in more than 30 countries and regions around the world have been actively looking for ways to control and treat COVID-19. As the latest and hottest technology in the 21st century, the Internet of Things system can realize data information, remote control and intelligent management and monitoring through real-time network. Therefore, it is of great significance to apply IoT technology to clinical medicine, especially in this context, to apply IoT technology to COVID-19 epidemic prevention and control work.

The use of IoT prevention and control systems to control COVID-19 is not only for patients, but also for the general population. People can use wearable devices to self-monitor, observe and record their respiration rate, heart rate, daily temperature and other physiological values, so as to make self-judgment. Even in isolation, they can quickly detect changes in vital signs. More noteworthy is that the application of the Internet of

Things in clinical medicine can promote the development of modern intelligent medical care, especially in remote screening, intelligent diagnosis and other aspects. Remote screening eliminates the need to screen large numbers of people in the emergency room, thereby reducing the possibility of exposure. Intelligent diagnosis solves the problem of slow speed and low accuracy of manual reading image scan report. Smart diagnostics serve as an auxiliary tool to help frontline physicians quickly determine if a patient is infected with COVID-19, so that patients can be isolated for further treatment in the first place. The application of Internet of Things technology in epidemic prevention and control not only avoids person-to-person contact and reduces the possibility of cross-infection, but also achieves precise treatment and prevention and control.

The second part below introduces the new epidemic prevention and control methods: wearable devices with Wi-Fi, 5G and Bluetooth technology as the main technologies; The third part introduces the method of remote screening. The fourth part discusses the methods of intelligent diagnosis, and the last part summarizes and looks into the future of medical technology in the Internet of Things. With the help of artificial intelligence, these new IoT technologies will drive the development of modern epidemic prevention and control. This review mainly discusses the advantages and disadvantages of different methods in remote screening and intelligent diagnosis based on COVID-19 related studies before April 31, 2021, hoping to provide some help and guidance for doctors and researchers in overcoming COVID-19.

2   New Methods for COVID-19 Prevention and Control - Wearable Devices

Wearable technology, first developed at a MIT lab in the 1960s, embedding technologies such as multimedia, sensors and wireless communications into people's clothing. Novel coronavirus wreaked havoc worldwide, wearable devices based on IoT systems are used to measure COVID-19-related signs, such as respiratory rate and body temperature. M. N. Mohammed et al. [1] propose the use of a smart helmet with a mounted thermal imaging system to automatically detect coronavirus through the thermal imaging system, thereby reducing the contact between people. There will be a positioning system in the smart helmet. When a temperature higher than normal is detected, the system responds, and the location system module immediately marks and determines the geographic location, while sending a notification to the designated smartphone via GSM. That way, health care workers can get timely data on people's temperature. However, as the second-generation mobile communication system, the Global System for Mobile Communications (GSM) has fallen behind the new era of 5G technology. Barbara Fyntanidou's team [2] has developed a wrist-worn wearable device that uses advanced digital signal processing algorithms to continuously extract heart rate, oxygen saturation, body temperature estimates, and more. Like smart helmets, wrist-worn wearables can transmit processed data to an app or designated third-party mobile port in a timely manner through mobile communication protocols. Muhammad Usman Ashraf et al. [3] propose an intelligent limbic system, as shown in Fig. 1, which is based on wearable devices to detect those at risk of infection. The system makes use of a wearable module equipped with infrared and pulse sensors to calculate body temperature and pulse rate in real time,

and an unwearable module to monitor respiratory and blood pressure data of suspected patients by placing them in crowded places such as airports and large shopping malls. The two modules work with each other and alert each other when a suspected case occurs in any public place. But it's hard to guarantee that the unwearable modules are installed in the public places suspected patients have visited. Meanwhile, the placement of modules also brings some security risks. Luca Lonini et al. [4] use a soft wearable sensor device attached to the body, placed in the sternal incision, to receive respiration, heart rate and other data. Wearable sensors provide a novel way to detect COVID-19, which depends on recording changes in heart rate, respiration, cough, and body temperature over a long period of time to determine the coronavirus infection. These technologies, as well as data networks, mainly rely on Wi-Fi, 5G, Bluetooth technology, but only relying on such a network for continuous monitoring may lead to low accuracy and data loss problems.
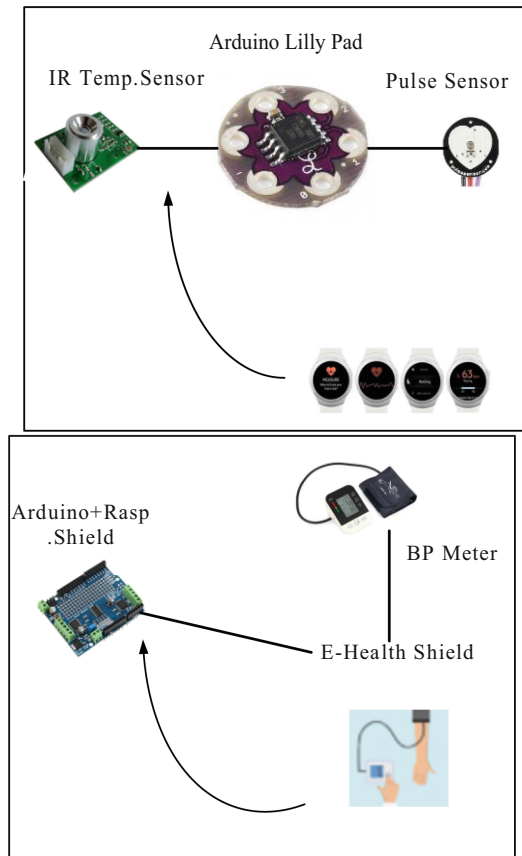
Fig. 1. Wearable device and non-wearable device module device. Wearable gadget on the left and Non-wearable gadget on the right.

Wearable devices are widely used on COVID-19 to measure the health status of potential infected people and self-detect physiological changes during isolation. Wearable devices use GPS data to track location, so doctors can easily track a patient's condition, for example, the intelligent helmet proposed by M. N. Mohammed et al. [1], which uses the intelligent helmet with a mounted thermal imaging system to automatically detect coronavirus from thermal images without the need for human intervention. This saves time and reduces human interactions that could cause the coronavirus to spread more quickly. Wearable devices such as smartwatches can also be accustomed to collect self-reported symptom-tracking data to distinguish between negative and positive cases among infected people. Wrist-worn devices could be used specifically for emergency situations, such as prioritizing and classifying emergency patients. The 3D wearable prototype design proposed by Nizar Al Bassam et al. [5] includes wearable body sensors, network application programming interface layer and mobile frontend layer, which is also the use of a portable wearable device to build an automated healthcare system. One of the advantages of the healthcare system is that COVID-19 can be detected at an early stage. Figure 2 is the architecture of the system. The wearable sensor layer is accustomed to measure temperature, heartbeat, oxygen saturation and cough technology. Compared with the intelligent limbic system of Muhammad Usman Ashraf et al. [6], this system improves the GPS location data of patients to medical units in real time, which makes it more convenient and effective to monitor patients and suspected patients, and is more suitable for responding to emergency situations. When the patient violates the self-isolation rules, the device will immediately send an alarm and notification signal. In the future, wearable devices based on the Internet of Things technology will provide various methods to identify, perceive and monitor coronavirus, which has great potential and hope in the healthcare system.

3   Remote Screening Based on IoT

In the context of the COVID-19 pandemic, timely screening infected people as soon as possible is an important measure to prevent and control the epidemic. Obviously, manual screening is very slow, and remote screening technology, can improve the speed and efficiency of screening. Therefore, remote screening based on the Internet of Things system must be developed. Timo Schinkothe team [8] has built a free, IoT-based caregiver cockpit (C19CC), one user is first connected to a healthcare worker, who quickly categorizes the user according to the user's description and color-codes the severity of the examiner. For example, users with red codes are automatically screened and their personal information and recent range of activity are marked. Similarly, patients also have multiple ways to contact the doctor, such as scanning the doctor's QR code. C19CC can not only be used in remote screening, but also can be widely used in remote monitoring, hospital wards, etc. C19CC prescreens patients in a contact-free manner and quickly learns about those who urgently need treatment. Because COVID-19 patients often have fever, in most cases the facial recognition system will also be used to detect the patient and alert the hospital platform via the Internet or mobile devices so that the patient can be isolated and further diagnosed. Weijun Tan et al. [9] propose a basic method based on facial recognition. The main function of this method is to remotely screen out suspicious

patients through thermal images, and then retrieve people who have close contact with them through facial recognition system, so as to isolate people in time and prevent the spread of the virus. Although facial recognition can be used for initial remote screening, it is extremely inefficient compared to other methods [29]. The portable scanner proposed by Dennis Hou et al. [10] will input the collected lung ultrasonic images to the platform. After the data is processed by the platform, it will classify the data in the quilt-space network. This sequence of steps allows the screening of patients with and without COVID-19. The classifier could be used on a large scale in nursing homes with adequate budgets, completely avoiding the possibility of infection in elderly people who go to hospital for screening. Cristian Chilipirea et al. [12] design a lightweight Web-based platform that could run on a low-end server infrastructure. The platform consists of a screening application, a data collection module, a machine learning module and a user notification module. The filter application process is designed for smartphone applications and Web pages. This can meet the need for simultaneous screening of multiple people. The data collection module is also available to a large number of simultaneous users. Machine learning modules are used to answer questions about the risk of possible infection. The user notification module is used to send specific messages on demand to inform the user of the outcome, and for those with COVID-19 symptoms, the notification module will inform them of the progression of the disease. However, whether there are malicious attacks on the platform and user information stolen, and how to prevent the abuse of the platform, this is an urgent problem to be solved. Table 1 summarizes the remote screening methods or examples mentioned in this article.
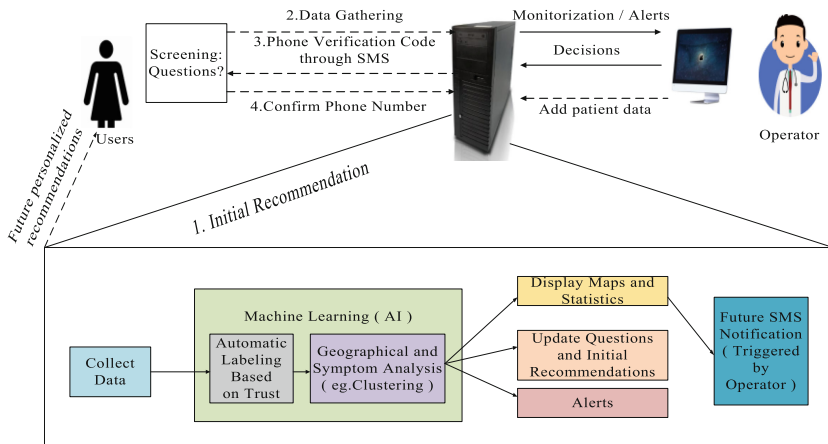
Fig. 2. Shows the architecture of the platform.

Table 1. The methods and results of remote screening mentioned.

Authors	Methods	Results	Features
Timo Schinköthe et al. [8]	The cockpit-C19CC	Identifying telemedicine options can quickly improve patient care	Improve care and safety for people living with COVID-19
Weijun Tan et al. [9]	Facial recognition system using thermal imaging	Ability to identify and track patients	To help control the spread of COVID-19
Dennis Hou et al. [10]	Portable scanner + Subspace multilayer networks	More than 96% of test data accuracy can be obtained	Improve energy efficiency
Cristian Chilipirea et al. [12]	Scalable COVID-19 screening platform	Data can be collected and analyzed for more than 200,000 people per minute	The main feature is the use of distributed, lightweight design

4   Intelligent Diagnosis of COVID-19

It is well known that X-ray and CT scans are the two standard methods for diagnosing COVID-19. But in the context of the outbreak of novel coronavirus, there are signs of cross-infection between actual CT and X-rays, which can put suspected patients and doctors at risk. At the same time, only reading a large number of image scans and manually drawing the outline of lung lesions will lead to the delay of COVID-19 diagnosis [30, 31]. Therefore, it is of great significance to develop an intelligent diagnostic system based on the Internet of Things system, which can assist front-line doctors and jointly fight against COVID-19.

Common chest CT findings of COVID-19 patients include ground glass opacity [11], consolidation, pleural effusion, etc. Figure 3 shows CT images of patients with COVID-19 at different periods and normal subjects. As shown in the figure, the CT manifestations of the lungs are often patchy or diffuse ground glass shadows in the early stage. In the progressive stage, the two lung diseases progress rapidly, multiple lesions fuse into large sheet consolidation, and the lesion density increases [13]. In the absorption phase, the lesion range was slightly reduced, the density was reduced, and the fibrous cord shadow was visible. X-ray images of patients with COVID-19 are typically characterized by large and blurred lungs, as shown in Fig. 4. It may be accompanied by thickening of the cleft system at night and a small amount of pleural effusion. When the disease is more severe, diffuse consolidation shadows can be seen in both lungs, with white lungs and pleural effusion.
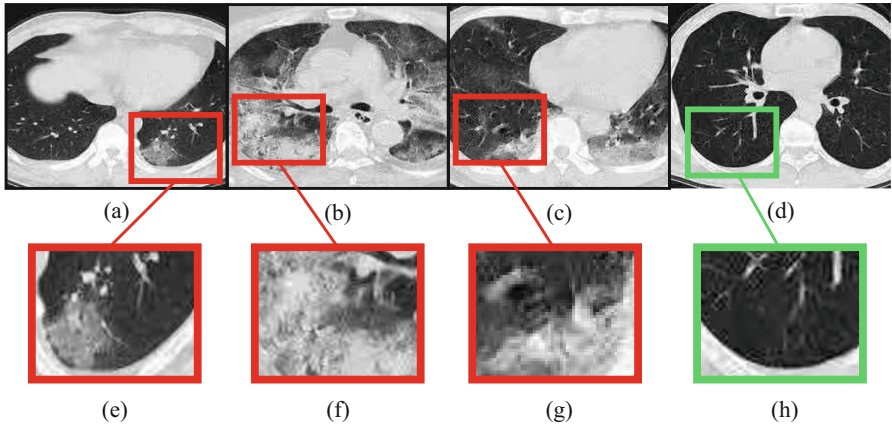
Fig. 3. Chest CT images of laboratory-confirmed COVID-19 patient and 1 healthy subject. Patients' GGOs are highlighted in red borders and normal subjects' GGOs are highlighted in green borders. (a), (b) and (c) lung CT images of COVID-19 patients in the early, advanced and absorptive stages, respectively; (e), (f) and (g) are the parts of GGO in these three stages respectively. (d) is a CT scan of a normal lung, and (h) is a local magnification of a normal lung.
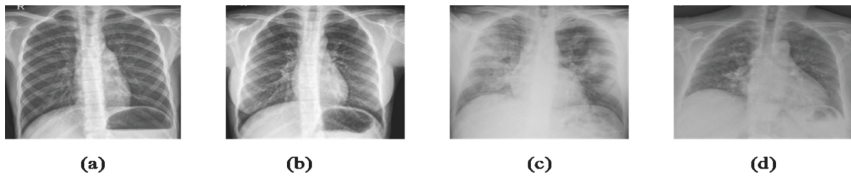

Fig. 4. (a) (b) is the CXR image of normal people, (c) (d) is the CXR image of COVID-19 patients [7, 14].

The premise of intelligent diagnosis by X-ray and CT scanning is image segmentation. The goal of segmentation is to separate the area or object of interest from the rest of the body for fixed-point measurement. Kai Gao et al. [15] develop a dual-branch combinatorial network (DCN) for combinatorial segmentation and classification. Ramin Ranjbarzadeh et al. [16] propose a dual-path convolutional neural network for detecting and classifying COVID-19 infections by extracting global and local features. Fan et al. [17] use the Deep Network for Pulmonary Infection Segmentation (INF-NET) to automatically segment infected tissue in sections. Naveen Paluru et al. [18] propose a lightweight CNN (ANAM-NET) based on deformation deep embedding for the segmentation of anomalies in COVID-19 chest CT images. All of these methods can automatically segment the scanned image. As lung segmentation technology becomes more and more mature, intelligent diagnostic COVID-19 will also become more reliable.

Qian Tang et al. [19] use neural network and digital image processing technology to design a lightweight classification model based on inter-section and intra-section attention mechanism, which integrates early screening, lesion assessment, lesion segmentation, and histogram of pixel distribution of lung and lesion. The average diagnosis

time per person in this algorithm model is only 0.4 s. Zhengfeng Jiang's team [20] trains a VGG-16 Convolutional Neural Network migration to build an intelligent COVID-19 diagnostic model, which is based on a small sample data set and uses CT scan images to distinguish which stage of COVID-19 patients are in early stage, advanced stage, and severe stage. However, Buyut Khoirul Umri et al. [21] express that CNN had more significant advantages compared with VGG-16, and they proposed to combine CNN and Clahe to detect novel coronavirus in X-ray images. Contrast-constrained adaptive histogram equalization and convolutional neural network were used to analyze the data sets. Juliana C. Gomes et al. [22] propose an intelligent system to support X-ray scan image diagnosis and developed IKONOS (a desktop application) to diagnose COVID-19 via X-ray image, this is shown in Fig. 5. After receiving the image, the doctor uploads the image to the app, which uses texture and shape descriptors and classical classifiers for feature extraction, which is analyzed by the intelligent system to identify COVID-19. Similarly, Ali Nevin et al. [23] also use the RESNET-50 model of Convolutional Neural Network (CNN) to carry out diagnostic research. With the help of the supervised learning method based on statistical learning theory (SVM algorithm), they could also directly extract features to determine whether there is a disease [32]. The sensitivity of their experiment is higher than that of Juliana C.Gomes et al. [22], which will make it easier for doctors to reduce the rate of missed tests. Krishna Kant Singh et al. [24] put forward an improved deep convolutional neural network for automatic diagnosis of COVID-19. The advantages brought by the combination of wavelet transform and deep network can be widely used to diagnose COVID-19 from chest X-ray images. The performance of this method is superior to the common methods, and it can be used for the early treatment and diagnosis of COVID-19 disease [28]. Sivaramakrishnan Rajaraman et al. [25] demonstrate how an iteratively built set of deep learning models could be used to intelligently diagnose COVID-19 via chest X-rays. Their combined use reduces the complexity of the model and the variance of the prediction values, thus promoting the adoption of digital chest radiograph detection COVID-19. Diego Hernandez et al. [26] introduce the trained deep learning model in advance, and then designed two models (RESNET-50 and VGG-16) to extend CNN. Although there were many mistakes in the actual prediction, this study also provided a new idea of intelligent diagnosis. Jun Wang et al. [27] propose a CT image scanning based model for diagnosing COVID-19. Specifically, it is a prioritized attention residual learning (PARL) module that training the 3DRESNET branch as a binary classifier for lung images. Its greatest advantage is that it can highlight the lesion area in the lung. With this advantage, this framework model can not only be widely used in the early intelligent diagnosis of COVID-19, but also can be applied to other computer-aided detection, such as glaucoma and skin lesions in retinal fundus images. The intelligent assisted diagnosis model proposed by Zhengfeng Jiang et al. [20], with its extremely high sensitivity and high reliability in small data based on learning transfer technology, can quickly provide doctors with reference information and improve efficiency during epidemic prevention and control period. However, this method still has the problem of insufficient sample size and needs to be expanded.
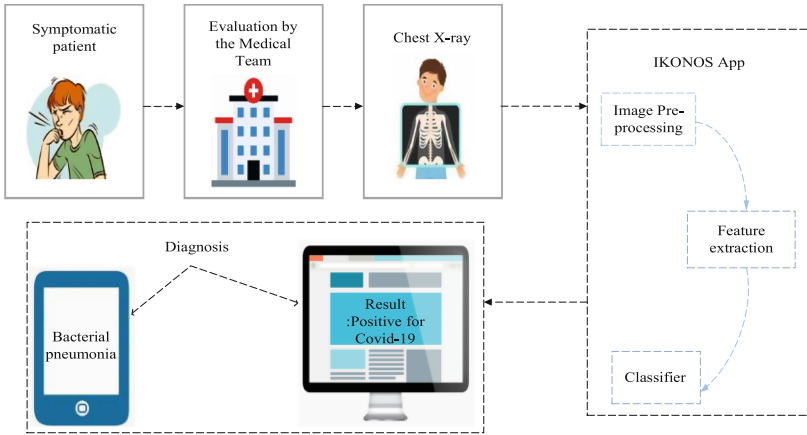
Fig. 5. Chest X-rays of symptomatic patients can be loaded into IKONOS. The application consists of an intelligent system capable of extracting features and classifying images. The results can be viewed on the computer on which the software is installed.

5   Conclusion and Future

Although COVID-19 has affected more than 30 countries and regions around the world, the good news is that the 21st century is an advanced scientific era. Researchers are working to make new breakthroughs in COVID-19 screening, diagnosis, and monitoring. Intelligent technology based on IoT systems has proven to be an extremely valuable resource, with applications ranging from wearable technology to COVID-19 remote screening and intelligent diagnostics. They are also discussed separately in this article. Wearable devices make it easy to measure an individual's health status and provide timely feedback to the hospital unit through a positioning system. Remote screening can avoid the mutual contact between medical personnel and screeners, improve screening efficiency, and realize early screening and early isolation. Smart diagnostics can quickly identify COVID-19 through X-ray and CT scans, reducing the workload of front-line doctors and enabling patients to receive treatment as quickly as possible. From then on, artificial intelligence image acquisition will enable patients to obtain high-quality images even at low radiation levels.

In the future, there will be more intelligent technologies based on the Internet of Things to fight human epidemics. For example, the use of cloud technology to achieve COVID-19 screening, diagnosis, monitoring and other integration. It is hoped that by integrating algorithms to predict risk events and take quick actions, the processing effect can be improved at the fastest and most efficient speed. Meanwhile, some intelligent systems of the Internet of Things can be built to automatically repair loopholes and problems.

References

1. Mohammedl, M.N., Hazairin, N.A., et al.: 2019 novel coronavirus disease (Covid-19): detection and diagnosis system using IoT based smart glasses. Int. J. Psychosoc. Rehabilit. **24**(7), 2296–2303 (2020)

2. Fyntanidou, B., Zouka, M., et al.: IoT-based smart triage of Covid-19 suspicious cases in the Emergency Department. In: 2020 IEEE Globecom Workshops, pp. 1–6. IEEE (2020)

3. Ashraf, M.U., Hannan, A., Cheema, S.M., et al.: Detection and tracking contagion using IoT-edge technologies: confronting COVID-19 pandemic. In: 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), pp. 1–6 (2020)

4. Lonini, L., Shawen, N., et al.: Rapid screening of physiological changes associated with COVID-19 using soft-wearables and structured activities: a pilot study. J. Transl. Eng. Health Med. **9**, 1–11 (2021)

5. Bassam, N.A., Asif, H.S., Qaraghuli, A., et al.: IoT based wearable device to monitor the signs of quarantine remote patients of COVID-19. Inf. Med. Unlocked **24**, 100588 (2021)

6. Ashraf, M.U., Hannan, A., Cheema, S.M., et al.: Detection and tracking contagion using IoT-edge technologies: confronting COVID-19 pandemic. In: 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), pp. 1–6. IEEE (2020)

7. Tabik, S., et al.: COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images. J. Biomed. Health Inf. **24**(12), 3595–3605 (2020)

8. Schinköthe, T., Gabri, M.R., et al.: A web- and app-based connected care solution for COVID-19 in- and outpatient care: qualitative study and application development. JMIR Publ. Health Surveill. **6**(2), e19033 (2020)

9. Tan, W., Liu, J., et al.: Application of face recognition in tracing COVID-19 fever patients and close contacts. In: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE (2020)

10. Hou, D., Hou, R., Hou, J., et al.: Interpretable Saab subspace network for COVID-19 lung ultrasound screening. In: 2020 11th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON). IEEE (2020)

11. Jiang, Y., Chen, H., et al.: COVID-19 CT image synthesis with a conditional generative adversarial network. IEEE J. Biomed. Health Inf. **25**(2), 441–452 (2020)

12. Chilipirea, C., Morogan, L., et al.: A scalable COVID-19 screening platform. In: 2020 IEEE Globecom Workshops, pp. 1–6. IEEE (2020)

13. Miao, Q., Ma, P., et al.: Pulmonary CT findings of COVID-19. J. Xuzhou Med. Univ. **201**, 306–309 (2020). In Chinese

14. Ashour, A.S., Eissa, M.M., Wahba, M.A., et al.: Ensemble-based bag of features for automated classification of normal and COVID-19 CXR images. Biomed. Signal Process. Control **68**, 102656 (2021)

15. Gao, K., Su, J., Jiang, Z., et al.: Dual-branch combination network (DCN): towards accurate diagnosis and lesion segmentation of COVID-19 using CT images. Med. Image Anal. **67**, 101836 (2021)

16. Ranjbarzadeh, R., Ghoushchi, S.J., et al.: Lung infection segmentation for COVID-19 pneumonia based on a cascade convolutional network from CT images. BioMed Res. Int. **16**, 5544742 (2021)

17. Fan, D.P., Zhou, T., et al.: Inf-Net: automatic COVID-19 lung infection segmentation from CT images. IEEE Trans. Med. Imag. **39**(8), 2626–2637 (2020)

18. Paluru, N., Dayal, A., et al.: Anam-Net: anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images. IEEE Trans. Neural Netw. Learn. Syst. **32**(3), 932–946 (2021)

19. Tang, Q., Du, B., et al.: Intelligent diagnosis system for COVID-19 CT images. Geomat. Inf. Sci. Wuhan Univ. **45**(06), 846–853 (2020)
20. Jiang, Z., Xu, X., et al.: Intelligent assisted diagnosis of COVID-19 based on CT images. J. Molecul. Imaging **43**(02), 264–269 (2021). In Chinese
21. Umri, B.K., Akhyari, M.W., Kusrini, K., et al.: Detection of Covid-19 in chest X-ray image using CLAHE and convolutional neural network. In: 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS), p. 9320806 (2020)
22. Gomes, J.C., Barbosa, V., Santana, M., et al.: IKONOS: an intelligent tool to support diagnosis of Covid-19 by texture analysis of x-ray images. In: Pattern Recognition, Metaheuristics and Machine Learning for Image Diagnosis (2020)
23. Narin, A., et al.: Detection of Covid-19 patients with convolutional neural network based features on multi-class X-ray chest images. In: 2020 Medical Technologies Congress (TIPTEKNO), vol. 9320806 (2020)
24. Singh, K.K., Singh, A., et al.: Diagnosis of COVID-19 from chest X-ray images using wavelets-based depthwise convolution network. Big Data Mining Analyt. **4**(2), 84–93 (2021)
25. Rajaraman, S., Siegelman, J., et al.: Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays. In: IEEE Access: Practical Innovations, Open Solutions, p. 1 (2020)
26. Hernandez, D., Pereira, R., Georgevia, P., et al.: COVID-19 detection through X-ray chest images. In: 2020 International Conferencatics and Informatics (Ie AutomCAI), p. 32206 (2020)
27. Wang, J., et al.: Prior-attention residual learning for more discriminative COVID-19 screening in CT images. IEEE Trans. Med. Imaging **39**(8), 2572–2583 (2020)
28. Xin, Z., et al.: Diagnosis of COVID-19 pneumonia via a novel deep learning architecture. J. Comput. Sci. Technol. (2021)
29. Zhang, Y., et al.: Introduction to the special issue on computer modelling of transmission, spread, control and diagnosis of COVID-19. Comput. Model. Eng. Sci. **127**(2), 385–387 (2021)
30. Wang, S.H., et al.: PSSPNN: PatchShuffle stochastic pooling neural network for an explainable diagnosis of COVID-19 with multiple-way data augmentation. In: Computational and Mathematical Methods in Medicine, p. 6633755 (2021)
31. Wang, S.H., et al.: AVNC: attention-based VGG-style network for COVID-19 diagnosis by CBAM. IEEE Sens. J. 3062442 (2021)
32. Zhang, Y., et al.: ANC: attention network for COVID-19 explainable diagnosis based on convolutional block attention module. Comput. Model. Eng. Sci. **127**(3), 1037–1058 (2021)

Review of Covid-19 Diagnosis Techniques Combined with Machine Learning and AI Analysis

Xiao-Wei Gu[1], Shu-Wen Chen[1,2(✉)], Xuan Tong[1], Hui-Shen Yan[3], Lu Chen[4], and Si-Ye Wu[5]

[1] School of Math and Information Technology, Jiangsu Second Normal University, Nanjing 211200, China
chenshuwen@126.com

[2] State Key Laboratory of Millimeter Waves, Southeast University, Nanjing 210096, China

[3] Department of Medical Science, Yangzhou Polytechnic College, Yangzhou 225000, China

[4] Department of Medical Imaging, Zhongda Hospital Affiliated to Southeast University, Nanjing 210096, China

[5] School of Acupuncture and Massage Health and Rehabilitation, Nanjing University of Chinese Medicine, Nanjing 210096, China

Abstract. The pandemic of coronavirus disease 2019 (COVID-19) is rapidly spreading all over the world. In order to reduce the workload of doctors, chest X-ray (CXR) and chest computed tomography (CT) scans are playing a major role in the detection and following-up of COVID-19 symptoms. Artificial intelligence (AI) technology based on machine learning and deep learning has significantly upgraded recently medical image screening tools, therefore, medical specialists can make clinical decisions more efficiently on COVID-19 infection cases, providing the best protection to patients as soon as possible. This paper tries to cover the latest progress of automated medical imaging diagnosis techniques involved with COVID-19, including image acquisition, segmentation, diagnosis, and follow-up. This paper focuses on the combination of X-ray, CT scan with AI, especially the deep-learning technique, all of which can be widely used in the frontline hospitals to fight against COVID-19.

Keywords: COVID-19 · Medical image · Chest CT · CXR · Deep learning

1 Introduction

Coronavirus disease 2019 (COVID-19), which was first reported in December 2019, has rapidly spread to more than 200 countries, becoming a serious threat to human lives. But now it goes on with some kinds of variation in 2021, and has already developed into a global health crisis. By May 13, 2021, more than 161 million cases of COVID-19

X.-W.G. and S.-W.C. — These authors contributed equally to this work. Xiao-Wei Gu and Shu-Wen Chen are co-first authors.

have been confirmed, including 3,300,000 deaths. The common symptoms of COVID-19 include fever, dry cough, and some respiratory problems. The control of COVID-19 largely depends on timely diagnosis. It is generally known that the reverse-transcription polymerase chain reaction (RT-PCR) test is the standard method for screening suspected cases [1]. However, the lab testing has some drawbacks. Firstly, the shortage of medical equipment and high demand of testing environments will limit the rapid screening of suspected cases. Secondly, the RT-PCR test is time-consuming because of laboratory processes, usually 24–48 h. Thirdly, comparing to RT-PCR test, chest X-ray and chest CT scan have a high sensitivity for COVID-19 infection [2], and in practice, their imaging equipment is more easily accessible and operated. Therefore, the listed advantages make medical imaging methods a necessary complement for early screening stage, providing a huge help to clinicians. The latest China's diagnosis and treatment protocol for COVID-19 (trial version 8) also highlights the value of imaging for detecting COVID-19. For example, in Wuhan, if the image features of X-ray or CT scan are observed [3], many suspected cases can be classified as suspected cases of COVID-19 as soon as possible, thereby receiving timely treatment. The suspected patients, though without clinical symptoms like coughing, fever, dyspnea and muscle aches, also need to be quarantined for further lab tests.

However, there is the phenomenon of cross-infection in the actual chest X-ray or CT scans. This is likely to cause COVID-9 to infect doctors who are not sick or patients with other diseases. Not only is there no way to save the patient, but there is also the risk of sacrifice. Meanwhile, due to the increment of confirmed and suspected cases of COVID-19, it is a labor-intensive task for radiologists to manually contour lung lesions and analyze a large number of scanning reports timely and accurately, which will lead to a delay in the diagnosis of COVID-19, miss the best time for treatment. To speed up diagnosis, it is important to develop fast-automatic segmentation and diagnosis method for COVID-19 combined with deep learning and AI technique.

Chest CT features of COVID-19 patients usually include ground glass opacity (GGO), consolidation, and other rare ones such as pleural or pericardial effusion [5], where GGO is a universal feature of all the chest CT findings. A study [6] points out identifying these common features is important for better identification of COVID-19.

However, CT imaging features of COVID-19 patients varied in different stages of the disease. CT images of COVID-19 patients with different disease stages and normal subjects are shown in Fig. 1. In the early stage, both lungs often exhibit patchy or diffuse GGO, with thickened small blood vessels in the lesions. Besides, there is nodular and patchy high-density opacity under the pleura of bilateral inferior lobes. In the progressive stage, the two lung diseases change quickly, multiple lesions fuse into huge sheets of consolidation, with the lesion density raising. In the absorption phase, the lesion area is lightly reduced and the density is slowly decreased [35]. However, due to the facts of age, autoimmunity of patients, the analysis and classification of COVID-19 medical image will also be influenced on a certain extent.
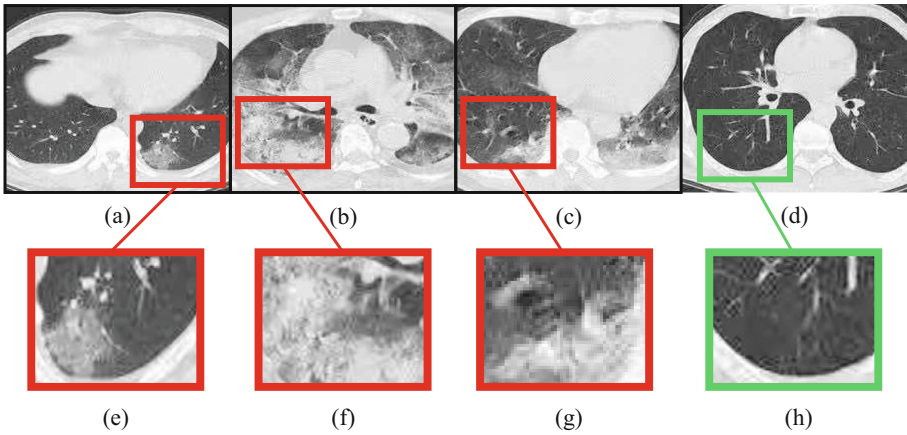
Fig. 1. [4] Chest CT images of laboratory-confirmed COVID-19 patient and one healthy subject. GGOS is highlighted with a red border in patients and compared with a green border in normal subjects. (a), (b) and (c) are lung CT images of COVID-19 patients in the early, progressive and absorption stages, respectively; (e), (f) and (g) are local magnification images of GGO in those three stages, respectively. (d) is a CT lung scan of a normal person, and (h) is the local magnification of a normal person. (Color figure online)
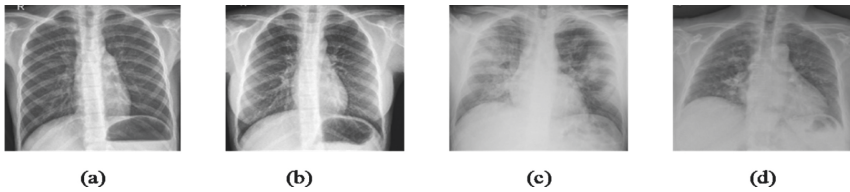
Fig. 2. [34] (a) (b) is the CXR image of normal people, (c) (d) is the CXR image of COVID-19 patients [10]. Images (c) (d) represent blurred lungs with a shadow of diffuse consolidation compared with those of normal people.

For the chest X-ray images of COVID-19, the representative characteristic is large blurred lungs, shown in Fig. 2, which probably be accompanied by thickening of cracks at night and a small amount of pleural effusion. When the patient is badly ill, there is a shadow of diffuse consolidation in both lungs, white lungs may appear, at times with pleural effusion.

Through reading of the COVID-19 image diagnosis literatures, people discover that major COVID-19 image diagnosis be based on chest CT images, with fewer references to CXR. This paper starts from the progress of medical image segmentation technique, including segmentation methods and their applications, then followed by a discussion of COVID-19 diagnosis methods by chest CT and X-ray imaging combined with machine learning and artificial intelligence techniques. The last part is the summary and looking forward to the future of medical imaging diagnosis. It is hoped that this review will provide some help and guidance for doctors and researchers to carry out COVID-19

treatment. This review is primarily based on medical imaging studies related to COVID-19 prior to March 31, 2021.

2   Medical Image Segmentation of COVID-19 and Its Applications

Segmentation refers to the division of an image into different regions according to the adjacent similar features. Image segmentation is complex and challenging in biomedical engineering tasks, which is influenced by many aspects, including noise, low contrast, illumination, and object irregularities. The goal of segmentation is to separate the area or object of interest from the rest of the body for quantitative measurement. Now segmentation methods have been divided into three categories: manual, semi-automatic and automatic. It takes time and effort for manual segmentation. Semi-automated segmentation is prevalent, which is normally integrated with open software. The automatic segmentation method does not require user intervention. The three methods have their strengths and weaknesses. Even now, the segmentation problem remains challenging. The paper [10] points out that although a large number of AI systems have been used to assist diagnosis in clinical practice, there is yet little work on the segmentation of infection in chest CT and X-ray scans.

In this section, Table 1 lists the latest existing medical image segmentation methods and summarizes some examples of COVID-19's application of image segmentation classification methods in this section, as well as the results obtained from each method.

Table 1. COVID-19 applies age segmentation classification methods and some examples specifically discussed in this section, along with the results obtained from each method.

Authors	Database used	Obtained results	Highlights
Gao et al. [11]	1202 subjects	The internal data reached 96.74%	Sensitivity to COVID-19 has significantly increased, especially for minor lesions
Fan et al. [12]	1600 unlabeled images were from the COVID-19CT acquisition dataset	Each suspected COVID-19 patient is diagnosed in 22 s by classification model and in 1 s for non-infected cases	Suitable for diagnosis of COVID-19
Yan et al. [13]	165,667 annotated chest CT images	The sensitivity is 0.986, the accuracy is 0.990	Good for early screening

(continued)

Table 1. (*continued*)

Authors	Database used	Obtained results	Highlights
Paluru et al. [14]	929 patients with COVID-19	The average image processing time was 0.4 s	Achieve full automation
Wang et al. [15]	Computed tomography images of 558 patients with COVID-19	The dice coefficient reached 90.68%	Improve the screening performance of COVID-19
Wu et al. [16]	144,167 chest CT images from 400 COVID-19 patients and 350 uninfected patients	The average sensitivity was 95.0% and the specificity was 93.0%	High sensitivity and high specificity
Ramin Ranjbarzadeh [17]	https://github.com/ UCSD-AI4H/ COVID-C	The accuracy rate was 96% and the recall rate was 97%	High accuracy

2.1   Performance Index of Medical Image Segmentation

In general, the method of image segmentation based on CT scanning technique detects its own performance through the statistical average value of performance indicators such as sensitivity, accuracy, robustness and special effect. Sensitivity represents the proportion of pairs among all positive examples, and it has the ability of measuring the algorithm model. The degree of special effect represents the proportion of pairs among all negative examples, which measures the capacity of the algorithm model to identify negative examples. Accuracy represents the proportion of positive examples in the examples divided into positive examples, which is the most common evaluation index. If the accuracy is higher, the performance of the method will be better. However, it is so far from comprehensive and scientific to estimate an algorithm model solely by the accuracy. Robustness has the ability to deal with missing values and outliers. The function of the algorithm model can be measured by these index values, which can aid doctors and researchers in selecting appropriate methods for CT image segmentation.

2.2   Segmentation Methods

In all the references, there are many lung segmentation techniques for different purposes. In paper [7], a dual branch combined network (DCN) for COVID-19 diagnosis is proposed. Individual level classification and lesion segmentation are achieved by DCN. Figure 3 reveals the general framework of the method. The whole method consists of three parts: 1) Lung segmentation based on U-NET; 2) Slice-level combinatorial segmentation and classification using DCN; 3) Individual level classification with three layers of fully connected network. This method realizes the simultaneous classification and segmentation of COVID-19, but it may be hard to distinguish from other organs

because of the serious lesions in some patients. In consideration of the low accuracy of the experimental results, especially for the CT images of COVID-19 patients. The paper [12] proposes to segment the deep network with INF-NET infection, which applying a parallel partial decoder to aggregate advanced features and create a global map. Explicit edge attention and implicit reverse attention are used to model the boundary and enhance the representation. A semi-supervised segmentation framework based on random selection propagation strategy is also used to alleviate the deficiency of labeled data. The semi-supervised segmentation framework enhances ability and attains better performance. Compared with unsupervised detection and segmentation, semi-supervised model can better identify the target region, which is suitable for COVID-19 detection. The paper [9] presents a new deep neural network for automatic segmentation of the COVID-19-infected area and the entire lung from chest CT images, extending the compression and excitation units [14] to deal with the weakened or blurred boundaries. The shape changes of the COVID-19 infected zone are handled by a Progressive Atru Space Pyramid Pool (PASPP). In the paper [10], researchers develop a novel lightweight convolutional neural network based on deep learning, known as ANAM-NET. Design for abnormal COVID-19 segmentation in chest CT images, introduce a label-based weighting strategy of network cost function. Figure 4 demonstrates the key steps involved in segmenting COVID-19 anomalies. ANAM-NET is fully automated, ensuring full turnaround time for segmentation. In the paper [11], an anti-noise framework is proposed to learn segmentation tasks from noise labels. The COVID-19 pneumonia lesion segmentation network is used to treat lesions with different scales or appearance. But collecting perfectly clean labels for medical image segmentation is a challenge. The paper [16] constructs a large-scale COVID-19 classification and segmentation dataset, COVID-CS, developed a joint classification and segmentation system for COVID-19 diagnosis. Using this system, each suspected COVID-19 patient is diagnosed in 22 s by classification model and in 1 s for non-infected cases. Although the mean time is not as good as that in paper [10], the system data set can achieve up to around 93% specificity and 95% sensitivity, providing a convincing visual explanation. Medical image segmentation technology is supposed to be from multiple angles to improve performance and sensitivity. When patients run into a wide range of lesions, the shape of lesions is unsure, and the scanning boundary is fuzzy, it is evident that the multiple COVID-19 segmentation techniques U-NET-based has lots of drawbacks, and the sensitivity, stability and time performance will be greatly reduced.
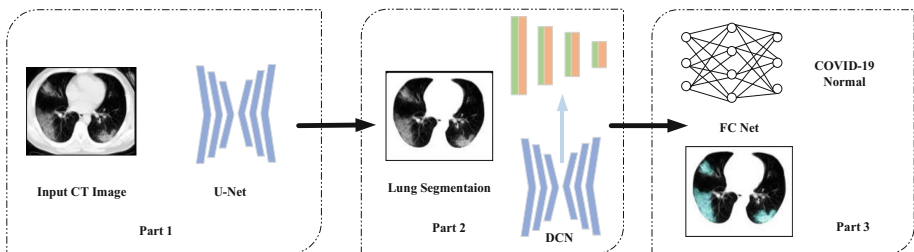
Fig. 3. The frame composition of the lung segmentation method using CNN.
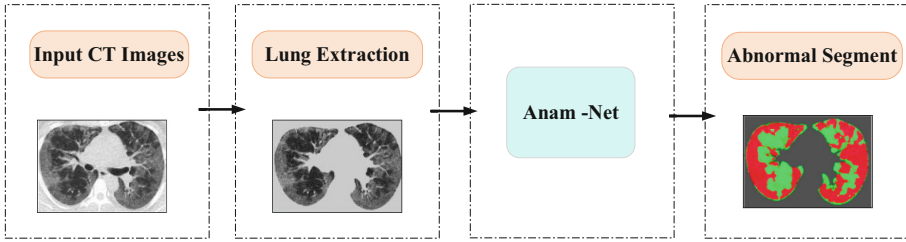
Fig. 4. Key steps of the proposed approach for automated segmentation of abnormalities in chest CT images.

2.3   The Application of COVID-19 Segmentation

Although the paper [8] achieves ideal results with INF-NET for segmentation of infected areas, there are a number of limitations in the current model. In clinical practice, it is frequently essential to classify COVID-19 patients and then segment the infected area to facilitate treatment. INF-NET mainly focuses segmentation of pulmonary infection. However, it is also good at monitoring longitudinal disease changes and performing large-scale screening treatments, and there is great potential for evaluating the diagnosis of COVID-19 by quantifying the area of infection. The paper [10] proposes ANAM-NET segmentation of abnormal COVID-19 images. Compared with other models, it has a very low computational complexity, therefore can be widely deployed in clinical Settings, enabling rapid evaluation of abnormal COVID-19 images [13]. A dual-channel CNN pipeline is implemented to automatically segment COVID-19 pulmonary infection tissue from CT images by making up three unequal input images. Compared with end-to-end learning, supervised and unsupervised methods, segmentation is more flexible and efficient, which is fully capable of detecting abnormal areas with low intensity contrast between diseased and healthy tissue. In the case of infection with fuzzy boundary or wide range of lesions, the utilization of dual CNN pipeline will have more benefits, which can reasonably simplify the procedure and decreases the time cost. The focus attention module LA is based on the development of a dual-branch composite network DCN [7]. LA module can amass the network on the infected sites, which can be applied in the early screening of COVID-19.

In short, the image segmentation is playing a very important role in automatically screening and detection of lesion area of COVID-19 CT or X-ray images, and is helping radiologists to make a quick judgement for the next step of the diagnosis and treatment in a short time.

3   Medical Image Diagnosis of COVID-19 Based on Machine Learning

At present, due to imaging technology's advantages of low cost and high sensitivity, X-rays and chest CT images have been widely used in the diagnosis of pulmonary infection by radiologists. However, medical images, especially chest CT scans, contain

hundreds of sections, which can be difficult for radiologists to handle. It's better to use a machine than to waste human resources on it. Therefore, computer vision method based on deep learning has broad application prospects in medical images, and it is very necessary to use artificial intelligence to assist diagnosis. Machine learning is a method in which a computer uses existing data to train an algorithmic model, and then uses the model to predict the future. Usually, machine learning will use regression algorithm, neural network, support vector machine, clustering and other methods to build models. Although deep learning is a kind of machine learning, it complicates the model by using deep learning networks. Deep learning networks are mechanisms that mimic the human brain for interpreting images, sounds, text and data. Deep learning has driven the rapid development of artificial intelligence. Deep learning not only gives machine learning more practical applications, but also extends the overall scope of artificial intelligence, such as medical imaging.

3.1   Chest X-ray Diagnosis of COVID-19 with Deep Learning

Controlling COVID-19 largely depends on correct diagnosis. AI-based X-ray equipment is readily available around the world, and results are obtained soon. Although it is less sensitive than chest 3D CT imaging, it can be used for the early detection of COVID-19, minimize the amount of treatment time patients waste at the diagnosis stage. Applying artificial intelligence technology, chest X-ray images can be analyzed [16].

Because of the limited computing power and lack of available data for identifying cases of infection through lung X-ray images, it is likely to construct a custom convolutional neural network (CNN) from scratch using the large number of historical lung X-ray images provided by ChExpert, and retrain a deep learning CNN model to examine whether lungs are healthy, adjust the final L-layer using COVID-19 X-ray images. One of the biggest superiorities of CNN model is that data are immediately used to realize the model in the process of training [17].

The paper [18] proposes to use the RESNET-50 deep learning model to extract features from chest X-ray images. All the operations are displayed in the flow chart in Fig. 5. Three different virus types of COVID-19, normal pneumonia and viral pneumonia are studied using SVM algorithm. The experimental results prove that the deep learning-based model can directly extract characteristic and generate results even with limited COVID-19 data.

In order to study the pattern of chest images of COVID-19 patients, it is a good choice to take CNN as a deep learning algorithm. CNN is a development of the traditional Neural Network proposed in 1968 [19]. In addition, the study uses a pre-trained model as a dataset [20], which included 100 COVID-19 chest radiographs and 100 normal chest radiographs. The COVID-19 chest radiograph dataset can be acquired through GitHub [21] and the normal chest radiograph data set also can be gained through Kaggle [22]. Contrast-Restricted Histogram Equalization (CLAHE) is used to improve image quality. Target recognition and image enhancement in medical images are realized by combining CNN and CLAHE, especially for COVID-19 virus detection in chest radiograph images. The purpose of the study is to compare and measure the accuracy of techniques used to improve image quality. Researchers carry out the experiment with the aid of HE, GC and CLAHE. The outcomes indicate that GC has prime sensitivity and CLAHE has prime
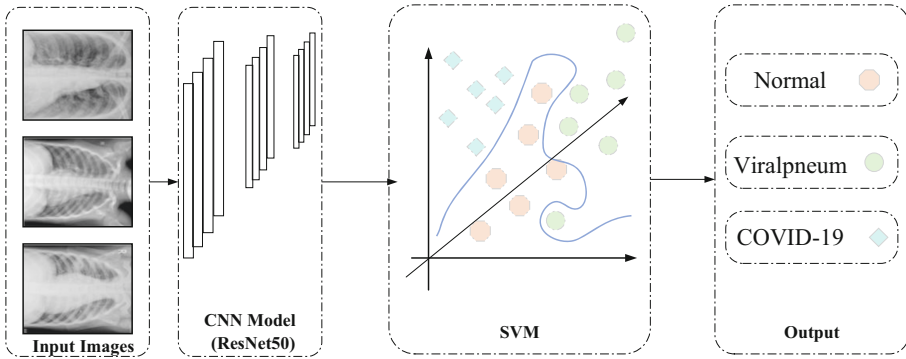
Fig. 5. Flow diagram

accuracy. The key factors of Clahe are the clip limit (CL) and the number of textures (NT) [23]. The experimental data demonstrates that the accuracy rate of the rule set is 99% the verification rate is 97%, which signifies that the basic model test method of Clahe is improved by applying CNN. The implementation of CLAHE assists to boost the quality of the classified image and enhance the contrast of the image. On the side, the new CCSHnet [24] for COVID-19 diagnosis in CCTs uses a group of computer-processed X-ray images taken from disparate angles to generate a cross-sectional image of the scanned area. This way can offer three-dimensional volume data, highlighting additional spatial features and anomalies, can also is used to generate high-quality, detailed images.

The paper [25] derives wavelets from varying frequency and finite duration. An automatic diagnosis method of COVID-19 in chest X-ray images is proposed which based on wavelet theory. This way integrates multi-resolution analysis into the disease recognition network. The dataset consists of chest X-ray images of healthy individuals, COVID-19 and patients with viral pneumonia. In this dataset, there are 1439 images from these three categories, 678 normal cases, 132 COVID-19 cases and 629 viral pneumonia cases. The results will be compared with four latest technologies (DarkCovidNet, Flat-EfficientNet B3, Hierarchy-EfficientNet B3, Detate-Resnet18). The performance of the four different technologies is displayed in Table 2. These techniques use deep learning models to diagnose COVID-19 from chest X-ray images. DarkCovidNet uses YOLO network to detect COVID-19 in chest radiograph images [26] with overall accuracy of approximately 87%. The second and third methods are both based on efficiency networks [27], and plane method and hierarchical method are proposed. The combined accuracy of the two methods is around 93%. This makes known that the accuracy has been increased. In the future, it is envisaged to use the improved deep convolutional network to automatically detect chest radiography images.

On the whole, most studies use X-ray images to classify COVID-19 patients and non-COVID-19 patients. Aim at obtaining more accurate and reliable data, the researchers propose X-rays based on convolution and deep learning strategies to detect COVID-19 virus patients. As a result of the restricted image data of COVID-19, it is hard to gain a large amount of data and conduct more experimental tests, so its clinical suitability still has doubts. In the next, people will put emphasis on early detection of COVID-19 virus diagnosis [40].

Table 2. Performance comparison of four different deep learning net [28]

Method	DarkCovidNet	Flat-EfficientNet B3	Hierarchica EfficientNet B3	DeTraC-ResNet18
Accuracy	87.02	93.34	93.51	95.12
Precision	89.96	93.93	93.93	93.36
Sensitivity	85.35	93.96	93.55	97.91
F1-score	87.37	93.94	93.73	95.58

3.2 Chest CT Diagnosis of COVID-19 with Deep Learning

Convolutional neural network is an extraordinary deep neural network model, generally divided into one dimensional, two dimensional, three dimensional. As the simplest, One-dimensional convolution is a linear space. In essence, it is the convolution of a word vector. One-dimensional convolutional neural network is usually used in low-dimensional space such as natural language processing and sequence model. Two-dimensional convolution is a spread of one-dimensional, which is often used in computer vision and image processing. The specific idea of three-dimensional convolution is the same as the previous two. However, three-dimensional convolution puts time dimension into the input of neural network which makes neural network extract temporal and spatial features and identifying behaviors. The training method of 3D convolutional neural network is similar to convolutional neural network. Therefore, 3D convolutional neural network is the most advanced, so it is frequently used in the medical field, such as CT image and video processing.

It is difficult for humans to distinguish COVID-19, because its appearance on CT images likes other types of viral pneumonia. Therefore, there are a lot of studies about how to distinguish COVID-19 patients from non-COVID-19 patients. This includes a new PSSPNN model [29] that can be used to classify COVID-19, secondary tuberculosis, community-captured pneumonia and healthy subjects, but the method still needs to be improved. In addition, Li Xu et.al have attempted to develop a CAD system for screening COVID-19 CT images. The paper [30] presents a 2DCNN, which takes a series of CT slices as input and 2D-RESNET50 as backbone to extract CNN features from each slice of the CT series. They then use the maximum pooling operation to combine these features, and the resulting map is provided to a fully connected layer, which can generate probability scores for each class. This paper [31] firstly uses a three-dimensional segmentation model, namely VNET [32], to segment lesion candidates from CT images. Then, the 2D-Resnet18 model is used to classify the COVID-19 or Influenza A viral pneumonia.

Besides direct segmentation and 3D networking, the authors also propose a new multi-task primitive attention learning strategy to realize COVID-19 screening in volume chest CT images. To be specific, by devising an up-front focus on residual learning (PARL) block, the aim is integrating two RESNET-based branches into an end-to-end training model framework. Within these data blocks, hierarchical attention information from the lesion area detection branch is transferred to the COVID-19 classification branch

to learn a more discriminating representation. This approach is getting more accurate in locating the lesion area and allowing additional surveillance information to heighten the performance of the COVID-19 classification task. Experimental results manifest that this method exceeds other state-of-the-art COVID-19 screening methods [39].

With the rapid growth of COVID-19, chest CT images have become the main basis for radiologists to collect information about patients because of their high sensitivity and low cost. However, due to the high infectious nature of the epidemic, medical personnel face great risks when collecting COVID-19 CT data. Therefore, even though the image data is relatively common, it cannot meet the large amount of data required for the training of the deep learning model. To meet the data requirements of COVID-19 CT imaging, researchers propose a CT image synthesis method based on conditional generation antagonistic network, which can effectively generate high quality and realistic COVID-19 CT images for deep learning-based medical imaging tasks. Results reveal that this way is beneficial to the synthesis effort of COVID-19 CT images. For good measure, a new AVNC model [33] has emerged which unites attention mechanisms and improves multiplexing data enhancement to extend data sets.

Since COVID-19 is highly infectious [10], for the sake of accelerating the process of COVID-19 CT data acquisition in deep learning-based CT imaging and preventing possible infection of healthcare workers when they contact COVID-19 patients, researchers put forward novel CGAN structure that contains a global-local generator and a multi-resolution discriminator. Both the discriminator and the generator adopt dual network to plan to emulate the local and global information of CT images respectively. Simultaneously, this dual structure has a communication mechanism of information interchange, which is helpful to generate lively CT images with stable global structure and various local details [36].

4 Summary and Future Work

Medical imaging methods are widely employed in medical diagnosis, especially chest X-ray and chest CT. X-ray is the most extensively used diagnostic X-ray examination in medical practice, which is of great significance for early clinical research and diagnosis. The quantitative analysis of medical images demands the segmentation and diagnosis of the boundary of the object of interest. Traditionally, doctors complete medical image segmentation. But when the number of divisions is massive, the need for storage is very urgent. Given the shape and difficulty of various organs, it is hard for doctors to segment the right areas accurately and there is nothing they can do about the large number of images that need to be segmented. Therefore, advanced automated segmentation technology is especially needed to solve this problem.

This paper reviews COVID-19 image segmentation and diagnosis techniques, introduces medical image segmentation methods and applications in detail, as well as COVID-19 diagnosis based on CT scan and X-ray scan. It illustrates the significant role of rapid automatic segmentation and AI diagnosis technology in COVID-19 prevention and control management.

With the raising demand for COVID-19 diagnosis, medical image segmentation technology complemented by dual CNN channels, deep neural network (CoVIDSEGNET)

or DCN model has greatly improved the accuracy and sensitivity of segmentation technology [38]. The intelligent diagnosis has the capacity of automatic screening and timely analysis, avoiding the possibility of misdiagnosis, and improving radiologists' working efficiency. It is important to note that the current study still has many limitations. The CT or X-ray images of COVID-19 patients have blurred boundaries, large lesion area, or consolidation, which makes it difficult to segment medical images accurately. It is expected that in the future, the image scanning will be clearer and the resolution will be higher, enabling medical image segmentation technology to integrate AI technology into the process of image acquisition and synthesis. It will also attempt to apply AI to the overall segmentation and diagnosis of COVID-19, realize automatic screening, analysis and report generation to make quick judgments by using artificial intelligence [37]. In particular, images of small lesions such as COVID-19 can be flagged and drawn to conclusions. In the future, AI image acquisition and synthesis workflow will enable patients to obtain high-quality images even at low radiation levels.

References

1. He, J.L., Luo, L., et al.: Diagnostic performance between CT and initial real-time RT-PCR for clinically suspected 2019 coronavirus disease (COVID-19) patients outside Wuhan, China. Respirat. Med. **168**, 105980 (2020)
2. Tabik, S., et al.: COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images. J. Biomed. Health Inf. **24**(12), 3595–3605 (2020)
3. Wang, J., Chen, Z., Lang, X., et al.: Quantitative evaluation of infectious health care wastes from numbers of confirmed, suspected and out-patients during COVID-19 pandemic: a case study of Wuhan. Waste Manag. **126**, 323–330 (2021)
4. Elaziz, M.A., et al.: An improved marine predators algorithm with fuzzy entropy for multi-level thresholding: real world example of COVID-19 CT image segmentation. IEEE Access **8**, 125306–125330 (2020)
5. Niu, R., Ye, S., Li, Y., et al.: Chest CT features associated with the clinical characteristics of patients with COVID-19 pneumonia. Ann. Med. **53**(1), 169–180 (2021)
6. Pan, F., Ye, T., Sun, P., et al.: Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia. Radiology **295**, 715–721 (2020)
7. Miao, Q., Ma, P., Yuan, Y., et al.: Pulmonary CT findings of COVID-19. J. Xuzhou Med. Univ. **201**, 306–309 (2021). In Chinese
8. Liang, Z., Huang, J.X., et al.: Enhancing automated COVID-19 chest X-ray diagnosis by image-to-image GAN translation. In: International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1068–1071. IEEE (2020)
9. Degerli, A., et al.: COVID-19 infection map generation and detection from chest X-ray images. Health Inf. Sci. Syst. **9**(1), 1–16 (2021)
10. Shi, F., Wang, J., et al.: Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. Rev. Biomed. Eng. **14**, 4–15 (2021)
11. Gao, K., Su, J., Jiang, Z., et al.: Dual-branch combination network (DCN): towards accurate diagnosis and lesion segmentation of COVID-19 using CT images. Med. Image Anal. **67**, 101836 (2021)
12. Fan, D.P., Zhou, T., et al.: Inf-Net: automatic COVID-19 lung infection segmentation from CT images. IEEE Trans. Med. Imaging **39**(8), 2626–2637 (2020)
13. Yan, Q., et al.: COVID-19 chest CT image segmentation network by multi-scale fusion and enhancement operations. IEEE Trans. Big Data **7**(1), 13–24 (2021)

14. Paluru, N., et al.: Anam-Net: anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images. Trans. Neural Netw. Learn. Syst. **32**(3), 932–946 (2021)
15. Wang, J., et al.: Prior-attention residual learning for more discriminative COVID-19 screening in CT images. Trans. Med. Imaging **39**(8), 2572–2583 (2020)
16. Wu, Y.H., et al.: JCS: an explainable COVID-19 diagnosis system by joint classification and segmentation. Trans. Image Process. **30**, 3113–3126 (2021)
17. Ranjbarzadeh, R., Ghoushchi, S.J., et al.: Lung infection segmentation for COVID-19 pneumonia based on a cascade convolutional network from CT images. BioMed Res. Int. **16**, 5544742 (2021)
18. Narin, A.: Detection of Covid-19 patients with convolutional neural network based features on multi-class X-ray chest images. In: Medical Technologies Congress (TIPTEKNO), pp. 1–4. IEEE (2020)
19. Hubel, D.H., et al.: Receptive fields and functional architecture of monkey striate cortex. J. Physiol. **195**, 215–243 (1968)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Comput. Sci. (2014)
21. Software source code hosting service platform. https://github.com/
22. Data science, machine learning competitions and sharing platforms. https://www.kaggle.com/
23. Campos, G.F.C., Mastelini, S.M., Aguiar, G.J., Mantovani, R.G., Melo, L., Barbon, S.: Machine learning hyperparameter selection for contrast limited adaptive histogram equalization. EURASIP J. Image Video Process. **2019**(1), 1–18 (2019)
24. Wang, S.H., et al.: COVID-19 classification by CCSHNet with deep fusion using transfer learning and discriminant correlation analysis. Inf. Fusion **68**, 131–148 (2021)
25. Mehrotra, A., Singh, K.K., et al.: Detection of 2011 Tohoku tsunami induced changes in Rikuzentakata using normalized wavelet fusion and probabilistic neural network. Disast. Adv. **7**(2), 1–8 (2014)
26. Abbas, A., Abdelsamea, M.M., Gaber, M.M.: Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. Appl. Intell. **51**(2), 854–864 (2020)
27. Luz, E., Silva, P., Silva, R.P., et al.: Towards an effective and efficient deep learning model for COVID-19 patterns detection in X-ray images. Res. Biomed. Eng. (2021). https://doi.org/10.1007/s42600-021-00151-6
28. Singh, K.K., Singh, A.: Diagnosis of COVID-19 from chest X-ray images using wavelets-based depthwise convolution network. Big Data Mining Analyt. **4**(2), 84–93 (2021)
29. Wang, S.H., et al.: PSSPNN: PatchShuffle stochastic pooling neural network for an explainable diagnosis of COVID-19 with multiple-way data augmentation. Comput. Math. Methods Med. 6633755 (2021)
30. Li, L., Qin, L., et al.: Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. Radiology **296**(2), 200905 (2020)
31. Butt, C., Gill, J., et al.: Deep learning system to screen coronavirus disease 2019 pneumonia. Appl. Intell. (2020)
32. Milletari, F., Navab, N., et al.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016)
33. Wang, S.-H., Fernandes, S., Zhu, Z., Zhang, Y.-D.: AVNC: attention-based VGG-style network for COVID-19 diagnosis by CBAM. IEEE Sens. J. (2021). https://doi.org/10.1109/JSEN.2021.3062442
34. Alghamdi, H.S., Amoudi, G., et al.: Deep learning approaches for detecting COVID-19 from chest X-ray images: a survey. IEEE Access **9**, 20235–20254 (2021)
35. Wang, S., et al.: A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). Eur. Radiol. **31**(8), 6096–6104 (2021)

36. Yang, G., et al.: Automatic segmentation of kidney and renal tumor in CT images based on 3D fully convolutional neural network with pyramid pooling module. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3790–3795 (2018)
37. Elkorany, A.S., Elsharkawy, Z.F.: COVIDetection-Net: a tailored COVID-19 detection from chest radiography images using deep learning. Optik **231**, 166405 (2021)
38. Zhao, H., Shi, J., et al.: Pyramid scene parsing network. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230–6239. IEEE (2016)
39. Wu, X., Huang, S., et al.: The application of artificial intelligence in medical imaging and cancer treatment decision. China School Med. **201**, 235–238 (2020)
40. Ashour, A.S., Eissa, M., Wahba, M.A., et al.: Ensemble-based bag of features for automated classification of normal and COVID-19 CXR images. Biomed. Signal Process. Control **68**, 102656 (2021)

Author Index