Shui-Hua Wang
Yu-Dong Zhang (Eds.)

Multimedia Technology and Enhanced Learning

4th EAI International Conference, ICMTEL 2022
Virtual Event, April 15–16, 2022
Proceedings

EAI

Springer

Lecture Notes of the Institute
for Computer Sciences, Social Informatics
and Telecommunications Engineering   446

More information about this series at https://link.springer.com/bookseries/8197

Shui-Hua Wang · Yu-Dong Zhang (Eds.)

Multimedia Technology and Enhanced Learning

4th EAI International Conference, ICMTEL 2022
Virtual Event, April 15–16, 2022
Proceedings

Springer

Editors
Shui-Hua Wang 🆔
University of Leicester
Leicester, Leicestershire, UK

Yu-Dong Zhang 🆔
University of Leicester
Leicester, Leicestershire, UK

Preface

We are delighted to introduce the proceedings of the fourth edition of the European Alliance for Innovation (EAI) International Conference on Multimedia Technology and Enhanced Learning (ICMTEL 2022). This conference aims to deliver a forum for scientific reports and debates on the state of the art and future perspectives in multimedia technology and enhanced learning.

The technical program of ICMTEL 2022 consisted of 69 full papers, which were selected from 188 submissions in a single-blind review process, with a minimum of 3.1 reviews per paper. The papers were presented in the following conference tracks: Track 1 - Internet of Things and Communication; Track 2 - Education and Enterprise; Track 3 - Machine Learning; Track 4 - Big Data and Signal Processing; Track 5 - Workshop of Data Fusion for Positioning and Navigation; and Track 6 - Workshop of Intelligent Systems and Control.

Aside from the high-quality technical paper presentations, the technical program also featured two keynote speeches given by De-Shuang Huang from Tongji University, China, and Juan M. Górriz from the University of Granada, Spain.

Coordination with the steering chairs, Imrich Chlamtac, De-Shuang Huang, and Chunming Li, was essential for the conference's success. 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 organizing and supporting the conference.

In particular, we are grateful to the Technical Program Committee, which completed the peer-review process for technical papers and helped to put together a high-quality technical program. We thank the publication chair, Shuai Liu, who helped us organize the special issues. We are also grateful to the conference manager, Karina Ogandjanian, for her support and to all the authors who submitted their papers to the ICMTEL 2022 conference and workshops.

We strongly believe that the ICMTEL 2022 conference provided a good forum for all researchers, developers, and practitioners to discuss all aspects of science and technology relevant to multimedia technology and enhanced learning. We also expect that the future ICMTEL conferences will be as successful and stimulating as this year's, as indicated by the contributions presented in this volume.

September 2022

Shui-Hua Wang
Yu-Dong Zhang

Organization

Steering Committee

Chair

Imrich Chlamtac	University of Trento, Italy

Co-chairs

Deshuang Huang	Tongji University, China
Chunming Li	University of Electronic Science and Technology of China, China

Members

Lu Liu	University of Leicester, UK
M. Tanveer	Indian Institute of Technology Indore, India
Huansheng Ning	University of Science and Technology Beijing, China
Wenbin Dai	Shanghai Jiao Tong University, China
Wei Liu	University of Sheffield, UK
Zhibo Pang	ABB Corporate Research Center, Sweden
Suresh Chandra Satapathy	KIIT Deemed to University, India
Yu-Dong Zhang	University of Leicester, UK

Organizing Committee

General Chair

Yu-Dong Zhang	University of Leicester, UK

General Co-chair

Ruidan Su	Shanghai Advanced Research Institute, China

Technical Program Committee Chair

Shui-Hua Wang	Loughborough University, UK

Local Chair

Yu Xiang   University of Leicester, UK

Workshops Chair

Yuan Xu   Jinan University, Lebanon

Publicity and Social Media Chair

Qinghua Zhou   University of Leicester, UK

Publications Chair

Shuai Liu   Hunan Normal University, China

Web Chair

Lijia Deng   University of Leicester, UK

Posters and PhD Track Chair

Shuhui Bi   University of Jinan, China

Panels Chair

Anqi Bi   Changshu Institute of Technology, China

Demos Chair

Shuwen Chen   Jiangsu Second Normal University, China

Tutorials Chair

Chenxi Huang   Xiamen University, China

Technical Program Committee

Abdon Atangana   University of the Free State, South Africa
Amin Taheri-Garavand   Lorestan University, Iran
Arifur Nayeem   Saidpur Government Technical School and
                             College, Bangladesh
Arun Kumar Sangaiah   Vellore Institute of Technology, India
Carlo Cattani   University of Tuscia, Italy
Dang Thanh   University of Economics Ho Chi Minh City,
                             Vietnam
David Guttery   University of Leicester, UK
Debesh Jha   Chosun University, South Korea

Frank Vanhoenshoven	University of Hasselt, Belgium
Gautam Srivastava	Brandon University, Canada
Gonzalo Napoles Ruiz	University of Hasselt, Belgium
Hari Mohan Pandey	Edge Hill University, UK
Hong Cheng	First Affiliated Hospital of Nanjing Medical University, China
Juan Manuel Górriz	University of Granada, Spain
Liangxiu Han	Manchester Metropolitan University, UK
Mingwei Shen	Hohai University, China
Nianyin Zeng	Xiamen University, China
Pengjiqiang Qian	Jiangnan University, China
Praveen Agarwal	Anand International College of Engineering, India
Raymond F. Muzic, Jr.	Case Western Reserve University, USA
Roin Deng	University of Leicester, UK
Seifedine Kadry	Beirut Arab University, Lebanon
Shipeng Xie	Nanjing University of Posts and Telecommunications, China
Shiting Sun	University of Coventry, UK
Siyuan Lu	University of Leicester, UK
Sunil Kumar	National Institute of Technology Jamshedpur, India
Tianming Zhan	Nanjing Audit University, China
Vikrant Bhateja	SRMGPC, India
William Wang	University of Leicester, UK
Xiang Yu	University of Leicester, UK
Xianwei Jiang	Nanjing Normal University of Special Education, China
Xinhua Mao	Nanjing University of Aeronautics and Astronautics, China
Xujing Yao	University of Leicester, UK
Yuan Xu	University of Jinan, China
Yuankai Huo	Vanderbilt University, USA
Yuriy S. Shmaliy	Universidad de Guanajuato, Mexico
Zhimin Chen	Shanghai Dianji University, China
Zhuqing Jiao	Changzhou University, China
Ziquan Zhu	University of Florida, USA

Contents

Machine Learning

Workshop of Data Fusion for Positioning and Navigation

Workshop of Intelligent Systems and Control

Internet of Things and Communication

On Line Monitoring System of Power Optical Fiber Transmission Network Under Internet of Things Technology

Lu Liu[✉] and Lanxiang Wang

Beijing Kedong Electric Power Control System Company Limited, Beijing 100192, China
seg8856296@126.com

Abstract. The traditional OTD power optical fiber transmission network online monitoring system takes a long time to monitor the network data, and the monitoring effect increases. Therefore, the power optical fiber transmission network online monitoring system under the Internet of things technology is designed. The sensor and microcontroller are designed in hardware. Analyze the functional requirements of on-line monitoring system in software; The data acquisition module of power optical fiber transmission network is established based on the Internet of things, and then the system software function is realized. By means of comparative experiment, it is verified that the online monitoring effect of the new system is better and has great popularization value.

Keywords: Internet of things technology · Power optical fiber · Transmission network · Online monitoring

1 Introduction

The safe operation of power system is related to the national economy and the people's livelihood. With the development of economy, the demand for electricity is increasing. Modern power system is developing towards high voltage, large units and large capacity, and the requirements for power supply reliability of power system are becoming higher and higher. At the same time, with the rapid development of information today, The rapid development of science and technology and the improvement of people's living standards year by year also promote the development of power system towards intelligence. Digitization and networking have become an inevitable trend [1]. The State Grid Corporation of China first announced the development plan of "smart grid" in 2009. The smart grid is based on an integrated, high-speed two-way communication network and combines advanced sensing and measurement technology, information and communication technology, analysis and decision technology, automatic control technology and energy and power technology, And a new modern power grid formed by high integration with power grid infrastructure [2]. Strength, interaction, automation, informatization, environmental protection and economy are the main characteristics of smart grid. Smart

grid requires to be able to monitor the operation status of power grid in real time, find faults and recover itself in time with as little manual intervention as possible, so as to improve the safety and reliability of power grid operation. Smart grid includes six links of power generation, transformation, transmission, distribution, dispatching and power consumption in the power system. It can realize observability (monitoring the state of all equipment in the power grid), controllability (controlling the state of all equipment in the power grid), complete automation (adaptive and self-healing) and comprehensive optimal balance of the system (optimal balance between power generation, transmission and distribution and power consumption) So as to realize cleaner, efficient, safe and reliable operation of power system [3].

The intelligence of power grid puts forward higher requirements for the safe operation of power system. There are many factors affecting the safe and stable operation of power system, one of which is the operation safety of power equipment [4]. As the key to connect the six links of smart grid, substation is an important power facility in the power system to transform voltage, receive and distribute electric energy, control the flow direction of power and adjust voltage. It connects the power grid at all levels and plays a key role in the power system. The safe and normal operation of power equipment in substation is one of the most important factors to ensure the safe transmission of power in power system.

In the power field, the implementation of the periodic maintenance system of equipment based on time cycle provides a certain guarantee for the safe operation of power equipment. However, this maintenance method is affected by both objective factors (such as fixed cycle, maintenance experience, different operating environments of different equipment, etc.) and subjective factors (such as rigid work, lack of manual inspection, missed inspection, etc.), which makes this maintenance method expose a variety of problems: low reliability in the station, many power outages, large workload of personnel and high maintenance cost [5]. With the increasing contradictions exposed by the traditional maintenance methods and the rapid development of power industry and power supply technology, the on-line monitoring technology of substation equipment came into being. However, there are many problems in the early monitoring, such as poor stability, signal distortion, backward technology, non-specific monitoring standards, large installation workload and difficult maintenance. Although this monitoring technology can greatly improve the safe operation of power grid, due to the prevalence of the above problems, most online monitoring systems have the problems of unreliable data, nonstandard data and data can not truly reflect the equipment status, which can not play the role of monitoring the equipment status, thus hindering the popularization and Application of this new technology. At the same time, most of the on-line monitoring equipment of various manufacturers in the substation work independently, and carry out data acquisition, data analysis and result output respectively, which makes the data structure incompatible and lacks unified management. With the gradual development of technology and the gradual increase of problems, we realize the necessity of establishing a unified, comprehensive and intelligent intelligent online monitoring equipment online monitoring system.

Internet of things technology is in line with the needs of this actual development. The Internet of things has the ability to obtain, analyze, process, control and feedback information, and can realize the integration of human society and physical system [6].

The application of Internet of things in the construction of intelligent substation is the inevitable result of the development of the information age to a certain extent, which can realize the real-time management and control of equipment. Using the Internet of things technology, through the perception of the external world, an infinite sensor monitoring network is constructed to monitor the operation status of online monitoring equipment in an all-round and real-time manner.

In the existing power system equipment monitoring applications, they are basically connected through wired media. In the implementation process of wired lines, the quantities are large, the lines are not easy to expand, and in some special geographical environments and sites, the wiring will be limited and can not be implemented. Wireless sensor networks communicate wirelessly, which not only gets rid of the constraints of wired, but also has the advantages of strong mobility, easy expansion and upgrading [7]. With the continuous development of sensor technology and communication technology, wireless sensor network has been gradually applied to the field of on-line monitoring of substation power equipment.

The Internet of things connects all kinds of power supply and electrical equipment, and has been widely used in the field of smart grid. It is the main trend and direction of power system development. The application of Internet of things technology to the on-line monitoring of power equipment in intelligent substation is of great significance to ensure the safe and stable operation of power system. At the same time, it is of great significance and practical value to ensure the safe and stable operation of on-line monitoring equipment, improve the automation level of substation and promote the digital and networked development of intelligent power network.

2 Hardware Design

2.1 Sensors

The sensor designed in this paper has the functions of current sensing, ultrasonic sensing, digital temperature sensing, micro water sensing and pressure sensing. Electromagnetic BCT-2 type is adopted for current sensing function, which can realize on-line monitoring of insulation characteristics of most high-voltage electrical equipment by measuring leakage current. The sensor has strong electromagnetic field anti-interference ability and excellent temperature characteristics. The iron core is made of platinum nickel alloy with high initial permeability and low loss, and the iron core is fully automatically compensated through deep negative feedback to ensure that the iron core works in an ideal zero flux state [8]. The sensor adopts active zero flux technology to ensure the high accuracy and stability of low current detection, and can detect leakage current signals as small as milliampere (mA). It can be conveniently used to accurately measure the insulation characteristics of CT, CV, OY and TB with small leakage current. Its core structure can also ensure the sampling safety of electrical equipment and signals. Wireless temperature sensing capability is a new type of temperature sensing module that integrates the functions of wireless RF chip and digital temperature sensing. The temperature monitoring is collected by the temperature and humidity sensing module HMP155 of Vaisala company. HMP155 temperature sensing module has automatic calibration function and

RS485 communication interface. It can send temperature signals through its own communication protocol. In addition, it can also set serial port parameters, data acquisition method, query historical data, etc. Users can transplant ZigBee protocol into nodes to form simple star topology networks and complex tree and mesh networks. The sensing module can be applied to the fields of high-voltage substation equipment and high-voltage switchgear temperature measurement in intelligent substation. The ultrasonic sensing function is made by using the principle that piezoelectric ceramics will produce mechanical deformation varying with voltage or frequency under the action of voltage, and piezoelectric ceramics will produce charge when receiving vibration. When making, two pieces of piezoelectric ceramics or piezoelectric ceramics and metal sheets can be used to make a vibrator (double piezoelectric chip element). When ultrasonic vibration acts on the double piezoelectric chip element, an electrical signal will be generated. The type of ultrasonic can be judged by the electrical signal.

Ultrasonic sensing function, with convenient installation (it can be fixed to the equipment surface during installation), and the ultrasonic wave can pass through the metal shell and will not affect the surrounding equipment. Therefore, it can be widely used for partial discharge measurement inside closed high-voltage metal equipment such as GIS, which also provides guarantee for the safe and reliable operation of closed substation equipment. Therefore, the sensor is used for partial discharge of transformer and partial discharge of GIS Discharge measurement.

2.2 Microcontroller

In the microcontroller designed in this paper, the microcontroller with low external clock frequency can effectively reduce the noise and improve the anti-interference ability of the system. For square wave and sine wave with the same frequency, the high-frequency component of square wave is much more than sine wave. The amplitude of the high-frequency component of the square wave is smaller than that of the fundamental wave, but the higher the frequency, the easier it is to be emitted and become a noise source. The most influential high-frequency noise generated by the microcontroller is about three times the clock frequency. Therefore, the lower the frequency of the selected microcontroller, the lower the frequency of its three times, and the lower the probability of developing into a noise source. In addition, decoupling design and shielding technology are added to this system. Reasonable decoupling design can filter the electromagnetic interference caused by peak current jump and improve the anti-interference ability and reliability of the system [9]. The decoupling capacitor directly connected between the ground of the device and the power supply can reduce the power impedance and remove the high-frequency components as high as 1GHZ. When designing a printed circuit board, a decoupling capacitor should be added between the power supply and ground of each integrated circuit. The decoupling capacitor has two functions: on the one hand, the energy storage capacitor of the integrated circuit provides and absorbs the charge and discharge energy at the moment of opening and closing the door of the integrated circuit; On the other hand, the high-frequency noise of the device is bypassed. Shielding is an effective anti-interference measure, which can reduce the outward or inward penetration of electromagnetic field. It is often used to isolate and attenuate radiated interference. Shielding technology divides space into two regions through metal objects. Its purpose

is to control the diffusion of electric field from one region to another. Shielding technology can be divided into electrostatic shielding, magnetic shielding and electromagnetic shielding. The function of electrostatic shielding is to eliminate the electromagnetic interference caused by distributed capacitance coupling between two circuits. Magnetic shielding is used to prevent the interference of low-frequency magnetic field. The high-voltage switchgear is mainly interfered by high-frequency electromagnetic field, so electromagnetic shielding is adopted, which can suppress the interference of electric field and magnetic field at the same time and prevent electromagnetic wave from entering. The sensor probe is placed in a shielding body made of low resistance metal material aluminum, and the shielding metal is used to absorb and reflect the electromagnetic field to achieve the purpose of shielding, The shielding technology is shown in Fig. 1:
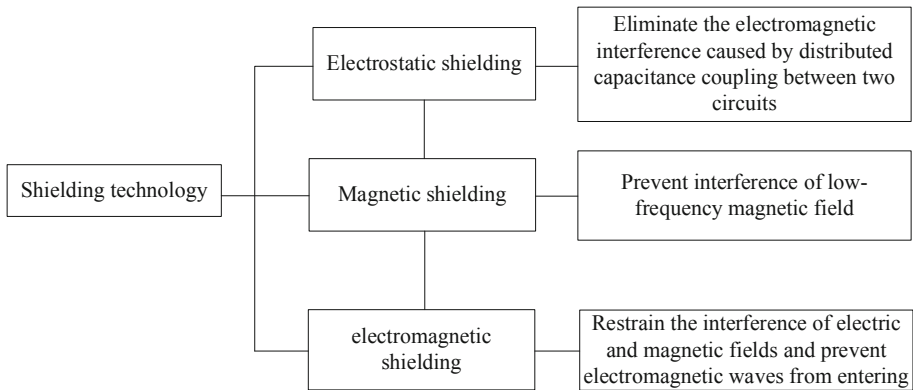
Fig. 1. Structure diagram of shielding technology

The microcontroller is mainly manufactured by high-speed CMOS technology. The static input current at the signal input end is about 1mA and the input capacitance is about 10pF. The output end of high-speed CMOS circuit has considerable load capacity, that is, considerable output value. If the output end of a gate is led to the input end with high input impedance through a long line, the reflection problem is very serious, which will cause signal distortion and increase system noise. When Tpd is greater than Tr (standard delay time), it becomes a transmission line problem. Problems such as signal reflection and impedance matching must be considered. The delay time of the signal on the printed circuit board is related to the impedance characteristics of the lead, that is, to the dielectric constant of the printed circuit board material. It can be roughly considered that the transmission speed of the signal on the lead of the printed board is about 1/3 to 1/2 of the speed of light. The tr of common logic circuit elements in the system composed of microcontroller is between 3-18ns. On the printed circuit board, the signal passes through a 7 W resistor and a 25 cm long lead, and the on-line delay time is about 4-20 ns. In other words, the lead of the signal on the printed circuit is very short, no more than 25 cm. And the number of vias is small. The rise time of the signal is faster than the signal delay time. The impedance of the transmission line is matched. For the signal transmission between the integrated blocks on the brush circuit board, Td > Trd

will not occur. The wiring on the printed board shall make the wiring between circuits as short as possible under possible conditions, which will greatly reduce the interference caused by wiring.

The microcontroller manufactured by CMOS process has high input impedance, high noise and high noise tolerance. Even if the digital circuit is superimposed with 100–200 mv noise, its work will not be affected. If the printed circuit board is a four layer board, one of which is a large area of ground, or a double-sided board, and the reverse side of the signal line is a large area of ground, the cross interference between signals will become smaller. The characteristic impedance of the signal line is reduced in a large area, and the reflection of the signal at the D end is greatly reduced, which also reduces the cross interference between the signal lines.

3 Software Design

3.1 Analyze the Functional Requirements of the Online Monitoring System

Network transmission monitoring condition monitoring and fault diagnosis system is an important content of smart grid construction. Network transmission monitoring condition monitoring technology is the key supporting technology to realize the construction of Smart Substation and the core content of Smart Substation construction. The network transmission monitoring status monitoring and evaluation system at the station control layer shall be able to evaluate the working status and remaining life of power equipment according to the obtained power equipment status information, adopt the comprehensive evaluation model based on multi information fusion technology, and combine the structural characteristics and parameters of equipment, operation history status records and environmental factors; Analyze, judge and predict the faults that have occurred, are occurring or may occur, clarify the nature, type, degree and cause of faults, point out the trend and consequences of fault occurrence and development, and put forward effective countermeasures to control fault development and eliminate faults, so as to avoid power equipment accidents and ensure safe, reliable and normal operation of equipment [10]. The software system of station control layer shall provide an overall solution for on-line monitoring of intelligent substation. The system can control the temperature and load of power equipment, dissolved gas in oil, micro water in oil, bushing insulation, iron core grounding current, partial discharge, auxiliary equipment (cooling fan, oil pump, gas relay, on load tap changer, etc.), SF6 gas density and micro water in circuit breaker and GIS, GIS partial discharge, action characteristics of circuit breaker, SF6 gas leakage in GIS room The insulation of current transformer and capacitive voltage transformer, coupling capacitor insulation and lightning arrester insulation shall be comprehensively monitored. The specific functional requirements are shown in Fig. 2.
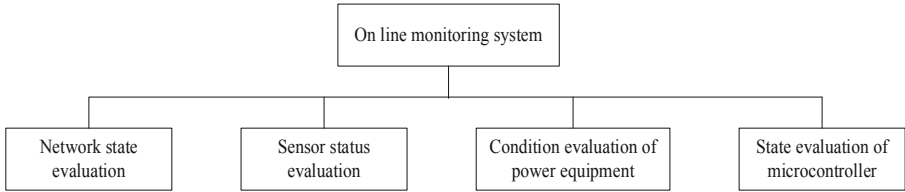
Fig. 2. System functional requirements

As shown in Fig. 2, in the system designed in this paper, its functional requirements are set as power equipment status evaluation, sensor status evaluation, network status evaluation, microcontroller status evaluation, etc. With the rapid development of intelligent substation and the wide application of computer technology in substation system in recent years, the original closure of substation system has been broken. The communication security of intelligent substation and its supporting network monitoring equipment have received great attention from power industry and relevant research departments at home and abroad. In recent years, the communication network between intelligent electronic devices in substations has become more and more perfect. The monitoring equipment grasps the operation of intelligent substations by capturing and analyzing the communication messages between substation networks, and timely finds problems, locates the root causes of faults and solves problems through efficient and reliable diagnosis algorithms, so as to avoid malignant power accidents, Ensure the safe and stable operation of power grid.

3.2 Data Acquisition Module for Establishing Power Optical Fiber Transmission Network Based on Internet of Things

The Internet of things is an important part of the new generation of information technology. The English name of the Internet of things is "the Internet of things". As the name suggests, the Internet of things is "the Internet connected with things" [11]. It mainly has two meanings: (1) the foundation and core of Internet of things technology is still the Internet, which is a network extended and expanded on the basis of the Internet; (2) Its client extends to the information exchange and communication between any object. Therefore, the definition of Internet of things is to connect any object with the Internet according to the agreed protocol through information sensing equipment such as radio frequency identification, infrared sensor, global positioning system and laser scanner, so as to realize the intelligent identification and positioning of objects Simultaneous interpreting, monitoring and managing a network [12]. With the development of sensor technology and network communication technology, the definition of the Internet of things will also be deepened. The Internet of things has distinct characteristics compared with the traditional Internet: it widely uses various sensing technologies. The Internet of things has deployed massive different kinds of sensors, each sensor is an information source. Data is real-time [13]. It collects environmental information periodically according to a certain frequency and constantly updates data. It is a network based on the Internet. The important foundation and core of Internet of things technology is still

the Internet. It is an extension and expansion based on the Internet. Through the integration of various wired and wireless networks with the Internet, the information of objects can be transmitted in real time and accurately. Internet of things It not only provides the connection of sensors, but also has the ability of intelligent processing and intelligent control of objects.Analyze, process and process meaningful data from the massive information obtained by sensors, so as to meet the different needs of different users and find new application fields and application modes. Therefore, this paper uses the Internet of things to design the data acquisition module of power network. Firstly, the operating power η of power grid equipment is calculated:

$$\eta = 1 - \left(\frac{\cos \Phi_1}{\cos \Phi_2} \right)^2 \tag{1}$$

In formula (1), η is the operating power of power grid equipment; $\cos \Phi_1$ and $\cos \Phi_2$ are the power factors of power grid equipment and distribution equipment respectively. Considering the target node of power grid monitoring, the system monitoring design equation is as follows:

$$P_i = \prod_{n=1}^{K} \alpha_n \tag{2}$$

In formula (2), P_i represents the monitoring results from the monitoring source node to the destination node; K represents the total number of monitoring nodes; α_n represents the monitoring probability of node n. The collection formula of network monitoring status is as follows:

$$a_{1k} = \frac{d_k}{d_1 + d_2 + \ldots + d_N} \tag{3}$$

In formula (3), a_{1k} is the network monitoring status; d_1 is the acquisition node corresponding to error free state 1; d_2 is the acquisition node corresponding to error free state 2 d_N is the acquisition node corresponding to the error free state d_k; k is the acquisition node corresponding to the burst state. The online monitoring data after using the Internet of things technology are:

$$P(X_i | \pi(X_i)) = P(X_i | X_1, X_2, \cdots, X_{i-1}) \tag{4}$$

In formula (4), X_i is the online monitoring power grid data, and the online monitoring data node related to X_i is recorded as $\pi(X_i)$. The collected pseudo-random monitoring data are as follows:

$$X_{n+1} = \lambda X_n (1 - X_n) \tag{5}$$

In formula (5), $0 \le \lambda \le 4$ and λ are random monitoring parameters; X_{n+1} is pseudo-random monitoring data; X_n is the random sequence of power grid monitoring of the Internet of things. After multiple iterations, X_n is transformed into pseudo-random sequence S_n The conversion formula is as follows:

$$S_n = f(X_n) \tag{6}$$

In formula (6), S_n is a pseudo-random sequence; Where f is the random function of pseudo-random sequence.

$$\bar{x} = \mathrm{a}(y - x) \tag{7}$$

$$\bar{y} = -xz + cy + (c - a) \tag{8}$$

$$\bar{z} = xy - bz \tag{9}$$

In formulas (7) to (9), $\bar{x}$, $\bar{y}$ and $\bar{z}$ are on-line monitoring coefficients; a, b and c are acquisition parameters; x, y and z are the spatial coordinate coefficients of $\bar{x}$, $\bar{y}$ and $\bar{z}$. The monitoring sequence of power optical fiber transmission network at this time is calculated as follows:

$$F = \frac{\left(\sum_{i=1}^{52} K[i] * 2^{i=1} \right)}{2^{52}} \tag{10}$$

In formula (10), F is the initial information of the system; $K[i]$ is the i monitoring data in 256 bit random sequence;Firstly, this paper can query the monitoring equipment status of the power grid, and display and set various status parameters in the monitoring parameter setting interface, including monitoring mode, number of monitoring lines, monitoring wavelength of each monitoring optical path, monitoring threshold value, and Jianli line working channel of manual monitoring mode. Secondly, the monitoring equipment can be initialized and the monitoring mode parameters can be set. In the automatic monitoring mode, the monitoring optical power threshold and monitoring wavelength of each detection optical path can be set. In the manual monitoring mode, the detection wavelength and working channel of Jianli line can be set. Finally, connect the optical tail fiber on the hardware equipment, select the monitoring line from the existing line list, automatically set the line monitoring wavelength parameters, store the line parameter information in the database, and monitor the multi-channel working optical fiber in real time according to the set threshold value. The system can manually disconnect the pigtail connection according to the illumination power threshold, or manually reduce the optical power of optical input, so as to realize the low threshold of optical power and automatically display the line status in the line monitoring status area of the main interface.

3.3   Realization of System Software Functions

In order to realize the system software function, this paper designs the monitoring data model of power optical fiber transmission network under the above environment, as shown in Table 1.

Table 1. Model table of system monitoring data

Parameter name	Parameter code	Character type
Monitored identification	LinkedDevice	character
Monitoring device	Device Code	character
Monitoring time	AcquisitionTime	date
System transmission volume	DischarageCapacity	character
Transmission location	Discharge Position	number
Number of pulses	Pulse Count	number
Transmission microwave	DischargeWaveform	Binary stream

As shown in Table 1, according to the database file, the optical fiber line, landmark information and the fault information of test analysis are corresponding, and the fault detection conclusion is displayed in the form of simple landmark map, that is, the section type of fault, accurate fault location, section landmark number, distance from front and rear landmarks, maintenance suggestions and other information. It can assign the line number, default wavelength and IOR parameter value to the newly added optical fiber line, and store the filled information in the database; It can update the parameter information of existing lines; Lines with landmark information cannot be deleted. It can add line landmark information to the existing optical fiber line, including the line number, landmark name, landmark type, landmark location, in reservation and out reservation of the landmark. It can automatically assign a new landmark number, support the addition of a new landmark before the existing landmark location, and realize the automatic calculation and update of the subsequent landmark location, The newly added landmarks can be correctly displayed in the landmark map: the line landmark information of existing optical fiber lines can be updated, including landmark name, landmark type, in reservation and out reservation. For the updated reserved landmarks, one can realize the automatic calculation and updating of subsequent landmark positions, The updated landmarks can be correctly displayed in the landmark map: the line landmark information of existing optical fiber lines can be deleted, and the reservation can be input and output according to the deleted landmarks, so as to realize the automatic calculation and update of subsequent landmark positions.

The detection event display can be realized, including the geographic location of the event, event type, event point loss, event point interval loss and other information. The test result curve can be displayed correctly, and the test curve can be enlarged and reduced arbitrarily; It can accurately locate on the detection curve and display the accurate position (accurate to m) and loss of the positioning point. The function of querying fault detection history can be realized according to single conditions. The conditions can be set as fault detection date, fault detection personnel, fault detection line name and fault handling status. The function of querying fault detection history can also be realized according to combined conditions. The conditions can be set as fault detection date, fault detection personnel The combination of fault detection line name and fault handling status. All fault detection personnel and fault handling status can be

selected. According to the detection history selected by the user, the detection records can be redisplayed, including the detection result curve and event list of the detection history in the main interface, and the comprehensive fault detection conclusion interface can also be displayed.

4   System Test

In order to improve the level of automatic maintenance and management in optical fiber communication network, improve the response ability of fault handling, and facilitate the operation and monitoring of maintenance personnel, the system designed in this paper adopts intuitive interactive interface and modular design, the specific simulation environment is as follows: the graphics card and memory are igame geforce RTX 3080 Vulcan 10 g and 16 GB respectively, the main frequency processor and operating system are Intel (R) core (TM) i7-2430 m @ 2.41ghz and windows respectively, and MATLAB 2017A is selected to complete the software programming. The simulation environment is shown in Fig. 3:
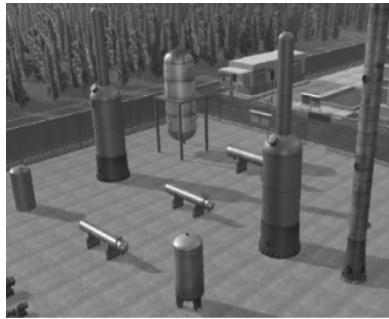
Fig. 3. Simulation experiment environment

According to the simulation experiment environment in Fig. 3 above, the experimental verification is carried out. The process and results are as follows.

4.1   Test Process

On the basis of completing the hardware and software of the dual-mode urban power optical fiber line on-line detection system, test the hardware and software of the system. The specific steps are as follows: first, use a multimeter to detect whether there is a circuit short circuit in the hardware, then power on the hardware, use a multimeter to test the power circuit voltage, and the test voltages are $+$ 24V, $+$ 12V, $+$ 5V respectively, $+$ Whether the output voltage of 3.3V and $-$5V circuits is normal. Based on the completion of power supply test. The software and Ethernet switching module are added to the hardware, and the software and hardware are connected according to the overall scheme of the system. The network transmission format in the system is shown in Table 2.

Table 2. Network transmission format

Frame header (HEX)	2byte
Lenth(hex)	1byte
Type(hex)	1byte
Mode(hex)	1byte
Data(hex)	nbyte
SUM(hex)	2byte

As shown in Table 2, bytes 0 and 1 are fixed data frame headers. Oxee starts with 0x55. The second byte (lenth) is the frame length. The third byte (type) is the function definition byte. The 4th byte (mode) is the working mode. The 5th to 4th + nth bytes (data) are the corresponding function parameters. The S + N and 6 + n bytes (sum) are checksum, overflow and discard the high bit. At this time, the normal operation of the hardware and software of the system can be guaranteed.

4.2   Test Results and Discussion

Under the above experimental environment, the system designed in this paper is compared with the traditional OTD online monitoring system. The experimental results are shown in Table 3.

Table 3. Experimental results

Number of experiments	Network monitoring time of traditional OTD online monitoring system / MS	The network of the system designed in this paperMonitoring time / MS
1	30	10
2	35	11
3	32	08
4	36	12

As shown in Table 3, the traditional OTD online monitoring system has a long network monitoring time, which directly affects the transmission effect of network data; The online monitoring system designed in this paper has shorter network monitoring time and better network data transmission effect, which is in line with the purpose of this paper.

In order to further verify the effectiveness of the method in this paper, the risk prediction of comprehensive intelligent training simulation of power communication transmission network is carried out, and the convergence curve of prediction output is obtained, as shown in Fig. 4.
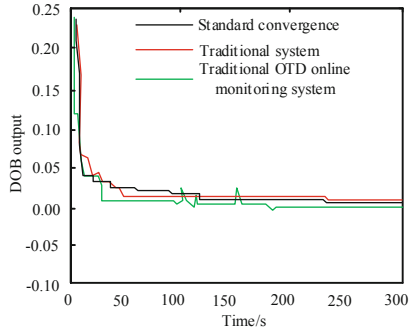
Fig. 4. Convergence curve of risk prediction of integrated intelligent training simulation for power communication transmission network

According to the analysis of Fig. 4, the convergence of risk prediction of comprehensive intelligent training simulation of power communication transmission network by this system is consistent with the actual convergence effect, while the convergence of risk prediction of comprehensive intelligent training simulation of power communication transmission network by traditional OTD online monitoring system is quite different from the actual convergence effect, It shows that the risk prediction effect of the system in this paper is better than that of the comprehensive intelligent training simulation of the power communication transmission network of the traditional system.

To sum up, the designed system has good performance.

5   Conclusion

This paper analyzes the current situation of on-line monitoring of power optical fiber transmission network, and completes the demand analysis of on-line monitoring system of power optical fiber transmission network based on Internet of things. The equipment environment in the substation is complex. How to avoid the interference factors in the substation is the key in the practical application of the scheme. This paper analyzes the interference sources and gives the corresponding network structure design to avoid interference. The on-line monitoring system of intelligent substation based on Internet of things is programmed. Although there is some research on on-line monitoring of power optical fiber transmission network based on Internet of things, it still faces many difficulties in practical application and has not been applied in practice. The next research work needs to fully consider the interference of electromagnetic signals in the substation and carry out anti-interference design through the actual operation in the substation. It is also necessary to optimize the software interface to further expand the functions.

References

1. Li, J., Zhang, A., Zhou, G., et al.: A large-core microstructure optical fiber for co-transmission of signal and power. J. Lightwave Technol. **4**(99), 1–10 (2021)

2. Fukai, C., Abe, Y., Takaya, M., et al.: Investigation into the influence of high-power optical transmission on fiber withdrawal from optical connector. J. Lightwave Technol. **38**(18), 5128–5135 (2020)

3. Kikuchi, N.: Multilevel signaling technology for increasing transmission capacity in high-speed short-distance optical fiber communication. IEICE Trans. Electron. **102**(4), 316–323 (2019)

4. Liu, Y., Li, Y,L.: Signal analysis and processing method of transmission optical fiber hydrogen sensors with multi-layer Pd-Y alloy films. Int. J. Hydrogen Energy **44**(49), 27151–27158 (2019)

5. Tong, H.T., et al.: Mid-infrared transmission by a tellurite hollow core optical fiber. Optics Express **27**(21), 30576–30588 (2019)

6. Naumov, A.V., et al.: Two-way transmission of time and frequency signals over optical fiber communication lines with the help of SATRE modems. Measurement Techniques **61**(10), 1009–1017 (2019)

7. Lin, T.C., Jun-Zhe, Y., Yu, C.S., et al.: Development of a transmission network fault location platform based on cloud computing and synchrophasors. IEEE Trans. Power Delivery **6**(99), 1–12 (2019)

8. Zhang, D., et al.: 4×4 MIMO fiber-wireless transmission based on an integrated four-channel directly modulated optical transceiver. Photonics Res. **7**(12), 101–107 (2019)

9. Lou, C., Yang, J., Li, T., et al.: New phase-changing soft open point and impacts on optimising unbalanced power distribution networks. IET Gener. Transm. Distrib. **14**(23), 5685–5696 (2020)

10. Yahav, I., Sheffi, N., Biofcic, Y., et al.: Multi-gigabit spatial-division multiplexing transmission over multicore plastic optical fiber. J. Lightwave Technol. **39**(8), 2296–2304 (2021)

11. Liu, S., Fu, W., He, L., Zhou, J., Ma, M.: Distribution of primary additional errors in fractal encoding method. Multimedia Tools and Appl. **76**(4), 5787–5802 (2014)

12. Liu, S., Chen, X., Li, Y., Cheng, X.: Micro-distortion detection of lidar scanning signals based on geometric analysis. Symmetry **11**, 1471 (2019)

13. Zhang, L.J., Liu, Z.L., Zhou, H.L., et al.: Simulation of repeated game theory intrusion detection model in mobile internet of things. Computer Simulation **36**(5), 337–340 (2019)

Research on Fault Monitoring Device of Highway Bridge Expansion Joint Based on Internet of Things Technology

Youjun Xu[(✉)]

College of Rail Transit, Nanjing Vocational Institute of Transport Technology, Nanjing 211188, China
xuyoujun569@yeah.net

Abstract. Among various large-scale engineering structures, highway bridge structures are characterized by a large number, long service period, and huge investment. Therefore, any construction quality, safety accidents, natural disasters or structural aging, insufficient bearing capacity, etc. may cause the bridge as a whole or Natural cumulative damage and accidental damage occurred in expansion joints of highway bridges, which affected the safety of bridge operation. Therefore, a research on a fault monitoring device for highway bridge expansion joints based on the Internet of Things technology is proposed. Design the fault signal acquisition equipment of the Internet of Things technology, first analyze the basic characteristics of the road bridge expansion joint fault, design a signal collector based on the Internet of Things technology, and design a signal extractor based on the Internet of Things technology. Design a data acquisition card for the failure monitoring of highway bridge expansion joints. Finally, through example analysis, it is proved that the detection accuracy of the highway bridge expansion joint failure monitoring device combined with the Internet of Things technology is very high, and the fault monitoring accuracy is good, which meets the basic requirements of highway bridge expansion joint failure monitoring.

Keywords: Internet of Things technology · Highway bridges · Expansion joints · Fault monitoring

1 Introduction

The Internet of Things refers to a variety of terminal equipment and facilities, in the private network or Internet environment, through wireless or wired means to achieve interconnection and intercommunication, adopt appropriate information security guarantee mechanisms, and provide personalized real-time remote control, remote maintenance and other management and Service function [1], to realize the integration of equipment management and control. The ZigBee technology is listed as one of the ten new technologies with the fastest development and the broadest market prospects in the world today, and one of the most critical technologies for Internet of Things applications." It is

a short-distance, low-complexity, low-power, Low data rate, low-cost two-way wireless communication technology [2]. Its Z-Stack protocol stack has a clear hierarchical structure, mainly composed of physical layer (PHY), medium access control layer (MAC), network layer (NWK) and Application layer (APL) composition. In the ZigBee network, users can easily define the role of each unit by compiling the above protocol.

In the process of applying the Internet of Things technology to the fault monitoring of highway bridge expansion joints, the development of the Internet of Things technology in the monitoring of road bridge expansion joint faults has been greatly restricted due to the difficulty of mining the operating data of the highway bridge. The reason why the operation data of road bridge is difficult to mine, the difficulty lies not in the technical level [3], but in the closed nature of the construction industry. In fact, judging from the current technical level, it is enough to realize the mining of highway and bridge data, but it is difficult to obtain the relevant data information, which causes many tasks to be carried out smoothly. At present, my country has gradually begun to be applied to highway bridge monitoring platforms, but because highway bridge developers have not disclosed the core operating data of road bridge, the failure warning function of expansion joints is still difficult to achieve smoothly. Therefore, the Internet of Things companies must work closely with highway bridge manufacturers to fully utilize the advantages of both parties, so that this problem can be effectively solved [4].

The failure of expansion joints of highway bridges brings great difficulties to the maintenance of highway bridges, and also brings many inconveniences to people's production and life. The monitoring device designed in this paper uses wireless sensing technology to realize precise positioning and intelligent monitoring of each node. The staff can easily understand the working status of highway bridges only through mobile phones, computers and other network terminals, reducing the investment of manpower and material resources., Which makes the fault monitoring more humane [5].

2   Design Fault Signal Acquisition Equipment for IoT Technology

2.1   Analyze the Basic Characteristics of the Failure of Highway Bridge Expansion Joints

Expansion joints of highway bridges are one of the important components of bridges. They are generally installed at the junction of the beam end and the road surface at the bridge abutment, between the two beam ends of the bridge beam body, and various expansion devices or joints at the junction of the bridge. The general term of structure, its main purpose is to meet the needs of bridge structure deformation and enable vehicles to pass the bridge comfortably and safely. The expansion joints of bridges are one of the most easily damaged components of the bridge structure [6]. Like the bridge deck system, they all bear the direct action of vehicle load and are exposed to the natural environment for a long time, so it is easy to cause The load-bearing system and anchoring system of the expansion joints are fatigued and damaged, and the repair and maintenance of the damaged expansion joints are difficult due to various reasons. The fatigue damage of expansion joints is the fall of the welded joints of embedded steel bars and anchoring steel bars, the fatigue cracking of concrete in the anchoring area, and the fatigue damage of the system including various forms of expansion joint devices. As shown in Fig. 1:

Fig. 1. Schematic diagram of highway bridge expansion joints

As the bridge continues to be used and the damage to the expansion joint device cannot be maintained and repaired in time, the damage to the expansion joint device continues to accumulate, which further increases the direct role of the vehicle in driving. Such a vicious circle not only accelerates The aging and damage of the expansion joint device will also directly affect the operational safety of the bridge structure.

A large number of studies have shown that long-span bridge structures are huge and easily affected by factors such as the natural environment and operating environment. Under the influence of environmental factors, such as wind load, temperature load, and vehicle load, etc. [7]. For a long-span bridge structure under normal conditions, the characteristic parameters of the bridge structure will float in a wide range. It is precisely because of this fluctuation that the bridge structure is affected by real local damage. The characteristic parameters and signals of the damage caused by the bridge structure are masked or submerged in such fluctuations. Therefore, based on the long-term health monitoring data of the bridge, an objective physical model is established to describe the relationship between the environmental conditions of the bridge and the structural damage characteristic parameters, and on this basis, the monitoring and damage of the overall state of the structure is established for "normalized environmental conditions" Early warning method, this is one of the main research issues of long-span bridge structural health monitoring technology for practical engineering applications.

2.2 Design a Signal Collector Based on the Internet of Things Technology

Aiming at the poor monitoring effect of abnormal signals of the original wireless robot communication fault monitoring device, the design of this communication fault monitoring device was launched. In this research, ZigBee technology is used to optimize the design principle of the original wireless robot communication fault monitoring device [8]. This technology has the characteristics of low power consumption, low cost, high safety and high capacity. Using this technology can effectively improve the poor signal monitoring effect of the original equipment. In order to ensure the effectiveness of this equipment design, the design framework of the communication fault monitoring device is set as shown in Fig. 2:
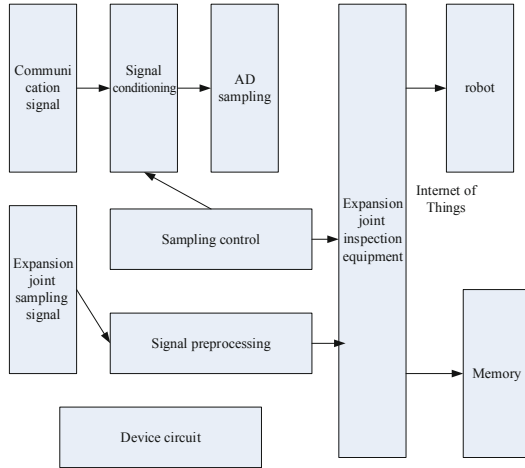
Fig. 2. Wireless device fault detection design framework

It can be seen from the above design framework that this device is mainly composed of a signal data acquisition part and a signal data analysis part. This device mainly uses a powerful ARM chip as the core of the design, and gradually realizes independent storage of signal data and signal data transmission, and realizes the task of detecting communication failures. The device is small in size, low in power consumption, and high in integration, which can effectively make up for the deficiencies of the original device.

In view of the characteristics of the Internet of Things technology, the communication between the signal data acquisition module of this device and the device is set through QTcpSocket. Therefore, in this design, an abnormal signal acquisition module suitable for the Internet of Things is set to realize the processing of abnormal signals. The robot integrated monitoring sensor is used to display the real-time communication status and data of the wireless robot in a unified device, and the data acquisition module designed this time is connected with the robot monitoring device. To ensure the effectiveness of this module design, set the hardware device parameters of this module as shown in Table 1:

Table 1. Signal collector parameter setting

Content	Parameter setting	content	Parameter setting
Signal driving	Digital signal	Interface Type	RS485
Interface form	Dry contact	Baud rate	1200–110 000 b/s
Form of protection	PPTC + TVS	Data bit	6
Protection level	600W	Stop bit	1
Overcurrent protection	30V/50 mA	Check Digit	even, odd
Collection form	1 kHz	Transmission distance	500 m

In this part, the LPC2131ARM microprocessor is used to realize the real-time acquisition and processing of robot signals. The device designed this time adopts the form of signal interruption and query to start collecting 00–11 of the signal sent by the robot. Develop the corresponding pre-processing of the collected signal, use wavelet transform to filter the signal data, and set it to a unified form to facilitate the subsequent processing and analysis of the signal data. To ensure the effect of signal acquisition, set the acquisition frequency of the acquisition device as follows:

$$y = a * b \tag{1}$$

In formula (1), y is the signal acquisition frequency, a is the robot communication signal transmission frequency, and b is the robot signal acquisition period. The frequency of signal acquisition is controlled by this formula. The memory interface of the monitoring device is set to a 64-bit width to improve the data storage capacity of the monitoring equipment, and the collected signals are uniformly transmitted to the memory of the monitoring device and processed in a unified manner.

2.3   Design a Signal Extractor Based on IoT Technology

The above-mentioned collected signals are preprocessed as data samples of abnormal communication signals, and the corresponding technology is used to complete the extraction of fault signals. In this part, by setting the form of the database analysis chip, the ability of data analysis and sorting is improved. This part adopts high-speed data computing chip as the carrier of equipment operation. Use the C-means clustering calculation method to complete the processing and analysis process of the fault signal [9], the algorithm is fast, simple, efficient for large datasets and scalable, set the signal of the device in the Internet of Things technology as the feature vector in the signal extraction process as y and set the signal data in the database as the signal analysis sample. $A = a_i$, $y = 1, 2, ..., n$, Set the signal category to be analyzed as z, y_i is z cluster center, and u_{iz} is the membership function of p signal, then the objective function of the communication signal classification is:

$$p_n = \sum_{i=1}^{z} \sum_{n=1}^{n} u_{iz} |a_i - y_i| \tag{2}$$

Set the elements in the above formula to be constant and greater than 1, then:

$$\sum_{n=1}^{n} u_{iz} = 1 \tag{3}$$

Setting the above formula as a constraint condition, the membership degree and clustering center of the collected signal can be expressed as:

$$u_{iz} = \frac{\left(1/|a_i - y_i|^2\right)^{1/(z-1)}}{\sum\limits_{n=1}^{z} \left(1/|a_i - y_i|^2\right)^{1/(z-1)}} \tag{4}$$

Through the above formula, the central membership degree of the signal data sample can be obtained, and the signal sample can be divided into one or more data types. In this part of the design, the typical characteristics of communication signal failure types are set as the cluster center. When a data sample belongs to multiple communication data characteristics at the same time, this signal data is a failure signal. In order to ensure the effectiveness of clustering, the robot communication fault category is pre-selected and set in this chip.

3   Design a Data Acquisition Card for Fault Monitoring of Highway Bridge Expansion Joints

According to the output voltage, data acquisition channel and data analysis requirements of the current transmitter of the highway bridge expansion joint fault monitoring device, the PCI-6034E multi-function data acquisition card (DAQ). The PCI-6034E multifunctional data acquisition card has a resolution of up to 16 bits, with 16 single-ended or 8 differential analog input ports, and a sampling rate of up to 200 kS/s, which can meet the resolution requirements of data acquisition and data analysis; The read and write speed of the hard disk can be up to 200 kS/s, which can ensure the simultaneity of information collection and information processing; it can meet the requirements of voltage signal collection and fault detection and positioning accuracy of signal collectors and signal extractors: compatible with real-time System, Linux, Mac OS and Windows operating systems, integrated LabVIEW, LabWindows/CVI and Measurement Studio software development environment for Visual Basic and Visual Studio.NET, provide a platform for the monitoring device to implement fault detection and location algorithms; highly integrated It has a minimum volume specification, which can meet the requirements of industrial computer design integration.

In addition, the PCI-6034E multifunctional data acquisition card has a sample storage capacity of 512 [10], a single-channel current drive capacity of 24 mA, a maximum signal source frequency of 20 MHz, and a digital trigger and synchronization bus (RTSI). The performance provides a good hardware condition for the non-intrusive low-voltage power failure monitoring device.

GPRS DTU can realize the transparent transmission of serial device data through the GPRS wireless network, that is, just like a mobile phone, just insert the GPRS-enabled

SIM card and simply set the parameters, then the data of the external serial device can be transparently transmitted to the public network fixed domain name or IP At the same time, it can also receive feedback commands from the server on the host computer, which is especially suitable for the transmission of multiple points, multiple points, inconvenient wiring, and real-time data requirements. It is especially suitable for the transmission of multi-point scattered, inconvenient wiring, and real-time data transmission. However, the installation location of the device in this project is not convenient for wiring and connecting to the network, so the use of GPRS DTU communication equipment can meet the needs of this article [11–13]. The process of the data card transferring data to the software through the device is shown in Fig. 3:
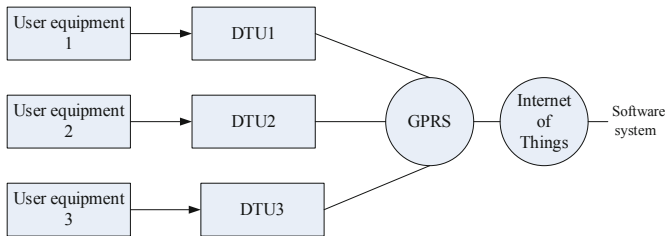
Fig. 3. The process of the data card transferring data through the device to the software

Considering the requirements of the acquisition card, the communication module uses the IQ1000 GPRS DTU communication equipment developed by Xiamen Lingqi Communication Co., Ltd. LQ1000 GPRS DTU provides a standard RS232/485 data interface, and has a built-in industrial-grade GPRS wireless module, which can be easily connected to RTU, PLC, industrial computer and other equipment, and only needs to complete the initial configuration at a time, and the user equipment can pass through the data center The GPRS wireless network establishes a connection and realizes the fully transparent transmission of data.

4   Case Analysis

In order to verify whether the road bridge expansion joint failure monitoring device based on Internet of Things technology designed in this article can meet the requirements of road bridge expansion joint failure monitoring, an example analysis method is designed, and the device designed in this article is used in actual detection to observe the monitoring effect. Verify the superiority of the method designed in this paper.

4.1   Project Overview

The span structure of Bridge A selected in the example is arranged as (south bank) approach bridge (5 × 40 m) + main span suspension bridge (820 m) + (north bank) approach bridge (6 × 40 m + 30 m). The bridge level is level 4, and the design safety level of the extra-lane expressway bridge is I level; the design load level is highway-I level, the design driving speed is 80 km/h; the crowd load: 2.5 kN/m2.

The bridge is a single-span steel box girder suspension bridge with a main span of 820 m, with a central buckle in the middle of the span; the sag-span ratio of the main cable in the completed state is 1/10, and the center-to-center spacing of the main cable is 29.1m; bridge deck width It is 29.78 m, and its layout is 29.78 m = sling anchorage area 0.84 m + sidewalk 1.78 m + anti-collision guardrail 0.52 m + carriageway 10.75 m + central divider 2.0 m + carriageway 10.75 m + collision guardrail 0.52 m + sidewalk 1.78 m + sling anchorage area is 0.84 m.

Under the action of temperature change, the long-span bridge structure is affected by the structural deformation and internal force caused by the temperature change. The research and analysis of the bridge temperature distribution through the health monitoring system can provide the original basis and data for the bridge design considering the influence of temperature, and the health monitoring system can provide temperature compensation and correction reference for the original data of other sensors. A study of temperature effects. Analyze and compare the operating conditions of bridges under the action of temperature changes, such as bridge structural disturbance changes, displacement changes, and component stress distribution changes, to verify and improve bridge design theory accurately and timely under the influence of actual temperature changes. Carrying out a safety assessment is of positive significance.

The highway bridge expansion joint failure monitoring device uses a structural temperature sensor to monitor and measure the structure temperature, and uses a temperature and humidity sensor to monitor and measure the temperature and humidity of the atmospheric environment. The structure temperature measurement points and data acquisition sequence are shown in Table 2:

The expansion joints are mainly set at the junction of the side span and the main span. Due to the large external environment effects such as vehicle load and temperature load, the Nanxi Yangtze River Bridge adopts a unit type multi-directional displacement bridge expansion device (RBQF1600). Expansion joint device is aimed at the problems and deficiencies of traditional modular expansion devices and comb expansion devices on large bridges such as suspension bridges and cable-stayed bridges. It has multi-directional displacement functions such as lateral displacement, longitudinal displacement and torsion in large bridges. And the new generation of bridge expansion devices developed under the influence of vehicle load on the expansion device itself. The service performance of this new type of expansion joint device still needs to be verified in actual engineering operations.

The monitoring of vehicle load is through the measurement of traffic flow, vehicle wheelbase, vehicle speed, vehicle weight, etc. of the operating bridge, and the measurement of these data is mainly carried out through vehicle speed and axle meters. By monitoring the vehicle load, it can provide data for the establishment of the vehicle load model of the bridge, and can provide a basis for evaluating the structural state. In addition, the analysis of the vehicle load spectrum can provide load parameters for the formulation of the fatigue load spectrum and the analysis of the bridge structure.

4.2 Analysis of Application Results

In this section, the influence of the change of ambient temperature on the temperature of the beam section of a bridge suspension bridge under normal operating conditions. In

Table 2. Structure temperature measurement points and data acquisition sequence

Collection location	Storage order	Measuring point number	Location
South Bank Approach Bridge	1	B05-DTH-001	Upstream
	2	B05-DTH-002	Upstream
	3	B05-DTH-003	Upstream
	4	B05-DTH-004	Upstream
	5	B05-DTH-005	Upstream
Main span suspension bridge	6	B05-DTH-006	Midstream
	7	B05-DTH-007	Midstream
	8	B05-DTH-008	Midstream
	9	B05-DTH-009	Midstream
	10	B05-DTH-0010	Midstream
North Shore Approach Bridge	11	B05-DTH-0011	Downstream
	12	B05-DTH-0012	Downstream
	13	B05-DTH-0013	Downstream
	14	B05-DTH-0014	Downstream
	15	B05-DTH-0015	Downstream

order to investigate the effect of the environmental temperature change of the bridge on the temperature of the beam section in the four seasons of the year, one day was selected in each of the four seasons, namely October 30, 2013 (autumn), January 16, 2b14 (winter), On April 16, 2014 (spring) and August 2, 2014 (summer), we analyzed the temperature data records of these four days. It should be noted that the structure temperature data is calculated with 1-min as the calculation interval, and the average value of all temperature sensors shown in Fig. 3 is taken as the effective value, then 1440 structure temperature data can be recorded every day. The structural temperature of the top, bottom, and web of the bridge is the average value of the sensors at their respective positions. The comparison chart of ambient temperature monitoring and actual temperature is shown in Fig. 4:

The experimental results shown in Fig. 4 show that the actual temperature detection result of the device in this paper is roughly consistent with the actual temperature change, indicating that the device designed in this paper has a high accuracy in detecting the temperature environment of the expansion joint failure. Temperature is a crucial factor for the expansion joints of Guiping Road Bridge. The change of temperature is directly linked to the failure of the expansion joints. Therefore, the accurate detection of temperature is very important for the failure monitoring of the expansion joints of highway bridges.
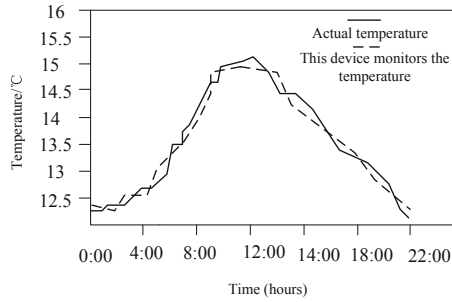
Fig. 4. Experimental results

5 Conclusion

This paper designs a fault monitoring device for highway bridge expansion joints based on Internet of Things technology. This model contains two main devices. The company's models and algorithms have been verified by on-site data, the fault signal acquisition equipment based on the Internet of Things technology is designed, the signal collector based on the Internet of Things technology is designed, and the signal extractor based on the Internet of Things technology is designed. Design for failure monitoring of highway bridge expansion joints Data acquisition card. Finally, through example analysis, it is proved that the detection accuracy of the highway bridge expansion joint failure monitoring device combined with the Internet of Things technology is very high, which meets the basic requirements of highway bridge expansion joint failure monitoring.

References

1. Yang, L., Wang, G., Qin, S., Wang, Q.: A study on reconstruction algorithm of multi-bearing vibration signal based on DCS. Computer Simulation **36**(5), 316–319,324 (2019)
2. Liu, Y., Xiao, F.: Intelligent monitoring system of residential environment based on cloud computing and internet of things. IEEE Access, PP(99), pp. 1–1 (2021)
3. Zeng, J., Peng, C., He, P., Liu, C.: Research on current monitoring and fault diagnosis technology for control rod drive mechanism. Hedongli Gongcheng/Nuclear Power Eng. **40**(1), 172–175 (2019)
4. Gao, D., Wang, L.: An implementation of intelligent substation monitoring system based on internet of things. J. Comput. Methods in Sciences Eng. **19**(S1), 219–226 (2019)
5. Naman, H.A., Hussien, N.A., Al-Dabag, M.L., Alrikabi, H.: "Monitoring the consuming of electrical energy based on the internet of things applications" has been accepted for publishing after double blind review. Int. J. Interactive Mobile Technologies (iJIM) **15**(2), 172–183 (2021)
6. Ahouandjinou, M.H., Medenou, D., Pecchia, L., Houessouvo, R.C., Jossou, T.R.: Modeling an integrated network for remote patient monitoring, based on the internet of things for a more preventive and predictive health system in West Africa. Global Clinical Eng. J. **3**(2), 19–31 (2020)

7. Irawan, Y., Wahyuni, R., Muzawi, R., Muhardi, M., Hamzah, M.L.: Real time system monitoring and analysis-based internet of things (iot) technology in measuring outdoor air quality. Int. J. Interactive Mobile Technologies (iJIM) **15**(10), 224–240 (2021)

8. Qin, L., Xie, Y.: Real-time monitoring system of exercise status based on internet of health things using safety architecture model. IEEE Access **9**, 2733–27345 (2021)

9. Musyafa, A., Abadi, I., Noriyati, R.D., Mukromin, R., Asy'Ari, M.K.: Design and implementation monitoring system based internet of things (iot) on battery charging - photovoltaic power plant using flc. Int. J. Mechanical Mechatronics Eng. **20**(4), 22–30 (2020)

10. Kareem, O.K., Adekitan, A.I., Awelewa, A.: Power distribution system fault monitoring device for supply networks in nigeria. Int. J. Electrical Computer Eng. **9**(4), 2803–2812 (2019)

11. Liu, S., Liu, D., Muhammad, K., Ding, W.: Effective Template Update Mechanism in Visual Tracking with Background Clutter. neurocomputing, online first (2020). https://doi.org/10.1016/j.neucom.2019.12.143

12. 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)

13. Liu, S., Li, Z., Zhang, Y., Cheng, X.: Introduction of key problems in long-distance learning and training. Mobile Networks Appl. **24**(1), 1–4 (2018)

Research on Security Access Technology of Power Internet of Things Gateway Equipment Based on Artificial Intelligence

Yanming Li[1(✉)], Ying Fan[2], and Siyao Xu[2]

[1] Guangdong Power Grid Co., Ltd., Guangzhou 510030, China
lymflyingfish@163.com
[2] Electric Power Research Institute of Electrical Guangdong Power Grid Co., Ltd., Guangzhou 510080, China

Abstract. Aiming at the problem that the security access of traditional power Internet of things gateway equipment is vulnerable to network attack, resulting in the decline of security, a security access technology of power Internet of things gateway equipment based on artificial intelligence is proposed. Analyze the functional requirements of power Internet of things equipment, and carry out data protection processing of gateway equipment. Based on artificial intelligence technology, network attack characteristics are extracted to realize the secure access of power Internet of things gateway equipment. By means of comparative experiment, it is verified that the safety performance of the new technology is better and has great popularization value.

Keywords: Artificial intelligence · Power Internet of things · Gateway equipment · Secure access technology

1 Introduction

With the wide popularization of the Internet and the deepening of network applications, people have been used to using the services provided by the network to participate in various network activities, especially e-government and e-commerce [1]. Because the sensitive information stored and processed on the network is increasing day by day, network security management has become the primary problem to be solved in the computer network gateway equipment. Traditional network security management technologies include firewall, intrusion detection, security audit, network monitoring, security evaluation, authentication and authorization [2]. Reference [3] proposed a secure and universal wireless communication solution for the distribution Internet of things in smart grid. This paper mainly studies the secure ubiquitous wireless communication solution of distribution network Internet of things (pd_iot) in smart grid. The detailed topology of secure universal wireless communication network is given, and the integrated encryption and communication equipment is developed. The scheme supports a variety of state secret encryption algorithms, including SM1 / SM2 / SM3 / SM4 and forward

and reverse isolation functions, so as to realize PD_ Secure wireless communication of Internet of things services. With the emergence of new network mode, especially the emergence of distributed gateway devices, the traditional network security management methods gradually show the following shortcomings: at present, the commonly used authentication mechanisms are based on user identity, which is known, but in large-scale and open distributed gateway devices, users are not necessarily familiar with gateway devices.

The traditional security mechanism has no delegation mechanism, but in the distributed gateway device, the delegation mechanism can improve the flexibility of the gateway device and reduce the management workload of the gateway device. Traditional security mechanisms cannot handle new access conditions and restrictions, many security policy elements cannot be described directly, and their expressibility and scalability are poor [4]. There are multiple management domains in large-scale distributed gateway devices, and different management domains should adopt different security mechanisms, which can not enforce unified policy and trust relationship, while the current security mechanism can not manage domains. The traditional security mechanism makes the server fully realize access control, which increases the burden of the server, and the security is limited by the security of the server itself. Once the security of the server fails, the whole access control policy will not work [5].

In order to solve the above problems, trust management came into being. Trust management is a security management mechanism suitable for large-scale, open distributed gateway devices. Compared with the traditional security management mechanism, it has the characteristics of flexibility, reliability and scalability. Therefore, trust management is a new stage in the development of network security management. Intrusion detection and trust management play a "mainstay" role in network security management [6]. This paper studies the security access technology of power Internet of things gateway equipment based on artificial intelligence. By analyzing the functional requirements of power Internet of things equipment, the data protection processing of gateway equipment is carried out. Based on artificial intelligence technology, network attack characteristics are extracted to realize the safe access of power Internet of things gateway equipment. This method has more network throughput, can effectively shorten the response time and ensure better security performance.

2 Design of Secure Access Technology for Power Internet of Things Gateway Equipment Based on Artificial Intelligence

2.1 Analyze the Functional Requirements of Power IoT Devices

According to the objectives of network users, its production mode is mainly for multiple different factories distributed in different regions to jointly carry out the production process of products, which involves important links such as production commissioning and storage of products. These links have certain restrictions on the requirements of the on-site environment. Therefore, in the general control (regional headquarters) area, it is necessary to monitor the environment and production process, respond to emergencies for the first time, reduce unit losses, and ensure the safety of production products [4].

According to the goal of the network user unit [7], the network user unit has deployed different types of sensor acquisition equipment in the production workshop and warehouse (generally unattended workshop or warehouse) in its plant area, and deployed a gateway equipment that can collect relevant data and preprocess at a certain level in several areas (generally 3–5 production workshops or warehouse), after the gateway equipment is connected to the main network of the plant area and incorporated into the public network, the preprocessed data is transmitted to the regional headquarters. After the regional headquarters service network optimizes the data through its own algorithm and graphically processes it, it is presented to managers or leaders in a visual way for them to make early warning and decision-making in case of emergencies [8]. Therefore, network user units put forward special considerations on data security. Its overall framework is shown in Fig. 1.
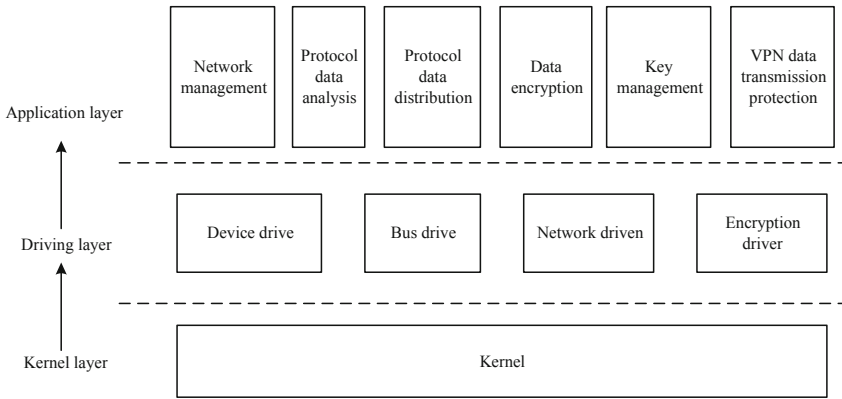
Fig. 1. Overall framework

As shown in Fig. 1, according to the overall framework diagram, the gateway device mainly completes the network kernel, data encryption and decryption, network management, and VPN data transmission protection. From the overall structure, in order to achieve network controllability and strong compatibility, the kernel still uses a tailored Linux kernel with a version of 3.18.17 [9]. The driver layer includes common devices, bus drivers and network drivers, which includes a variety of heterogeneous network protocol stacks to facilitate the analysis and distribution of protocol data. The encryption driver mainly provides a communication method (USB2.0) with the on-board hardware encryption chip (localized chip with encryption engine). When called by the application layer, the driver interacts with the hardware encryption chip to realize encryption and decryption. Function. The development of application layer adopts C + +, and the development of drive adopts C language.

According to the above description, it can be seen that the main functions of the gateway device include network management function, protocol data analysis and distribution, data encryption and decryption, key management and data transmission protection. At the same time, there are certain performance requirements for data encryption and decryption and data transmission protection [10]. Therefore, it is necessary to design the

function first, complete the module corresponding to the function and the relationship between these function modules, and then design the corresponding business process in the follow-up. The device initialization module will initialize network parameters, load protocol types, and initialize various functional modules and interfaces according to the current network environment. The user obtains the corresponding network management authority through authentication login (USBKey + password), and adjusts and sets the network strategy of the device through a friendly interface. The network management module determines the mode of key generation, the configuration type of the VPN function, and the type of data filtering according to the strategy generated by the user [11]. When the externally collected data enters the device, it will first enter the data analysis and transceiver module, and then screen the data according to the packet filtering strategy set by the network management module. After that, the data will be sent to the data encryption and decryption module, while the VPN module obtains the network quintuple of the data and other information, and establishes a dedicated transmission channel in real time. When the data enters the encryption and decryption module, the module will generate and select the key according to the key management module, and perform encryption and decryption operations on the data. The encrypted and decrypted data is sent to the data analysis and transceiver module, and the module pushes the data to the destination according to the dedicated channel established by the VPN management module. After the entire transmission process is over, the VPN management module must remove the established dedicated transmission channel, the key management module will clear the key used for this transmission, and the network management module will clear the internal data cache without retaining any data information.

According to the division of functional modules and the relationship between modules, the software of the whole gateway equipment can be divided into several key processes. The whole gateway equipment software achieves the purpose of cooperative work through the interaction of these processes [12]. These key processes include equipment initialization, user authentication and login, authority control, firmware upgrade, factory setting recovery, key management, security policy management, protocol data analysis and distribution, data protection, status monitoring, etc. The design of each process will be described in detail below. The gateway device startup is mainly divided into two stages. The first stage is the bootloader startup and completes the inspection of the hardware interface or peripherals, such as wireless network interface, wired network interface, USB interface, hardware encryption and decryption chip, etc. [13]. Once it is found that the interface or peripherals cannot be started, the entire device initialization process will be stopped, and the fault indicator will be lit to remind the user or management personnel to repair. If the peripherals or interfaces are successfully started, the boot program will start the kernel and transfer the control of the device to the kernel. After the kernel obtains the disposal right, it will load the kernel modules, such as the algorithm interface module responsible for encryption and decryption, and the interface module for data transmission and reception and the gateway device management module responsible for management, etc., and automatically configure parameters for the startup of the services of these modules or processes, but when the module is found to be faulty and cannot be effectively started or configured, the kernel will stop the device start the

initialization process, and light up the fault indicator to remind users or managers to repair the equipment.

The gateway device logs in to the device by providing wired Ethernet network and internal serial port. If you log in to the device from a wired Ethernet network, a web-based interface method is provided to log in. If you log in from the internal serial port, you will be provided with a printed character interface to log in. At the same time, the internal serial port is generally not open to the outside world and is mainly used by relevant technical personnel for production, debugging, and maintenance. The user logging in to the device is mainly to query, configure and audit the services of the gateway device. When you need to log in to the device, you need to first insert the USBKey into the USB-Host of the device and enter the password. If the device detects that the USBKey is not inserted, the user will not be able to log in to the device. The gateway device provides an SM2-based identity authentication method, so a password is required. The device uses the password as a factor to generate an SM2-based key pair, and performs two-way authentication of user identity together with the key pair pre-prepared in the Key. And return the authentication result. If the authentication result is wrong, it proves that the user password does not match the Key, and the user authentication login fails. If the authentication is successful, the key matches the user's identity. At this time, the device will read in the initialization information of the hidden area in the key, confirm the user's identity (management user, audit user, ordinary user), and assign permissions to it. In this way, the safety performance of the equipment can be initially guaranteed.

2.2 Perform Gateway Device Data Protection Processing

Data protection mainly provides a combination of privacy protection (password protection) and channel protection. First, the gateway device receives the collected data periodically reported by each node, and then classifies and preprocesses the data according to the strategy, and encapsulates the data according to the protocol (application protocol) negotiated with the management center, and transmits it after calling the transmission module. The transmission module will first consider encrypting the data after receiving the data, and establish a VPN transmission channel for this business, and report the encapsulated and encrypted data to the management center through the channel. After the service is completed, the VPN channel will be closed, and real-time connection to the VPN channel is not supported. After receiving the data, the management center obtains the actual node collection data through decryption, classifies it, and stores it in its database. When the collected data is abnormal, the management center will send an alarm to the administrator by means of short messages. After the device is started, it will automatically start a set of monitoring service processes, responsible for regularly collecting the running status of the gateway device, including the CPU running status of the device, the running status of the gateway device process, the network connection status, the packet filtering interception status, and the peripheral access status, the relevant information of whether the nodes under the gateway are operating normally or not, and report it to the management center, which is convenient for the management center to audit. In addition, in case of an emergency, such as abnormal network connection and the addition of an unauthorized authentication node, the gateway equipment will first send an alarm locally through audible and visual means. At the same time, it will report

the alarm information to the management center according to the preset configuration strategy. After obtaining the alarm information, the management center will alarm the administrator through short messages. The external interface of the gateway device is shown in Fig. 2.
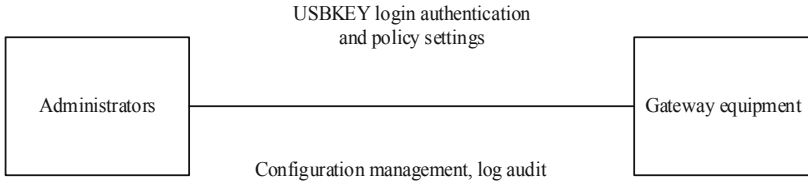
Fig. 2. External interface of gateway equipment

As shown in Fig. 2, an external interface diagram is provided for the gateway device. The external interface of the gateway device is mainly for the administrator. The administrator needs to perform login authentication, policy setting, network configuration, log audit and other functions on the device. It mainly involves the management of gateway equipment. The device provides a unified login interface based on artificial intelligence. The functions mentioned in the above figure are integrated in the intelligent interface. The administrator needs to use the host to connect to the management network port and log in to the device through USBKey to carry out relevant management operations. Login authentication. After the administrator inserts the USBKey and enters the password according to the prompts, the interface is responsible for identity authentication and returns the authentication result. Policy settings. After logging in, the administrator enters the policy setting interface and configures relevant information on the interface, including firewall settings, IP packet filtering principles, supported IoT protocols, encryption methods, algorithms, etc. After that, the interface is responsible for generating configuration files and activating policies. Network settings. After the administrator logs in, enter the network setting interface. Configure device network parameters, including local address, node ID, background management center destination address, etc., and then the interface is responsible for generating configuration and activating. Log audit. This interface is responsible for collecting the policy execution status and network data filtering and forwarding alarm information of the local device for nearly a week, and presents it to the front-end interface after the administrator enters the log audit interface.

The internal interface mainly refers to the interface between the various modules inside the gateway device. The device is initialized. The first process executed after the gateway device is started needs to check the network configuration, whether the security policy configuration is normal, the service interface related to the gateway device, whether the main external hardware interface is normal, etc., and the device background service process will be started in order. Data reception. Data processing process call. The interface establishes two sets of threads. These two groups of sub-processes will receive data in two directions, one is from the management center to the gateway device, and the other is from the node to the gateway device, and respectively open up receiving resource pools for the data in the two directions, waiting for the data analysis and forwarding module to process. Data transmission. Data processing process call. The

processed data is distributed in two directions. This interface establishes two groups of sending threads, and extracts data from the processed data sending queue and distributes it to the corresponding destination address. Intelligent management. The main process is called this interface. This interface is responsible for analyzing the administrator's configuration of the machine, including network configuration and security configuration, and after optimizing the configuration data, it is distributed to each execution module to complete activation. At the same time, this interface is also responsible for HTTPS service management. Key generation. The main process calls the interface. After the interface is called, it is responsible for interacting with the key management gateway device, and finally generates a key that can be used for encryption or authentication. SM2 certification. The gateway device management process is called. This interface is responsible for completing the integrity check of the received data and the two-way authentication of the user identity when the administrator logs in. Login authentication. The gateway device management process is called. After the administrator inserts the USBKey and enters the password according to the prompts, the interface is responsible for identity authentication and returns the authentication result. Protocol analysis. Data analysis and forwarding process call. This interface is responsible for the analysis and load stripping of the IoT protocol of the node.

Data confidentiality protection uses encryption algorithm to encrypt user data. Even if unauthorized users get the data, they can't know the content. The gateway equipment supports a variety of encryption algorithms in line with international standards, and also supports encryption algorithms independently developed and approved by the national competent department. At the same time, various encryption algorithms are implemented efficiently, which not only ensures the security of data, but also ensures the efficient processing of gateway equipment. In the design of gateway equipment, confidentiality protection is mainly applied in business data transmission protection, firmware upgrade, data packet transmission and so on. Integrity protection data is not modified by illegal users during transmission, or if a data packet transmitted on the network is modified by an illegal user, the receiver of the data can find that this is an illegally modified data packet. The non-repudiation of the data is used to prevent the user from refusing to acknowledge that a specific data packet has been sent. The processing method of this gateway device is to add a special message digest to the related IP datagram and encrypt the digest with a private key. After the data is transmitted to the destination, the corresponding public key is used to decrypt the digest, and the decrypted data is compared with the digest of the original data. If the comparison results are consistent, it can be determined that the data packet was sent by the user. The reason is that the encrypted private key is only held by the corresponding user, so it is effective to use this method to solve the non-repudiation of the data. The user specifies which network communication needs what kind of security protection. This is the VPN security strategy. The security policy selects data packets according to the source address, destination address, transport layer protocol, port, data transmission direction [14], etc. of the data packet, and encrypts, clears, and discards the selected data packet according to the needs of users. Security policy management is mainly embodied in the use of a central distribution mechanism based on the encryption and protection of configuration data in the gateway design, and the authority control mechanism used in the local configuration management of the gateway device. Through

the above security protection mechanism, the access of the IoT gateway device is more securely protected.

2.3   Extracting Network Attack Characteristics Based on Artificial Intelligence

Artificial intelligence is a branch of computer science. It is a science that studies machine intelligence, that is, using artificial methods and technologies to develop intelligent machines or intelligent gateway devices to imitate, extend, and expand human intelligent behavior. Since the birth of the Dartmouth Conference in 1956, artificial intelligence has made gratifying progress on bumpy roads, especially in machine learning, data mining, computer vision, expert gateway equipment, natural language processing, pattern recognition and robotics and other fields. In order to extract the characteristics of network attacks more truthfully, this paper conducts an intelligent analysis on them as follows.

$$\mu_{X_i}(x_l) = \mu_{il} = \begin{cases} 1, x_l \in X_i \\ 0, x_l \notin X_i \end{cases} \tag{1}$$

$$M_h = \{\mu_{il} | \mu_{il} \in \{0, 1\} \tag{2}$$

$$M_f = \left\{ \mu_{il} | \mu_{il} \in \{0, 1\}, \sum_{j=1}^{k} \mu_{il} = 1 \right\} \tag{3}$$

$$\hat{M}_f = \frac{M_h}{\sqrt{\mu_{X_i}(x_l)}} \tag{4}$$

In formula (1–4), $\mu_{X_i}(x_l)$ is the network sample subset; μ_{il} is the function set of the network sample; x_l is the membership function; X_i is the feature vector; M_h is the unit function; M_f is the unit basis vector; $\hat{M}_f$ is the normalization the unit basis vector to be processed. From that we get:

$$D = (x_l - p_i)^T (x_l - p_i) \tag{5}$$

$$(x_l - p_i) = \frac{(x_l - p_i)^T (x_l - p_i)}{D^2} \tag{6}$$

$$C = \sum_{i=1}^{k} D \tag{7}$$

$$U = \sum_{i=1}^{k} \mu_{il}^m \tag{8}$$

$$D(x_l - p_i) = \sum_{i=1}^{k} \sum_{l=1}^{n} U \tag{9}$$

$$P^* = \{p_i | 1 \le i \le k\} \tag{10}$$

In Eq. (5–10), D is the dissimilarity measure; p_i is the optimal intelligent analysis effect; T, m, n and k are constants; C is the objective function of network attack; U is the optimal intelligent analysis prototype; $D(x_l - p_i)$ is distortion; P^* is a fuzzy function.

This technology is mainly aimed at the processing of gateway equipment after upgrade failure. After the upgrade fails, the most ideal state is to monitor the startup process when the gateway device boots, and automatically select whether the fallback version is required through the state judgment. The key technical point is how to supervise the whole startup process, because the control of CPU will be transferred and the life cycle of boot software will end. By establishing the start state base in an address segment of the memory, the automatic fallback of the version can be realized to a certain extent. After the boot software is started, read the status base recorded in the previous startup process to judge whether version fallback is required. In addition, the timing of rollback is very important. The following situations need to be considered, which may cause misoperation in the version rollback process. During equipment startup, the hardware is powered off. This situation will make the state base unable to update, triggering version fallback. When there is no ups (standby power supply), the CPU hardware register cannot record any status or fill in the status base after power failure. After repowering up, it is difficult to judge whether the startup failure is due to hardware power failure or upgrade failure. After power on again, the boot software will still perform version fallback, resulting in misoperation. The upgraded file is incorrect, but it will not affect the basic operation of the operation gateway device. After the upgrade, the configuration of the software has problems, which will affect the business operation, but will not affect the normal startup and operation of the operation gateway equipment. Such errors will not cause the software to report errors or exit abnormally, and it is difficult for the boot software or kernel to be notified. Finally, the temporal and spatial randomness of such errors is strong, so it is difficult to monitor them by fixed inspection methods. Therefore, the boot software will not judge such errors as upgrade failure, and will continue to load the software gateway device with business problems. To sum up, the determination of the fallback time of gateway equipment is a complex and comprehensive determination process. At present, there is no very perfect fully automatic fallback scheme in the field of operating gateway equipment.

Based on artificial intelligence, this paper extracts the characteristics of network attack, which can maximize the security access effect of gateway equipment and ensure the normal operation of equipment.

2.4 Realize the Secure Access of Power IoT Gateway Equipment

After completing the detailed design of the Internet of things gateway software, the corresponding verification and testing work is carried out on the gateway equipment to verify whether the gateway equipment meets the needs of practical application. The verification and test work is carried out in the corresponding test environment built by using the actually developed gateway equipment. The purpose of the test is to verify the compliance of the gateway equipment with the requirements in the actual application scenario and whether it meets the design security. The testing work includes two parts: verification of management and business functions and verification and testing of security. The first part of the function verification work is as follows: the verification of

the gateway device management function and service function, such as device initialization, user login, authority control, firmware upgrade, verifies the function design and interface design of the gateway software; The verification of protocol data analysis and distribution function verifies the basic business functions of Internet of things gateway equipment. The second part of the test work is as follows: the national secret algorithm encryption and decryption function verification, encryption and decryption performance test, key management function verification and data transmission protection function verification of the gateway equipment. Verify whether the actual gateway equipment meets the expected security design requirements. This paper will introduce the environment, process and conclusion of verification and testing in detail.

The administrator opens the management interface, enters the device IP address, and the gateway device checks whether the device has been initialized, prompts the user and starts the initialization operation. Select the standard that the device follows and execute the next step. The gateway device prompts the user to confirm the standard. Choose the authentication method of the administrator: if the administrator chooses to use the key authentication, check whether the user inserts the USBKey, otherwise prompt the user to insert the USBKey; then enter the PIN code of the USBKey; verify the correctness of the PIN code, the wrong PIN code prompts the user, when the number of PIN code errors exceeds 5 times, the USBKey is locked, the gateway device verifies the validity of the USBKey, and the illegal key refuses to generate a certificate request and prompts the user; the gateway device administrator, security administrator, and security auditor are initialized in turn. The gateway device prompts that the initialization is complete and records the log, and enters the login interface.

The administrator enters the gateway device login interface. Enter the administrator name, password, verification code or USBKey + PIN. The gateway device verifies the legitimacy of the administrator. An illegal user refuses the next operation and prompts the user. If the number of consecutive incorrect password entries exceeds the set threshold, the user will be locked. The gateway device needs to convert the IoT protocol of the perception layer into a standard Ethernet frame and send it to the application layer service gateway device. The standard command issued by the service gateway device can be converted into the IoT protocol format and forwarded to the perception layer node. The verification of this function is mainly done with the help of a user protocol analyzer. According to the above method, the secure access of the gateway device is realized.

3   Experiment and Analysis

In order to verify whether the technology designed in this article has the effect of use, this article conducts experiments on the above methods. The experimental process and results are shown below.

3.1   Experimental Process

This experiment first collects network node data, and then uses the method designed in this article for data processing, and applies the obtained data to the experiment. At this time, the internal interface of the gateway device is shown in Table 1.

Table 1. Internal interface description

Interface	Name	Describe
wst_dev_init()	Device initialization	The first process executed after startup needs to check the network configuration, whether the security policy configuration is normal, the service interface related to the gateway device, whether the external main hardware interface is normal, etc. and the device background service process will be started in order
wst_data_receive()	Data reception	Data processing process call. The interface establishes two sets of threads. These two groups of sub-processes will receive data in two directions, one is from the management center to the gateway device, and the other is from the node to the gateway device, and respectively open up receiving resource pools for the data in the two directions, waiting for the data analysis and forwarding module to process
wst_data_send()	Data sending	Data processing process call. The processed data is distributed in two directions. The interface establishes two groups of sending threads, and extracts data from the processed data sending queue, and distributes it to the corresponding destination address
wst_sm4_verify()	VPN management	Data analysis and forwarding process call. Before sending data, a VPN tunnel needs to be established. This interface is responsible for establishing the tunnel and closing the tunnel after the service is completed

As shown in Table 1, data confidentiality protection uses encryption algorithms to encrypt user data, even if unauthorized users get the data, they cannot know the content. The gateway equipment supports a variety of network encryption effects that comply with international standards, and also supports self-developed encryption algorithms approved by the national competent authority. At the same time, various network encryptions are also efficiently implemented, which not only guarantees the security of data, but also guarantees the efficient processing of the system. At this time, the resource configuration of the gateway device is shown in Table 2.

As shown in Table 2, for the resource configuration of the gateway device, because the role of the Internet of Things gateway is to connect the perception layer network and the Internet network as an intermediary, it must adapt to various perception layer network protocols, convert data between different protocols, and quickly connect enter the Internet network to control and manage various sensors. Therefore, under this configuration condition, the security of the gateway device can be guaranteed.

Table 2. Gateway device resource configuration

Resources name	Quantity	Model
IoT gateway	2	Operating System: Linux Network port: 5, of which 4 are switching ports
Switch	1	Huawei Quidway S5700
Network tester	1	Spirent Test Center 9000
Test PC	2	Pentium(R)Dual-CoreE5300/2 g/500G Windows XP, IE8

3.2 Experimental Results and Discussion

Under the above experimental environment, the network security access technology designed in this paper is tested. The network security access can be detected by the network packet loss rate and response time. Therefore, this paper tests the above two indicators, and the test results are shown in Table 3.

Table 3. Test results of throughput

Throughput(Mbps)	Reference [3] technical network throughput	Technical network throughput designed in this paper
No package	93.976	98.961
Pause time between two private rooms 10 ms	93.732	99.895
Pause time between two private rooms 1 ms	93.601	99.253
Full speed contracting	32.536	18.426

The meanings of the four evaluation indicators in Table 3 are, respectively, the throughput when no packets are sent, the throughput when the pause is 10 ms, the throughput when the pause is 1 ms, and the throughput when the packets are sent at full speed. As shown in Table 3, through the four test environments of No package, 10ms pause time between two packets, 1ms pause time between two packets, and Full speed contracting, the reference [3] technology has less network throughput, below 94.0, when Full speed contracting, the throughput is 32.536, the throughput effect is poor, and it does not meet the security requirements of gateway equipment; Under the same conditions, the technology designed in this article has a large network throughput, above 98.0, when Full speed contracting, the throughput is 18.426, and the throughput is better, which meets the security requirements of gateway equipment. In addition, this article tests the security response time of the network, and the test results are shown in Table 4.

As shown in Table 4, in the four test environments of No package, 10ms pause time between two packets, 1ms pause time between two packets, and Full speed contracting,

Table 4. Test results of response time

Response time (s)	Reference [3] technology network security response time/s	This paper designs the technical network security response time/s
No package	0.851	0.085
Pause time between two private rooms 10 ms	0.854	0.085
Pause time between two private rooms 1 ms	0.855	0.085
Full speed contracting	2.459	0.814

the reference [3] technology network security response time is longer. Above 0.8 s, Full speed contracting. When the network security response time is 2.459, the network security effect is not good; Under the same conditions, the technology designed in this paper has a short network security response time. Within 0.1s, when Full speed contracting, the network security response time is 0.814, and the network security effect is better. Through the above test, this paper compares the number of attacks on the reference [3] technology gateway device network with the number of attacks on the technology gateway device network designed in this paper. The experimental results are shown in Table 5.

Table 5. Experimental results

The amount of data	The number of attacks/times of the reference [3] technology gateway device network	The number of attacks/times of the technical gateway device network designed in this article
1000	248	24
2000	556	35
3000	827	42
4000	1516	51
5000	1612	95
6000	2745	106
7000	3871	132
8000	4987	163

As shown in Table 5, under the same experimental environment, this article conducted 8 experiments, and obtained the reference [3] technology gateway equipment under the conditions of the data volume of 1000, 2000, 3000, 4000, 5000, 6000, 7000, and 8000. The network has been attacked more frequently. As the amount of data increases, the number of attacks will also increase, and the network security performance is poor; and under the same conditions, the technical gateway device designed in this article has fewer attacks on the network, and the network is safe. High performance. This meets the research purpose of this article.

4   Conclusion

This paper studies the security access technology of power Internet of things gateway equipment based on artificial intelligence. By analyzing the functional requirements of power Internet of things equipment, the data protection processing of gateway equipment is carried out, so as to effectively ensure the security of network being attacked. Based on artificial intelligence technology, network attack characteristics are extracted to realize the safe access of power Internet of things gateway equipment. This method has more network throughput and shorter network security response time. This article analyzes the structure of the gateway. The role of the Internet of Things gateway is to connect the perception layer network and the Internet network as an intermediary. It must adapt to various perception layer network protocols, convert data between different protocols, and quickly access the Internet network. Sensors for control and management. In order to ensure the integrity of the gateway device of the Internet of Things, it can also provide various peripheral interfaces for different types of devices. The dynamic loading module is used to ensure that the software module can be dynamically loaded under different environmental requirements. It is applied in actual scenarios, so that the IoT gateway device has the characteristics of flexibility, generality, and scalability. At the same time, the gateway studied in this paper uses hardware encryption technology and VPN technology to protect the data and improve the integrity and reliability of the data.

References

1. Zhan, K.: Design of computer network security defense system based on artificial intelligence and neural network. J. Intelligent and Fuzzy Syst. **9**, 1–13 (2021)
2. Chu, M., Song, Y.: Analysis of network security and privacy security based on AI in IOT environment. In: 2021 IEEE 4th International Conference on Information Systems and Computer Aided Education (ICISCAE). IEEE, 390–393 (2021)
3. Chen, L., Suo, S., Kuang, X., et al.: Secure ubiquitous wireless communication solution for power distribution internet of things in smart grid. In: 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE). IEEE, 780–784 (2021)
4. Kou, G., Wang, S., Tang, G.: Research on key technologies of network security situational awareness for attack tracking prediction. Chin. J. Electron. **28**(01), 166–175 (2019)

5. Chemouil, P., Hui, P., Kellerer, W., et al.: Special issue on artificial intelligence and machine learning for networking and communications. IEEE J. Selected Areas in Communications **37**(6), 1185–1191 (2019)
6. Padmaja, M., Shitharth, S., Prasuna, K., et al.: Grow of artificial intelligence to challenge security in IoT application. Wireless Personal Communications, pp. 1–17 (2021)
7. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mobile Networks Appl. **24**(1), 5–17 (2018)
8. Li, G.: DeSVig: decentralized swift vigilance against adversarial attacks in industrial artificial intelligence systems. IEEE Trans. Industrial Informatics **16**(5), 3267–327 (2019)
9. Zhou, C., Liu, Q., Zeng, R.: Novel defense schemes for artificial intelligence deployed in edge computing environment. Wirel. Commun. Mob. Comput. **2020**(8), 1–20 (2020)
10. Sikora, P., Malina, L., Kiac, M., et al.: Artificial intelligence-based surveillance system for railway crossing traffic. IEEE Sensors J. **21**(14), 15515–15526 (2020)
11. Zhang, Z., Yang, Y.: Design of remote monitoring system for a mechanical equipment based on internet of things. In: 2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC). IEEE, pp. 54–59 (2021)
12. Foubert, B., Mitton, N.: Lightweight network interface selection for reliable communications in multi-technologies wireless sensor networks. In: 2021 17th International Conference on the Design of Reliable Communication Networks (DRCN). IEEE, pp. 1–6 (2021)
13. Namasudra, S., Chakraborty, R., Kadry, S., et al.: FAST: fast accessing scheme for data transmission in cloud computing. Peer-to-Peer Networking Appl. **14**(4), 2430–2442 (2021)
14. Zhang, J., Hou, X.: Simulation of layered filtering method for multi-channel false data in sensor networks. Computer Simulation **37**(02), 339–342,364 (2020)

Construction of a Gateway Boundary Security Protection Platform Based on the Internet of Things and Cloud Computing

Chen Cheng[✉], Siyao Xu, Mingyang Peng, Ziying Zhang, and Yan Li

Electric Power Research Institute of Electrical Guangdong Power Grid Co., Ltd.,
Guangzhou 510080, China
cchustdky@163.com

Abstract. In view of the problems of long warning time and poor protection effect of traditional gateway boundary security protection platform, a gateway boundary security protection platform based on Internet of things and cloud computing is designed. Net FPGA chip is used for the verification and development of network communication equipment, connecting the ATA serial port connection line port of multiple boards. Combined with the register host computer, the read-write operation of the registers inside each module in the hardware is completed through PCI bus, and the hardware design of the gateway boundary security protection platform is completed. Establish the gateway border security protection module and complete the software design of the gateway border security protection platform. Based on the Internet of things and cloud computing technology, match the network security link, so as to realize the security protection of the network boundary. The experimental results show that the security protection effect of the platform constructed in this paper is better, and can effectively shorten the security early warning time.

Keywords: Internet of Things · Cloud computing · Gateway boundary · Security protection platform

1 Introduction

Network boundary refers to the boundary between our network and other networks. At present, the widely used network boundary security device is firewall [1]. The firewall can control the traffic in and out of the network boundary according to the preset security policy, but the firewall lacks the analysis and detection function of the traffic in and out. Although the intrusion detection platform proposed later makes up for this deficiency, the intrusion detection platform is a passive security device. It only analyzes whether the incoming and outgoing traffic contains attack messages, but can not process the messages containing attack information. Corresponding measures can be taken only when the network administrator sends the alarm information of the intrusion detection platform, at this time, the intruder may have already completed the network intrusion

[2]. Network boundary security protection platform is an intelligent security platform, which can not only detect the occurrence of intrusion, but also stop the occurrence and development of intrusion in real time through closed-loop response, so as to protect the information platform from substantive attacks [3].

Network boundary security protection platform is an active and active intrusion prevention and blocking platform. It can detect and intercept intrusion activities and aggressive network traffic in real time to avoid any loss [4]. The platform is deployed at the boundary of the network. When an attack attempt is detected, it will automatically throw away the attack packet or take measures to block the attack source. Although the network boundary security protection platform is similar to Intrusion Detection System (IDS) and firewall in some aspects, it is a new security technology integrating detection and access control. The analysis and detection function of the network boundary security protection platform is similar to that of IDS, but it is connected to the network in series, and the detection methods and strategies are adjusted according to the special requirements of protection, balancing the characteristics of false positives and false positives [5, 6]. Reference [7] proposed to build a zero trust security protection system in the environment of power Internet of things. The characteristics of boundary protection model commonly used in network security protection are analyzed. Aiming at the problem of insufficient security protection ability of the model, a security protection model of power Internet of things network based on zero trust security architecture is proposed. The application of zero trust in power Internet of things is analyzed and studied. This method has certain effectiveness, but the safety protection effect is poor. Reference [8] proposed an endogenous security protection framework suitable for 5G MEC in power industry. According to the characteristics of power 5G MEC, this paper proposes an endogenous security protection framework suitable for power 5G MEC, which can effectively resist various security threats, but the security early warning time is long.

To solve the above problems, this paper constructs a gateway border security protection platform based on Internet of things and cloud computing. This paper designs the hardware of gateway boundary security protection platform through net FPGA chip and register. Based on the Internet of things and cloud computing technology, the gateway boundary security protection module is established, and the gateway boundary security protection platform software is designed to match the network security link, so as to realize the security protection of the network boundary. The platform can effectively improve network security, shorten the security early warning time, and provide conditions for the development of network security.

The Internet of Things (IoT) is a variety of terminal perception devices with some perception, processing and control capabilities installed in real life entities. It uses the network to complete information interaction, processing and coordination in order to achieve large-scale data exchange between things and people, things and things. Its ultimate goal is to use the network to complete the mutual communication and communication between people and things, things and things, and all objects, so as to facilitate identification, control and management [9]. Its related technologies have considerable development prospects in national defense and military industry, industrial control, public facilities, medical assistance, smart grids, smart cities, smart transportation, and

environmental monitoring. Cloud computing is a delivery and usage model of software/hardware services. Cloud computing service providers use the network to provide software/hardware resources as services to users in accordance with the user's platform requirements. Among them, "cloud" serves as the software/hardware resources on the server cluster on the network, such as hardware resources (central processing unit, Storage, Servicer, etc.) and software resources (Application Program, operating platform or compilation platform, etc.). The client device only needs to after sending the request, the cloud computing platform can integrate the software/hardware resources on the Internet of Things to provide corresponding services and return the final calculation results to the client device. In this way, the client device can get far beyond its own computing power [10]. In recent decades, due to the emergence of IoT sensing technology and cloud computing technology, it is possible to install a variety of terminal sensing devices with recognition capabilities, processing capabilities, and control capabilities in physical entities, which will sense data through the network. Transfer to the server for storage, identification, analysis, management and control. For all walks of life, there are many types and numbers of terminal sensing devices, and massive amounts of sensing data are collected. Computing these massive amounts of data requires computing devices with huge processing capabilities. Cloud computing, as the basis for the Internet of Things technology to complete massive data processing, provides processing support for the massive information data collected by the Internet of Things platform [11]. There are many possibilities and problems to be solved in the combination of Internet of things and cloud computing. Scale: scale is a prerequisite for the combination of cloud computing and the Internet of things. Only when the scale of the Internet of things is large enough can it be combined with cloud computing, such as industry applications: smart grid, seismic network monitoring and so on. For general, local and home network IOT applications, it is not necessary to combine cloud computing. How to make them develop to the corresponding scale remains to be solved.

2   Hardware Design

2.1   Net FPGA Chip

Net FPGA chip is mainly used for verification and development of network communication equipment, and it has abundant resources for researchers to develop high-speed Ethernet equipment. There are four RJ45 high-speed Ethernet ports, supporting Cat5E or Cat6 standard cables, two 2.25MB SRAM, two 32MB DDR2 DRAMs, and two ATA serial ports that can connect multiple boards. The user is the logic core, an international test port, and the IP core of the PCI interface is solidified in the chip. The chip parameters are shown in Table 1.

As shown in Table 1, Net FPGA chips have high-speed and large-capacity peripheral storage space and programmable Gigabit physical network interfaces, and provide reference design projects such as four-port network cards, routers and switches [12]. Description language engineering package. The design code in the engineering package configures the main chip through the PCI interface after the integrated wiring, so as to realize the designed platform function. Among them, stanfor's DEMO has three

Table 1. Chip parameters

Resources name	Remark	Quantity
Slices	Logic piece	23616
SliceFlipFlops	Deposit	47232
4 input LUT	Input lookup table	47232
Bonded IOBs	External IO	692
BRAMs	RAM on chip	232
GCLKs	Global clock	16
DCMs	Clock management	8

projects, namely four-port network card and gigabit switches and Linux-based hardware-accelerated routers with simple routing protocol (PW-OSPF), these three projects fully reflect the idea of modular design, that is, develop within the same framework and use the same data and register pipeline [13], the reference design only needs to re-develop individual modules to achieve different platform functions. This unified basic design framework is also the design basis for the subsequent development of Net FPGA chips. The platform designed in this article is designed with reference to the router project as the basic framework. Among them, the physical resources of the Net FPGA chip have four physical ports, corresponding to four Gigabit Ethernet ports, which are called MAC0MAC3 in platform engineering. The other four CPU queues, the CPU0CPU3 queue ports correspond to the device drivers of the operating platform, and they mainly realize the transfer of data packets to the host computer through the PCI bus. In the User_data_path module, the core processing module of the platform, the Input_arbiter module completes the round-robin reception of data packets in eight different queues. The Output_port_lookup module is mainly responsible for determining the output port of the transmitted data packet. The Output_queue module is responsible for the round-robin transmission of data packets in eight different queues. The principle is shown in Fig. 1.
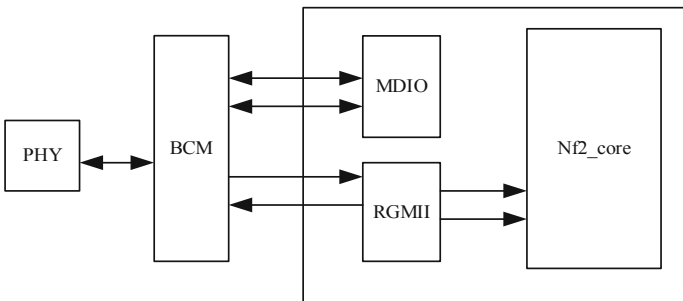
Fig. 1. Chip principle

As shown in Fig. 1, in the physical layer interface module, in order to reduce the internal clock of the platform, an RGMII interface module is added inside the FPGA. This module combines the original 4-bit network data of the network into 8-bit data and then goes to the next stage. Module transfer, MDIO module is responsible for the configuration of Gigabit Ethernet data input and output functions. The platform configures the Net FPGA chip and the internal registers of the platform realizes the programming of the external chip, so as to achieve the control of the data transmission function.

2.2 Register

The function of the register pipeline is to enable the host computer to complete the read and write operations of the internal registers of each module in the hardware through the PCI bus, so as to realize the functional configuration of the hardware module or read the relevant data information of the hardware module. The internal registers of the hardware module are divided into two types in the design of this platform: The first type is software registers, which can be read and written by the upper computer, while the hardware modules can only read operations on them. This type of register is mainly used for module functions. The second type is the hardware register, which can be read and written by the hardware, while the software can only perform read operations on it. This type of register functions mainly to realize the hardware counting function and is mainly used to record the flow information through the module. The register interfaces are all bidirectional bus signals, which can simultaneously receive the register data signal from the previous module and complete the backward transfer of the register signal of this module. In terms of register configuration, the Net FPGA project provides users with upper computer device driver functions, including hardware register read and write operation functions (regread, regwrite). Developers can use these two functions to implement hardware registers. The entire operation process of the read and write access operation is as follows: first, the user sends out a read and write signal, the addressing bit width of the read and write signal is 27 bits, and 32 bits can be read or written each time, and then the data passes through the operating platform the device driver function converts the user request into a register pipeline signal and sends it to the Net FPGA chip. Each module of the register is transferred until the register module with the corresponding address is found. After the write operation is completed, the data is sent back to the operating platform along the same path.

3 Software Design

3.1 Establish a Gateway Border Security Protection Module

In the platform designed in this paper, besides the hardware design, the software design is very important. Based on this, this paper designs a security protection module for the gateway boundary. The working principle of the simple message filtering module is to judge whether the network message is allowed to enter and leave the network by checking and comparing some information in the header of the network message, without

checking the content of the network message [14]. The simple message filtering module uses the information in the header of TCP / IP protocol message, such as protocol type, source IP address, destination IP address, TC or UDP port, to determine whether the network message can pass through the network boundary. A simple message filtering function is embedded in the filter table of Netfilter in Linux 2.4 kernel. A simple packet filtering module can block some unsafe network access at the network boundary.

Stateful message filtering module. Connection tracking (CONNTRACK) is mainly used to track and record the connection state, and realize the transition between states, and it also saves the communication information of each session. CONNTRACK does not register any rule table, indicating that it does not need rules to determine whether to do connection tracking. The connection here does not only refer to the TCP protocol connection, it also includes the UDP protocol and the ICMP protocol. Of course, this is only the protocol included in the standard implementation of the kernel. The connection trace of other protocols can be added to the kernel itself. Connection tracking is the basis of address translation. This module must be loaded when using the address translation function. The platform only processes the first message of a connection (TCP or UDP), and then relies on the connection tracking mechanism to complete the processing of subsequent messages. Connection tracking is a mechanism that can be used in conjunction with NAT to process actions related to higher-level protocols at the transport layer (or even the application layer). There are many components in CONTRACK to handle TCP, UDP or ICMP protocols. These modules extract detailed and unique information from network messages, so they can keep track of every data stream. This information also determines the current state of the CONNTRACK stream. For example, UDP streams are generally uniquely determined by their destination address, source address, destination port, and source port. In the control table, the message is related to the four different states of the traced connection. They are NEW, ESTABLISHED, RELATED, INVALID, the following is the description of these states. NEW—It means that this message is the first message of a connection. ESTABLISHED-Indicates that a connection has message transmission in both directions. The connection in the ESTABLISHED state is very easy to understand. As long as the response is sent and received, the connection is in the ESTABLISHED state. To change a connection from NEW to ESTABLISHED, it only needs to receive a response message. Using the above state information to filter network messages, the state-based filtering function can be realized. This will make the network border protection platform very strong and effective. For example, when there is no state mechanism, it is often necessary to open all ports above 1024 to release the response data. Using the state mechanism, you can only open those ports that have response data, and all other ports can be closed, which enhances the function of platform protection.

The network address translation (NAT) module is used to realize the conversion between internal network addresses and public network addresses, and the NAT table of Iptables implements this function. This module can perform one-to-one, one-to-many, and many-to-many network address conversion tasks for network messages. NAT was originally a solution proposed to solve the shortage of P address space, but it hides the internal network topology from the outside, forming an invisible boundary between the internal network and the external network, so that external hosts cannot actively access the network. Internal nodes, thereby improving the security of the internal network.

The protocols supported by the NAT module include IP, ICMP, UDP, TCP, etc. There are 3 types of NAT: static address translation, dynamic address translation, and port address translation. The NAT module in the Netfilter framework contains three built-in hook chains: NF_P_LOCAL_OUT, NF_IP_PRE_ROUTINGNF_IP_POST_ROUTING. The NAT table supports the following NAT types: SNAT: Change the source address of the data packet. DNAT: Change the destination address of the data packet. M.ASQUERADE: Belongs to a special form of SNAT. Realize the P address camouflage function. REDIRECT: A special form of DNAT. It is used to change the destination IP address of a qualified network message to the I address of the network interface when the network message enters the platform.

The analysis and detection module obtains network messages from the network protocol stack and analyzes whether there is an attack event. Analyze the content and characteristics of the attack through techniques such as feature matching, traffic analysis, protocol analysis, and session reconstruction. The results obtained through analysis and detection are submitted to the response module, and the response module executes the corresponding response action, and submits the message containing the attack information, the analysis result of the attack event, and the response strategy to the log platform for storage, so that the network administrator can view it afterwards.

The function of the closed-loop response module is to judge whether the network message passes through the network boundary in real time according to the detection results of the analysis and detection module. Because there are some false alarms in the analysis and detection module, in order to reduce the impact on the normal network operation, the response of the closed-loop response module must be classified. This platform is divided into three categories. One is to only send out alarm (Alert) information, the second is to drop the message when it is more dangerous, and the third is to send a response message when it is the most dangerous. Such as Tcp-reset packet to cut off the connection (Reject). Each specific detection rule controls whether to perform a closed-loop response. The specific method is to use the first keyword of the rule to distinguish. If the keyword in this part is Drop, it means that the rule is matched and the report is discarded. The text and send out the alarm information; Alert means that the rule is matched and only send out the alarm message; Reject means that the rule is matched and send out a response message and send out the alarm message. The closed-loop response module is implemented by the Target module in Netfilter. As the analysis and detection module is Netfilter's Match module, Netfilter calls the Target module to respond to the message according to the detection result of the analysis and detection module.

The analysis and detection module generates corresponding alarm information when a network attack is detected. These alarm messages are generated in the Linux kernel, and they need to be transmitted to the user space through a certain mechanism, and then the user space daemon processes the alarm messages. The usual practice for these alarm information is to store these alarm information in the log file of the platform so that the network administrator can view and analyze the alarm information. There are many ways to exchange information between the Linux kernel and the user space. This platform uses the Linux Netlink mechanism to complete the transmission of the alarm information of the analysis and detection module.

After adding the relevant kernel module options to the kernel, the Iptables software used in the user space must provide the relevant command line options. The analysis and detection module in this platform is implemented in the way of Match module expansion, so the program used in user space must also be provided. In order to make each extension module use a version of Iptables software without having to write a specific software version of the related extension, Iptables adopts a shared library to solve this problem. The shared library has the function of init (), which is similar to the function of the kernel module. It is called automatically when loading. This function calls register according to the new match added_match(), a shared library can provide the functions of initializing data structures and providing related options. The main data structure used in writing shared libraries is iptables_ Match, which is passed to register as a parameter_match() registers the relevant command-line matching options to let iptables recognize the new match. Through the establishment of the protection module of the platform, the safety performance of the platform can be initially guaranteed.

3.2 Matching Network Security Links Based on the Internet of Things and Cloud Computing

Cloud computing enables computer system resources, especially storage and computing capabilities, to be provided on demand without the need for users to directly manage them actively. Cloud computing is usually used to indicate that many users can use the data center through the Internet of Things. The large clouds that dominate today have the ability to distribute from a central server to multiple locations. If the connection to the user is relatively close, it can be designated as an edge server. The cloud can only serve a single organization (private cloud), it can serve many organizations (public cloud) or a combination of the two (hybrid cloud). Therefore, this article combines cloud computing with the Internet of Things to create a secure protection link for the network boundary.

In order to better improve the security of the network, this paper uses the network similarity measure for cloud computing, and uses the Euclidean distance to unify the calculation results, as shown below:

$$d = \sum_{i=1}^{N} \sqrt{(X_{1i} - X_{2i})^2} \tag{1}$$

$$A = \sum_{p} \left\| I_1^p - I_2^p \right\| \tag{2}$$

$$d' = \max(|x_1 - x_2|, |y_1 - y_2|) \tag{3}$$

$$J = \frac{dx_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}} \tag{4}$$

In formula (1–4), d is the Euclidean distance from the network information transmitting end X_{1i} to the receiving end X_{2i}; N is a constant; A is a unified vector parameter; I_1^p and I_2^p are different boundary coordinates; p is the attribute measurement index; x_1,

x_2, y_1 and y_2 are coordinate values of different network nodes; d' is the distance after unification. J is the included angle of space vector; x_i and y_i are network node similarity indicators. Taking $d(i, j)$ as the network boundary node, the ratio of network security is as follows:

$$D = \frac{\min_{1 \leq i \leq j} d(i, j)}{\max_{1 \leq k} d'(k)} \tag{5}$$

$$\delta_i = \min d_{ij} \tag{6}$$

In formula (5–6), δ_i is the network sensitive value; i and j are two network nodes, and the ratio of the maximum and minimum values is the network security ratio. In order to strengthen the real effect of cloud computing, this article calculates the Rand index of the network:

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

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

$$RI = P + R \tag{9}$$

$$RI' = RI + P + R \tag{10}$$

In formula (7–10), P is the accuracy rate; TP is the number of real nodes; FP is the number of false positive nodes; FN is the number of false negative examples; R is the recall rate; RI is the similarity measure of network nodes; RI' is the desired measure of network security.

3.3   Realize the Security Protection of the Network Boundary

In order to realize the security protection of the network boundary, this paper designs the above methods. The platform is composed of many sub-parts, and each part has its own specific function, and the combination of its functions. The platform has many elements, and each element is related to each other, showing a clear purpose, level, and being able to continuously adjust itself and adapt to the environment. Platform protection is a structured multidisciplinary approach used to define the concepts and requirements of complex and multi-faceted issues, including concepts, implementation and verification. Platform protection uses mathematics, physics and related scientific disciplines, as well as the principles and methods of protection design and analysis to formulate, predict and evaluate the vulnerability of the platform to security threats. The application fields of platform protection are very wide, including platform security in risky industries such as mining protection and tunnel protection. By combining multiple disciplines, analyzing and designing a platform suitable for the protection, the various parts of the platform can be optimized and controlled. The working principle of platform protection makes it have the characteristics of integration, integration and optimization. As an important object

of platform security protection, platform protection is essentially the use of a series of platform protection methods to assess whether the platform is safe, and organize suitable solutions for unsafe factors to avoid dangers, and minimize the hidden dangers in the platform. Specifically, the main purpose of platform security protection is to identify, minimize and control vulnerabilities and related risks. It is a multidisciplinary work that uses scientific principles, principles and methods to identify and assess vulnerabilities. Security analysis is the most important part of platform security protection. Only when security issues are analyzed comprehensively, correctly, and accurately, can we make feasible plans for security issues in the platform and avoid risks. For the analysis of platform security protection, different levels of analysis can be determined according to the needs of the project, including preliminary analysis, detailed analysis, etc. Analyze the existing security risks, and make scientific and reasonable predictions for each issue using platform security methods and principles.

4   Platform Test

In order to verify whether the platform designed in this article has practical effects, this article tests the above design. The test process and results are shown below.

4.1   Test Process

Connect the hardware and software of the platform to check whether there is a short circuit or short circuit. At this time, the hardware parameters of the protection platform are shown in Table 2.

Table 2. Hardware parameters

Parameter	Is it dangerous	Filter results
And	No	-
Exec	Yes	Success
Insert	Yes	Success
Select	Yes	Success
Delete	Yes	Success
Update	No	-
Count	No	-

As shown in Table 2, through the addition of different parameters, the degree of risk is relatively compared. The platform designed in this paper can filter the dangerous parameters, and the hardware operation effect is better at this time. After the hardware debugging is completed, this article debugs the software, and this article adds related cases, as shown in Table 3.

Table 3. Software test

Case study	Problem	Test effect
Physical security	Divide the area for management, set up standby power supply equipment in different areas, and the platform operation effect	Run successfully
Cyber security	Allow or deny the network access of portable and mobile devices according to the security policy. Terminate the network connection after the session is inactive for a certain period of time or after the session ends	The network is successfully connected or disconnected
Platform security	Use two or more combinations of authentication technologies for the same user to achieve user identity authentication; set sensitive marks for important information resources and all subjects who access important information resources	Mark the main resource successfully

As shown in Table 3, in terms of physical security, network security, platform security, etc. [15], the test results are all successful, indicating that the platform's software is operating normally. After the hardware and software are debugged, the login interface of the platform is shown in Fig. 2.

As shown in Fig. 2, after the hardware and software are debugged, the platform runs successfully at this time.

4.2 Test Results

In the above-mentioned test environment, this paper tests the performance of the two-way SM1 at the gateway boundary, and the test results are shown in Table 4.

As shown in Table 4, the performance results of two-way transmission of SM1 can be seen, the performance can reach 11.465 when the large packet of 1024 bytes is transmitted, the performance is only 1.109 when the small packet of 64 bytes is transmitted, and the performance is good when the large packet of 1518 bytes is transmitted. It reaches 9.031, so it can meet the encrypted transmission requirements of IoT gateways. Based on this, this paper tests the two-way SM4 at the gateway boundary, and the test results are shown in Table 5.

As shown in Table 5, during the SM4 performance test, the lntended Load, Offered Load, Throughput and other indicators can be kept consistent. When the 1518 bytes large packet is transmitted, the performance is 8.412. At this time, the security effect of network data transmission better, therefore, can meet the security of the gateway boundary. Through the above test environment, the platform designed in this paper is compared with the traditional PB security protection platform and AP security protection platform. The test results are shown in Table 6.

Table 6. Comparison results of safety warning time of different methods

Number of experiments	Security warning time of traditional PB security protection platform/ms	Security warning time of traditional AP security protection platform/ms	The safety warning time of the safety protection platform designed in this paper/ms
1	0.11	0.15	0.01
2	0.23	0.24	0.03
3	0.26	0.31	0.04
4	0.28	0.36	0.05
5	0.30	0.42	0.05
6	0.31	0.44	0.05
7	0.35	0.46	0.05
8	0.36	0.48	0.05

As shown in Table 6, the security warning time of the traditional PB platform and AP platform is longer, both above 0.1ms, the response is slow, the protection effect is poor, and it does not meet the security protection requirements of the IoT gateway boundary. The platform designed in this paper has a shorter warning time, the longest warning time is 0.05ms, the response is faster, and the protection effect is better. It can meet the security protection requirements of the IoT gateway boundary and is of great promotion value.

5   Conclusion

With the continuous development of Internet technology and mature as well as the increase in Internet equipment, Internet of things has gradually spread to every aspect of People's Daily lives, and security is a big problem in the daily life can not be neglected. Therefore, many researchers gradually began to combine the Internet of things and security early warning theory, make full use of their respective advantages. It is widely used in transportation, industrial manufacturing, medical and health, disaster emergency

response and other fields. Therefore, an in-depth study of the security early warning system in the context of the Internet of Things is of great significance to all areas of people's daily production and life. The intelligent system, efficient management, and convenient monitoring it brings will also produce significant economic benefits. This article starts from the perspective of the security early warning system architecture of the Internet of things, based on the Internet of Things system architecture, first expounds and studies the three-layer network architecture of the perception layer, network layer and application layer, in order to meet the requirements of reducing delays and increasing transmission processing speeds. Improve the demand for system response, introduce cloud computing methods, and establish a cloud-side collaborative IoT security protection platform architecture, aiming to improve the security protection effect of the network and create conditions for the development of the Internet.

References

1. Vajjha, H., Sushma, P.: Techniques and limitations in securing the log files to enhance network security and monitoring. Solid State Technol. **64**(2), 1–8 (2021)
2. Snehi, J., Bhandari, A., Snehi, M., et al.: Global intrusion detection environments and platform for anomaly-based intrusion detection systems. In: Proceedings of Second International Conference on Computing, Communications, and Cyber-Security. Springer, Singapore, pp. 817–831 (2021)
3. Zhao, S.: Simulation of scheduling fault tolerant control of big data cluster for security monitoring of cloud platform. Computer Simulation **38**(7), 486–490 (2021)
4. Sun, Q.Y., Liu, X.J., Sun, Y.M., et al.: A security wireless monitoring and automatic protection system for CCEL. Wirel. Commun. Mob. Comput. **2021**(1), 1–14 (2021)
5. Ma, H., He, J., Liu, Y., et al.: Security-driven placement and routing tools for electromagnetic side channel protection. IEEE Trans.on Computer-Aided Design Integrated Circuits Syst. **40**(6), 1077–1089 (2020)
6. Wang, C., Yu, L., Chang, H., et al.: Application research of file fingerprint identification detection based on a network security protection system. Wireless Commun. Mobile Comput. 1–14 (2020)
7. Zeng, R., Li, N., Zhou, X., et al.: Building a zero-trust security protection system in the environment of the power Internet of Things. In: 2021 2nd International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT). IEEE, pp. 557-560 (2021).
8. Xuesong, H., Wei, L., Tao, Z., et al.: An endogenous security protection framework adapted to 5G MEC in power industry. In: 2021 China Automation Congress (CAC). IEEE, pp. 5155–5159 (2021)
9. Miki, T., Nagata, M., Sonoda, H., et al.: Si-backside protection circuits against physical security attacks on flip-chip devices. IEEE J. Solid-State Circuits **55**(10), 2747–2755 (2020)
10. Yen, C.C., Ghosal, D., Zhang, M., et al.: Security vulnerabilities and protection algorithms for backpressure-based traffic signal control at an isolated intersection. IEEE Trans. Intelligent Transportation Syst., 99, 1–12 (2021)
11. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mobile Networks and Appl. **24**(1), 5–17 (2018)

12. Xu, S., Qian, Y., Hu, R.Q.: Edge intelligence assisted gateway defense in cyber security. IEEE Network **34**(4), 14–19 (2020)
13. Robinson, T., Harkin, J., Shukla, P.: Hardware acceleration of genomics data analysis: challenges and opportunities. Bioinformatics **37**(13), 1785–1795 (2021)
14. Chen, B., Kim, H., Yim, S.I., et al.: Cybersecurity of wide area monitoring, protection and control systems for HVDC applications. IEEE Trans. Power Syst. **36**(1), 592–602 (2020)
15. 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

Comprehensive Intelligent Training Simulation System for Power Communication Transmission Network

Lu Liu[✉]

Beijing Kedong Electric Power Control System Company Limited, Beijing 100192, China
seg8856296@126.com

Abstract. The simulation effect of AE system in power communication network data transmission is poor, and the effect of intelligent training decreases accordingly. For this purpose, a comprehensive intelligent training simulation system for power communication transmission network is designed. The POWER chip designed on the hardware has 30 channels, and the AD sampled values are forwarded to the opposite side CPU board to realize the AD re-sampling of the dual CPU boards; 10 independent Ethernet control effects can be realized by using the GOOSE processor; the training simulation of the power grid is analyzed on the software According to the requirements, the functional modules of the system are established to realize the intelligent simulation of power network data. The system test method is used to verify that the simulation effect of the designed comprehensive intelligent training simulation system for power communication transmission network is better, which is of great promotion value.

Keywords: Power communication · Transmission network · Comprehensive · Intelligent training · Simulation

1 Introduction

With the continuous and rapid development of power industry and the gradual standardization of power market, power system puts forward higher requirements for the security, stability and economy of power grid operation. The quality of power system management and operation personnel is an important factor to ensure the stable and safe operation of power grid. Therefore, training the technical quality of power system operators is an important topic of power system [1]. Since the 1990s, with the rapid development of many high-tech fields such as computer, network image technology, information processing, software engineering, database technology and expert system, China's relevant scientific research institutions and colleges and universities have successively developed and issued characteristic power communication transmission network operator training simulation software system in cooperation with operation units, Raise the technology of simulation training system to a new course [2]. In order to train power system operation and management personnel, colleges and universities and scientific

research institutions have developed dispatcher simulation training system, centralized control station simulation training system and power communication transmission network simulation training system according to different needs to train the technical ability of operators [3]. At present, the simulation training of power communication transmission network, power grid dispatching system and centralized control station have made great development, but there are many problems. Although the simulation system has been continuously developed and improved with the wide application of computer, multimedia and database technology, the application of simulation system in college teaching is very limited [4]. Therefore, if the combination of simulation training system and theory can be used in college teaching, give full play to the advantages of simulation training system, so that electrical related professional employees can quickly integrate into their work roles after entering their posts, meet the needs of the unit and become qualified professional operators, it is a major breakthrough in teaching work.

Today's simulation training system has been widely used in the training of operators of power enterprises and employees majoring in electrical engineering in Colleges and technical secondary schools [5]. Based on the teaching requirements, I want to use the existing simulation system framework and functions to establish a regional power grid power communication transmission network system. However, in the process of project implementation, there are some problems, such as slow calculation speed after switching operation and non convergence of power flow, which can not meet the simulation requirements of normal operation conditions of power grid and power communication transmission network. Secondly, the existing simulation system also needs to be upgraded and improved in terms of operation ticket rule maintenance, operation training mode and evaluation. Finally, the data model of the existing simulation system is converted from the EMS data model by manual + software semi-automatic mode, so it is difficult to add new functions, and more work needs to be done to transplant it into the data model [6]. Therefore, it is urgent to re research and develop the intelligent evaluation function of normal operation condition simulation and operation training of power communication network.

2 Hardware Design

2.1 POWER Chip

The POWER chip designed in this paper has 30 channels (15 channels AD sampling and 80 point sampling rate for a single CPU board). At the same time, the AD sampling value (the first 15 channels) is forwarded to the opposite CPU board using the State Grid extended FT3 motherboard to realize ad re acquisition of Dual CPU boards. Receive the State Grid extended FT3 message forwarded by the opposite CPU board through the motherboard (AD recovery of Dual CPU board) [7]. 2-way point-to-point FT3 reception, receiving l OMbps national network extended FT3 protocol (80 point sampling rate) or SMbps standard FT3 protocol (80 point sampling rate) according to the configuration. 2-way point-to-point SV reception (80 point sampling rate) and message analysis. After synchronous processing of time interpolation, ad local sampling, ad re sampling, FT3. SV sampling values are interpolated into an 80 point sampling value data set. In addition to sending it to PowerPC, the data set is also assembled into SV (only one ASDU) message

and FT3 message (1 OMbps national network extension FT3 or SMbps standard FT3) according to the configuration and sent to other chips in groups (Note: the data source of FT3 message does not contain SV sampling value). The AD local sampling and AD opposite side re acquisition data contained in the SV and FT3 messages sent to other chips need to be de zeroed. PowerPC is responsible for calculating the zero drift value, and FPGA is responsible for deducting the zero drift value from the sampling value [8]. Receive the external time synchronization signal to realize time synchronization and synchronization. At the same time, convert the time synchronization signal on this side into 1PPS pulse and forward it to the CPU board on the opposite side through the motherboard [9]. The external timing synchronization signal source includes optical 1PPS, optical B code, 1588 timing 1PPS and 1PPS forwarded by the opposite CPU board. If the timing synchronization signal is optical B code, the FPGA will store the absolute time in the register after decoding, and generate a second pulse signal with a pulse width of 1 OMS, and the falling edge of the second pulse signal corresponds to the whole second time. PowerPC can realize accurate timing according to the absolute time and second pulse. It has synchronous punctuality function and meets the accuracy requirements of 10 min and 4 microseconds [10]. Therefore, a high-precision constant temperature crystal oscillator needs to be provided on this board. It has PT parallel and PT switching functions (when SV and FT3 message framing, copy the sampling values of some sampling channels to specific sampling channels according to the configuration). It has the function of zero sequence self production (when SV and FT3 message framing, vector sum the data of three sampling channels according to the configuration and fill in the specific sampling channel).

2.2   GOOSE Processor

The GOOSE processor designed in this paper can realize 10 independent Ethernet control effects, of which 8 are connected with 8 Ethernet optical ports of the small panel, one is connected with the fast bus and the other is connected with the internal bus of the board [11]. The eight Ethernet on the small panel are used for GOOSE transceiver with external devices, the Ethernet connected to the fast bus is used for GOOSE transceiver between two CPU boards in the device, and the Ethernet connected to the internal bus of the board is used for GOOSE transceiver with PowerPC [12]. According to the principle of shared memory switch, the translucent forwarding of GOOSE message between any Ethernet controller is realized ("translucent" means that the source address of GOOSE message can be modified according to the configuration during GOOSE message forwarding without changing other contents of the message). When processing Comtrade data, because the sampling message sent by the process layer equipment is of fixed frequency, interpolation processing is required [13]. The usual interpolation methods include Lagrange interpolation, Newton interpolation, difference quotient interpolation. Different interpolation methods have different effects on the operation speed and accuracy of the program. This paper adopts the linear Lagrange interpolation method [14]. Assuming that the original sampling data of power communication data is the original simulation signal, then:

$$S = \sum_{k=0} A^{kw} \sin(kwt + \theta_k) \tag{1}$$

$$x(nT'_s) = x\left(n\frac{f_s}{f'_s}T_s\right) \tag{2}$$

In formulas (1) and (2), S is the original simulation signal; k and w are constants; t is the sampling time; θ is the operation effect index; T'_s is the signal acquisition cycle; $x(nT'_s)$ is the signal acquisition sequence; f_s and f'_s are two adjacent sampling points respectively; n is the number of samples, so $n\frac{f_s}{f'_s}T_s$ is the processing cycle of communication data and can only be an integer. Assuming that n' is the maximum integer less than $n\frac{f_s}{f'_s}T_s$, the processing interpolation effect of the hardware is as follows:

$$P_n(x) = \sum_{k=0}^{n}\left(\prod_{\substack{j=0 \\ j\neq k}}^{n}\frac{x-x_j}{x_k-x_j}\right)y_k \tag{3}$$

$$R_n(x) = \frac{f_n+1(\xi)}{(n+1)!} \tag{4}$$

$$\omega_{n+1}(x) = (x-x_0)(x-x_1)(x-x_2)...(x-x_n) \tag{5}$$

$$y_n = x(m)(m+1-n') + x(m+1)(n'-m) \tag{6}$$

In formulas (3) to (6), $P_n(x)$ is a data processing expression; x, x_j, x_k and y_k are linear interpolation parameters; $R_n(x)$ is the remainder expression of data processing; f_n is the n data sampling point; $\omega_{n+1}(x)$ is the multiple expression of communication data; Assuming that data interpolation processing is performed in x_m and x_{m+1}, y_n is an interpolation linear expression; n' is a constant; Let $n = 1$, from which:

$$R_1 = \frac{f^2(\xi)}{2!}\omega_2(x) \tag{7}$$

$$\omega_2(x) = \left(n\frac{f_s}{f'_s}T_s - m\right) \tag{8}$$

$$y_2 = \left(n\frac{f_s}{f'_s}T_s - m - 1\right)*T_s^2 \tag{9}$$

In formulas (7) to (9), m is a constant; The resulting interpolation error is as follows:

$$R = \frac{x(\xi)}{f_s}T_s^2 \tag{10}$$

In formula (10), R is the interpolation processing error; Through the processing of the processor, the simulation effect of the system will be improved to the greatest extent.

3   Software Design

3.1   Analyze the Training and Simulation Requirements of Power Grid

According to the demand analysis of the system, the functions and requirements of the power grid normal operation and operation simulation training system are described as follows: the simulation function and requirements of normal operation conditions can simulate the following situations under normal conditions. The load of power system changes with time, and the power distribution, power loss and power supply power in the power grid change with the change of load; During the maintenance of power equipment in the power system, manually operating the switchgear, exiting the maintenance equipment or putting the maintenance equipment into operation will change the operation mode of the power system; In case of fault of power equipment, the protection acts, the switchgear will automatically disconnect and automatically exit the fault equipment; The regulation of reactive power distribution in power system involves the input and exit of regulating transformer tap and reactive power compensation equipment. In the above cases, the power flow distribution of the power grid changes after operating the equipment, and the measured values of the system change accordingly. The normal operation condition simulation function shall meet the following requirements under the data test of Beijing power grid. Under each normal operation condition, the calculation time response is between 0.51.5 s. This calculation time includes system wiring analysis, power flow calculation, measured value calculation and interface measured value display; The response time of operating the switchgear is also between 0.51.5 s; Under each normal operation condition, the power flow of the system can converge. Switching operation training is a very important part of the simulation training function. It is mainly a series of operations that comply with the operation procedures and operation procedures according to the operation tasks, combined with the operation mode of the system and the operation state of electrical equipment. Compiling correct operation ticket is the core content of switching operation training and an important link to improve the professional quality of employees. In order to make the process of staff writing operation ticket can be tracked, guided, assessed and evaluated by computer, the system training process is designed in this paper, as shown in Fig. 1.
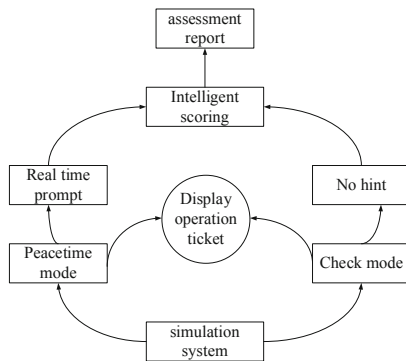
Fig. 1. System training process

As shown in Fig. 1, the training mode is divided into normal training mode and examination mode, which are set at the beginning of training. Set the operation task function in scheduling terms. Select the equipment to be operated, use scheduling terms such as operation / hot standby / cold standby / maintenance / operation on ## bus, set the operation mode of initial state and target state, and automatically generate the initial state, target state and connection relationship of all equipment related to the operation task according to different equipment types (circuit breaker, transformer, line and bus). Operation task management function. Including: delete and create operation tasks; View the evaluation report generated by the operated task; Check the specific details of the operation task, etc. According to the type and current status of switchgear selected according to the point diagram, in addition to providing general operation modes (on / off, check off / on, etc.), other possible operation modes are automatically provided. If the secondary plug-in is selected, two operation modes of "up" and "down" will be provided; Select the trolley switch to provide the operation modes of "working position", "test position" and "maintenance position". Real time intelligent checksum prompt function. This is a function used for employees' daily training. The system tracks and verifies each step of employees' operation in real time. If there is an error, it will prompt the employees to violate the specific rules. If it is correct, the operation will be output to the operation ticket currently being written. Display and management functions of operation ticket contents currently being written. You can modify and delete the steps that have been operated. In the examination mode, you can delete any operation item that has been operated. In the normal training mode, only the last operation item can be deleted. Intelligent scoring function. According to the content of violation of rules, the evaluation criteria set in advance according to the degree of harm shall be used for score evaluation. After the operation, the system generates an evaluation report. The evaluation report includes all operation steps of employees, contents of violation of rules, scores deducted according to the degree of harm, etc.

3.2   Establish the Functional Modules of the System

This paper designs a simulation training system for normal operation and operation of power grid. Under normal operation of power grid, the response of operating switchgear meets the requirements, power flow convergence and short output time. An intelligent evaluation function system for switching operation of dispatching power communication transmission network is developed by using Visual C + + and database, which can track and remind users of operating rules violated in real time, Give the final evaluation report score. The main functions of the power grid normal operation and operation simulation training system are divided into two parts: the data model conversion of the power grid normal operation and operation simulation training system. The lowest raw data comes from the model of EMS primary equipment data in dispatching automation. The data required in this system is only a part of it, so only the useful data is screened, One part of the secondary equipment data is generated by the program, such as CT and Pt tables generated in the program, and another part is expanded manually. The figure middle module is the core of the system. It is based on data and realizes the following two functions. When the equipment value on the interface changes, the program receives the changed equipment label and the changed value. The system performs topology

analysis, power flow calculation and measured value calculation, and returns to the measured value on the monitoring interface, thus, as the basis of the system functional module, the overall structure of the functional module of the system is shown in Fig. 2:
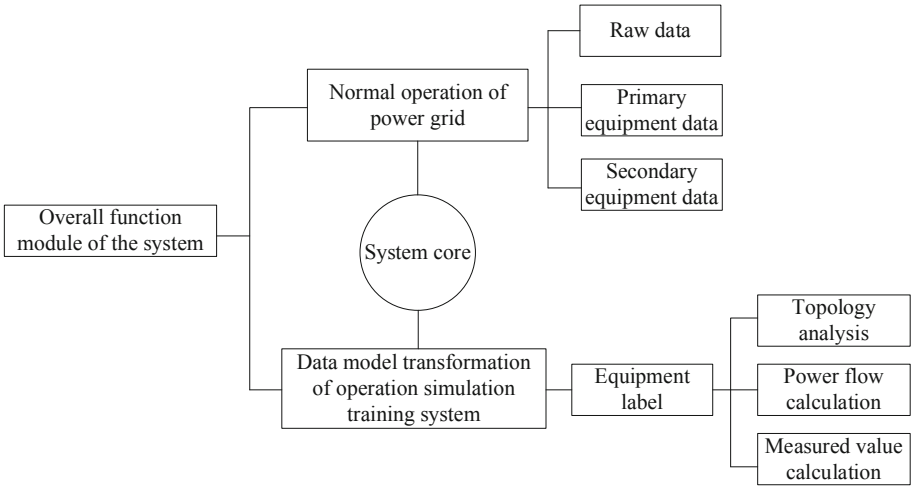
Fig. 2. Overall structure diagram of functional modules of the system

In addition to the Wizcon graphic interface, the operation interface of the system also has various dialog boxes designed by MFC. The maintenance and management of system data are completed by these dialog boxes. For example, operate the task attribute management window, which can add tasks, delete tasks and view task details. The operation is convenient and the human-computer interaction interface is friendly.

3.3   Intelligent Simulation of Power Network Data

In order to realize the intelligent simulation effect of power network data, the data sources of the simulation training system designed in this paper are obtained from the equivalent model. Manual + software semi-automatic equivalence is carried out from EMS. The manual equivalent data is huge and the workload is quite large. If this huge workload can be completed by the program, and the system can generate the equivalent model directly from the EMS data, it will enhance the universality of the simulation training system to various power grid data and improve the development efficiency of the system. This paper needs to develop such an automatic modeling software. The functions of EMS are different from those of the normal operation of power grid and operation simulation training system, so the required data information structure is also different. After the automatic modeling program obtains the EMS data, it needs to be transformed, filtered and expanded. The reasons are as follows: some equipment information required for the normal operation of power grid and operation simulation system cannot be obtained directly from the EMS data information, It can be used by the simulation program only after necessary data format conversion. For example, the branch

parameter information in the power grid normal operation and operation simulation system describes the equivalent power system model of primary equipment in the power system after equivalence and conversion. The "branch type" in the branch data table does not exist in the branch parameter attribute data of EMS data, so it must be based on other EMS data information, Generate data required for normal operation of the simulation system. The power grid normal operation and operation simulation system only needs part of the EMS data information, and other invalid data information to the system must be filtered and removed. For example, the data information of line parameters and line attribute information in EMS data are very comprehensive. For example, the "line status" and "related line name" information are line attribute information not required by normal simulation system, which should be removed. The normal operation and operation simulation system of power grid needs the basic information of transformer, but the configuration information of transformer is missing in the EMS data information, so it is necessary to expand the data not in the EMS data. When the interface operates the switchgear, the corresponding measured values change, so the system must configure the attribute information of the transformer.

The system designed in this paper can start directly from EMS, and the equivalence work is completed by program. Compared with manual equivalence, it shares data with EMS, and has the advantages of easy expansion and real-time updating data. Because the EMS data is different from the data structure required by the simulation function under normal operating conditions, the program will convert, filter and expand the EMS data after obtaining the EMS data. According to the analysis of operation simulation function under normal working conditions, most of the simulation data comes from EMS model data, and the program calculation needs to analyze, convert and modify the primary data to generate the required intermediate data. The extended switch knife switch parameter table is obtained by adding CT table to the original switch knife switch table. Adding current transformer equipment to the original main wiring is equivalent to a closed "switch". The basic information of the data item is consistent with the switch switch parameter table. The starting and pointing element codes of CT correspond to the left and right end element codes of the switch and are set to the closed state. The branch parameter table is obtained by synthesizing the line parameter table, transformer parameter table, reactance parameter table and capacitance parameter table. Before power flow calculation, as long as the branch is equivalent, the equivalent parameters are obtained according to the equivalent models of different equipment and filled in the branch parameter table. In power flow calculation, branch parameters and node voltage information are called to generate power flow results. Therefore, the simulation results of the system designed in this paper can be guaranteed to be good.

4   System Test

In order to verify whether the system designed in this paper has practical effects, the above system is tested. In order to ensure the fairness of the experiment, the method in this paper and the traditional method are set to the same experimental environment. The test process and results are as follows.

4.1   Test Process

In the test process, first test the hardware of the system, and the test parameters are shown in Table 1.

Table 1. Hardware test parameters

Data item	Data type
CT coding	Character
CT name	Character
Voltage level	Integer
Starting element code	Character
Pointing component code	Character
Starting equipment code	Character
Meter current value	Float

As shown in Table 1, under this condition, the hardware of the system can be ensured to operate normally. At this time, the software is tested, and the login interface of the system is shown in Fig. 3.
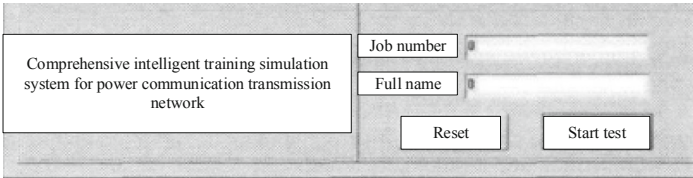
Fig. 3. System login interface

As shown in Fig. 3, when the above interface appears, it can indicate that the system is running normally.

4.2   Test Results and Discussion

Under the above test environment, compare the traditional AE system with the system designed in this paper to verify the simulation effects of the two systems, as shown in Table 2.

As shown in Table 2, the communication data transmission error of traditional AE system is more than 0.10, and the simulation effect is poor; The communication data transmission error of the system designed in this paper is less than 0.10, and the simulation effect is good, which is in line with the purpose of this paper.

Table 2. Test results

Number of experiments	Communication data transmission error of traditional AE system	The communication data transmission error of the system designed in this paper
1	0.67	0.06
2	0.36	0.03
3	0.28	0.02
4	0.12	0.01

5 Conclusion

The main achievements of this paper are as follows: the training mode of compiling operation ticket by point diagram in line with the actual workflow is designed, and the overall structure and function of power grid normal operation and operation simulation training are designed. The automatic data modeling function and normal operation condition simulation function based on EMS data and meeting the requirements of dispatching substation simulation training are studied and developed. The general judgment rules for automatic generation of operation tasks and the general verification method of hierarchical operation rules suitable for various wiring forms are studied and developed, which greatly reduces the system maintenance workload of intelligent operation training software. The intelligent operation evaluation function based on the above general rules is studied and developed to automatically generate the evaluation report, assist the superior leaders to objectively and comprehensively evaluate the employees' operation, and reduce the workload of the superior leaders. The function of power grid normal operation and operation simulation training system is verified by using an actual power grid data. The test results show that the simulation results of normal operation conditions are reasonable, the response time of operating equipment meets the design requirements, the switching operation can track and prompt in real time, and the final results are given in the form of report. The superior leaders can evaluate the employees according to the report information, which improves the training efficiency.

References

1. Wu, Y., Chen, J., Ru, Y., et al.: Research on power communication network planning based on information transmission reachability against cyber-attacks. IEEE Systems J. **10**(99), 1–12 (2020)
2. Yi, X., Milstein, L.B.: Design and performance analysis for short range, very low-power communications. IEEE Trans. Commun. **68**(9), 5938–5950 (2020)
3. Juwono, F.H., Reine, R., Liu, J., et al.: BITFCM-OFDM scheme for power-line communication systems. AEU – Int. J. Electronics Commun. **105**(6), 116–123 (2019)

4. Yu, L., Wu, Q., Xu, Y., et al.: Power control games for multi-user anti-jamming communications. Wireless Netw. **25**(5), 2365–2374 (2019)
5. Kong, P.Y.: Optimal configuration of interdependence between communication network and power Grid. IEEE Trans. Industrial Informatics **15**(7), 4054–4065 (2019)
6. Shi, F., Chen, R., Shen, H., et al.: Energy-efficient power allocation for D2D communication underlaying cellular networks. Mobile Networks Appl. **6**(1), 1–9 (2021)
7. Hu, B., Liu, X., Zhao, J., et al.: A packet scheduling method based on dynamic adjustment of service priority for electric power wireless communication network. Wireless Commun. Mobile Comput. 9, 1–13 (2020)
8. Qian, Y., Shi, L., Li, J., et al.: An edge-computing paradigm for internet of things over power line communication networks. IEEE Network **34**(2), 262–269 (2020)
9. Hao, S., Zhang, H.Y.: Theoretical modeling for performance analysis of IEEE 1901 power-line communication networks in the multi-hop environment. J. Supercomputing **76**(4), 2715–2747 (2020)
10. Abd, A., Alsarhan, F.F., Safaa, A., et al.: Applying AI in LTE wireless network for power-system communications load balancing. Solid State Technol. **63**(2), 83–92 (2020)
11. Dong, J., Wu, G.W.: Serial communication interface chip overvoltage tolerance detection simulation. Computer Simulation **36**(12), 327–331 (2019)
12. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mobile Networks Appl. **24**(1), 5–17 (2018)
13. Liu, S., Fu, W., He, L., Zhou, J., Ma, M.: Distribution of primary additional errors in fractal encoding method. Multimedia Tools and Appl. **76**(4), 5787–5802 (2014)
14. Liu, S., Chen, X., Li, Y., Cheng, X.: Micro-distortion detection of lidar scanning signals based on geometric analysis. Symmetry **11**, 1471 (2019)

Research on Intelligent Anti-collision Device of Heavy Truck Based on Wireless Communication Technology

Xiaolin Qian and Youjun Xu[✉]

College of Rail Transit, Nanjing Vocational Institute of Transport Technology, Nanjing 211188, China
xuyoujun569@yeah.net

Abstract. The increase in the number of vehicles, rapid development of road traffic, traffic accidents have become increasingly serious. However, these sensors have some limitations, such as single detection direction, detection blind area, and can not realize long-distance transmission. So they can not meet the current requirements. Based on this, the intelligent anti-collision device of heavy truck based on wireless communication technology is put forward. The intelligent anti-collision research of heavy truck is realized by designing the display interface of intelligent anti-collision device, constructing anti-collision early warning database and modulating and demodulating algorithm based on wireless communication technology. Tests show that the wireless communication module and the collision warning and display function of the intelligent anti-collision device can run successfully, and the overall performance of the device is reliable and stable.

Keywords: Wireless communication technology · Heavy-duty truck · Intelligent anti-collision device

1 Introduction

There are many ways to achieve wireless communication technology, the current mainstream communication technology to DSRC and LTE-V mainly. The most widely used V2X communication technology is the Dedicated Short Range Communication (DSRC) technology based on the IEEE 802.11p and the IEEE 1609 series standards. LTE-V is an extended technology based on the fourth generation mobile communication technology. The current version of LTE-V is 4.5G technology [1].

At present, the domestic and foreign main such as Benz, Honda, Volvo and other automotive enterprises equipped with vehicles common collision warning system, as an assistant driver. The Mercedes-Benz Pre-Safe Brake system for vehicles uses millimeter radar to detect the vehicle ahead, is equipped with a microwave detector, and is equipped with a brake assist system that warns 2.6 s before a collision is identified, gives a second warning 1.6 s before a collision, performs braking control on the vehicle if the driver has not yet taken a substantive collision avoidance operation, and outputs a maximum

braking deceleration 0.6 s before a collision [2–4]. Toyota's Pre-Collision Active Collision Avoidance system combines millimeter-wave radar, cameras and infrared sensors to improve ambient detection accuracy, and infrared sensors enable vehicles to detect dangerous targets at night [5]. When it is determined that a collision is imminent, the system automatically tightens the seatbelt and performs auxiliary braking when the driver depresses on the brake pedal, so that the vehicle can achieve maximum braking force and minimize the impact force and damage [6–9].

The core of collision prevention and early warning system is early warning strategy [10]. Many universities and scholars at home and abroad have conducted in-depth research on collision prevention and early warning strategy. There are two kinds of anti-collision algorithms: the safe time algorithm and the safe distance algorithm. Select the vehicles that may collide in a certain range of time, predict the trajectory, calculate the collision time TTC, when the TTC in the set range, judge the potential collision risk. This paper designs an intelligent anti-collision device for heavy-duty trucks based on wireless communication technology. Through wireless communication technology, the transmission performance of information such as position, speed, distance and braking distance of heavy-duty trucks is improved. Collision warning model with braking distance as parameter. Design a risk assessment model based on braking distance, collision time and lane change safety distance, and calculate risk level and intervention time in real time.

2   Research on Intelligent Collision Avoidance Device for Heavy Trucks Based on Wireless Communication Technology

2.1   Design of Display Interface for Intelligent Anti-collision Device

Good user interface is very important to improve the usability of anti-collision device. The control of the login interface includes "user name", "password", "login" and so on.

Table 1. Interface structure of intelligent anti-collision device

Button	Features
Domain name	Enter the anti-collision warning IP address
Username	Enter the anti-collision device administrator account
Password	Enter the administrator password
Log in	Click the "Login" button to log in to the alert server with the administrator information entered

As shown in Table 1, it is the interface structure of the smart collision avoidance device. The main interface displays map information by default, and displays the location of the train after the map is matched on the map. The location of the train is updated every time the location data is updated. Ensure that the location update information of
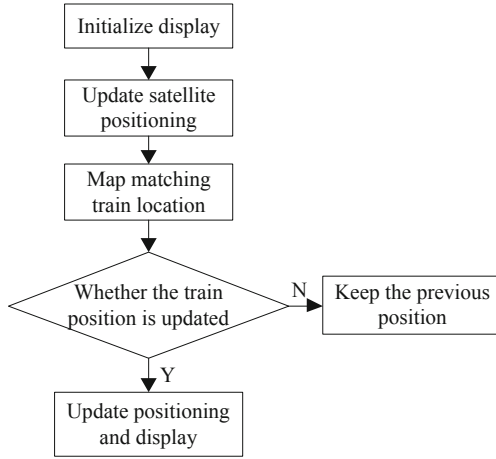
Fig. 1. Heavy truck location information update process

heavy trucks is displayed in the anti-collision device interface. The specific location update process is shown in Fig. 1.

As shown in Fig. 1, in the intelligent collision avoidance device interface of heavy trucks, the location information of the truck is updated in real time. First, the location of the truck is initialized to display the location of the truck. Based on the satellite positioning space position, combined with the map function in the interface, it is matched to the geographic information of the truck., The final update positioning is successful and displayed in the interface. The setting interface contains a record query menu, which can query the communication records in the heavy truck operating database, and record the record items contained in the query menu, as shown in Table 2.

Table 2. Record query menu item settings

Record item	Features
Early warning record	Can query historical warning records recorded in the database
Call log	Can query the historical call records recorded in the database
Recording	Can query the recording records of calls in the database

As shown in Table 2, the interface designed for this article records query menu item settings and features.

The Alert Settings Window sets up the alert parameters, which are opened by the button in the menu bar, and the Alert Time Edit Box on the left, which sets up the specified alert interval. You can also quickly select a preset warning interval by using a drop-down menu. The list of alert targets shows the ID list of the terminals that currently join the alert function. The terminals in the list can receive alert information in the alert area. You can edit the list of alert targets by removing and adding buttons.

The voice call in the display interface is very important to avoid the collision of heavy trucks. Voice call means that the early warning system can talk with the terminal directly. In the number box, enter the terminal number you want to call. Voice call calls have two properties: an emergency call and a broadcast call. To make an emergency call, check the box in front of Emergency Call. To make a broadcast call, check the box in front of Broadcast Call. When the number is entered, click the "PTT" button or press the space bar on the keyboard to initiate a voice call. The voice call has been successfully paged out and the call window will pop up [11]. Can view all the terminal information registered in the alarm intercom scheduling desk, can enter the number to query, in the list of selected terminal number in the main interface display terminal information and location. In addition to receiving the terminal to confirm the early warning message, but also to distinguish the general message warning icon for display. When the terminal receives the alarm information, it prompts with the alarm interface and exits the prompt after selecting the safety button. In the offline state, the terminal display interface status bar to exclamation icon reminder, while offline using the alarm window to remind, and every 10 s to remind the tone reminder, until the terminal online again.

2.2 Building an Early Warning Database for Anti-collision Devices

First, design an early warning database. Since the smart collision avoidance device uses handheld terminals, it is necessary to establish a terminal information table to record all the registered terminal information, which contains the terminal equipment number and the information of its users. Because heavy truck information is needed in the early warning, it is also necessary to establish a heavy truck number information table. The generated early warnings need to be recorded, so an early warning record form needs to be established, as shown in Table 3.

Table 3. Early warning information record table

Field name	Type	Constraint	Illustrate
ID	NUMBER	Not null	
Time of warning	VARCHAR2(14)	Not null	Complies with the valid date and time format, year, month, day, hour, minute, and second
Warning type	VARCHAR2(2)	Not null	Divided into: SMS alert, voice call
Warning station	VARCHAR2(6)	Not null	
Warning end time	VARCHAR2(14)	Not null	
Target confirmation	VARCHAR2(2)	Not null	Divided into: confirmed, unconfirmed
Heavy truck number information	NUMBER	Not null	

As shown in Table 3, an early warning configuration information table is established to record the early warning parameters, including the early warning time, the early warning type, and the early warning end time. Since the intelligent anti-collision device also uses the location data of the construction site, it is also necessary to establish a construction area data table. In addition, there are recording information tables for recording early warning calls, attendant registration tables for recording station attendant login information, and reporting information tables for recording early warning reports. In order to be able to use the map matching method to correct the positioning, it is also necessary to establish a digital map database, which contains the segment point calibration table used to correct the coordinates of the segment point, and the curve information table used to identify the segment curve (including all the divided curves Km coordinates and radius of curvature of the start and end points) [12].

This paper designs the E-R relationship of the database, as shown in Fig. 4 and Fig. 5. Due to the large number of data tables, this article only describes the design of representative data tables.

Table 4. On-duty officer login information table

Field Name	Type	Constraint	Illustrate
ID	NUMBER	Not null	
Username	VARCHAR2(10)	Not null	Login user name, 2–10 characters
User type	VARCHAR2(2)	Not null	Divided into: system administrator, attendant on duty, assistant attendant on duty
Station name	VARCHAR2(12)	Not null	
IDStation ID	VARCHAR2(6)	Not null	
User rights	VARCHAR2(20)	Not null	Divided into: administrators, attendants on duty
Password	VARCHAR2(10)	Not null	6–20 characters, composed of letters and numbers

Table 4 is mainly responsible for recording the information of the attendants who log in to the early warning walkie-talkie dispatching station.

Table 5. Configuration table of anti-collision warning rules

Field name	Type	Constraint	Illustrate
ID	NUMBER	Not null	
Warning time	VARCHAR2(4)	Not null	00 (min) 00 (s)–59 (min) 59 (s)
Collision avoidance distance	NUMBER	Not null	Unit: 100 m
Longitude and latitude of anti-collision starting point	VARCHAR2(8)	Not null	Complies with the latitude and longitude format, such as east longitude 23°27′30″ means east longitude 232730
Longitude and latitude of collision avoidance terminal	VARCHAR2(4)	Not null	
Urgent escalation	VARCHAR2(8)	Not null	Divided into: report all, report failure early warning, never report
Collision avoidance warning target confirmation	VARCHAR2(2)	Not null	Divided into: mandatory, optional

As shown in Table 5, it is primarily responsible for recording early warning configuration information. Through the above pre-warning information record table, the log-in information table of duty personnel and the pre-warning rule configuration table, the pre-warning database of anti-collision device is constructed.

2.3 Modulation and Demodulation Algorithm Based on Wireless Communication Technology

The core algorithms of this paper include modulation and demodulation algorithm used in DMR wireless communication technology and map matching algorithm used in location correction [13]. Modulation is to use the baseband signal to control the change of one or several parameters of the carrier signal, and load the information on it to form a modulated signal for transmission, while demodulation is the inverse process of modulation. The change will restore the original baseband signal. The modulation and demodulation algorithm mainly accomplishes the modulation and demodulation of 4FSK signal, and its performance has great influence on communication distance and speech quality. In order to improve the precision of map matching, the paper uses the method of tangent line to improve the precision of map matching on curve orbit. The model of digital wireless communication includes source codec, channel codec, modulation and demodulation, etc. The intelligent model structure of wireless communication technology is shown in Fig. 2.

As shown in Fig. 2, source coding, also known as band compression coding, can reduce the rate of code elements and the number of code sources without distortion recovery. The source decoding transforms the information in the opposite way to the source encoding process. Channel coding is to counter the noise and attenuation in the
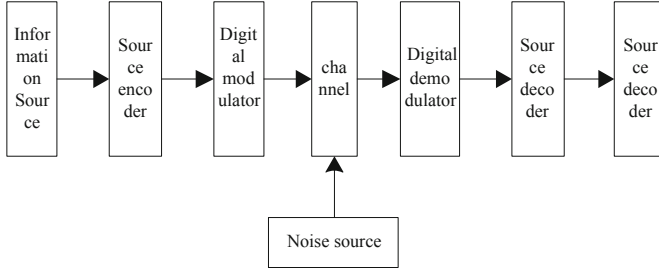
Fig. 2. Intelligent model structure of wireless communication technology

channel by adding redundancy codes to the transmitted information codes in a certain way to form new code words, and the receiver then detects or corrects errors in a corresponding way so as to achieve the purpose of improving the anti-interference and error correction capabilities [14]. Channel decoding is the process of restoring the code elements after channel coding. In order to improve the transmission efficiency of the signal, the spectrum of the digital baseband signal is moved to the high frequency, so that it can improve the utilization of spectrum resources.

DMR standard chooses 4FSK as its modulation and demodulation mode. 4FSK is a constant envelope modulation technology, whose characteristic is that the peak value of pulse envelope is constant in a certain time, and this modulation technology has high spectrum utilization. The process of 4FSK modulation is that the signal is filtered by shaping filter after level transformation, then the filtered signal is processed, and finally filtered by first-order RC filter. According to Nyquist criterion, it is possible to eliminate ISI only when the symbol rate is less than 2 times the channel bandwidth. Consider natural filters, the frequency offset expression for rectangular filters is:

$$h(t) = \begin{cases} 1 \le t \le \dfrac{T_s}{2} \\ 0 \ge t \ge \dfrac{T_s}{2} \end{cases} \tag{1}$$

Among them, $h(t)$ represents the frequency deviation of the rectangular filter; T_s is the constant of the filter index. The bandwidth of this rectangular pulse is infinite, the out-of-band power decays very slowly, and out-of-band leakage occurs. In practical applications, it is generally not used as a shaping filter. In order to solve the problem of out-of-band leakage, consider a rectangular function in the frequency domain to completely limit the power within the band. The calculation formula for the impulse response is:

$$h(m) = \frac{\sin(\pi t / T_s)}{\pi t / T_s} \tag{2}$$

This article uses a root raised cosine filter, this filter realizes the rapid attenuation of the time domain signal at the cost of widening the frequency band, and at the same time, in order to reduce the loss of the band widening, a roll-off is added on the basis of the raised cosine filter. Coefficient α, the value interval is [0, 1]. The smaller α is, the

closer to the rectangular filter, and the larger the α is, the closer to the cosine filter. By adjusting the roll-off coefficient, the occupied bandwidth and attenuation speed can be adjusted.

To realize this filter, we must first generate a set of filter coefficients corresponding to the order of the filter. This article uses Matlab functions to generate filter coefficients. The function expression is as follows:

$$functionB = r \cos fir(r, N - T, rate + 1, T) \tag{3}$$

Among them, parameter r represents the roll-off coefficient, $N - T$ represents scalar or vector; together with $rate$, determines the length of the filter output, $rate$ represents the sampling rate; T represents the sampling period of the input signal, in seconds, used to generate ordinary raised cosine filters Coefficient, when $N - T$ is a scalar, B returns a one-dimensional array of length $rate + 1$, which is the coefficient of the required root raised cosine filter. In order to facilitate the calculation in the DSP, the set of coefficients need to be further fixed-point to 16-bit integer processing, that is, each number is multiplied by 2^{15}, so that the accuracy of the coefficient can be maximized when the coefficient is used, and the transformed coefficient is put into the array. After calculating the coefficients of the filter, the operational relationship between the input and output of the filter can be established. The output of the FIR filter is obtained by the convolution operation between the input data and the tap coefficient of the filter. The formula is as follows:

$$y(n) = \sum_{k=0}^{n} x(k) \cdot h(n - k) \tag{4}$$

Among them, k is the FIR filter coefficient; $y(n)$ represents the input of the filter at time n; $x(n)$ represents the output of the filter at time n. This kind of convolution operation is to add or subtract the input data (according to the positive or negative of the input data) to the product of the previous filtering operation result and the corresponding filter coefficient, and output the operation result as the input of the subsequent stage. According to the DMR wireless communication protocol, a frame of data contains 288 bits. After serial-parallel conversion, 144 symbols are obtained, which is a four-level signal. Then the four-level signal is interpolated by 8 times to obtain the input data of the root raised cosine filter.. Next, the characteristics of the root raised cosine filter are tested by adding Gaussian white noise to the input signal. The entire inspection process is completed by matlab simulation. First, grab a set of input data with a length of 1152, the data type is a 16-bit unsigned integer, the value range is 0 to 4096, as the original input signal.

3   Experiment and Analysis

3.1   Experimental Preparation

In order to verify the effectiveness and performance of the intelligent anti-collision device for heavy truck based on wireless communication technology, and to avoid the danger and

high cost of real vehicle test, this paper chooses Prescan and MATLAB/Simulink joint test. The use of feature-oriented parametric modeling, including the setting of vehicle dynamic parameters, can reduce the experimental cost and shorten the development cycle. It can be used for simulation verification related to unmanned and active safety, such as automatic emergency braking system, orbit deviation early warning system, etc. Track has rich traffic scene elements, including a variety of straight lines, curved sections, all kinds of vehicles, pedestrians, obstacles, and so on, and can detect the surrounding environment of a variety of sensors. Input the measured data, can also be set parameters to get the required data, into the control module algorithm, output test results. In the user interface (GUI) operation, the establishment of the intersection confluence and follow the same road these two traffic scenarios for experimentation, the possibility of collision collision and side collision early warning performance comparison. The vehicle simulation model is established and the state parameters of the vehicle are set up. After setting up the speed and acceleration (spot) of the vehicle, setting up the speed, acceleration, trajectory and other parameters of the vehicle in the process of movement, it is necessary to add corresponding sensors to all vehicles within the communication scope. Version 8.4.0 of Pre Scan supports the DSRC communication protocol and the SAE J2735 protocol for V2X communication experiments. Drag the V2X wireless communication device onto the vehicle body on the GUI interface, configure the V2X plug-in parameters, preprocess the acquired data, establish the pretreatment module, and then build the corresponding MATLAB function module according to the ITTC model and the TTW model. The input of the model is the state and position of the vehicle, and the output is the warning information. The 3D visual interface is used to observe the road and the warning result.

3.2 Analysis of Results

Assuming that the vehicle starts from the initial position 0 at a speed of 60 km/h, and accelerates to 100 km/h after 5 s, the change of acceleration is constant, and the simulation records the motion state of the vehicle for 20 s. The algorithm in this paper is used to predict the total travel distance of the vehicle after 20 s of driving, and the error between the predicted position and speed and the actual position and speed of the vehicle is used as the performance standard. The result of the vehicle position prediction is shown in Fig. 3.

As shown in Fig. 3, the simulation value and predicted value curve of the horizontal position and the longitudinal position of the vehicle position prediction result are basically fitted, and the vehicle position prediction error is expressed by calculating the difference between the predicted position of the vehicle and the actual position of the vehicle. This test case, The prediction error of the horizontal position is -0.015 ± 0.252 m, and the prediction error of the longitudinal position is 0.089 ± 0.207 m. The result of vehicle speed prediction is shown in Fig. 4.

As shown in Fig. 4, the prediction error of vehicle speed is calculated by predicting vehicle speed and actual vehicle speed. The results show that the prediction error of lateral speed is -0.028 ± 0.279 m/s and that of longitudinal speed is -0.064 ± 0.231 m/s. The experimental results show that the prediction data of the intelligent anti-collision
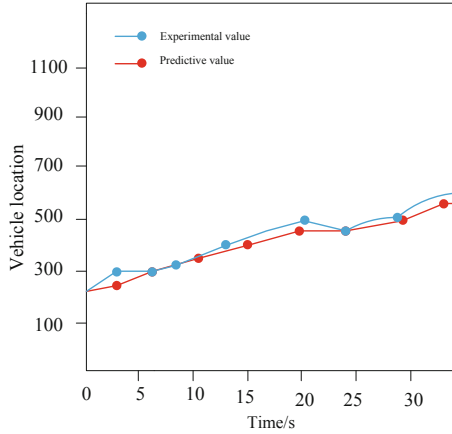
Fig. 3. The result of vehicle position prediction
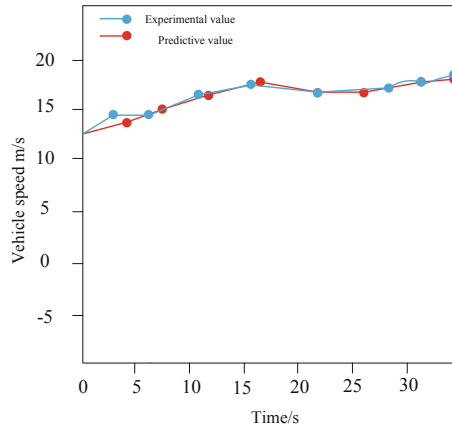
Fig. 4. The result of vehicle speed prediction

device based on wireless communication technology is less than the real data, and the prediction results can be used in anti-collision scheme.

The test of early warning and display module is to test whether the data received by the early warning system can be correctly pre-processed, whether the data input early warning algorithm can correctly judge the collision risk of vehicles, and can trigger the correct early warning information display according to different scenes. Because the communication function and information collection function of the system have been tested, considering the actual test environment and feasibility, the simulation value is used to test the vehicle state information.

In the user terminal interface, the relative positions of the front and rear vehicles can be visually displayed according to the vehicle information in the radar chart, and the corresponding information of the vehicle can be displayed correctly. It can judge that

the type of collision between the front and rear vehicles is a rear collision, and display the warning information in the warning information status bar. Early warning success rate test: By setting the heading angle and trajectory parameters of the vehicle, setting the rear-end collision and side collision scenarios of two vehicles, the terminal runs 100 times for each collision, and runs 200 times to verify the success rate of the early warning function. The success rate of the early warning function is shown in Table 6.

Table 6. Results of early warning success rate of anti-collision device

Collision type	Number of successful warnings	Success rate
Rear-end collision	100	100%
Side collision	100	100%

As shown in Table 6, the test results of the early-warning and display module show that the early-warning and display functions of the user terminal are realized normally, and the position, speed, relative distance and other information of the vehicles can be displayed correctly, and the relative position of the two vehicles can be displayed visually in the radar chart, and different warning information can be displayed according to different collision types. The success rate is 100%, which shows that the performance of the anti-collision pre-warning function is stable, and the pre-warning and display modules meet the requirements of intelligent anti-collision.

4   Conclusion

Based on the research of wireless communication technology and the analysis of existing anti-collision early warning algorithms, an intelligent anti-collision device for heavy trucks is designed and implemented. Based on the analysis of vehicle movement in both transverse and longitudinal directions, the impact of vehicle velocity and acceleration parameters on collision risk assessment is considered. The test results show that the wireless communication module and the function of collision warning and displaying can be operated successfully, and the overall performance of the device is reliable and stable.

References

1. Kang, S.: Research on quantitative calculation of radio wave coverage in wireless communication system. Comput. Simul. **36**(09), 202–205, 281 (2019)
2. Wang, X., Liu, J., Qiu, T., et al.: A real-time collision prediction mechanism with deep learning for intelligent transportation system. IEEE Trans. Veh. Technol. **PP**(99) 1 (2020)

3. Ulrich, M., Dolar, C., Marbach, C., et al.: Collision warning system for forklift trucks. ATZheavy Duty Worldw. **13**(4), 16–21 (2020)

4. Wang, J.J., Song, Y.C., Wang, W., et al.: Evaluation of flexible floating anti-collision device subjected to ship impact using finite-element method. Ocean Eng. **178**(APR.15), 321–330 (2019)

5. Ni, S., Liu, Z., Cai, Y., et al.: Coordinated anti-collision path planning algorithm for marine surface vessels. IEEE Access **PP**(99), 1 (2020)

6. Sorokin, A.V., Shepeta, A.P., Nenashev, V.A., et al.: Comparative characteristics of anti-collision processing of radio signal from identification tags on surface acoustic waves. Inf. Control Syst. **2019**(1), 48–56 (2019)

7. Su, J., Sheng, Z., Liu, A.X., et al.: A group-based binary splitting algorithm for UHF RFID anti-collision systems. IEEE Trans. Commun. **PP**(99), 1 (2019)

8. Huang, Z., Su, J., Wen, G., et al.: A physical layer algorithm for estimation of number of tags in UHF RFID anti-collision design. Comput. Mater. Continua **61**(1), 399–408 (2019)

9. Pérez-Carabaza, S., SchErer, J., Rinner, B., et al.: UAV trajectory optimization for Minimum Time Search with communication constraints and collision avoidance. Eng. Appl. Artif. Intell. **85**, 357–371 (2019)

10. Walambe, R., Patwardhan, N., Joshi, V.: Development of auto-parking and collision avoidance algorithms on car type autonomous mobile robots. IFAC-PapersOnLine **53**(1), 567–572 (2020)

11. Bachmann, M., Morold, M., David, K.: On the required movement recognition accuracy in cooperative VRU collision avoidance systems. IEEE Trans. Intell. Transp. Syst. **PP**(99), 1–10 (2020)

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**(5), 2123–2126 (2021). https://doi.org/10.1007/s11036-021-01777-7

13. 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

14. Liu, S., Sun, G., Fu, W. (eds.): e-Learning, e-Education, and Online Training, pp. 1–386. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63952-5

Education and Enterprise

Design of Diversified Evaluation System of College Teaching Quality Based on Deep Data Mining

Dong Yang[1](✉) and Jun Yang[2]

[1] Nanchang Institute of Technology, Nanchang 330108, China
yd11214@163.com
[2] College of Marxism, Fuyang Normal University, Fuyang 236041, China

Abstract. At present, the mainstream teaching evaluation is dominated by students, supplemented by experts in or out of school. Under this evaluation method, only simple quantitative evaluation results can be obtained, and the evaluation data can not be analyzed. Therefore, it is proposed to design the diversified evaluation system of college teaching quality based on deep data mining. In the hardware design of the diversified evaluation system of college teaching quality, the ARM microprocessor based on diversified evaluation is designed, and the data manager based on data mining is designed. In the software design of the diversified evaluation system of college teaching quality, the diversified evaluation model of teaching quality is established, Design evaluation model database. The test shows that the evaluation accuracy coefficient of the system is more than 95, which is much higher than the traditional two systems.

Keywords: Data mining · Deep mining · College teaching · Teaching quality · Evaluation system

1 Introduction

The enrollment scale of colleges and universities has continued to expand in recent years, and the annual growth rate has remained at about two percentage points. As the scale of enrollment increases, the quality of students will inevitably decline. The faculty and teaching facilities of colleges and universities have grown relatively slowly, and insufficient teaching resources will inevitably affect the quality of education [1]. In recent years, various reports have often appeared, reflecting the decline in the quality of college graduates and poor work ability, and various aspects of society often comment and question the teaching quality of colleges and universities. Educational departments at all levels of the country continue to adhere to the scientific development concept and carry out in-depth education and teaching reforms [2]. In particular, we will continue to increase investment in college education, strengthen the construction of the teaching staff of colleges and universities, and pay particular attention to the evaluation of the quality of colleges and universities. Among them, the assessment of student abilities is the

focus of teaching quality assessment. Data Mining Technology is a cross-discipline that integrates technologies and achievements in many different fields [3], bringing together multiple disciplines such as machine learning, pattern recognition, expert systems, data visualization, statistics, and high-performance computing. Data mining technology can automatically convert the massive data we collect in our work and life into various styles of information, providing a conversion channel between people and data, and providing us with a very important basis for judgment and decision-making.

Data mining technology has been successfully applied in retail, finance, telecommunications, scientific research and other fields. Provide customers with data analysis, predict potential customers, formulate sales strategies, avoid business risks and improve service quality. In the field of Education [4], data mining technology also has certain applications, mainly in network education. A large number of researchers obtain the information of teachers and students and their interaction information through the network online education platform, obtain relevant information through data mining technology, and then put forward suggestions to teachers, use more appropriate teaching methods [5], and constantly improve the teaching level and quality. However, this is only limited to network teaching, but data mining technology is rarely used in most classroom teaching.

Therefore, this paper designs a diversified evaluation system of college teaching quality based on deep data mining. Main design route of the system:

Step 1: in the hardware design of the diversified evaluation system of teaching quality in Colleges and universities, the ARM microprocessor based on diversified evaluation is designed;

Step 2: design the data manager based on data mining, establish the diversified evaluation model of teaching quality and design the evaluation model database in the software design of diversified evaluation system of teaching quality in Colleges and universities.

Step 3: experimental analysis.

Step 4: conclusion and future outlook.

2   The Hardware Design of the Diversified Evaluation System for Teaching Quality in Colleges and Universities

2.1   Design an ARM Microprocessor Based on Diversified Evaluation

The ARM architecture is the first RISC microprocessor designed for the low-budget market. In addition to some features of RISC, the ARM architecture also uses some special technologies to minimize the chip area and reduce the chip area while ensuring high performance. Power consumption [6]. The ARM microprocessor has the following main features: small size, low power consumption, low cost, high performance; support Thumb (16-bit) and ARM (32-bit) dual instruction set; a large number of registers are used, and the instruction execution speed is faster; Most data operations are completed in registers;—The addressing mode is flexible and simple, and the execution efficiency is high;—The instruction length is fixed; ARM microprocessors are mainly used in industrial control, wireless communication, network applications, consumer electronics, imaging products, and security products, Storage products, and the automotive industry.

The operating frequency of the system largely determines the processing power of the ARM microprocessor. The typical processing speed of ARM7 series microprocessors is 0.9 MIPSMHz, and the common ARM7 chip system main clock is 20 MHz–133 MHz. Different chips deal with clocks differently. Some chips only need one master clock frequency, and some chip internal clock controllers can provide clocks of different frequencies for the ARM core and USB, UART, DSP, audio and other functional components. Most of the on-chip memory capacity of ARM microprocessors is not too large. Users need to expand the memory when designing the system. However, some chips have relatively large on-chip storage space. Simplify the design of the system.

Serial interface circuit is used for short-distance two-way serial communication between S3C4510B system and other application systems. One serial port is a full-function serial port, and the other serial port is shared by RS232/RS485; the reset circuit can complete system power-on Reset and user button reset when the system is working. It also has a watchdog function, and can reset the system board core, system board and JTAG respectively;-The power supply circuit is a 5 V to 3.3 V DC-DC converter, which is S3C4510B And other peripheral circuits that require 3.3 V power supply; 10 MHz active crystal oscillator provides the working clock for the system, and is multiplied by the on-chip PLL circuit to 50 MHz as the working clock of the microprocessor; - FLASH memory can store the debugged user applications, Embedded operating system or other user data that needs to be saved after the system is powered off; SDRAM memory is the main area of the system when it is running, the system, user data, and stack are all located in the SDRAM memory; - 10M/100M Ethernet interface is the system Provides a physical channel for Ethernet access. Through this interface, the system can access Ethernet at a rate of 10M or 100 Mbps;-The USB interface provides two downstream ports and one upstream port that comply with the USB1.1 specification [7]. It can be used as a USB host to connect to a USB device, it can also be used as a USB device to communicate with a PC, and it can also work in a USB bridge mode. The JTAG interface is used to access all the components inside the chip, through which the system can be debugged, programmed, etc.; IIC memory can store a small amount of user data that needs to be stored for a long time;—LED digital display is used as the digital information of the system parameters Display; The configuration circuit is mainly used by the user to configure the operating state of the system by adjusting the level of some pins of the ARM microprocessor [8];—The keyboard module is an expansion board of the system, using 64 key functions and 8 A 7-segment digital display provides more human-computer interaction functions. LCD expansion board is mainly used to display Chinese and Western characters and simple graphics; CPLD expansion board is used to implement some user-defined logic and complete specific logic operations; a real-time clock expansion board is mainly used as a system clock reference;–ADC expansion board completes the conversion of analog quantity to digital quantity for further processing by ARM microprocessor; DAC expansion board converts the digital quantity calculated by ARM microprocessor into analog quantity, which acts on the controlled object; IDE expansion board is used Connect IDE hard disk to realize large-capacity storage of data;-Audio input and output expansion board can realize the collection, processing, amplification and output of audio information;-Video input expansion board

mainly completes the collection and compression of video information for ARM micro-processing The device is further processed; the system bus expansion leads to the data bus, address bus and necessary control bus, which is convenient for users to expand the peripheral circuit according to their own specific needs.

2.2 Design Data Manager Based on Data Mining

In addition to the ARM7TDMI core, the more important on-chip and off-chip function modules of S3C4510B include: 2 buffer descriptors HDLC channel, 2 UART channels, 2 GDMA channels, 2 32-bit timers and 18 programmable I/O ports. The on-chip logic control circuit includes: an interrupt controller DRAM/SDRAM controller ROM/SRAM and flash controller system manager, an internal 32-bit system bus arbiter and an external memory controller. The structural block diagram of S3C4510B is shown in Fig. 1:
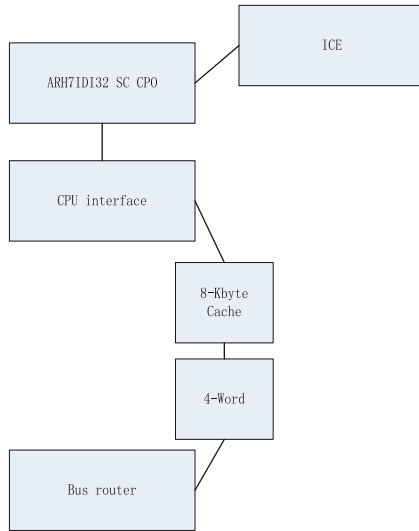
Fig. 1. Data manager structure

This manager is used in an integrated system for embedded Ethernet applications-full 16/32 RISC architecture-supporting large and small endian modes. The internal architecture is big-endian mode, and the external memory can be big-endian or small-endian mode. A high-efficiency and powerful ARM7TDMI processor core is included-a cost-effective, JTAG interface-based debugging solution, boundary scan interface.

The System Manager of the S3C4510B microprocessor plays a vital role in the work of the entire system. It mainly has the following functions: Based on a fixed priority [9], it arbitrates the system bus access requests from several main functional modules.—Provide necessary memory control signals for accessing external memory. For example, if the DMA controller or CPU wants to access a certain address of the DRAM group, the DRAM controller of the system manager will generate the necessary normal/EDO

or SDRAM access signals. The signal to access normal/EDO or SDRAM can be set by SYSCFG. It provides necessary signals for bus communication between S3C4510B and ROM/SRAM and external IO group. The difference in bus width is coordinated for the data flow between the data bus of the external memory and the internal data bus. For external memory and IO devices, S3C4510B supports both little-endian mode and big-endian mode access. By generating an external bus request signal, the peripheral can access the external bus of the S3C4510B. In addition, S3C4510B can access low-speed peripherals by inserting a wait period (WAIT signal). The WAIT signal is generated by the peripheral, which can extend the memory access cycle of the CPU.

2.3   Design System Storage Mapper

S3C4510B uses unified addressing to map the system's off chip memory, on-chip memory, special function registers and external I/O devices to a 64 MB address space. At the same time, in order to facilitate management, the address space is divided into memory groups as shown in Fig. 1. If there are multiple memory groups, they can be configured to include Base Pointer and End Pointer Set the size and position of each memory group. Users can use the base pointer and tail pointer to set continuous memory mapping. The specific operations are as follows: set the address of the base pointer of a memory group to the address of the tail pointer of the previous memory group. Please note that when setting the control register of the memory group, the address of each two connected memory groups Address spaces must not overlap, even if these groups are disabled [10]. The starting physical address of each group is "base pointer moves 16 bits left" and the physical address at the end of each group is "tail pointer moves 16 bits left - 1".

The starting address of external I/O group 1 is equal to the starting address of external I/O group 0 + 16 KB. Similarly, the starting address of external I/O group 2 is equal to the starting address of external I/O group 0 + 32 KB, the starting address of external I/O group 3 is equal to the starting address of external IO group 0 + 48 KB. Therefore, the total continuous addressable range of the four external groups is defined in the starting address of external I/O group 0 + 64 KB of address space. In the entire addressable address space, the start address of the external I/O group is not fixed. By setting the base pointer of the group, a specific group start address can be set, but the total address space is continuous 64 KB.

After power-on or system reset, the address pointer registers of all groups are initialized to their default values. At this time, all group pointers (except ROM/SRAM/Flash group 0 and special function register group) are cleared. This means: Except ROM/SRA MF lash group 0 and special function register group, all other groups are undefined when the system is started. When the user is designing a program, it is generally necessary to first define the storage space of the system by configuring the corresponding register.

The reset values of tail pointer and base pointer of ROM/SRAM/flash group 0 are 0x200 and 0x0 respectively. This means that after system reset, the address space of ROM/SRAM/flash group 0 will be automatically defined as 32 MB, and the actual address range is 0x00000000–0x02000000-1. This initialization definition of ROM/SRAM/flash group 0 enables the system to hand over the control of the system to the startup code written by the user after power on or reset. Of course, these startup codes should be stored in the external ROM and mapped to ROM/SRAM/flash group 0.

When the startup code is executed, it performs various system initialization tasks, and reconfigures the memory mapping of the system according to the actual situation of the external memory and equipment of the application system.

The base address pin of the special function register group is initialized to 0x3ff0000 during system reset, and generally will not be changed. The mapper is described in Table 1:

Table 1. System storage mapper parameters

Mapper element	Offset	Operate	Describe	Reset amount
SYSCFG	0.0000	Read, Write	System configuration Register	0×37FFFF91
CLKCON	0.3000	Read, Write	clock control register	0×37FFFF91
EXTACONO	0.3008	Read, Write	Register 1	0×37FFFF91
EXTDBWTH	0.3000	Read, Write	Register 2	0×37FFFF91
ROMCON1	0.3018	Read, Write	Controller	0×37FFFF91
ROMCON2	0.3011	Read, Write	Controller	0×37FFFF91
ROMCON3	0.3024	Read, Write	Controller	0×37FFFF91

The S3C4510B microprocessor can detect and respond to the bus request signal (Ext MREQS) generated by the external bus master. When the CPU sends out the external bus response signal (Ext MACK), the bus control is handed over to the external bus master, and the external bus request signal should continue to be valid at this time. When the external bus response signal of S3C4510B is valid, its memory interface is in a high-impedance state so that the external bus master can drive the external memory interface. When the S3C4510B does not control the bus, it will no longer perform DRAM refresh operations. Therefore, when the external bus master obtains bus control and it will last for a long time, it must be responsible for completing the DRAM refresh operation.

3   Software Design of the Diversified Evaluation System for Teaching Quality in Colleges and Universities

3.1   Establish a Diversified Evaluation Model for Teaching Quality

The mathematical model of fuzzy comprehensive evaluation is composed of index set T, judgment set P and judgment matrix R. Establish a hypothetical set of known indicators:

$$T = \{t_1, t_2, ...t_m\} \tag{1}$$

Among them, $\{t_1, t_2, ..t_m\}$ represents the composition of known indicators and m is the number of known indicators.

Judgment set:

$$P = \{p_1, p_2, ...p_m\} \tag{2}$$

Among them, $\{p_1, p_2, ...p_m\}$ represents the composition of the evaluation index data. The weight of each indicator is the fuzzy subset Q on T:

$$Q = (q_1, q_2, ...q_m) \tag{3}$$

Under the traditional teaching method, the assessment method of students' ability is mainly through the "examination + usual" mode. "Examination" is based on the final exam of each semester, and an exam is used to assess the degree of mastery of a certain course. "Usually" is mainly based on the teacher's subjective impression evaluation of the students' learning process during a semester. This kind of evaluation is highly subjective and arbitrary, and cannot objectively evaluate the students' learning process. In order to enhance the objectivity of "normal grades", many teachers will subdivide their grades, adding subdivision grades such as "homework", "attendance", and "experiment". There are also many shortcomings in the way of subdividing the grades, such as the inability to confirm whether the homework and experiment are completed independently, and so on. Based on the above situation, q_i in formula (3) is the weight corresponding to i indicators, then:

$$\sum_{i=0}^{m} q_i = \frac{Q}{P} \tag{4}$$

Suppose the evaluation of the i index is the fuzzy relationship from T to P:

$$R_i = (r_{i1}, r_{i2}..., r_{im}) \tag{5}$$

When the evaluation model is established, it is integrated into the evaluation model. The mining requirements are random and usually have no special requirements. The output of data mining may be hidden and possibly related information; Data mining has certain prediction function. By mining the information in the database, we can find the possible accidental events and predict the accidental events; Fast response to data changes. Because the changes of information and demand are relatively rapid, data mining needs to respond quickly to these needs and changes and respond in time. Data mining is more about maintaining and updating rules and requirements; The accuracy of data mining depends on a large amount of data. Only after analyzing a sufficient amount of data, the results are scientific and the laws are credible. The steps of data mining in the evaluation model are:
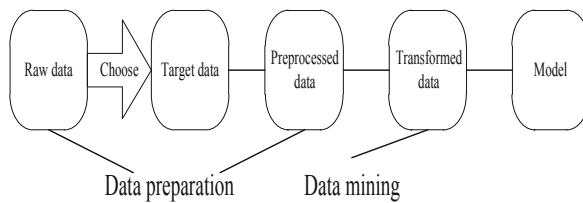
Fig. 2. Steps of data mining

In Fig. 2, in the data mining of the evaluation model, first select the original data, then determine the key data to be mined, preprocess according to the determined data, and transform the transformed data into patterns to complete the in-depth data mining.

Data preparation is similar to requirements analysis in the software development process. First, determine the source of the data, and preprocess the data so that the system can identify and operate, and then the data needs to be transformed and subtracted. Data selection refers to selecting useful data to be processed from a large amount of data stored in the database. In this step, the selected data will be processed preliminarily, so that these data can be recognized and manipulated by the data mining system. The main operation in this step is to check for deficiencies and delete data that does not meet the requirements. The evaluation matrix of m indicators is established as:

$$R = \bigcup_{i=1}^{m} R_i = \begin{bmatrix} r_{11}, r_{12}, ...r_{1n} \\ r_{21}, r_{22}, ...r_{2n} \\ ... \\ r_{m1}, r_{m2}, ...r_{mn} \end{bmatrix} \tag{6}$$

Therefore, the result of comprehensive evaluation is the synthesis of Q and R

$$E = Q^{\circ}R = (e_1, e_2, ...e_n) \tag{7}$$

It is a matrix with m rows and n columns, which is a fuzzy subset of P. After computer simulation and comparison, the combined operation of the teaching quality evaluation in the system adopts the weighted average type. If the system is more complex, there are many indicators to consider, and there are levels of division between the indicators, then you can also use the indicator set T to comprehensively evaluate each category in some way, and then perform the evaluation results. The high-level comprehensive evaluation between "classes" is the so-called multi-level comprehensive evaluation. At this time, in the indicator set T, t consists of k sub-indicators:

$$t = \{t_{11}, t_{12}, ..., t_{1k}\} \tag{8}$$

t2 consists of j sub-indicators:

$$t_2 = \{t_{21}, t_{22}, ..., t_{2j}\} \tag{9}$$

By analogy, we can get:

$$t_m = \{t_{m1}, t_{m2}, ..., t_{ml}\} \tag{10}$$

In the formula, k, j, 1 etc. can be taken as $1, 2 ...$, etc. according to the actual situation. Therefore, according to the relationship between each indicator, an indicator relationship table can be divided and constructed, as shown in Table 2 (in the Table 2 indicators and 2 layers are taken as an example), where qi and qij respectively represent the weight of each main indicator and sub-indicator:

This step is the main step of data mining and the embodiment of the core technology of data mining. At this stage, the processed data are mined, and the result is to find

Table 2. The relationship table of the divided structure index

Second floor			Level one		
Serial number	Evaluation index	Weights	Serial number	Evaluation index	Weights
One	t1	q1	1	t11	q11
			2	t12	q11
Two	t2	q2	1	t21	q21
			2	t22	q22
			3	t23	q23
			4	t24	q24
			5	t25	q25

the hidden rules in the data. In the data mining stage, there are also the following steps: determine the specific methods of data mining: different data formats and types, different methods; Different methods are adopted for different results. Therefore, at this stage, the method selection should be made for different data. Select mining algorithm: for different formats of data and different needs, different mining algorithms should be selected. Different algorithms have different accuracy and different mining results. Mining implementation: use appropriate methods and algorithms to mine data on data sets.

3.2 Design Evaluation Model Database

The target objects of this system are mainly students, teachers and administrators. The target objects of this system are mainly students, teachers and administrators. After students log in to the system, they can complete the following operations on the corresponding interface. Viewing personal information: Students can enter this page to view personal information. Online teaching function: students can enter this interface to perform related operations such as course selection and evaluation. Others: students can perform account-related operations such as login and logout. The school's teaching quality evaluation system has its own unique features. After summary analysis, it mainly includes: There are three main types of users of the system: students, teachers, and administrators of the evaluation system; in the actual operation of the school, the relationship between teachers, students and courses It is a one-to-many relationship. The system will be recorded in this system. Since there are three main types of users, these three types of users correspond to three entities: students, teachers, and administrator entities, which are mainly related to courses, etc. In addition, it is also necessary to record the information of students and teachers. Course information.

After the physical structure design of the database is completed, these physical designs need to be transformed into logical designs. Scientific database design can effectively organize data, save storage space, ensure data integrity and improve data access speed.

In the teaching quality evaluation system, users are divided into teachers, students and administrators. Among them, the relationship between teachers, students and courses is one to many. A teacher can teach multiple courses, and students can also choose multiple courses. At the same time, students need to score a course and its corresponding teachers. Teachers can view the student scores of the courses they teach. According to the above analysis, the data tables to be created include:

(1) Student information table: The fields in this table include basic information such as student ID, name, password and class department, among which the student ID is the main key, and the rest of the fields cannot be empty.
(2) Teacher information table: The fields of this table include personal information such as the teacher's number, name, etc., as well as information such as the course department, and the teacher's number is the primary key.
(3) Administrator table: The fields of this table include personal information such as the administrator number. The number is the primary key.
(4) Course table: The fields of this table include information such as course number, name, teacher, and teaching time and place.
(5) Course selection record table, including data items: course number, course name, student number.
(6) Student evaluation record table: The fields of this table include information such as student number, teacher number, course number, and evaluation score.
(7) Teacher score record table: The fields contained in this table include information such as the teacher's number, name, course number, and score.
(8) Message form: number, message, teacher number, teacher name. According to the above data table, when the actual database design is carried out, it is necessary to restrict the fields in the data table and the relationship between the table and the table, so as to ensure that the database design conforms to the standard paradigm. The structure of each data table in the database is shown in Table 3, 4 and Table 5:

Table 3. Teacher number table

Serial number	Field name	Type	Illustrate
1	T_number	Char(8) not null	Teacher ID
2	T_name	Varchar(8)	Teacher's name
3	T_tech	Varchar(2)	Teacher title
4	T_major	Varchar(20)	Teacher professional direction
5	Deptnumber	Varchar(8)	Faculty number of the teacher

The user's model is independent of any kind of data model and any specific database management system (DBMS). Therefore, it is necessary to convert the conceptual model to a data model supported by a specific DBMS, and then establish the database that the user needs. This system uses SQLServer2020 as the database management system to establish the evaluation system database (teaching).

Table 4. Student table

Serial number	Field name	Type	Illustrate
1	S_number	Char(8) not null	Student ID
2	S_name	Varchar(8)	Student name
3	S_sex	Varchar(2)	Student gender
4	deptnumber	Varchar(20)	Student's department number
5	S_pawd	Varchar(8)	System login password

Table 5. Evaluation index data table

Serial number	Field name	Type	Illustrate
1	Deptnumber	Varchar(8) not null	Department number
2	Tarnumber	Char(4) not null	Evaluation index number
3	Tarname	Varchar(12)	Evaluation index name
4	tarkind	Varchar(8)	Evaluation index category (theory, experiment, sports)
5	tardate	Varchar(9)	Semester of the school year using evaluation indicators

4 Test Experiment

In order to verify the practicability of the system designed in this paper, the system test experiment is designed. Software test is the key step to ensure software quality. It is the final review of software specification, design and coding. The purpose of software testing is the process of executing the program in order to find the errors in the program. A good test scheme is a test scheme that is very likely to find the errors that have not been found so far.

4.1 Test Preparation

In order to make the evaluation results of teaching quality universal, the courses with a large number of students are selected as the test objects, including advanced mathematics, College English, college physics, computer foundation and program design. In order to analyze the performance of the university teaching quality evaluation system integrated with neural network, the simulation test is carried out. The parameter settings of the simulation test environment are shown in Table 6:

Under the above experimental environment setting conditions, the experimental analysis is carried out. In the experiment, the English Majors of a certain school are selected as the sample data, and the student quality of 100 students of the major in the first half

Table 6. Test environment

Environment type	Parameter	Parameter value setting
Hardware	CPU	Intel Core i510500
	RAM	Kefu DDR4 2666 8 GB
	Motherboard	MSI B450M MORTAR
	Hard disk	Western Digital Blue Disk 1TB64MB
Software	Programming tools	Java
	Operating system	Linux

of the semester is analyzed for research and analysis. The number of experimental itera-
tions is 120, and the evaluation accuracy of the experimental sample is analyzed through
iteration. The specific sample teaching scenario is shown in Fig. 3:

Fig. 3. Specific sample teaching scenario.

Choose RBF neural network, BP neural network, this paper design college teaching
quality evaluation system to conduct comparative experiments, and use college teaching
quality evaluation correct rate as the result evaluation index. Randomly select 1000
sample data from the sample data in Table 2 to form the verification sample set, and the
others are used as the training sample set to obtain the correct rate of college teaching
quality evaluation of various systems as shown in Fig. 4:

The test results are shown in Fig. 4. The evaluation accuracy of the evaluation system
designed in this paper is more than 95%, which is higher than the evaluation coefficient
of the other two methods.
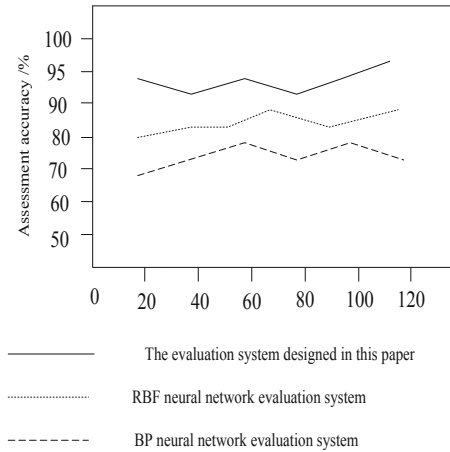
Fig. 4. Test results

The experiment analyzes RBF neural network, BP neural network and the university teaching quality evaluation system designed in this paper to compare the evaluation time. The results are shown in Fig. 5:
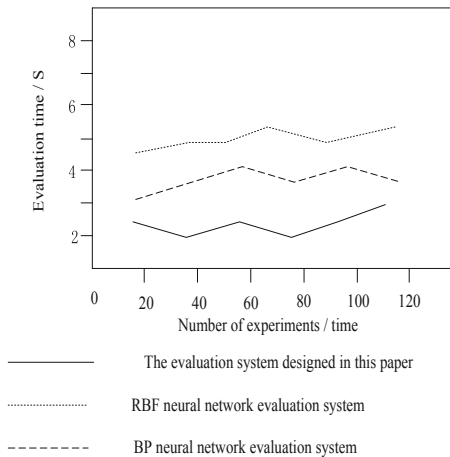
Fig. 5. Time consuming analysis of teaching quality evaluation system in Colleges and universities with different methods

By analyzing the experimental results in Fig. 5, it can be seen that there are some differences in the results obtained by comparing the evaluation time using RBF neural network, BP neural network and the college teaching quality evaluation system designed in this paper. Among them, the evaluation using RBF neural network and BP neural network takes a long time, and the evaluation using the university teaching quality evaluation

system designed in this paper takes a short time, which verifies the effectiveness of the proposed method.

5 Conclusion

This paper proposes to design a diversified evaluation system of teaching quality in Colleges and Universities Based on data mining technology. In the hardware design of the diversified evaluation system of teaching quality in Colleges and universities, the ARM microprocessor based on diversified evaluation and the data manager based on data mining are designed. In the software design of the diversified evaluation system of teaching quality in Colleges and universities, the diversified evaluation model of teaching quality is established and the evaluation model database is designed. Realize the design of diversified evaluation system of teaching quality in Colleges and universities, find these potential relationships and summarize the laws according to the association rule technology in data mining, so as to better support the management of teaching quality. The test shows that the evaluation accuracy coefficient of the system is more than 95, which is much higher than the traditional two systems. Although the existing method research at this stage is feasible, there are some deficiencies in determining the key degree of indicators in the selection of quality evaluation indicators, which need to be improved.

References

1. Bao, L., Yu, P.: Evaluation method of online and offline hybrid teaching quality of physical education based on mobile edge computing. Mob. Netw. Appl., 1–11 (2021)
2. Dong, Q.W., Wang, S.M., Han, F.J., Zhang, R.D.: Innovative research and practice of teachers' teaching quality evaluation under the guidance of 'innovation and entrepreneurship.' Procedia Comput. Sci. **154**, 770–776 (2019)
3. Bao, L., Yu, P.: Evaluation method of online and offline hybrid teaching quality of physical education based on mobile edge computing. Mob. Netw. Appl. **12**(01), 1–11 (2021)
4. Dong, Q.W., Wang, S.M., Han, F.J., Zhang, R.D.: Innovative research and practice of teachers' teaching quality evaluation under the guidance of "innovation and entrepreneurship." Procedia Comput. Sci. **154**(02), 770–776 (2019)
5. Jian, Q.: Multimedia teaching quality evaluation system in colleges based on genetic algorithm and social computing approach. IEEE Access **7**(14), 15–20 (2019)
6. Jin, X.: Deep mining simulation of unstructured big data based on ant colony algorithm. Comput. Simul. **37**(11), 329–333 (2020)
7. Jiang, L., Wang, X.: Optimization of online teaching quality evaluation model based on hierarchical PSO-BP neural network. Complexity **14**(7), 1–12 (2020)
8. Chen, Y.: College english teaching quality evaluation system based on information fusion and optimized RBF neural network decision algorithm. J. Sens. **14**(5), 1–9 (2021)

9. Liu, S., Chen, X., Li, Y., Cheng, X.: Micro-distortion detection of lidar scanning signals based on geometric analysis. Symmetry **11**(07), 1471 (2019)

10. Liu, S., Bai, W., Srivastava, G., et al.: Property of self-similarity between baseband and modulated signals. Mob. Netw. Appl. **25**(4), 1537–1547 (2020)

Design of a Lightweight MOOC Teaching System for Online Learning in Colleges and Universities

Yangbo Wu[1,2](✉), Edmund Ng Giap Weng[1], and Ying Lin[3,4]

[1] Faculty of Computing and Informatics, University Malaysia Sabah, 88000 Sabah, Malaysia
wuyangbo@126.com
[2] College of Mathematics and Computer, Xinyu University, Xinyu 338000, China
[3] Faculty of Psychology and Education, University Malaysia Sabah, 88000 Sabah, Malaysia
[4] College of Foreign Languages, Xinyu University, Xinyu 338000, China

Abstract. In order to improve the teaching function and operation performance of the MOOC teaching system in colleges and universities, under the condition of online learning, a lightweight MOOC teaching system in colleges and universities is designed from three aspects: hardware, database and software. In terms of hardware, the program operation controller, teaching data acquisition and transmission module of the system are mainly modified and optimized. Collect the user and related teaching resource data in the teaching system, install it in a certain format to form a database table, and obtain the design result of the system database through the logical relationship between the database tables. This process keeps the system running in a lightweight state. With the support of hardware devices and databases, the identity of the users entering the system is determined, and students and teachers are given different functional rights. The software teaching function of the system is realized through the design and development of functional modules such as interactive practice of lightweight MOOC teaching content, uploading MOOC teaching resources, online testing and score management in the course management module. Through the system test experiment, it is concluded that the functional operation success rate of the designed system is higher than 99%, and the operation performance of the system in the two aspects of response speed and concurrency has been significantly improved.

Keywords: Online learning · College teaching · Lightweight · MOOC teaching system

1 Introduction

In recent years, the new educational concepts of "Flip Classroom", "MOOC" (Massive Open Online Course) and "Micro-class" are promoting the continuous reform of education. The Outline of the Ten-Year Development Plan for Education Informatization (2011–2020) of China points out that the development of education informatization

requires the innovation of educational concepts, the construction of high-quality educational resources and information-based learning environment, and the innovation of learning methods and modes. Other relevant policies such as the Outline of National Informatization Development Plan (2006–2020) and the Outline of National Medium and Long-term Education Reform and Development Plan (2010–2020) have raised the importance of information technology to education to a new level [1]. Massive Open Online Course (MOOC) refers to a web-based class designed to support large number of participating students. It can deliver learning content online to any person who wants to take a course, with no limit on attendance. It is the product of "Internet education". Due to the open teaching resources, the integration of various teaching methods, and the spread of the Internet, the teaching system can effectively improve students' initiative, and is widely loved by teachers and students. Especially in the new corona- pneumonia epidemic environment, online learning in the MOOC class teaching model has become one of the major universities teaching methods.

According to the present research, the flow design of the teaching system is unreasonable, such as the storage, audit, management and backup of the video data is imperfect. At the same time, the MOOC system seldom involves student credit management, and lacks real-time communication between students and teachers, and the popularity of MOOC management system is not high. And when a large number of users online learning, online interactive discussion, online evaluation, the traditional server is difficult to provide the corresponding technical support. It can be seen that the existing teaching system of MOOC courses in colleges and universities has some problems, such as slow response speed, poor functional performance and poor concurrency. In order to improve the function and operation performance of the system, the light-weight teaching system of online learning in colleges and universities is designed optimally.

2 Design of Lightweight MOOC Teaching Hardware System in Colleges and Universities

Under the traditional model, the design and construction of university mathematics MOOC teaching platform is usually based on specialized information technology, the development threshold is high, and the functions are solidified, which is not conducive to university mathematics teachers to independently create MOOC teaching platform. In order to improve the above-mentioned problems encountered in the construction of the university mathematics MOOC system, and enhance the teaching and application experience of the university mathematics MOOC, the lightweight MOOC teaching system in universities has been optimized from the three aspects of hardware, database and software. The system should have the teaching attribute functions of the conventional MOOC platform, and should also have the characteristics of convenient creation and low maintenance threshold, so that college mathematics teachers who do not have the professional knowledge of information systems can build and maintain a personalized MOOC teaching platform. The system structure designed this time is B/S. Based on the characteristics of the B/S structure, the university teacher-student exchange platform studied in this paper consists of three layers: application layer, service layer and data

layer. The application layer provides users, including teachers, students, and administrators, with interfaces to use system functions. The service layer provides support for the implementation of the application layer functions. The data layer provides the data required to realize the functions of the application layer and completes the management of system data.

System design requirements analysis:

(1) Input and output requirements: ensure that users can input new, error-free data or update data information to the system without errors, and enable users to complete the input work conveniently and easily; the main requirement for output is to ensure that the system can be used in appropriate provide the right information to the right people at the right time and place.
(2) Reusability requirements: When designing a system, try to make the code written can be applied to future program development, which requires standardizing the code, simplifying it as much as possible, reducing redundancy, and saving future system development. Human and material resources.
(3) Management requirements: The manageability of the system includes the manageability realized under daily working conditions and the manageability realized in emergencies or major changes. In order to realize the manageability of the system, the hardware such as the host and database involved in the system should be effectively supervised and deployed. The essence is to ensure the controllability of the system to further ensure the effectiveness of the system by constantly monitoring and managing the information passed by the system itself.

2.1 PLC Selection

PLC is the product of the combination of microcomputer technology and the conventional control concept of relays. It is a special computer mainly used for digital control with a microprocessor as the core [2]. Therefore, its hardware configuration is similar to a general microcomputer device. The hardware of a PLC is mainly composed of a central processing unit, a memory, an input unit, an output unit, a communication interface, a power supply, and an expansion interface power supply, as shown in Fig. 1.
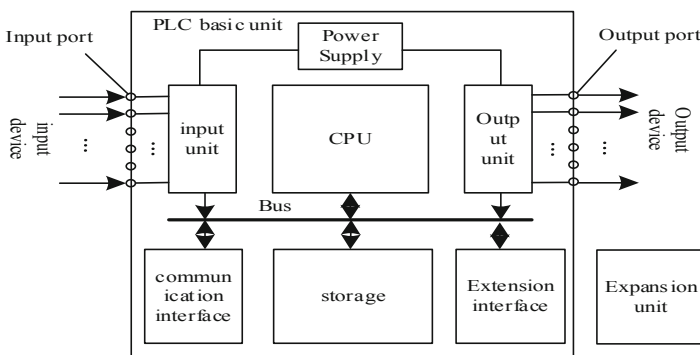
Fig. 1. PLC hardware configuration diagram

The processor in PLC is divided into two parts: word processor and microprocessor. Word processor is the main processor, which is used to execute the interface function of programmer, process byte operation instruction, control system bus, internal timer, internal counter, monitor scan time, coordinate bit processor and input and output, etc. The microprocessor is a slave processor, which is mainly used to process the bit operation instruction and realize the conversion from PLC programming language to machine language. There are two main types of memory: read-write random access memory RAM and read-only memory ROM, PROM, EPROM, and EEPROM. In PLC, the memory is mainly used to store system program, user program and working data. The lightweight MOOC teaching system in colleges and universities needs to complete function subroutine call management, logic operation, communication and parameter setting, etc. Therefore, the space of memory in PLC hardware is more than 1 TB, and the peripheral memory is connected by the external interface. In the ROM, PROM or EPROM of the internal read-only memory during the use of PLC, the user cannot access and modify. The input/output unit is the bridge between the PLC and other parts of the system. PLC through the input interface can detect the current system of various data, these data as PLC control of the object of information. At the same time, PLC sends the processing result to the controlled object through the output interface to drive the function module to realize the control goal.

2.2   Teaching Information Collection and Transmission

In order to make the teaching images clear and legible on the students' terminal display, the display resolution is generally set to 1280 * 1024, and 4096 scanning lines are generated on the students' terminal display, and at least 1024 video points are collected on each scanning line [3].

Suppose the system collects 4096 video points at each azimuth, each point is quantized by 8 bits, the data of a teaching image is 16 MB, so the system needs at least 16 MB of storage space. Figure 2 is the hardware structure of teaching information collection module.
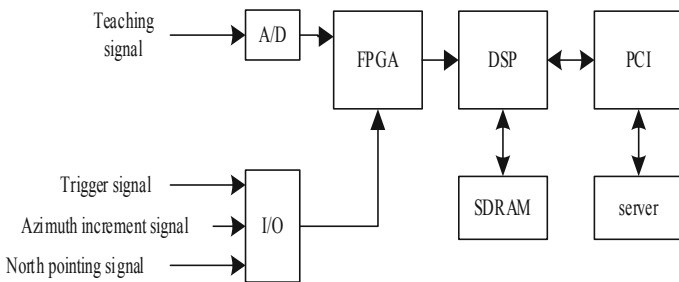
Fig. 2. Hardware structure diagram of teaching information acquisition equipment

The A/D converter in Fig. 2 is responsible for the conversion of teaching video to digital video, and performs filtering processing of clutter. FPGA has good logic operations

and data processing capabilities, and is fully adapted to the real-time and fast processing of the collected teaching image data. DSP is the core part of the system, which mainly completes the processing and transmission control of teaching image information. PCI bus is an advanced high-performance 32-bit/64-bit address data multiplexing local bus. The bus clock frequency is 33 MHz, which can effectively solve the real-time transmission of data and facilitate the real-time processing of data. SDRAM has the advantages of large capacity, low cost, and high speed, and its access efficiency to continuous address storage space is very high. High-performance DSPs are provided with SDRAM controllers to realize the efficient management of massive amounts of data such as teaching signals by the DSP. Teaching image data is transmitted from the server of the acquisition module board to the student terminal through the local area network. The usual local area network file transmission protocols include transmission control protocol and user datagram protocol. The transmission control protocol requires a three-way handshake, has functions such as timeout retransmission, data inspection and flow control, and has good data transmission reliability. However, when using the transmission control protocol, a TCP connection must be established between the student terminal and the server before the student terminal and the server can exchange data with each other.

2.3   ARM11 Processor

ARM11 processor is a new generation of RISC processor launched by ARM, with high-performance processing capabilities. The ARM11 processor uses a 5 V DC power supply. The 5 V input DC power supply generates a stable 5 V output voltage through a special voltage regulator, and then the obtained 5 V stable voltage is converted into the required 3.3 V through a corresponding voltage regulator. Use an external crystal oscillator circuit to provide an accurate clock for it, and an external 32.768 kHz crystal oscillator supported by the backup battery when the system is powered off. In this way, in the case of a system power failure, the backup power supply is used to enable the clock to continue to run, so that time information will not be lost. The ARM11 processor has four A-type USBHost1.1 interfaces, one of which is a USBHOST interface for connecting to a multimedia classroom speaker controller [4]. The NAND FLASH memory, K9Fl208, manufactured by Samsung, has a storage capacity of 512 Mb. The memory operates at a voltage of 2.7–3.6 V. The memory enables automatic program erasure, block erasure, page erasure, and intelligently read/write and erase operations [5]. The memory can read/write at one time or erase four pages or four blocks of content. It has multiple registers inside it.

2.4   Lightweight Teaching Server

The Apache lightweight server works by the client using the URL to request the appropriate resource, and when the requested resource is found, the resource returns to the client, completes the request, and disconnects the next time the request is made. Apache server can support a large number of concurrent access and data throughput, MOOC teaching system can meet the needs of many people simultaneously online applications.

3   Software Function Design of Lightweight MOOC Teaching System in Colleges and Universities

3.1   Divide User Roles in the Teaching System

Before starting MOOC teaching, person concerned need to determine the identity of the user who enters the system. Generally, users are divided into three identities: student, teacher, and administrator. Among them, when students use the system, the main process has three levels. The first level is mainly for browsing and viewing information, the second level is mainly for the modification and feedback of information on the web page, and the third level is for uploading and downloading of homework. Among them the content of level 2 must be to undertake an operation after the user is logged in, the premise that evaluates operation inside group is to hand in oneself job hind. Teachers are the main users of the whole online auxiliary teaching system. All the users' information management, teaching information, teaching courseware, teaching video, homework uploading, homework marking, grading and so on need teachers' active participation. The main functions of teachers are divided into four levels: the first level is similar to the students' users, mainly the viewing and browsing of information [8]. The second layer is the modification of information and background entry. The third layer is the background management of the curriculum, including basic settings, student management, class management, project management, task management, operation management, etc. The fourth layer is the editing of teaching resources and the uploading of homework. The administrator is mainly used to give different users different functional permissions.

3.2   Course Management Module

Curriculum information management function is operated by teachers. Curriculum information is used to explain to students an overview of the curriculum, including curriculum name, curriculum overview, curriculum arrangement, etc. This information may vary from semester to semester, so it needs to be managed. The course information management implementation activity diagram is shown in Fig. 3.

   The management process in Fig. 3 is the sequence and criteria for teachers to input information on the teaching system. As shown in Fig. 5, the details are as follows. Teachers open the curriculum information management interface, in this interface has the previous curriculum information, may also not. If there is a teacher, the curriculum will be revised according to the schedule of the semester. Teachers submit the revised information, the application server to determine whether the information is standard, that is, whether there cannot be empty content is empty, and whether the information format is correct and so on. If these are valid, the data is stored in the database. If the storage is successful, the server page returns the modified success screen, which the instructor can see in the browser [9]. If the information format is illegal, then return to the information input page to fill again, and then repeat the verification process, if the data is not stored successfully, then the modification activity. In the lightweight MOOC teaching webpage designed this time, there are page interaction modules such as login and logout, course introduction, course video, courseware, course tasks, and course chapter options.
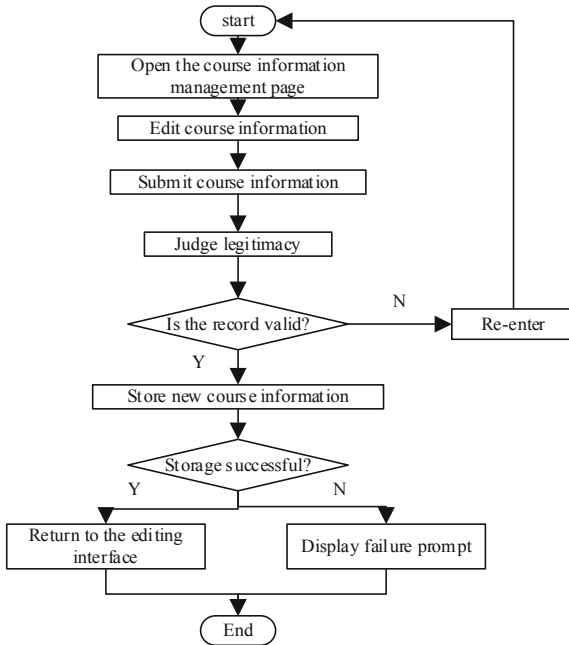
Fig. 3. Course information management flowchart

3.3 Lightweight MOOC Teaching Content Interactive Exercises

There are two parts in the interactive exercise of MOOC: teacher-student interaction and group interaction. The interactive module provides the scene of online teaching, sharing learning experience and grade checking for teachers and students. Specifically, it is divided into three parts: teacher answer module, student message module and performance evaluation module. First, design the "discussion area", "notification bar", "@ specific people" and other functions, students can use these functions to ask teachers anytime and anywhere. Teachers can answer students' questions online instantly. Secondly, the design of the message board, with the help of the board, students can not only ask teachers questions, but also upload pictures, videos and other resources. In the mathematics curriculum some inconvenient with the words expression question by the picture or the video's form sends to the message area, facilitates the teacher to view answers. Finally, after each examination, the teacher can publish the finished student's score to the performance evaluation module, which can give the concrete evaluation grade according to the weight distribution designed by the teacher and the student's score, so that the student can get his own study score in time.

Group interaction learning is to divide batch students into several groups, and then realize the information exchange by the internal group. Automatic grouping algorithm aims to make each group of students are different aspects of the advantages of talent, so as to be able to complement each other's strengths, to better complete the teacher's task. First of all, the student data cleaning, and then analyze the gap between each student indicators, try to make different types of talent in the same group [10]. Assuming that

the value range of each index is 1–5 after data cleaning, the cosine similarity is used to judge the similarity of each student in all aspects of the overall quality.

Given that the two students to be compared have indexes x_i and y_i, their cosine similarity can be expressed as follows:

$$\cos\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{x_i^2}\sqrt{y_i^2}} \tag{1}$$

If there are N students in a certain class, then we can find a N * N similarity matrix to show the similarity between the two students in the class. The value of the cosine can only range from 0 to 1. The closer the value of $\cos\theta$ is to 1, the closer the comprehensive abilities of the x and y students are to each other. And the closer the value of $\cos\theta$ to 0, the greater the difference between the two students, the more they should be divided into the same group, complement each other. The students who are divided into a group are marked with the same symbols, and the data of the same group are shared and transferred.

3.4 Database Design of Lightweight MOOC Teaching System in Colleges and Universities

The database system is the basis of system development. A good database design can ensure that the manpower and financial resources invested in the later maintenance and upgrade of the system are smaller. The system database can ensure that the system has a stable data source [6]. Carry on database design to the system. Database design is an important part of software system design, mainly for database logical structure design and database physical structure design. According to the design of tables and indexes, design the size of the table space reasonably [7]. The relationship between entities is shown in Fig. 4.
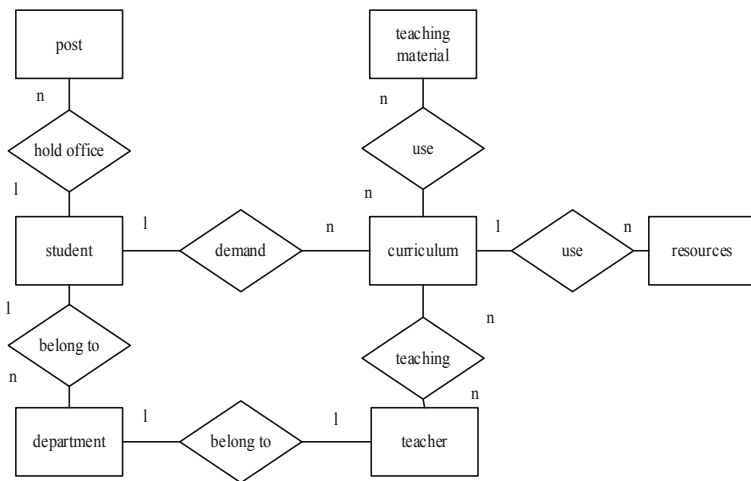
Fig. 4. Relationship between entities e-r diagram

Table 1. Student user basic information

Datasheet item name	Table item type	Word length	Specific meaning	Can the default value be null
student id	int	4	Student ID	Cannot be empty
student	char	20	Student name	Cannot be empty
name	int	4	Student gender	Can be empty
student sex	datetime	8	Student birthday	Can be empty
student —birthday	char	18	Student ID number	Cannot be empty
student_ identity	char	30	Contact address	Cannot be empty
student address	char	14	Student phone	Can be empty
student-phone	char	50	Student mailbox	Cannot be empty

Through the above E-R diagram, and then use tools to export the corresponding tables, such as course information table, teacher table, student table, transcript table, textbook table and so on. The basic information of student users is shown in Table 1.

From this, the construction structure of other data tables in the database can be obtained, and the connection between the database tables can be realized according to the relationship shown in Fig. 3.

3.5 Upload MOOC Teaching Resources

The process of uploading MOOC teaching resources is actually the process of transferring teaching resources from teachers to system terminals. If the distribution area of the system network is a rectangle, the distribution area is $L_1 \times L_2$, and the density of the node in the distribution area is μ, the identity ID of any sensor node is not repeated, and the initial energy of the node is E_0. In addition, if the maximum communication radius of the sensor node is R, the node within the transmission radius of node I is $S_1(R)$, and the node within the transmission radius of the next-hop node is $S_2(R)$, for node I, there is the following relationship:

$$\begin{cases} |S_1(R)| = \left\| (u_i, v_i) - (u_j, v_j) \right\| \\ |S_2(R)| = \left\| (u_i, v_i) - (u_j, v_j) \right\| \end{cases} \qquad (2)$$

In Eq. (2), (u_i, v_i) and (u_j, v_j) are the coordinates of nodes i and j, respectively. For any sensor node i in the network, if its influence region may interact with the influence region of any other node j, the degree of interaction can be described by the overlapping area coefficient ω of node i and node j:

$$\omega = \frac{L_i \cap S_1(R)}{L_i} \qquad (3)$$

In Eq. (3), L_i represents the maximum coverage area of node i. If node i and node j can influence each other, the cross-correlation factor $\gamma(i, j)$ of the radio frequency area of node i and node j satisfies the following relationship:

$$\gamma(i, j) = \frac{R^2 \arcsin \sqrt[3]{\|(u_i, v_i) - (u_j, v_j)\|}}{2R - 1} \tag{4}$$

When there are n nodes interacting with each other around node i, the RF region cross-correlation factor $\gamma(i, j)$ of node i and these nodes satisfies:

$$\gamma(i) = \frac{R^2 \arcsin \sqrt[3]{\sum_{j \in i} \|(u_i, v_i) - (u_j, v_j)\|}}{2R - 1} \tag{5}$$

The larger the value of $\gamma(i, j)$ is, the greater the impact of node i on other nodes. Once node i fails, the transmission path will be interrupted, causing serious transmission jitter in the network. When node i transmits data to node j with distance of l under the condition of transmission bandwidth of B, the energy consumption of i and j satisfies the following relations:

$$\begin{cases} E_{\text{srud}}(i) = Bl + P_0 l^3 \\ E_{rv}(j) = BP_0 l^2 \end{cases} \tag{6}$$

In Eq. (6), P_0 represents the current transmit power of the node, and there is a positive correlation between the energy consumption and l, which is the shortest distance between nodes i and j in Euclidean space. Therefore, the energy consumption of the communication network can be optimized by optimizing the energy consumption of the sensor nodes. Thus, the serious transmission blocking phenomenon of WSN nodes due to energy depletion is reduced, and the stable transmission of MOOC teaching resources is realized.

3.6 Online Test and Score Management

Student performance management functions include self-rating included, group rating included, teachers and comments included, the calculation of the total score of each homework, to achieve a certain score recommended for excellent works, excellent works in the display page. Students' achievement management is a vital part of daily teaching. It can feed back the students' mastery of classroom knowledge in time, adjust the progress of classroom teaching in time, and make up for the missing points. Open the grade management interface, use the query operation, by class query or view all, the system automatically invoke the relevant parameters in the background information displayed in the list, including student number, class, name, number of operations, homework score. After the homework is handed in, the other members of the group will grade the students' homework, and then the teacher will grade the students' homework. Finally, the teacher will calculate the total score of the three grades and decide whether to show the excellent works.

The overall design block diagram is shown in Fig. 5.

Hardware design

| PLC selection | Information collection and transmission module | ARM11 processor | Lightweight Teaching Server |

Software design

| User role design module | Course Management Module | Interactive contact module | Database design |

| Teaching resource upload module | Teaching test scoring module |
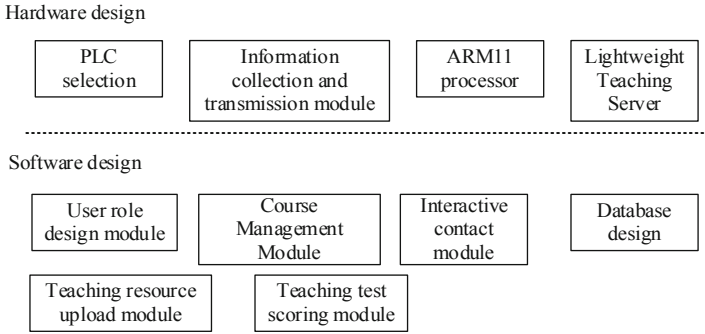
Fig. 5. System design general block diagram

4 System Test

In order to speed up the development speed and improve the accuracy of writing code, the PDT is chosen as the preferred development environment. Eclipse, on the other hand, is available on several major operating systems, including Windows, Linux, and Macosx, and is also an open source development platform, so there is no extra charge for using it. Although eclipse is primarily for Java development, among the many plug-ins for eclipse are two plug-ins for PHP development, eclipse foundation and PHPeclipse, installed in the large open source environment of lamp for PHP development. After automatically downloading and installing the lamp environment, install it using eclipse downloaded from the eclipse website, and you can then complete the foundation development environment for an online learning, lightweight, high-school MOOC instructional system. The programming language used in this test is the C# language. C# (C Sharp) is a programming language tailored by Microsoft for the .NET Framework. C# has the powerful functions of C/C++ and the easy-to-use features of VisualBasic. It is the first Component-oriented programming language. Like C++ and Java, it is also an object-oriented programming language.

After constructing a good system test and running environment, the system test case is designed to test the function module and running performance of the system, and many teachers and students are simulated. The function test index of the teaching system is the success rate of the system function, and the numerical results can be expressed as follows:

$$\eta_{success} = \frac{n_{success}}{n_{total}} \times 100\% \tag{7}$$

$n_{success}$ and n_{total} in Eq. (7) are the number of successful tasks and the total number of tasks set, respectively. According to the function of the teaching system, the task use case is divided into five parts: login management, course management, interactive practice, uploading and downloading of course resources, and grade query. The response speed can be obtained by reading the start time of the task and the output time of the result.

Concurrency is counting the number of people who are online at the same time in different time intervals. The design of online learning oriented lightweight MOOC

teaching system for colleges and universities into the program code into the experimental environment, the results of the system design. Figure 6 shows the course video playback interface of the lightweight MOOC teaching system in universities.

Fig. 6. MOOC remote video playback interface of teaching system

By the same reason, we can get the display interface of MOOC teaching system in different running state, and compare the difference between the actual display interface and the expected result to judge whether the system is successful. After many experiments, the test results of system functions are obtained, as shown in Table 2.

Table 2. System functional test results

System test case type	Set the number of use cases/pcs	Number of successful runs/pcs
Login management	200	200
Course management	200	199
Interactive exercises	200	200
Course resource upload and download	200	199
Result inquiry	200	199

The data in Table 2 are put into Formula 7, and the result shows that the success rate of the design of lightweight MOOC teaching system is 99.8%, higher than 99%, which shows that the teaching function meets the design requirements.

Moreover carries on the statistics to the system function run-time data, obtains the system run-time and the concurrency performance test result, as shown in Fig. 7.

It can be seen intuitively from Fig. 7 that the maximum response time of the system function is 0.38 s, which is lower than 0.5 s, and it can ensure that more than 350 people
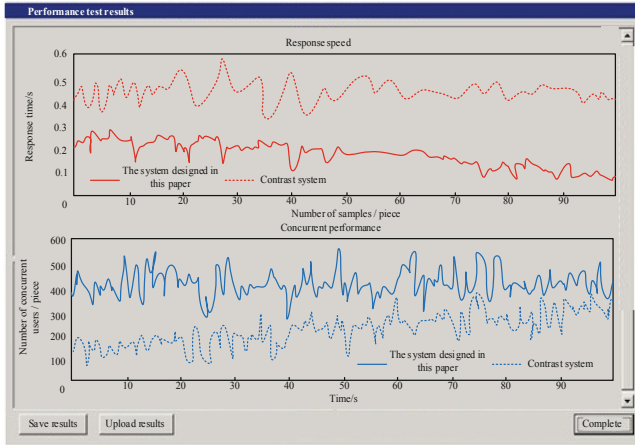
Fig. 7. System performance test results

are online at the same time. The response time of the system used for comparison is higher than 0.3 s, and the average number of people online at the same time is 210. This shows that the designed system has good concurrency performance.

5   Conclusion

In order to improve the teaching function and operation performance of the MOOC teaching system in colleges and universities, this paper designs a lightweight MOOC teaching system design in colleges and universities. In terms of hardware, the program operation controller, teaching data acquisition and transmission module, processor and server of the system are mainly modified and optimized. Collect users in the teaching system and related teaching resource data, and build a teaching resource database. With the support of hardware devices and databases, the identity of the users entering the system is determined, and students and teachers are given different functional rights. The software teaching function of the system is realized through the design and development of functional modules such as interactive practice of lightweight MOOC teaching content, uploading MOOC teaching resources, online testing and score management in the course management module. It can be seen from the system test results that the design system has obvious advantages in both function and performance. At the same time, this system still has certain shortcomings. For example, the function of the system is not perfect due to time reasons, and the interface of the system is not mature enough. These are all improvements that I need to make in the future, and I also hope to get criticism and guidance from the teacher.

References

1. Liu, S., He, T., Dai, J.: A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese. Mob. Netw. Appl. **26**(5), 1891–1903 (2021). https://doi.org/10.1007/s11036-020-01725-x

2. Xu, X., Li, D., Sun, M., et al.: Research on key technologies of smart campus teaching platform based on 5G network. IEEE Access **7**, 20664–20675 (2019)
3. Liao, J., Tang, J., Zhao, X.: Course drop-out prediction on MOOC platform via clustering and tensor completion. Tsinghua Sci. Technol. **24**(4), 412–422 (2019)
4. Wang, S., Mu, M.: Exploring online intelligent teaching method with machine learning and SVM algorithm. Neural Comput. Appl. (2021). https://doi.org/10.1007/s00521-021-05846-6
5. Duan, C.: Design of online volleyball remote teaching system based on AR technology. AEJ Alex. Eng. J. **60**, 4299–4306 (2021)
6. Li, M., Li, Y., Guo, H.: Research and application of situated teaching design for NC machining course based on virtual simulation technology. Comput. Appl. Eng. Educ. **28**(3), 658–674 (2020)
7. Scherer, R., Howard, S.K., Tondeur, J., et al.: Profiling teachers' readiness for online teaching and learning in higher education: who's ready? Comput. Hum. Behav. **118**, 106675 (2021)
8. Cobos, R., Jurado, F., Blázquez-Herranz, A.: A content analysis system that supports sentiment analysis for subjectivity and polarity detection in online courses. Revista Iberoamericana de Tecnologias del Aprendizaje **14**(4), 177–187 (2019)
9. Liu, S., Fu, W., He, L., et al.: Distribution of primary additional errors in fractal encoding method. Multimedia Tools Appl. **76**(4), 5787–5802 (2017)
10. Bao, L., Yu, P.: Evaluation method of online and offline hybrid teaching quality of physical education based on mobile edge computing. Mob. Netw. Appl. (2021). https://doi.org/10.1007/s11036-021-01774-w

"Online + Offline" Hybrid Teaching Model in the Post Epidemic Era Based on Deep Reinforcement Learning

Shaolin Liang[1](✉) and Pei Su[2]

[1] College of Mathematics, Sichuan University of Arts and Science, Dazhou 635000, China
uhdsfd545@sina.com
[2] Wuhan Railway Vocational College of Technology, Wuhan 430000, China

Abstract. In order to achieve students' in-depth understanding of the teaching content, in the post-epidemic era, an "online + offline" hybrid teaching model based on deep reinforcement learning has been designed. First, the basic data is preprocessed to remove interfering data and convert it into a form that can be directly used by the model. In the domain knowledge unit of the model, on the basis of determining the composition of the domain knowledge elements and their associated relationships, a structure in which the superordinate relationship and the subordinate relationship, the predecessor relationship and the successor relationship coexist is constructed; in the learner unit of the model, the deep reinforcement determines Based on the learning source, a block-based data management mechanism is established to jointly promote the operation of the model. The experimental results show that the "Online + offline" hybrid teaching model in the post epidemic era based on deep reinforcement learning has good performance and can achieve good teaching results.

Keywords: Deep reinforcement learning · Post epidemic era · "Online + offline" · Mixed teaching model · Domain knowledge unit · Learner unit

1 Introduction

A sudden outbreak of New Coronavirus disrupted people's normal study, work and life. As an emergency measure during the epidemic prevention and control period, "suspend classes and not stop learning" became a milepost event in the development of online education in the world. Today, the epidemic situation is gradually stable, and schools have resumed classes one after another, marking that online teaching has entered the "post epidemic era". When we re embrace the real classroom again, what has this unprecedented and the world's largest online education changed? What's left for us? Will our class go back to the past? Does the classroom teaching model return to the origin or seize the opportunity to change and transform? What is the future form of education? What direction will education develop in the future? These will become the problems that every educator should think and study. "Epidemic period" Online teaching has become

the main way to carry out education. Online teaching is a teaching form in which teachers and students use educational resources to realize interaction on the technical platform. Only by realizing the interaction among teachers, students, resources and technology can we build efficient online teaching. In the large-scale online education practice launched due to the epidemic, the majority of teachers and students from everything in good order and well arranged life, teachers have too much understanding, and students have different experiences. From teachers' and students' perspectives, it is of great significance to summarize the experience and problems of online teaching in the epidemic period. It has far-reaching significance for teachers to study the transformation of classroom teaching mode and promote the integration of online and offline teaching in the era of epidemic [1–3]. With the Internet plus education, With the deepening development of, online teaching based on MOOC, superstar learning link, wisdom tree and other teaching platforms is more and more widely used in colleges and universities. In recent years, the hybrid teaching mode of combining online and offline has gradually become a trend in college teaching [4, 5]. These large-scale open online course platforms not only break the constraints of learning time and place, but also provide learners with rich and diverse learning resources. Students produce a large number of learning behavior data in the process of online learning. By analyzing these learning behavior data, we can explore the learning laws behind the data, and teachers can improve teaching according to these laws. It provides personalized management and teaching for students with different learning situations, which is of great significance to both teachers' teaching and students' learning [6–8]. Based on MOOC and SPOC, the online and offline mixed teaching mode is discussed. The specific implementation steps of SPOC curriculum are described in detail. The design idea and practice process of mixed teaching mode and flipped classroom under line and online are given. The effectiveness and feasibility of the combination of teacher student evaluation and curriculum goal comparison are verified, and the results of questionnaires before and after the course are given. The analysis and reflection provide valuable experience for promoting teaching reform [9]. At present, the analysis and research on the relationship between relevant data in the learning process and teaching effect are basically aimed at the time and energy paid by learners in a certain learning behavior, that is, quantity and quality. These rough data can not accurately and completely reflect the actual situation of teaching, nor can they accurately define and analyze the teaching effect [10]. At the same time, compared with learning participation and learning input, the behavior sequence actually generated by learners in the learning process can better reflect learners' behavior path and cognitive process. With the help of the influence of learners' behavior sequence on learning effect, teachers can determine the key behavior sequence for learning process analysis, so as to monitor learners' learning status and implement teaching intervention in time, the purpose of improving the learning effect is to combine online and offline blended teaching as a teaching mode of university courses. In recent years, it is gradually moving towards campus. Especially under the influence of the New Coronavirus epidemic in 2020, teachers who are not yet ready are learning how to use MOOC, super star learning, intelligent tree, online teaching platform, QQ group classroom, sharing screen and so on. Online and offline hybrid teaching was originally a teaching mode combining the advantages of online teaching and traditional teaching. Through the organic combination of the two teaching

organization forms, learners' learning is led from shallow to deep to deep learning. The ultimate purpose of hybrid teaching is not to use the online platform, build digital teaching resources, or renovate teaching activities, but effectively improve the learning depth of most students.

Based on this, this paper designs a "Online + offline" hybrid teaching model in the post epidemic era based on deep reinforcement learning, and evaluates its teaching effect.

2   Data Preprocessing

In order to achieve the purpose of deep reinforcement learning, it is necessary to mine teaching resources and student data. Data preprocessing is an important step in the process of data mining, because the data in the real world is dirty, or incomplete (the attribute of interest has no value), noisy (there are errors or exceptions in the data, that is, data deviating from the expected value) and inconsistent (inconsistent data connotation) [11]. Data must be preprocessed before mining. The preprocessing of model data in this paper includes data preparation, data cleaning and data transformation.

2.1   Data Preparation

Whether data mining is successful or not, data preparation is very important, and it is the premise of realizing the application of data mining. Data preparation contents: first, determine the data source to mine data and collect the original data; second, merge and sort the data from different data sources into the same database. This model takes the original data in Teachers' basic information table, teachers' evaluation information table, student achievement table and so on as the research object. For example, in the data of teacher evaluation information table, a teacher may correspond to multiple teaching evaluation records. Therefore, it is necessary to average the attribute value of teaching evaluation score in the teacher's N records, so that the teacher has only one record in the database. According to the purpose of data mining, filter out valuable data and establish data source table.

2.2   Data Cleaning

Due to various data quality problems, the data may contain incorrect values. When integrating from multiple different data tables, we must pay attention to the consistency of data between different data tables. Data cleaning is to improve data quality by eliminating tuples such as errors, noise, defects and inconsistencies in the original data set. In this paper, the incorrect and inconsistent data are processed by manual correction. For example, for the record with empty score in the student's grade sheet, ① if the student has transferred or dropped out, the record will be deleted directly, and the student's record will be deleted in the student's basic information data sheet; ② If the teacher missed or mistakenly entered the score in the "positive teaching management model", the correct score will be entered. And the students' scores: first, the calculation of students' make-up examination scores. According to the relevant regulations of our school,

students who fail in the examination shall take a make-up examination, and the make-up examination result shall be included in the student file according to 60 points. There is a certain difference between the make-up examination result and the original result. In order to ensure the accuracy and rationality of data mining results, the result of failed students shall be subject to the original result rather than calculated according to the make-up examination result. Second, the calculation of the scores of students who do not normally take the exam. According to the regulations of our university, students may not take the examination with the approval of relevant departments of the University for special reasons, but they must take the make-up examination, and the make-up examination results shall be included in the student file according to the normal examination results.

2.3 Data Conversion

Data transformation is to transform data into a description form suitable for data mining. The transformation of this paper mainly includes the following contents.

(1) Smoothing: if the current data point is null or noisy data, take out the weighted average of K (K can be customized) data points before (after) the current point and replace them.
(2) Aggregation: summarize and summarize data. It mainly constructs the data side for multi granularity data analysis.
(3) Data generalization: data generalization is to replace low-level or data level data objects with more abstract or higher-level concepts.

In this paper, the method of data generalization is mainly used for data conversion. First, convert the attribute value of birth date in the data into the corresponding age segment, convert the attribute value of workload into the corresponding workload segment, convert the attribute values of middle school students' admission average score and students' school average score in the data into the corresponding grade, and convert the attribute value of teaching evaluation score in the data into the corresponding teaching evaluation grade. According to the principle that the teaching evaluation score is not less than 60 points, There are three grades of design evaluation: excellent, medium and qualified. In the specific transformation process, the nodes experienced by the data form a set of model meta nodes. There is a data transmission link between each two nodes. Under this condition, the model can be expressed as:

$$B = (b_{i,j})n * n \in \{0, 1\}^{n*n} \tag{1}$$

In the formula, $b_{i,j}$ represents the node location of the model data center. If there is link connectivity between the two nodes, the corresponding element is 1, otherwise it is 0. It is assumed that there is a non ring path link L between two nodes. If the link passes through the node, there is $l \in b_{i,j}$. At this time, the implementation process of the three

transformations is:

$$C_l = \sum_{i=a}^{n} \sum_{j=a}^{n} c_{i,j} \cdot l_{i,j}$$

$$D_l = \sum_{i=a}^{n} \sum_{j=a}^{n} d_{i,j} \cdot l_{i,j}$$

(2)

$$H_l \leq \sum_{i=a}^{n} \sum_{j=a}^{n} l_{i,j}$$

(3)

$$s(l_{i,j}) = h(l_{i,j}) - H_l \leq 0$$

(4)

In the formula, H_l represents the link effectiveness factor corresponding to the link represented by the individual, and C_l, D_l, $h(l_{i,j})$ and $s(l_{i,j})$ are different attributes of the data respectively. Through this condition, it provides a reference for the construction of hybrid teaching model.

3 Design of "Online + Offline" Mixed Teaching Model Based on Deep Reinforcement Learning

3.1 Framework Design of "Online + Offline" Hybrid Teaching Model

Based on the project response theory, social comparison theory and metacognition theory, and referring to the hypermedia general model of deep reinforcement learning education, this study designs a "Online + offline" hybrid teaching model based on deep reinforcement learning for learning experience. According to certain standards, the domain knowledge unit and learner unit are designed and developed. The basic laws of learning and teaching are as follows: first, learning is a process in which learners actively participate; Second, learning is a gradual process of experience accumulation. Third, different types of learning have different processes and conditions. Fourth, for learning, teaching is the external condition of learning. Effective teaching must be an activity that gives timely and accurate external support to learners according to the law of learning. Starting from the emphasis on "learners' individual characteristics and learning needs", dynamically track learners' knowledge status, knowledge level and learning behavior, and dynamically update learners' units by using coverage modeling technology and data-driven technology under the coordination of deeply strengthened learning mechanism according to the domain knowledge unit and learner unit driven teaching model, Present open learner unit, open social learner unit and good adaptive learning content for learners, and finally realize personalized service and trigger learners' learning experience in metacognition and social comparison.

3.2 Domain Knowledge Unit

Domain knowledge unit is the foundation of deep reinforcement learning hybrid teaching model. It points out the application field and learning content of deep reinforcement

learning hybrid teaching model, and provides domain structure and information that needs to be adapted. Domain knowledge unit describes the knowledge unit, knowledge point, learning object, association relationship between knowledge units, association relationship between knowledge units, association relationship between knowledge points, association relationship between knowledge points and association relationship between knowledge points and learning objects involved in the application field. Each knowledge point corresponds to multiple learning objects, and each learning object has text, video Test questions and other forms. By analyzing the learning object metadata celts-3.1 defined by the Educational Information Technology Standards Committee of the Ministry of education of China, this study outlines the structure diagram of domain knowledge units, determines the attributes of domain knowledge elements, gives the reference specifications of domain knowledge units, and constructs domain knowledge units.

3.2.1   Construction of Domain Knowledge Unit Structure

Domain knowledge unit is a collection of domain knowledge elements and their relationships. Domain knowledge elements have different names and relationships in different models. Therefore, in the domain knowledge unit, it is necessary to determine the composition and correlation of domain knowledge elements. Learning objects are the support of learning tasks and learning activities in the learning process. Learning objects present learners with learning content suitable for their personality characteristics with their rich and diverse types. The types of learning objects include text, video, audio, pictures, tests, test questions, examples, animation and demonstration, courseware, teaching cases, FAQs, etc. The learning objects designed in this research include text, video and test questions. Text, video and test questions are related to domain knowledge units through knowledge points. It can be seen that the domain knowledge elements contained in the domain knowledge unit include disciplines, primary knowledge points, secondary knowledge points, courses, chapters, sections, knowledge points, learning objects, texts, videos and test questions. In a broad sense, this study refers to disciplines, primary knowledge points, secondary knowledge points, courses, chapters and sections as knowledge units. The relationships between domain knowledge elements in a domain knowledge unit are defined as the following.

(1) *Superior relationship and inferior relationship*
     The domain knowledge element a at the upper level is more integrated, contains more knowledge content and expresses more abstract content; The lower domain knowledge element B is more localized and the content expressed is more specific. Generally, it only reflects one aspect of the upper knowledge point, that is, a contains B and B is A part of A. For example, disciplines, primary knowledge points, secondary knowledge points and knowledge points form a superior subordinate relationship in turn; Courses, chapters, sections and knowledge points form a superior subordinate relationship in turn; The learning object is the upper knowledge point of text, video and test questions, and the text, video and test questions are the lower knowledge points of the learning object.
(2) *Precursor relationship and follow-up relationship*

The antecedent relationship and follow-up relationship indicate that there is a logical sequence of domain knowledge elements. If you must master the premise domain knowledge element B before learning A domain knowledge element A, that is, B precedes A, then the antecedent of a is B and the follow-up of B is A. For example, the relationships between primary knowledge points and primary knowledge points, between secondary knowledge points and secondary knowledge points, and between knowledge points form precursor and follow-up relationships; Precursor and follow-up relationships are formed between courses, between chapters, and between sections. A more specific example is that before learning the knowledge point "rational division", we must master the knowledge point "reciprocal", and we must first learn the knowledge point "rational addition" before learning the knowledge point "rational subtraction". The core element of domain knowledge unit is knowledge points. The domain knowledge elements contained in disciplines, courses and learning objects are centered on knowledge points and represent well structured domain knowledge through top-down hierarchical relations. Such a domain knowledge structure is conducive to the reuse of knowledge and learning resources.

3.2.2   Attribute Division of Domain Knowledge Elements

After determining the domain knowledge elements and their relationship, it is necessary to refine the attributes of domain knowledge elements according to certain metadata standards.

The first is learning object metadata. The Educational Information Technology Standards Committee of the Ministry of education of China defines learning object metadata celts-3.1, which is conducive to educators, learners or automated software to find, acquire, use and evaluate learning objects, and in the process of knowledge representation in the field of design and development, It will enable developers to fully consider the cultural and linguistic diversity of learning objects and their metadata in the use context, so as to promote the exchange and sharing of learning objects. The data elements of learning objects can be described in nine different categories:

(1) General class refers to the general information that describes the learning object as a whole;
(2) Lifetime class refers to the attribute information, learning history and personal and organizational information related to the learning object in the current state;
(3) Meta metadata class refers to some information of metadata instance itself;
(4) Technology refers to the information on the technical requirements and technical characteristics adopted by the learning object;
(5) Education refers to the information about the learning object and the characteristics of education and teaching;
(6) Rights refer to the information of the learning object in terms of intellectual property rights and use conditions;
(7) Relationship class refers to the relationship information between learning objects and other related learning objects;
(8) Commentary refers to the information that evaluates the learning object in terms of teaching use, including the author and creation time of the learning object;

(9) Classification class refers to the relationship information between learning objects and one or some specific classifications.

3.3 Learner Unit

Learner unit is an important basis for deep reinforcement learning model to realize deep reinforcement learning. It records learners' personal information, knowledge state, knowledge level, learning behavior, knowledge point planning, learning object review and other attribute information. Here, except that personal information is static information, other information is dynamic information. Based on the domain knowledge unit and the information that learners interact with the model, this study uses coverage modeling technology and data-driven technology to track the dynamic information in real time to ensure that the dynamic information is always in the latest state. Dynamic information provides a basis for adaptive presentation of learning objects and learning contents.

3.3.1 Learner Unit Construction

In the mixed teaching model designed in this paper, learner unit construction mainly includes the following contents:

To determine the source of deep reinforcement learning, that is, to model the learner unit according to the learner's personality characteristics, which are called the source of deep reinforcement learning. According to the relationship between personality characteristics and domain, personality characteristics are divided into domain related and domain independent personality characteristics, as shown in Fig. 1.
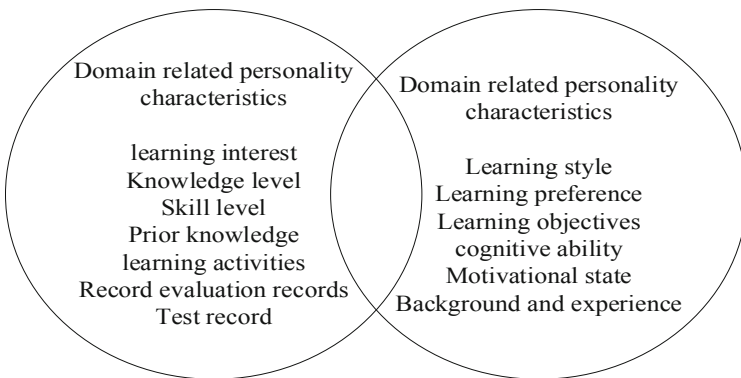
Fig. 1. Domain related and domain independent personality characteristics

It can be seen from Fig. 1 that the domain-related personality characteristics refer to the information recorded by the model is directly related to the learned domain knowledge, and is a reflection of the learner's knowledge level and skills in the learning situation, such as learning interest, knowledge level, skill level, prior knowledge. Knowledge, learning activity records, assessment records, and test records; domain-independent personality characteristics refer to those information that are not directly related to the

knowledge in the field of study, but have indirect guiding significance for the learner's learning process, such as learning styles, learning preferences, and goals., cognitive abilities, motivational states, background and experience. According to whether the personalized features change with time, the personalized features are divided into static and dynamic personalized features, as shown in Fig. 2.
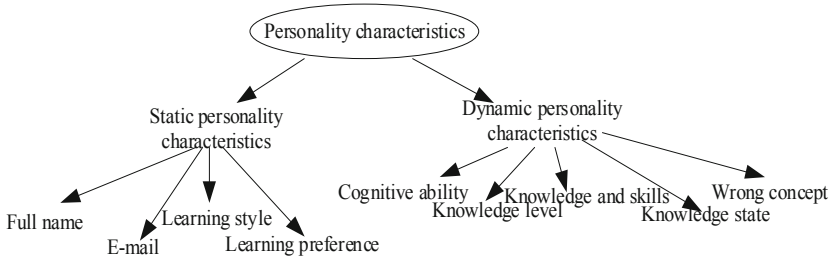
Fig. 2. Static and dynamic personality

According to Fig. 2, static personalization features are all set before the learning process occurs and usually remain unchanged throughout the learning process, such as name, email, age, native language, learning style, learning preference, etc., which can be determined by the learner Available directly through the options menu, and also through the use of a questionnaire. The dynamic personalization feature is the information collected in the learning process, such as the interaction between the learner and the model, the performance behavior and the learning history, which need to be determined by certain rules and algorithms, and constantly updated in the model, such as learning interest, cognitive ability, knowledge level, knowledge skills, knowledge mastery state, wrong and lost concepts, learning behaviors, emotional factors, metacognition and other personalized characteristics. The dynamic personalized characteristics of learners constitute the basis of each learner's needs in the model of in-depth reinforcement learning, which can be measured by questionnaire and the tests that learners must complete in the learning process. The deep reinforcement learning sources of this study include personal information, subject knowledge status, curriculum knowledge status, knowledge level, learning behavior, knowledge point planning, text and wrong question review.

3.3.2 Learners' Choice of Unit Modeling Technology

Learner unit modeling technologies include coverage modeling, lead plate modeling, perturbation modeling, machine learning technology, constraint based model, fuzzy learner modeling, Bayesian network, project response model, ontology based learner modeling and so on. Different modeling techniques are usually combined to infer and modify the dynamic personality characteristics in the learner unit, so as to continuously improve the learner unit and ensure the accuracy and diversity of information in the learner unit. This study uses coverage modeling technology, item response model and weight algorithm to construct learner units. Through the visual processing of the designed learner unit, the information of the learner unit in the model is transmitted to the learners in the

form of visualization. The designed visual view of learner unit helps learners understand the gap between what they have learned and what experts expect by presenting learners' knowledge status and learning progress, so as to promote learners to learn more knowledge.

3.3.3 Learner Unit Data Management

The purpose of data management of learner unit is to realize the sharing and migration of learner information between different models, improve the portability of learner information data, and ensure the privacy, security and integrity of learner data.

The first is the management of core data. The packaging of learner information core data in the model is to ensure that learner information is transmitted and exchanged among different learner information models, such as learning management model, human resource model, student information model, enterprise e-learning model, knowledge management model, resume database and so on. The main purposes of describing learners' personality characteristics based on the data model include: recording and managing learning history, learning objectives and learning achievements related to learning; Involve learners in the learning process; Discover learning opportunities for learners. In this paper, the core data of the model is divided into 11 categories: identity recognition; Accessibility; Qualification, certification and licensing; Objectives; Activities; Ability; interest; Affiliation; Performance report; Security key; Association relationship. The specific description is shown in Table 1.

Table 1. Model learner information core data management mechanism

Core data name	Core data description
Identification	Learner's personal information related to learning, such as ID, grade, E-mail, QQ number, wechat account, etc.
Target	Learning goals or strong desire to learn
Activity	In formal learning, informal learning, training, work experience, lifelong education, and all learning related activities in military or civil service, self-report, etc.
Qualification, certification and licensing	Certification certificates, licenses, qualification certificates, etc. issued by authoritative certification bodies
Ability	Describe the knowledge, skills and abilities acquired by learners in emotional, cognitive and psychological fields
Report card	The summary report of learners' academic achievements can be in a variety of forms

Table 1. (*continued*)

Core data name	Core data description
Interest	Describe learners' hobbies, recreational activities and other interest information
Subordination	Members of professional organizations
Accessibility	Accessibility to learner information such as language competence, qualifications and learning preferences, learning preferences include cognitive preferences (such as learning style preferences), physical preferences (such as liking large fonts), and technical preferences (such as liking specific computer platforms)
Security key	The password and security key set when the learner interacts with the learner information model and service
Association relationship	Represents the association between the core components of the model

The second is the management of public and private information, which aims to standardize the semantic grammar of learners' units and describe learners' information and knowledge. In this paper, public and private information are divided into learner personal information, learner performance, learner relationship, learner security and learner document; Learners prefer six types of learner information to realize the construction of "Online + offline" mixed teaching model.

4   Analysis of Teaching Effect

Taking college mathematics teaching as an example, the model is solved by SPSS software to obtain the multiple linear regression model of teaching effect. The data obtained from the questionnaire and sorted out are imported into SPSS software, and the multiple linear regression method is used to obtain the relevant results, that is, the college mathematics teaching effect under the online and offline mixed teaching model. Analyze the coefficient of each influencing factor in the model, fully discuss with teachers, and then readjust the proportion of each influencing factor online and offline, so as to analyze the teaching effect of the model in college mathematics courses.

4.1   Data Sources

This paper collects and arranges the relevant data of college mathematics courses taught by online and offline mixed model in three colleges and universities, and designs a questionnaire for the relevant influencing factors to obtain the relevant data. Through the preliminary investigation, the factors affecting the teaching effect are analyzed and summarized into the following 10 influencing factors: online teaching video (X1), online

teaching PPT (X2), online homework (X3), online test (X4), online Q & A (X5), offline classroom teaching (X6), offline homework (X7), offline test (X8), offline Q & A (X9), and participating in college students' mathematical modeling activities (X10), the dependent variable Y represents the total score. The questionnaire is prepared by using the questionnaire star app, as shown in Table 2.

Based on the questionnaire in Table 2, a survey was conducted in three universities that adopted the online and offline mixed teaching model for mathematics courses. 328 questionnaires were distributed and 328 questionnaires were recovered, of which 328 were valid. The obtained data were processed. Single choice questions corresponded options 1, 2, 3 and 4 to 100, 80, 60 and 40 respectively, and the data were statistically

Table 2. Questionnaire

Influence factor	Problem	Option 1	Option 2	Option 3	Option 4
X1	Whether to watch the online teaching video carefully and completely	Very serious	Quite serious	Commonly	Unclear
X2	Whether to watch online teaching ppt carefully and completely	Very serious	Quite serious	Commonly	Unclear
X3	Do you complete online work carefully and independently	Very serious	Quite serious	Commonly	Unclear
X4	Do you complete the online test carefully and objectively	Very serious	Quite serious	Commonly	Unclear
X5	Do you seriously participate in online Q & A	Very serious	Quite serious	Commonly	Unclear
X6	Whether offline classroom teaching is carried out seriously	Very serious	Quite serious	Commonly	Unclear

(*continued*)

Table 2. (*continued*)

Influence factor	Problem	Option 1	Option 2	Option 3	Option 4
X7	Whether the offline operation is completed carefully and independently	Very serious	Quite serious	Commonly	Unclear
X8	Have you completed the offline test carefully and objectively	Very serious	Quite serious	Commonly	Unclear
X9	Do you seriously participate in offline Q & A	Very serious	Quite serious	Commonly	Unclear
X10	Whether to participate in college students' mathematical modeling activities	Participate many times	Attend once	Understand	Unclear

analyzed, That is, each line represents a questionnaire, a total of 328 valid data, as shown in Table 3.

Table 3. Questionnaire data

Total score		Influence factor				
Serial number	Y	X1	X2	X3	…	X10
1	82	80	60	70	…	60
2	85	100	60	70	…	60
3	93	90	80	90	…	80
4	79	80	90	60	…	80
…	…	…	…	…	…	…
328	69	90	100	80	…	70

4.2 Analysis of Teaching Effect

It can be seen from Table 2 that 10 influencing factors are counted in the questionnaire, the data in Table 3 are standardized, and the following multiple linear regression model

is established:

$$Y = a_0 + a_1 X 1 + a_2 X 2 + a_3 X 3 + ... a_{10} X 10 + \varepsilon \tag{5}$$

In the formula, $a_i (i = 1, 2, ..., 10)$ and ε represent adaptability error. The stepwise regression method with SPSS software can be used to select independent variables. These six variables can be removed from the x 2 online teaching ppt, x 5 online Q & A, x 7 offline homework, x 8 offline quiz, x 9 offline Q & A and x 10 participating in college students' mathematical modeling activities, so as to eliminate the remaining four variables X 1 online teaching video, x 3 online homework, x 4 online quiz X 6 offline classroom teaching is analyzed to obtain the multiple linear regression model of the effect of online and offline mixed teaching model:

$$Y = a_0 + a_1 X 1 + a_2 X 2 + a_3 X 3 + ... a_6 X 6 + \varepsilon \tag{6}$$

The regression model was tested by F-test, and the $P < 0.05$ was obtained, so it passed the significance test. The parameters of this model are estimated by the least square method to obtain the following regression equation:

$$\hat{Y} = 35.25 + 0.225 X 1 + 0.213 X 3 + 0.075 X 4 + 0.575 X 6 \tag{7}$$

To sum up, the mixed teaching model designed in this paper has an obvious positive effect on the four factors affecting the teaching effect: online teaching video, online homework, online test and offline classroom teaching, and the offline classroom teaching has the greatest impact on the teaching effect of these four factors, followed by online teaching video, with a test $P < 0.05$, The analysis results clearly show that the model can give full play to the irreplaceable offline classroom teaching, and maximize the teaching effect in the new era network environment with the help of online teaching video.

5 Conclusion

This study proposes a hybrid teaching model of deep reinforcement learning for learning experience. In the process of demonstration and evaluation, the metacognition and social comparative learning experience triggered by the teaching model have a good positive response, and give good feedback to the teaching institutions provided. With the deepening of research and continuous reflection in the research process, it is found that there are still two aspects of research to be carried out:

(1) Optimization of domain knowledge unit
    In this study, domain knowledge unit is not the focus of the research, but it plays an important role in the research. The domain knowledge unit of this study only provides learning objects such as videos, texts and test questions of each knowledge point in the rational number chapter. In order to conduct more in-depth and lasting research, we need to further improve the learning resources of the remaining chapters, which is more conducive to demonstration and evaluation.

(2) Expansion of deep reinforcement learning source

The learner unit of this study focuses on knowledge status, knowledge level and learning behavior. Although it evaluates the metacognition and social comparison learning experience caused by two open learner units and teaching strategies, it does not take metacognition and social comparison as the two dimensions of the learner unit. On the basis of this study, It is necessary to expand the source of deep reinforcement learning to psychological personality characteristics such as metacognition and social comparison, so as to provide more comprehensive learning services for the deep reinforcement learning model.

References

1. Zalat, M.M., Hamed, M.S., Bolbol, S.A.: The experiences, challenges, and acceptance of e-learning as a tool for teaching during the COVID-19 pandemic among university medical staff. PLoS ONE **16**(3), e0248758 (2021)
2. Dias, S.B., Hadjileontiadou, S.J., Diniz, J., et al.: DeepLMS: a deep learning predictive model for supporting online learning in the Covid-19 era. Sci. Rep. **10**(1), 19888 (2020)
3. Schlesselman, L.S.: Perspective from a teaching and learning center during emergency remote teaching. Am. J. Pharm. Educ. **84**(8), 8142 (2020)
4. Snook, A.G., Schram, A.B., Jones, B.D.: Faculty's attitudes and perceptions related to applying motivational principles to their teaching: a mixed methods study. BMC Med. Educ. **21**(1), 188 (2021)
5. Doukanari, E., Ktoridou, D., Efthymiou, L., et al.: The quest for sustainable teaching praxis: opportunities and challenges of multidisciplinary and multicultural teamwork. Sustainability **13**(13), 1–21 (2021)
6. Giddens, J., Curry-Lourenco, K., Miles, E., et al.: Enhancing learning in an online doctoral course through a virtual community platform. J. Prof. Nurs. Off. J. Am. Assoc. Coll. Nurs. **37**(1), 184–189 (2020)
7. Jia, Y.: Research on the practice of college English classroom teaching based on Internet and artificial intelligence. J. Intell. Fuzzy Syst. **2021**(1), 1–10 (2021)
8. Rodriguez-Rodriguez, E., Sanchez-Paniagua, M., Sanz-Landaluze, J., et al.: analytical chemistry teaching adaptation in the COVID-19 period: experiences and students' opinion. J. Chem. Educ. **97**(9), 2556–2564 (2020)
9. Baghal, T.A.: The effect of online and mixed-mode measurement of cognitive ability. Soc. Sci. Comput. Rev. **37**(1), 89–103 (2019)
10. Pp, A., Kr, B.: Implicit adversarial data augmentation and robustness with Noise-based learning. Neural Netw. **141**(9), 120–132 (2021)
11. Xu, N., Fan, W.H.: Research on interactive augmented reality teaching system for numerical optimization teaching. Comput. Simul. **37**(11), 203–206+298 (2020)

Design of Corpus Based Comprehensive Ability Evaluation System for College English Teaching

Ying Yu[1(✉)] and Shengzuo Lin[2]

[1] Panjin Vocational and Technical College, Panjin 124010, China
yuying12012@126.com

[2] Guangdong Polytechnic of Environmental Protection Engineering, Foshan 528000, China

Abstract. At present, the evaluation accuracy of the comprehensive ability evaluation system of College English teaching is low, and the stability of the evaluation process is poor, resulting in greater risks in the evaluation process. In order to solve the above problems, a new comprehensive ability evaluation system for College English teaching is designed based on corpus, and the hardware and software of the system are optimized. The system hardware mainly designs microprocessor, signal modulator, power module, evaluator, single chip microcomputer and collector. The processor selects sep83 microprocessor and introduces sd63c84 chip. At the same time, en, A1 and A0 are designed as the input ports of the program-controlled amplification circuit of the regulator. The main control chip of the single chip microcomputer is ti7392 and the acquisition is realized through snt5428. By introducing corpus, the software workflow is realized through information collection, quantitative evaluation, designing practical and effective evaluation contents of comprehensive ability of College English teaching, reducing evaluation errors. The experimental results show that the corpus based College English teaching comprehensive ability evaluation system can effectively improve the evaluation stability, reduce the evaluation risk and enhance the accuracy of the evaluation process.

Keywords: Corpus · College English · English teaching · Comprehensive ability · Evaluation system

1 Introduction

With China's development entering a new era, the degree of economic and cultural globalization is deepening. With the support of the national foreign policy, the number and scale of multinational groups and foreign enterprises continue to increase. Therefore, the demand for foreign language talents with strong ability is increasing in China. In recent years, the research on College English has also made many achievements. The state has also issued relevant laws and regulations to support the development of College English in China. At present, China is still in the primary stage of development, has not formed a complete development system, and lacks special laws and regulations

protection. There are many problems such as loose system and lack of educational resources [1, 2].

With the deepening of globalization and more and more frequent exchanges in various fields of society, College English education is very important. For these problems, we should first consider the quality and ability level of English teachers. Teachers are an important driving force to promote the stable development and progress of the industry. How to improve teachers' knowledge, skills and professional quality has attracted the attention of the education sector [3–5].

In order to test and improve English teachers' professional competence, relevant scholars put forward the framework of teachers' technical teaching content knowledge and ability in terms of teachers' teaching skills and knowledge content as early as ten years ago. This paper designs a new college English teaching comprehensive ability evaluation system based on corpus, designs the hardware and software of the system, and verifies the effectiveness of the evaluation system through experiments.

2   Design of College English Teaching Comprehensive Ability Evaluation System

2.1   Microprocessor Design

The processor is the driving device of the system hardware system, which mainly completes the interaction between the system terminal and the control terminal. The microprocessor has stronger performance and higher processing accuracy than the processor. In this paper, SEP83 microprocessor is selected to complete the design of the system. SEP83 microprocessor adopts 1.sv core power supply mode, which can reduce the power consumption of the chip, reduce the load inside the system and improve the running rate of the microprocessor [6]. The chip of this microprocessor is equipped with a special chip management PMC unit. The clock state of the microprocessor chip can be controlled in real time through the unit module, so as to achieve the purpose of low power consumption of the microprocessor [7, 8].

The processor supports 80 gpios, and in order to protect the internal data security of the system, the external interrupt function is designed. The device supports the transmission of DMA linked list. 32-bit RISC core is embedded in the microprocessor, which is compatible with 720t arm. 8 kB instruction data set is used to complete the transmission of control instructions. In order to improve the link efficiency between the processor and other devices, this paper adopts the 16 bit false interface. When the device is started, this interface is also opened at the same time to avoid missing data frames. The internal data transmission protocol of sep83 microprocessor adopts 10m adaptive Ethernet. The DC / DC circuit diagram of sep83 microprocessor is shown in Fig. 1:

The microprocessor chip adopts SD63C84 chip newly launched by SD Company. The bus interface unit includes 8-byte input and output interface and 8-bit memory. The control unit includes 32-bit UART interface and 8-bit general register. In addition, the control unit also has a microcontroller, which can control the microprocessor's processing of performance evaluation data, Ensure the integrity and accuracy of the evaluation data [9, 10].
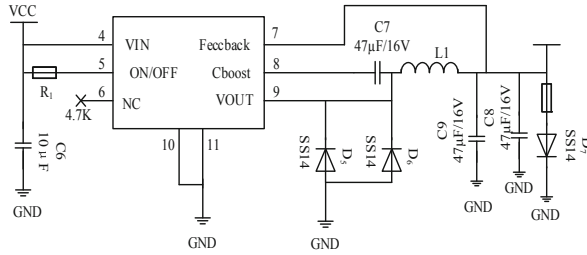
Fig. 1. DC/DC circuit diagram of microprocessor
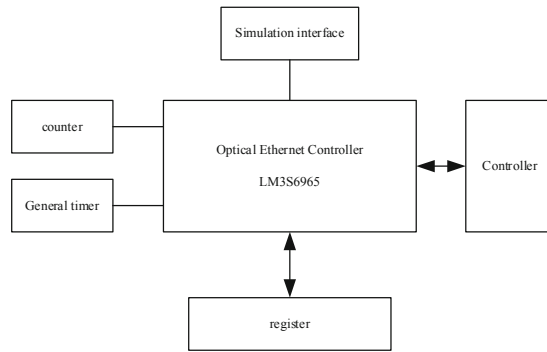
Fig. 2. Microprocessor structure

The microprocessor structure is shown in Fig. 2:

It can be seen from Fig. 2 that the peripheral circuit and power circuit can be used together to facilitate the processing of the system. The microprocessor chip has 10 basic registers to maintain the processing state of the microprocessor. The other six multi-purpose registers can store basic variables. The control unit is not directly connected with the processor bus. It can send read and write requests to the system with an address of 8 bits, Four commands can be controlled and transmitted at the same time. After the control unit controls all processor commands, it can send the interrupt vector to the CPU of the processor [11].

2.2 Design of Signal Modulator

The function of the signal modulator in the system hardware area is to ensure the operation stability of the information security risk assessment system. According to the functional considerations of the system designed in this paper, the signal modulator is designed with 8 regulation modules, which can carry out the interactive processing of 32 channels at the same time. Compared with the traditional controller, the signal modulator has the advantages of not only control, but also communication filtering and amplification process. The program-controlled amplification circuit is the control circuit of the signal regulator. The circuit design en, A1 and A0 are used as the input ports of the program-controlled amplification circuit of the regulator, and the amplification factor is

4 times [12]. The resistance of the signal regulator is divided into four levels, from high to low, which are 8.6 Ω, 91 Ω, 349 Ω and 732 Ω respectively. The four different levels of resistance can minimize the error of the signal regulator. The effective range bit of the signal regulator for the information signal is 50–200 kHz. The structural diagram of the specific signal modulator module is shown in Fig. 3:
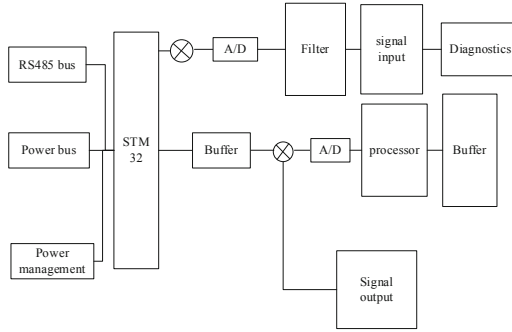
Fig. 3. Structure diagram of modulation module of signal modulator

Taking the broadband I/Q vector modulation signal as an example, the signal modulation is realized by frequency conversion in the transmission link. When the signal system is up and down converted, it is found that the modulation quality and EVM become worse, which are normal phenomena. In the modulation process, the controller error is controlled by different levels of resistance to make it successfully complete the signal modulation.

2.3 Power Module Design

In order to reduce the physical loss of the power block to other devices in the system hardware area, the power module of ub78vb is selected as the power supply device. The limit of the power module is 5 V DC power supply. For the work of small processes of the system, it can also provide 3.3 V or 1.8 V DC power supply. The power module of UB78VB is characterized by adding a 3.8 V filter capacitor on the basis of the traditional power structure, providing the input and output working core for the system and improving the anti-interference of the power module. In order to ensure the safety of the power module, a fuse is added in the device structure. When the power supply continues to supply power or the power circuit current is too large, the device will automatically power off to protect the power circuit board. The control circuit of the power module is shown in Fig. 4:

2.4 Evaluator Design

In order to reduce the complexity of the evaluation process, the device prohibits the low-level state from completing the change of the protection signal. The address bus width of the evaluation board is 19 bits, and the capacity for NOR flash is 2 mbyte.
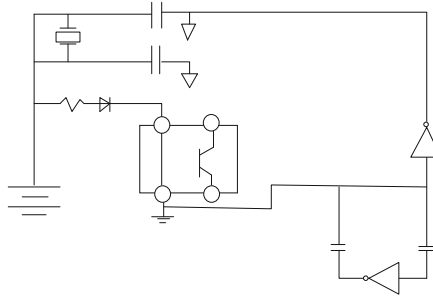
Fig. 4. Power structure circuit diagram of motherboard of power module
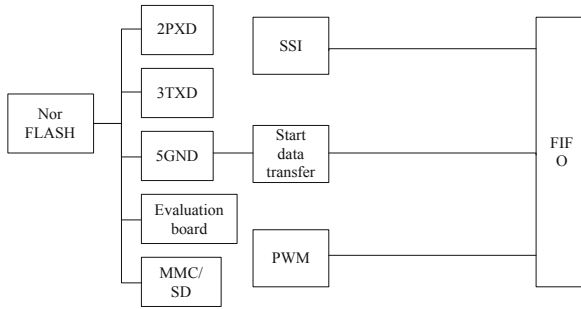
Fig. 5. Structure of evaluator

The structure of the evaluator is shown in Fig. 5:

According to Fig. 5, the evaluation board adopts 2pxd, 3txd and 5GND pins to complete the operations of receiving data, sending data and encrypting data respectively. The evaluation rate of the evaluation board is 11 Mbps. It supports MMC/SD dual card mode. The evaluation board is set with two channels of SSI and PWM, and supports ISO and microwire operating system protocols without random code. The reset voltage of the evaluation board is 3.3 V, the response reset voltage is 2.9 V–3.00 V, and the reset pulse time is 140 s. The operation of the evaluation board adopts 4 m passive crystal oscillator circuit for startup and shutdown control.

SXC765b4 produced by Samsung is selected as the evaluator chip, which is the core of the evaluator, and a microprocessor with efficient processing capacity is added to the periphery to assist the evaluator in processing the performance evaluation result data [13].

Abundant peripherals are set on the periphery of the evaluator. The frequency of one crystal oscillator is 38.728 kHz, and the two crystal oscillators are frequency divided in the chip of the evaluator, so as to improve the working frequency of the evaluator. The circuit diagram of the evaluator is shown in Fig. 6:
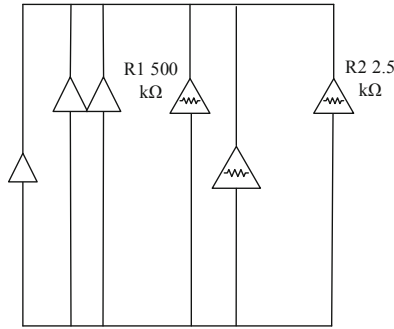
Fig. 6. Circuit diagram of evaluator

Stu5278, one of the peripherals of the evaluator, is a microprocessor with high performance and low power consumption. The working frequency of this microprocessor can reach 380 MHz at most. It adopts a full duplex core. It has simple structure, convenient operation and low power consumption. The periphery of the evaluator is equipped with peripheral circuits. The voltage of the peripheral circuits is controlled at 1.8–3.3 V and the current is 1.2–1.8 A, It is mainly used to assist the power circuit to supply power to the evaluator and other peripherals. When the voltage of the power circuit is unstable or cannot supply power normally, the peripheral circuit can directly supply power to the evaluator and its peripherals, which can effectively improve the voltage utilization and reduce the loss of the power circuit.

2.5 Single Chip Microcomputer Design

The main control chip of the evaluation system designed in this paper is ti7392, which contains 4 kbytes program memory and 64 Kbytes data memory. The structure of single chip microcomputer is shown in Fig. 7:
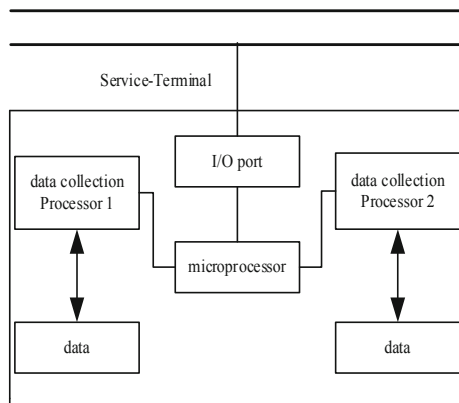
Fig. 7. Structure diagram of single chip microcomputer

It can be seen from Fig. 7 that the single chip microcomputer chip has 15 pins and 4 external and medium ports, which can count the performance data and talent ability data of the human resources management department. In addition, it also has 6 two-way parallel communication ports.

The clock frequency of single chip microcomputer chip is very low, with sleep mode and super power saving mode. I/O port and data memory can be used to wake up the single chip microcomputer. In the power saving mode of single chip microcomputer, the data memory is in shutdown state and stops storage. The working voltage of single chip microcomputer is 1.8–10 V, which can be operated statically, and the working frequency is 0–48 MHz, The size of data storage space is 256×8 bytes, with 12 interrupt sources, Serial uart channel and USB interface. The single chip microcomputer can receive and process the evaluation signal transmitted by the performance evaluation system.

2.6 Collector Design

The collector chip of the evaluation system is snt5428 produced by Samsung, which has strong acquisition capability. The peripheral circuit of the collector does not need to be expanded, which can improve the integration and reliability of the collector and make the data acquisition operation of the collector relatively simple.

The collector has a power circuit to prevent the USB interface from being powered due to accidents. It can collect resource data and evaluation data in the human resources performance evaluation system. The chip pin needs to be connected with an 8 Ω resistance. The collector is also equipped with a reset button to download and debug the evaluation system. The USB connecting line between the timer and the collector is a double-layer protection line, It can effectively protect the collector and other accessories from damage and interference.

The collector circuit diagram is shown in Fig. 8:
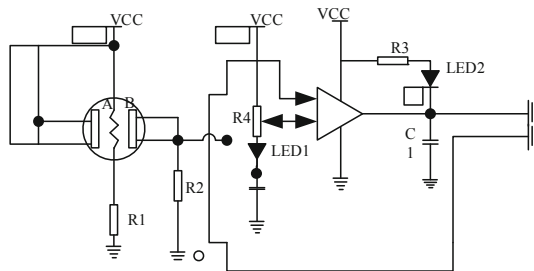
Fig. 8. Collector circuit diagram

3 Software Design of College English Teaching Comprehensive Ability Evaluation System Based on Corpus

College English teaching comprehensive ability evaluation system mainly calls a quantitative evaluation model jointly constructed by corpus to analyze abstract meaning

and judge College English teaching comprehensive ability. The design concept of the evaluation model is shown in Fig. 9:
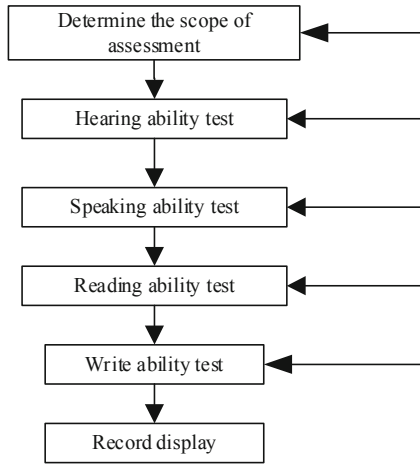
Fig. 9. Structure diagram of evaluation model

The set of evaluation factors is established by using the expected set. The variables stored in the set are the decisive variables of the model to analyze the elements to be evaluated. Establish an evaluation weight set, and the weight calculation formula is shown in formula (1):

$$b_i = \frac{1}{k} \sum_{j=1}^{n} b_{ij} + \partial a_{ij} \tag{1}$$

where, b_i represents the weight calculation result; b_{ij} represents the set of evaluation factors; k represents the serial number of the evaluation expert; a_{ij} represents the weight set; ∂ is the normalization coefficient.

According to the division specification of evaluation results, the mapping relationship and evaluation membership matrix are constructed, and the evaluation set of evaluation model is established. In order to shorten the error between model evaluation levels, the membership matrix is evaluated. The evaluation fuzzy membership matrix of data is shown in formula (2):

$$s = \begin{Bmatrix} S_{11} & S_{12} & \dots & S_{M1} \\ S_{21} & S_{22} & \dots & S_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ S_{m1} & S_{m2} & \dots & S_{mm} \end{Bmatrix} * \frac{S_r}{B} \tag{2}$$

where, S_{ij} represents the membership vector; S_r represents risk factors; B represents the weight allocation set.

After the successful construction of the evaluation model, the corpus is used to pre-process the data to be evaluated, input the data into the evaluation model for calculation, and the model outputs the evaluation vector results for many times. After weighted average calculation, the final comprehensive ability safety evaluation result of College English teaching is obtained, which can be the evaluation of comprehensive ability.

The software flow of corpus based College English teaching comprehensive ability evaluation system is shown in Fig. 10:
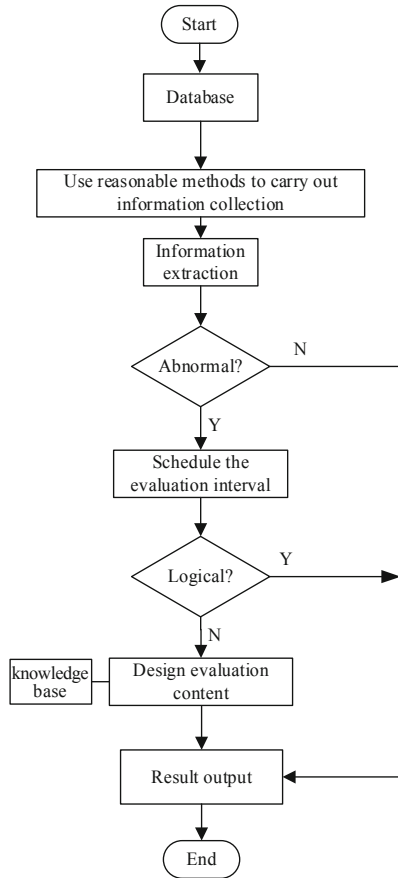
Fig. 10. Sysem software implementation process

According to Fig. 10, the corpus based College English teaching comprehensive ability evaluation system software can be divided into several steps.

Step 1: carry out information collection in a reasonable way. The designed collector is used to collect a large amount of information. After the last performance evaluation, the information will be collected by the system. The collected results will be modulated by the signal modulator and stored in the mobile intelligent terminal. After the first collection, analyze the College English teaching performance, determine the employee

performance through the evaluation of others, and evaluate the employee performance after the evaluation is synthesized. Before the evaluation meeting, Screening the information related to the performance of colleges and universities to ensure the fairness and objectivity of information.

Step 2: scientifically arrange the time interval of comprehensive ability evaluation of College English teaching. The time interval of College English teaching comprehensive ability evaluation will affect the results of College English teaching comprehensive ability evaluation. When arranging the time interval of College English teaching comprehensive ability evaluation, we should ensure the rationality of the arrangement.

Step 3: design practical and effective evaluation contents of comprehensive ability of College English teaching. The content of College English teaching comprehensive ability evaluation should be designed on the basis of fairness, impartiality and science. The evaluation content should be timely. The evaluation content should be adjusted according to the nature of teachers' work. The last evaluation results can not be used. The evaluation results should be consistent with teachers' comprehensive performance in a period of time.

Step 4: reduce the evaluation error of College English teaching comprehensive ability. In the evaluation of College English teaching comprehensive ability, although the errors caused by external factors can not be avoided, certain methods can be adopted to avoid the errors as far as possible. The methods to reduce the errors include: unified training for evaluators to ensure the fairness and impartiality of the evaluation work, and the evaluation standards of evaluators should be unified and coordinated, The assessment focuses on work ability and practical operation.

Step 5: feedback the evaluation results and strengthen communication. After the comprehensive ability evaluation of College English teaching is completed, the evaluation results shall be fed back to the teachers to understand their own evaluation results. If there are problems, they shall communicate with the evaluators in time to solve the problems in the evaluation process and discuss the solutions together.

4   Experimental Study

In order to verify the effectiveness and stability of the corpus based College English teaching comprehensive ability evaluation system designed in this paper, the traditional evaluation system is compared with this evaluation system.

The experiment refers to the normal evaluation criteria, analyzes the distribution rate, establishes the judgment matrix, establishes the evaluation result matrix according to the evaluation index weight, normalizes the evaluation result matrix according to the fuzzy evaluation method, and integrates the above process according to the 9 importance levels and their assignment given by Saaty. The score of the evaluation result is 80 points, while the score of the evaluation result obtained by the traditional evaluation system is 75 points, indicating that the evaluation effect of the evaluation system designed in this paper is better. The experimental results of evaluating process accuracy are shown in Fig. 11:

It can be seen from Fig. 11 that with the increase of time, the traditional evaluation system and the evaluation system studied in this paper extend towards two different
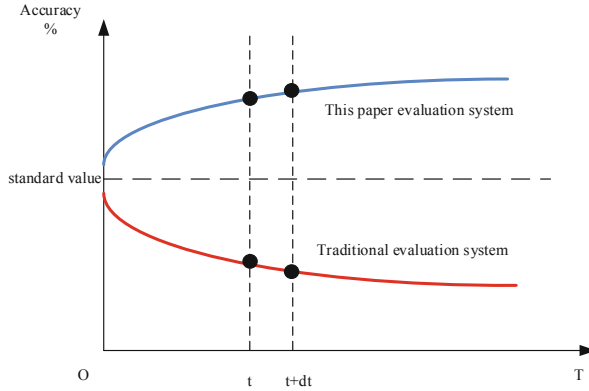
Fig. 11. Experimental results of evaluating process accuracy

trends. The evaluation results of the evaluation system provided in this paper become more and more accurate, while the evaluation results of the traditional evaluation system become more and more inaccurate. This is mainly because the traditional evaluation system integrates too many subjective factors, and with the increase of time, Subjective factors increasingly affect the accuracy, and the performance evaluation system studied in this paper has strong objectivity. Therefore, with the increase of time, the evaluation results become more and more accurate. It can be seen that the traditional human resources performance evaluation system has few evaluation schemes and inaccurate evaluation results, unscientific evaluation methods and low reliability and efficiency of evaluation results. The experimental results of evaluating system stability are shown in Table 1:

Table 1. Experimental results of evaluating system stability

Evaluation times/time	Evaluate system stability %	
	Traditional evaluation system	Paper evaluation system
1	70.25	97.25
2	75.22	98.25
3	74.28	99.44
4	76.34	97.83
5	74.22	98.96
6	73.86	99.84
7	74.85	98.25
8	75.96	97.86
9	78.51	98.69
10	76.15	97.39

According to Table 1, the evaluation system proposed in this paper has better stability in the evaluation process. The evaluation system designed in this paper combines the corpus with the human resources evaluation system, optimizes the means of evaluation, makes the human resources performance evaluation more intelligent and scientific, and makes the evaluation effect more obvious, so that the stability, reliability and feasibility of the comprehensive ability evaluation system of College English teaching are higher, which can be proved by the experimental data, The College English teaching comprehensive ability evaluation system designed in this paper is better than the traditional evaluation system, has higher stability and feasibility, the evaluation results are more accurate, fair and fair, and the evaluation means are more scientific.

The time spent in the evaluation process is shown in Fig. 12:
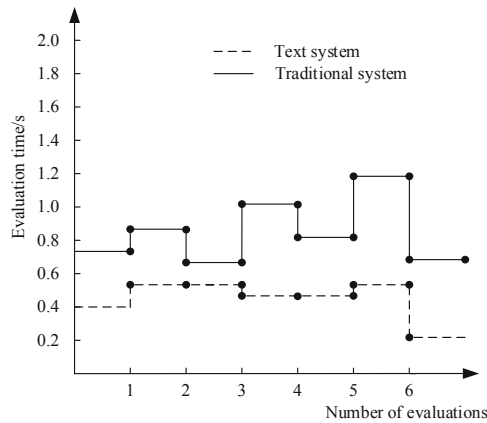
Fig. 12. Experimental results of time spent in the evaluation process

It can be seen from Fig. 12 that due to the influence of external factors, the time spent in the evaluation process will change irregularly, but the system will accumulate experience in each evaluation. When the number of evaluations reaches more than 6, the evaluation time will be greatly shortened. Although it is difficult to determine the time spent in the evaluation process, the evaluation system designed in this paper always takes less time than the traditional system. The evaluation time of the traditional system fluctuates between 0.6 s and 1.2 s. The evaluation time of the system in this paper fluctuates between 0.2 s and 0.6 s. The evaluation time is very short, which ensures the real-time performance of the evaluation.

The evaluation system studied in this paper displays the work results in a short time to effectively ensure the work effect. Once it is found that the anti-interference performance of the equipment is poor, the system will automatically put forward the solution strategy. It is an efficient and accurate evaluation system.

5   Conclusion

Efficient English education is the future development direction of higher English education in China. At present, the development of efficient English Teaching in China still

faces many difficulties and challenges. Therefore, this paper designs a corpus based College English teaching comprehensive ability evaluation system, evaluates the existing college English teaching comprehensive ability through the optimization of hardware and software, and provides reference for formulating English teaching strategies, which can play a certain guiding role in the training and development of College English teachers in China, It can provide reference for the development and training plan of English education for relevant educational institutions and promote the development and progress of English education in Colleges and universities in China.

References

1. Wu, C.C., Yang, C.Y., Hwang, M.S., et al.: The design and application of a web-based teacher evaluation system for STEM education. Int. J. Electr. Eng. Educ., 002072091985278 (2019)
2. Ulusoy, A.J., Pecci, F., Stoianov, I.: An MINLP-based approach for the design-for-control of resilient water supply systems. IEEE Syst. J. **PP**(99), 1–12 (2020)
3. Wu, X.L., Liu, L.X., Yang, L.Y., et al.: Comprehensive rehabilitation in a patient with corpus callosum syndrome after traumatic brain injury: case report. Medicine **99**(28), e21218 (2020)
4. Damme, P., Ungethuem, A., Hildebrandt, J., et al.: From a comprehensive experimental survey to a cost-based selection strategy for lightweight integer compression algorithms. ACM Trans. Database Syst. **44**(3), 1–46 (2019)
5. Ni, Q., Zhang, L., Zhang, B., et al.: Interdisciplinary method for assessing students' ability based on STEM projects. Int. J. Eng. Educ. **35**(2), 698–709 (2019)
6. Huang, J.: An Internet of Things evaluation algorithm for quality assessment of computer-based teaching. Mob. Inf. Syst. **2021**(13), 1–10 (2021)
7. Blondet, G., Le Duigou, J., Boudaoud, N.: A knowledge-based system for numerical design of experiments processes in mechanical engineering. Expert Syst. Appl. **122**, 289–302 (2019)
8. Cheng, F., Gao, Y., Guo, X., et al.: Fabrication of nanostructured skutterudite-based thermoelectric module and design of a maximum power point tracking system for the thermoelectric pile. IEEE Sens. J. **PP**(14), 1 (2019)
9. 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
10. Chen, H., Zhou, J.: Effects of sodium selenite on oxidative damage in the liver, kidney and brain in a selenite cataract rat model. Biol. Trace Elem. Res. **197**(2), 1–11 (2020)
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. Xi, T.: Design of English diagnostic practice sentence repetition recognition system based on matching tree and edge computing. Wirel. Commun. Mob. Comput. **2021**(3), 1–8 (2021)
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

Fuzzy Evaluation Model of Teaching Quality of Physical Education Course Based on Deep Reinforcement Learning

Zhiqiang Wang[✉] and Xiangyu Xu

Sports Center, Xian Eurasia University, Xian 710000, China
`wangzhiqiang8561@163.com`

Abstract. Because there are many factors affecting teaching evaluation, the evaluation results are difficult to reach a high level. Therefore, a fuzzy evaluation model of teaching quality of physical education course based on deep reinforcement learning is designed. On the basis of clarifying the requirements of teaching quality evaluation, the factor analysis method is used to preprocess the teaching data, the Bartlett spherical test is used to verify it, and the data meeting the verification requirements are K-mean clustered. Finally, based on the clustered data results, the fuzzy evaluation model is constructed according to the idea of minimum membership weighted average deviation. The test results show that the evaluation results of the design model have high adaptability with the actual results, and can meet the evaluation needs.

Keywords: Deep reinforcement learning · Physical education courses · Teaching quality · Fuzzy evaluation · Factor analysis · Bartlett sphere test · K-Mean clustering

1 Introduction

As the key task of education, the quality management of teaching is very important in the overall management of the school. The evaluation of teaching quality can reflect the results of running a school. In school management, evaluating teachers' teaching quality is conducive to school managers to accurately and comprehensively grasp the teaching progress, teachers' teaching work and the realization of teaching objectives, so as to improve the teaching quality. Teaching quality evaluation on teaching quality has become a very important issue in university management. Teaching quality evaluation can improve teaching quality and promote the reform of quality education, which is an important influencing factor of the whole education quality [1–3]. However, because the teaching process is constantly changing under the influence of many factors, it belongs to the content of spiritual labor and is displayed in the form of art, so there will be a large number of non quantitative factors in it, and the fuzziness is strong. If its quantitative evaluation is quite difficult, the above factors undoubtedly increase the complexity of teaching quality evaluation and strengthen its difficulty. The so-called teaching process

includes teaching and learning. Evaluating the quality of teachers' teaching is much more complex than evaluating the quality of a product [4, 5]. Teaching activities are not carried out by teachers alone, but need bilateral interaction with students. Because there are many factors affecting teaching and many factors affecting teaching quality, there are naturally many aspects of evaluation content. Building a sound teaching quality evaluation system can achieve considerable and fair evaluation standards, which is a meaningful topic [6–8]. Nowadays, the times are developing. In the face of the new situation, how to help the school's teaching level improve rapidly, the most key means is to improve the teaching quality. Then, where is the key to improve the teaching quality? The key is to improve the teaching quality evaluation system. Nowadays, the focus of the major colleges and universities in China is based on modern educational thinking. In addition, it must be consistent with the actual situation of their colleges and universities, so as to make the established teaching quality evaluation system scientific and reasonable. With China's development entering a new period, higher education has also ushered in new challenges and higher requirements for teaching quality. The main task of this work is teaching. If we want to promote the reform and development of colleges and universities, we must improve the teaching quality.

Constructing a scientific and reasonable teaching quality evaluation system can strengthen teaching management and promote the further improvement of teaching quality. Because different schools have different understanding and attention to teaching quality, there are also differences in the content of evaluation [9, 10]. The above teaching quality evaluation methods play a positive role in promoting teachers' teaching level and improving teaching quality, but these methods have defects: the first two methods are mixed with the influencing factors that they think will directly affect the evaluation results, and the third method only evaluates students' transcripts, which is one-sided. Teaching process is a dynamic process, which reflects the bilateral interaction between education and learning. There are many factors that can affect teaching quality, and the degree of value influence is different. Therefore, the evaluation results are also more complex, which can not be expressed by mathematical analytical formula alone. In essence, it is a more complex and nonlinear comprehensive decision-making problem. Therefore, it is unscientific to evaluate the teaching quality according to the above model, which is easy to have strong subjectivity and randomness, resulting in large deviation of the calculation results, that is, unreasonable. In recent years, artificial neural network has been born and become a new technology. Its new reason is that it contains characteristics different from other technologies, such as nonlinear mapping, real-time optimization and learning classification. Its characteristics reflect its unique advantages in recognition filtering and pattern recognition. Artificial neural network brings more advantages to teaching quality evaluation. Its characteristic is that it can mine laws from unknown and a large number of complex data. In this regard, it opens up a new path for teaching quality evaluation, especially when dealing with various types of data, it can fully approach any complex nonlinear relationship, Modeling can greatly solve the problem of comprehensive evaluation, which greatly reduces the influence of factors, which is far beyond the effect of traditional methods. Therefore, introducing the theory of artificial neural network into the teaching quality evaluation system, on the one hand, solves the dual problems of qualitative and quantitative indicators. In the traditional

evaluation methods, it is necessary to establish cumbersome mathematical models and analytical formulas, and the new method perfectly solves the difficulties in the traditional model; On the other hand, it avoids the influence of factors on the evaluation results, so as to ensure the effectiveness and accuracy of the evaluation. Therefore, the evaluation model based on neural network theory is the most effective means to evaluate teaching quality.

Based on this, this paper designs a fuzzy evaluation model of physical education teaching quality based on deep reinforcement learning, and verifies the effectiveness of the design method through experimental tests.

2   Analysis of Teaching Quality Evaluation Requirements

The evaluation of teaching quality in colleges and universities is not a simple evaluation work, but a more complex process, which contains a variety of factors, including teaching conditions, curriculum, teaching and learning effects, etc., and various factors interact with each other. In addition, teachers The relationship with students is complex, which increases the possible factors that affect teaching quality. Nowadays, the education circle still has not proposed a set of recognized teaching quality evaluation system. From the analysis of existing research, most of the research mainly focuses on the subject On issues such as evaluation and teacher-student relations, the method and method of establishing a teaching quality evaluation system, using it as a prerequisite, how to explore the method of evaluating teaching quality levels.

2.1   Analysis on the Requirements of Teaching Quality Evaluation of Physical Education

There is more than one method to evaluate the quality of teaching, and there are many different methods such as individual teachers, peers, administrative leaders, experts, etc. Due to the different roles of the evaluation subjects, their roles in the evaluation should be different. Each evaluation method and its results are only a part of the evaluation of the teaching quality of physical education, but not the whole of the teaching quality of physical education. From the point of view of the evaluation of college teachers, there are many teachers, and the number of evaluations increases. If the census evaluation method is time-consuming and laborious, it is inevitable, and at the same time, it is impossible to find out that it is caused by interpersonal relationships or unfamiliarity with the teaching process. Other influencing factors are difficult to operate in reality. Therefore, the method of taking students as the subject of evaluation is more dominant in reality and is adopted by most colleges and universities. Since the 1980s, Chinese colleges and universities have paid more attention to the evaluation of higher education, and gradually carried out evaluation activities with students as the main body, which strengthened the quality of colleges and universities to a certain extent. Analyzed from the perspective of students, it is the direct object of education, so it has the right to evaluate the teaching work of teachers, and there is authenticity. Due to the different classifications of universities in our country, the professional settings between schools are diverse and complex, and the basic conditions of students are quite different. The evaluation of teachers' teaching quality of physical education is naturally inconsistent.

2.2   Determination of Teaching Quality Evaluation Content of Physical Education Course

The key to designing a complete teaching evaluation system lies in its content setting. The first thing to consider is that learning is a process from point to surface. This is already the case for development. The common feature of learning and growth environments is that they are both diverse and almost impossible. To quantify one of the courses and the teaching role played by the teacher in a stage, usually the performance is not used as the main evaluation index, and the teaching effect is not used as the main evaluation index. Generally, the key points are placed in the teaching process. From another perspective, from the perspective of process management, school teaching is a combined process, which contains multiple factors and multiple links. Compared with the nature of the curriculum, the teaching structure, and the differences in disciplines, it is impossible to compare them one by one in evaluation.. Therefore, when designing the teaching quality evaluation system of physical education, the key is to consider the factors that directly affect the teaching level, especially the commonness between factors and design.Based on the analysis of the existing evaluation system, the design of more content in the system index design is summarized as the following six points:

① Teaching attitude: In the teaching process, whether the teacher is serious and responsible, the mental state in the class, whether there is sufficient preparation for the lesson, and whether the tutoring and answering questions after class are specific and patient.

② Teaching content: whether the selection is appropriate or not, whether the key points of the course are highlighted, the concept is clearly explained, whether students can understand deeply, and whether it is combined with reality.

③ Teaching ability: clear thinking, with language charm, neat writing on the blackboard, and whether there are breakthroughs in key and difficult points.

④ Teaching methods: Whether to teach students in accordance with their aptitude, flexible teaching, whether to maintain communication with students, pay attention to their real-time dynamics, whether to use flexible teaching methods, and pay attention to cultivating students' practical ability and innovative spirit.

⑤ Teaching and educating people: Whether to uphold rigorous teaching tasks and be a teacher; whether to strictly require students to treat each student to ensure fairness and justice.

⑥ Teaching effect: Is it guaranteed to improve students' performance while promoting the ability of students to think proactively? Does it ensure that students have a full grasp of the teaching content?

3   Design of Fuzzy Evaluation Model for Teaching Quality of Physical Education

3.1   Factor Analysis

In order to improve the credibility and reliability of evaluation, this paper makes factor analysis on students' teaching evaluation, so as to provide a set of real and detailed basic

data for the model. Factor analysis mainly studies how to condense a large number of original variables into a small number of factors through the least information loss, and how to make these factors have naming and explanatory multivariate statistical analysis methods.

Suppose the variables that affect the quality of teaching: $x_1, x_2 \ldots x_p$, and the standard deviation of the p parameters is set to 1, and the average value is 0, the original variable parameters at this time are k $(k < p)$ factors $f_1, f_2, \ldots f_k$ represents the linear combination, namely:

$$\begin{cases} x_1 = a_{11}f_1 + a_{12}f_2 + \ldots + a_{1k}f_k + \varepsilon_1 \\ x_2 = a_{21}f_1 + a_{22}f_2 + \ldots + a_{2k}f_k + \varepsilon_2 \\ \ldots \\ x_p = a_{p1}f_1 + a_{p2}f_2 + \ldots + a_{pk}f_k + \varepsilon_p \end{cases} \tag{1}$$

Equation (1) is the mathematical model of factor analysis, in which the parameter F is called a factor, and its name is also called a common factor. The name comes from because it frequently appears in almost all linear expressions, hence the name. The factor is the k coordinate axes that are perpendicular to each other in the high-dimensional space; the parameter A represents the factor loading matrix, and $a_{ij}(i = 1, 2, \ldots, k, j = 1, 2, \ldots, k)$ represents the load of the i original variable on the j factors, which is called the factor loading. If x_i is regarded as a vector in the k dimensional factor space, then the value of a_{ij} is the projection of x_i on the coordinate axis f_i; the parameter ε takes its value as 0, and its name is called a special factor, which is used to represent certain factor pairs The unexplainable part of the original variable.

In order to further adopt the factor analysis method to determine the relationship between the influencing factors and the original variables, and to clarify the effect of factor analysis and its importance, this paper sets the following concept definitions.

(1) **Factor loading**

It has been verified that when there is no correlation between factors, factor load a_{ij} implies that variable x and factor f_i reflect the correlation between x_i and f_i. Set the value range of the factor load to $a_{ij} \leqq 1$, and the value of its absolute value is closer to 1, then it proves that the correlation becomes stronger, and a_{ij} also verifies that the factor f_i is extremely important for x_i.

(2) **The degree of commonality of variables**

The variance of a variable is also known as the degree of commonality of variables. The definition of the degree of commonality affected by the teaching quality of physical education variable x_i is:

$$h_i = \sum_{i=1}^{k} a_{ij}^2 \tag{2}$$

As shown in formula (2), the value of the commonality of variable x_i is the sum of the squares of the elements in the i row in the load matrix A. If the common degree

of variables is almost all above 0.8, then the extracted factors can easily obtain the original variable information, and there is very little loss of information, thus obtaining a relatively good factor analysis effect. Therefore, the degree of commonality of variables is an extremely important measure of the effect of quantitative factor analysis.

Variance Contribution of Factors

The mathematical definition of the variance contribution of factor f_i is:

$$s_i = \sum_{i=1}^{p} a_{ij}^2 \qquad (3)$$

Equation (3) shows that the variance contribution of factor f_i is the sum of the squares of the elements in the i column in the factor loading matrix A. The variance contribution value of factor f_i shows how important the corresponding factor is.

On the basis of the above, the basic steps of factor analysis of the teaching quality of physical education courses are as follows.

First of all, it is necessary to clarify the key content of factor analysis: one is how to construct factor variables; the other is how to interpret factor variables. The basic steps are as follows:

Whether the original variables are suitable for the factor analysis method, in order to verify its effectiveness, it is necessary to ensure that the parameters of the original variables must have a strong correlation. If they are relatively weakly related to each other, they cannot fully reflect the common factor variables of the common characteristics of some variables. Whether a variable is suitable for factor analysis, the easiest way is to calculate the personality coefficient matrix between the variables. In the process of calculation and testing, most of them are < 3 and fail to pass the test, which means that this variable is not suitable for factor analysis. In order to make further judgments, this article uses Bartlett's sphere test (Bartlett Test of Sphericity). Its operation is mainly based on the correlation coefficient matrix of the variables, in which all elements on the diagonal of the correlation coefficient matrix of the null hypothesis correlation coefficient matrix are 1, and non-diagonal The elements on the line are all 0, and their values are all calculated based on the determinant of the correlation coefficient matrix. If the value is large, it is suitable for factor analysis; on the contrary, it is not suitable for factor analysis. Calculated as follows:

$$\chi = \frac{\sum_{i=1} r_{ij}^2}{\sum_{i=1} r_{ij}^2 + \sum_{i=1} t_{ij}^2} \qquad (4)$$

Among them, the simple correlation coefficient in the model is represented by r_{ij} between variable i and variable j; the partial correlation coefficient is represented by t_{ij} between variable i and variable j. The value of χ ranges from 0 to 1. If its value is closer to 1, then the sum of squares of simple correlation coefficients between variables is larger than the sum of squares of partial correlation coefficients, which means that it is more suitable for factor analysis. Conversely, the smaller χ is, the less suitable it is for factor analysis.

Based on this, this article sets an χ standard:

$0.9 < \chi$: very suitable; $0.8 < \chi < 0.9$: suitable; $0.7 < \chi < 0.8$: general; $0.6 < \chi < 0.7$: not suitable; $\chi < 0.5$: not suitable.

On this basis, the task of constructing factor variables of teaching quality evaluation model of physical education course is completed. There are many ways to determine factor variables in factor analysis, among which the principal component analysis method based on principal component model is the most frequently used factor analysis method. The research of this paper makes it more practical in the actual analysis work. By analyzing the value of load matrix, the relationship and logic between the original variables and factor variables are obtained, and finally the factor variables are named. In the load matrix, there are multiple large load coefficients in a row, which means that an original variable may have a large correlation with some factors at the same time. Similarly, there may be multiple relatively large load factors in a column, indicating that a factor variable may explain the information of multiple original variables. In this paper, the maximum variance method is used to realize the process:

$$\delta(x_{ij}) = \frac{1}{\varepsilon_1 h_{ij}/x_{ij} + \varepsilon_2 h_{ij}/x_{ij} + ... + \varepsilon_k h_{ij}/x_{ij}} \tag{5}$$

Among them, $\varepsilon_k h_{ij}/x_{ij}$ represents the ratio of the number of disturbed data to the total number of previous data, and finally calculates the score of the factor variable. After determining the factor variable, for each sample data, in order to obtain their detailed data value on each factor, the obtained value is called the factor score, which corresponds to the score of the original variable.

3.2 Factor Clustering

The purpose of clustering method in overall analysis is to cluster the same category of things and extract the factors related to each factor. This method is usually based on the actual data level. It needs to extract a certain number of samples from the basic data as the basis of the research method, and treat different categories of factors differently according to the similarity principle.

Usually, we use the clustering method as a metaphor for the classification of real things. Assuming that there are many kinds of items in front of us, we need to select some items from these items, and we need to classify these items first. The standard principle of distinction is to gather the goods with the same characteristics and divide them into one category according to the characteristics of the goods themselves.

This paper extends the clustering method to explore the process of factor clustering in physical education classroom. On the premise of exploration, we need to have certain data samples, and set different standards and groups according to different categories. The clustering methods can also be classified from different angles.

From the perspective of different levels, the diversity of categories endows the method with different levels of standards. Focus on observing whether there is a positive and negative correlation ratio in the close relationship among the sample factors, how to form the same group at what classification speed, whether its characteristics are representative, which can reflect the characteristics of things themselves, and what kind of chemical

reaction occurs when things are classified. From different speed division angles, the samples analyze different variable factors according to their characteristic principles, conduct observation experiments on these variable factors, observe whether the studied variable factors have rigorous characteristic indicators, and analyze whether the speed of the combination of variable factors meets the initial clustering attribute requirements, Whether the effectiveness of the correlation extraction factors of the specified data samples affects the data of the basic elements of the physical education classroom, we should make a simple adjustment.

In order to better observe the degree of change of the factor variables during the use of the clustering method, two types of clustering forms are obtained according to the above analysis from different angles, one is type Q and the other is type R, and type Q represents macro The nature of the research sample data clustering model, such as: the obvious characteristics of things, the length of the physical education class, the participants in the physical education class, exercise items, etc. Type R represents a clustering model of research sample data of microscopic nature, such as the changes that occur after things happen, the improvement of physical fitness, and the enthusiasm for sports classes. The clustering factors with the above several representative classifications combine their characteristics to perform rapid clustering, which is much more convenient. Specific steps are as follows:

First, create a working data file in the SPSS data window or directly open the file that needs to be clustered. Click the menu item in K-Means order to open the cluster analysis dialog box. In the dialog box, specify the variables involved in the analysis, select the clustering method, select the statistics required to be output, select the statistical chart, set the options for generating new variables, etc., and finally you can output the analysis and application of the results.

3.3 Building a Fuzzy Evaluation Model

After obtaining the data in the form of clustering, this paper constructs a fuzzy evaluation model based on the idea of the weighted average deviation of the minimum membership degree, as follows.

First, based on the clustering output results, build a set of objects with a top and bottom structure, where the number of layers depends on the number of factors. And establish the corresponding judgment set V:

$$V = \{v_1, v_2, ..., v_n\} \tag{6}$$

Among them, n represents the number of clusters.

Establish a single factor evaluation matrix R with the cluster unit as the target

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix} \tag{7}$$

Calculate the corresponding weight set

$$W = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{pmatrix} \tag{8}$$

in,

$$w_{mn} = \frac{\sum\limits_{n=1} (\chi_{mn})}{\sum\limits_{m=1}\sum\limits_{n=1} (\chi_{mn})} \tag{9}$$

$$\chi_{mn} = x_{11} \vee x_{12} \vee \ldots \vee x_{ij} \tag{10}$$

Among them, $\vee$ represents the fuzzy operator.
At this time, a comprehensive evaluation of all clusters is performed

$$T = C \sum_{i=1}^{n} \frac{W_i}{100} \tag{11}$$

Among them, C represents the correlation coefficient, and T is the final evaluation result.

4 Experimental Results and Analysis

4.1 Experimental Data Settings

According to the teacher's teaching practice and teaching plan, select the most appropriate evaluation result from the evaluation variables after each evaluation index in the evaluation table, and mark it with "√". Then, according to the scores corresponding to the evaluation results of each evaluation index, the total score is calculated, which is the evaluation score of the teaching quality of physical education.After the model is built, the collected data is processed, and 270 sets of data are used for training, and the remaining 30 sets of data are used for testing. Part of the data is shown in Table 1.

Table 1. Part of the experimental data

Factor	Sample number			
	1	2	3	…
Educational goals	0.85	0.59	0.59	…
Professional knowledge	0.92	0.78	0.68	…
Explanation level	0.46	0.82	0.82	…
Patience	0.82	0.83	0.89	…
Infectious	0.37	0.75	0.87	…
Manner	0.69	0.48	0.78	…
Conceptual theory	0.52	0.59	0.74	…
Content settings	0.54	0.59	0.71	…
Practice content	0.59	0.67	0.76	…
Depth of knowledge	0.68	0.63	0.62	…
Enlightenment	0.82	0361	0.64	…
Way to use	0.46	0.85	0.86	…
Teach students in accordance with their aptitude	0.53	0.81	0.83	…
Innovation and development	0.69	0.82	0.92	…
Learning interest	0.67	0.84	0.57	…
Self-study ability	0.69	0.75	0.62	…
Basic knowledge	0.54	0.56	0.63	…
Problem analysis	0.52	0.53	0.51	…
Problem solved	0.62	0.63	0.24	…
Comprehensive quality	0.71	0.44	0.91	…
Creativity	0.55	0.52	0.50	…

Among them, the 1–21 columns of each sample group are the data of 21 secondary indicators included in the five primary indicators of teacher quality, teaching attitude, teaching content, teaching methods, and teaching effect. Through the verification and testing of the model, the methods proposed in [9] and [10] are used as the control group.

4.2 Analysis of Test Results

In order to verify the effect of fuzzy evaluation on the teaching quality of physical education, the methods of reference [9], reference [10] and this paper are used to test its adaptability, and the test results are shown in Fig. 1.

According to the analysis of Fig. 1, when the experimental data is 10, the fitness of fuzzy evaluation of physical education teaching quality of reference [9] method is 0.60, the fitness of fuzzy evaluation of physical education teaching quality of reference [10] method is 0.56, the fitness of fuzzy evaluation of physical education teaching quality of
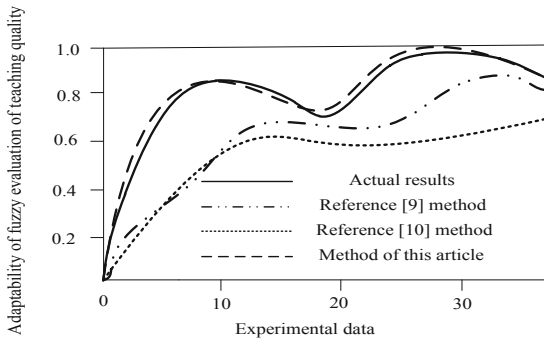
Fig. 1. Comparison of evaluation results of different methods

this method is 0.86, and the fitness of fuzzy evaluation of physical education teaching quality of the actual result is 0.86, The fitness of fuzzy evaluation of physical education teaching quality of this method is basically consistent with the actual value, which shows that this method can make a scientific evaluation of teaching quality.In this paper, the weighted k-clustering method is used to improve the effectiveness of the fuzzy clustering model, because the weighted k-clustering method is used to improve the fitness of the teaching data in this paper.

In order to test the evaluation error of teaching quality of physical education, the methods of reference [9], reference [10] and this paper are used to detect the evaluation error. The test results are shown in Table 2.

Table 2. Evaluation error statistics of different methods

Data number	Reference [9] method	Reference [10] method	Method of this article
1	0.25	0.35	0.05
2	0.48	0.45	0.07
3	0.51	0.47	0.02
4	0.42	0.41	0.03

According to the analysis of Table 2, when the data number is 1, the evaluation error of reference [9] method is 0.25, the evaluation error of reference [10] method is 0.35, and the evaluation error of this method is 0.05; When the data number is 2, the evaluation error of reference [9] method is 0.48, the evaluation error of reference [10] method is 0.45, and the evaluation error of this method is 0.07; When the data number is 4, the evaluation error of reference [9] method is 0.42, the evaluation error of reference [10] method is 0.41, and the evaluation error of this method is 0.03. The above data show that the evaluation error of physical education teaching quality of this method is far lower than that of other methods, and the evaluation effect of physical education teaching quality is better. This is because the model designed in this paper uses the factor analysis method to preprocess the teaching data, and realizes the verification of

the teaching data through the Bartlett spherical test method, so as to effectively reduce the quality evaluation error.

5   Conclusion

This paper constructs the fuzzy evaluation model of physical education teaching quality based on deep reinforcement learning, preprocesses the teaching data by factor analysis method, verifies the teaching data by Bartlett spherical test method, clusters the teaching data by K-mean clustering, and constructs the fuzzy evaluation model according to the minimum membership weighted average deviation method in deep reinforcement learning to realize the fuzzy evaluation of physical education teaching quality. The experimental results show that:

(1) When the experimental data is 10, the fitness of fuzzy evaluation of physical education teaching quality of this method is 0.86, which is basically consistent with the actual results, indicating that this method can make a scientific evaluation of teaching quality.
(2) The maximum error of the teaching quality evaluation of physical education course in this method is no more than 0.07, which shows that the teaching quality evaluation effect of physical education course in this method is better.

This paper only completes the vertical evaluation of teachers, and can also consider comparing all teachers in different indicators in the whole system.

References

1. Cho, J., Baek, W.: Identifying factors affecting the quality of teaching in basic science education: physics, biological sciences, mathematics, and chemistry. Sustainability **11**(14), 3958–3968 (2019)
2. Ali, J., Bashir, Z., Rashid, T.: Weighted interval-valued dual-hesitant fuzzy sets and its application in teaching quality assessment. Soft Comput. **25**(5), 3503–3530 (2020)
3. Liu, S.: Research on the teaching quality evaluation of physical education with intuitionistic fuzzy TOPSIS method. J. Intell. Fuzzy Syst. **40**(5), 1–10 (2021)
4. Yuan, T.: Algorithm of classroom teaching quality evaluation based on Markov Chain. Complexity **21**(21), 1–12 (2021)
5. Xie, Z., Su, Z.: Evaluation of college English classroom teaching quality dependent on triangular fuzzy number. Int. J. Electr. Eng. Educ. 002072092110020 (2021)
6. Liang, W.: Modeling and simulation of teaching quality in colleges based on BP neural network and training function. J. Intell. Fuzzy Syst. **37**(5), 6349–6361 (2019)
7. Peng, F.: Application of deep learning and cloud data platform in college teaching quality evaluation. J. Intell. Fuzzy Syst. **39**(4), 5547–5558 (2020)

8. Liu, S., Bai, W., Srivastava, G., et al.: Property of self-similarity between baseband and modulated signals. Mobile Netw. Appl. **25**(4), 1537–1547 (2019). https://doi.org/10.1007/s11036-019-01358-9

9. Bao, L., Yu, P.: Evaluation method of online and offline hybrid teaching quality of physical education based on mobile edge computing. Mob. Netw. Appl. 1–11 (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

Design of Online Teaching Mode Recognition System for Ideological and Political Curriculum Based on Hash Algorithm

Jun Yang[1] and Dong Yang[2(✉)]

[1] College of Marxism, Fuyang Normal University, Fuyang 236041, China
[2] Nanchang Institute of Technology, Nanchang 330108, China
wangdong1233211@163.com

Abstract. In the practical application process, there are many defects in the online teaching of ideological and political courses, leading to the inability to accurately identify the teaching content. In order to avoid this situation, this paper designs and analyzes the ideological and political course online teaching mode recognition system based on the hash algorithm. In the system hardware section, Design main control circuit and PCB board of teaching mode identification system; In the software part, the teaching recognition function module adopts hash algorithm, designs the multi-core recognition database under the hash algorithm, and then completes the design process of the ideological and political course online teaching pattern recognition system. The final test results show that the recognition rate of the designed system is high, which is more than 90%, indicating that its recognition effect is good, has strong stability and certain practical significance.

Keywords: Hash algorithm · Ideological and political courses · Online teaching · Teaching mode · Online teaching recognition · System design

1 Introduction

Under the new situation, in response to the national call of "continuous suspension of classes", universities across the country will make full use of network technology to build an online teaching platform, and the curriculum education has been fully completed online. In just a few months, teachers have experienced the rapid growth stage of online teaching from basically completing online teaching tasks, to the ability of online teaching to actively redesign courses [1]. During this period, it is worth noting is that in the daily teaching activities, most professional teachers combined with current affairs in the course, ideological and political exploration, did in the mission at the same time, combined with the real case in the society, encourage and guide graduate students advocate science, patriotic family, perfect practice of the trinity education concept, for "ideological" education background opened new ideas, expand the new path [2]. In fact, the concept of integrating ideological and political education into professional curriculum education has a long history, which can further strengthen and improve the effect

of ideological and political education for college students [3]. Online ideological and political education should be run through the talent training system, comprehensively promote the ideological and political construction of diversified courses in colleges and universities, give full play to the educational role of each course, and take it as an important guideline to improve the quality of talent training in colleges and universities, highlighting the fact that the Party and the state attach great importance to this work.But online courses: Ideological and political courses and ordinary ideological and political courses are two different concepts. Ideological and politics of online courses refers to the comprehensive education concept of combining all kinds of courses and ideological and political theory courses in the form of building a full, whole and whole political education pattern, to form a synergistic effect and take "cultivating people through virtue" as the fundamental task of education. Ideological and political online courses emphasize the integration and combination of ideological and political elements as implicit education [4]. Therefore, it is very necessary to establish the main battlefield and ideological and political promotion path of online and offline classes, and establish a scientific and reasonable curriculum. It has become an urgent topic to be solved and improved in the training of college students in China, which has irreplaceable value to improving the quality of talent training.

Different teaching modes also have certain differences in the effect of different teaching, and have great purpose and initiative. However, college students' understanding of ideological and political courses is relatively more relaxed, mainly because after this stage of learning, it has a certain basic knowledge reserve, and has a certain purpose and initiative to engage in scientific research, so it is very different compared with other students when choosing its online training mode. Moreover, the identification of teaching modes in ideological and political courses is usually related, and will have a great impact on the final teaching results. Teaching identification is mainly to sort out and summarize some problems and hot topics in the course, and combine with their own integration, to create a new extended teaching mode.

In recent years, many scholars have analyzed and studied pattern recognition methods. For example, Gao y et al. Proposed a pattern recognition algorithm based on container theory, which summarized the requirements of pattern classification into two different categories: one is the comparison and matching of input-output sequences, and the other is discrete data structure to realize effective pattern recognition [5]; Wang h et al. Designed a flatness pattern recognition method based on Legendre polynomial, which makes full use of the adaptive feature learning ability of deep learning network to complete the recognition of working pattern. Although the above scholars have studied the pattern recognition methods, they are not aimed at the teaching field, resulting in poor effect in the pattern recognition process of Ideological and political courses [6].

Hash algorithm is a more comprehensive and systematic calculation method, also known as hash. It mainly inputs data of any length into fixed length output through hash algorithm, and the output value is hash value [7]. In the calculation process, it is very strict and has almost no error. Therefore, it is widely used in the creation and adjustment of online teaching recognition system. Therefore, in order to solve the problems of the above methods, this paper designs an online teaching pattern recognition system based on hash algorithm. In the hardware part of the system, the main control circuit and PCB

board of the teaching pattern recognition system are designed; In the software part, the teaching recognition function module adopts the hash algorithm, and designs the multi-core recognition database under the hash algorithm. Here, complete the design process of the ideological and political course online teaching pattern recognition system, in order to provide some help to improve the pattern recognition effect of the ideological and political course.

2   Hardware Design of Online Teaching Mode Recognition System in Ideological and Political Courses

2.1   Main Control Circuit Design of the Teaching Mode Recognition System

It is necessary to design the hardware of the online teaching mode recognition system of ideological and political courses. In the design process, the master control circuit needs to be designed first. The main contol circuit is mainly composed of microcontroller, infrared thermo electric detection circuit, luminance detection circuit, DALI interface circuit, wireless communication module, code dial switch and knob circuit, power supply circuit, etc. The block diagram of the system hardware is shown in Fig. 1:
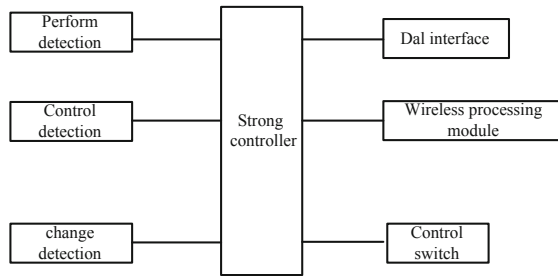
Fig. 1. Composition block diagram of main control system

After completing the design of the system hardware composition block diagram described above, next, the main control system DALI bus is required to supply the DC voltage, the working voltage of the microcontroller PIC16F884 is 5 V, and the working voltage of the wireless transceiver module RF905SE is 3.1−3.6 V, so the multi-stage DC-DC transformation needs to be designed. The first stage DC-DC transformation transforms the power supply voltage 25 V to 12 V through the three-end voltage stabilizer L7812, the second stage DC-DC transformation transforms the 12 V voltage to 5 V by L7805, and the third stage DC-DC transformation transforms 5 V to 3.3 V through LM317. The high-frequency filter capacitances C4, C5, C18, and C10 are selected as 0.1 uF, and the low-frequency filter capacitances C11, C12, and C19 are selected as 47 uF. According to the design of the LM317 circuit, the calculation of the output voltage of the LM317 is specific as shown in formula 1:

$$K = 2 - \frac{1}{3w} + 3\beta \qquad (1)$$

In Formula 1: *K* represents the output voltage, *w* represents the adjustment limit value, and *β* represents the actual control range. From the above calculation, the actual output voltage can finally be obtained. Using it as a microcontroller is the core standard voltage of the main control system, using Microchip PIC16F884. The PIC16F884 is the Microchip's high-performance 8-bit microprocessor, It uses a streamlined instruction set, Harvard bus structure, secondary water to take instructions, The chip is equipped with quite rich resources: high-precision internal oscillators, Enhanced low-current watchdog timer with an on-chip oscillator, High durability flash memory EEPROM unit, FLASH of 4K bytes, The 256 Bytes of the RAM, 35 IO pins with direction controlled control, 2 analog comparator modules, 14 A/D conversion channels at 10-bit resolution, 3 Time counters, Enhanced capture, comparison, and PWM modules, Enhanced USART module, Main Sync serial port, Supports a 3-line SPI and an IC primary/slave mode with an IC address blocking function, Greantly reduce the peripheral devices, And can be used for online serial programming with two pins, It facilitates debugging and subsequent development. When completed, the microcontroller also needs to be associated with other parts of the circuits.

The PIC16F884 has two timers that provide electric delay, an oscillator start timer to ensure that the chip is reset until the vibration reaches stability and an electric delay timer providing a fixed delay of 64 ms at each charge to ensure that the device is reset before the supply voltage stabilizes. If under pressure conditions occur, the under pressure reset circuit provides a reset delay of at least 64 ms. Because of these three functions, it no longer requires the assistance of external reset circuits in most applications, so that no additional external reset circuits are designed in this paper. And its oscillator module has a variety of clock sources and selection functions, through the software can choose the external or internal system clock sources. To reduce peripheral circuits and facilitate debugging, take the software to set the internal clock to 4 MHz [8]. Signal ADO, AD1, AD2, and AD3 are used for code dialing switch inputs. The dial switch is used to set the address of the master controller, the address of the RF905SE.AD0, AD1, AD2, and AD3 are directly connected to the I/O port of the MCU, and connected to the VCC by 10K resistance. When the switch is OFF, the pin level is high and a low level when the switch is ON [9]. After the MCU collects these pin levels, an address is obtained. As the receiving address of the RF905SE, only if the sending address of the NetUSB905SE wireless receiving module and the receiving address of the RF905SE match, the information can be successfully sent to the RF905SE through the NetUSB905SE, then read and decoded by the main control system, and transmits the DALI command to the next level through the DALI bus [10].

In addition, the circuit of the master control system has also designed two adjustable analog signals, implemented using two potentials, RT1 and RT2. Signal DTIME and LLD are the potentiometer voltage input. The potentiometer RT1 is used to set the shutdown time torr, where the presence of an abnormality is not detected by the topr time, and the execution command is issued. The potentiometer RT2 is used to set the desired recognition properties. In automatic dimming mode, the main control system issues an identification command by comparing the illumination and LLD values of the environment. The DTIME and the LLD are connected to the A/D functional pin

RBO/RB1 of the microcontroller, respectively. The potentiometer and receives a small capacitance to prevent voltage mutations that can act as filtering.

The signals ICSPCL, ICSPDA, VPP and GND and VCC, Implement the connection to the development tool, MPLAB ICD2, It has the capabilities of online debugging and online serial programming, Easy for development and commissioning; Signal PWR_UP, AM, TRX_CE, SCK, MISO, MOSI, TX_EN, DR, CD, CSN connect to the wireless communication module, Complete the wireless communication function; Pins COM1-, COM2-, DALIR, DALIRXEN, DALITXEN, and DALITX are used to design the DALI interface circuits, Complete the sending and receiving of DALI signals; The LIN is the luminosity detection signal, To detect the illumination degree of the external environment; HW is the input signal of infrared detection, For realizing automatic sensing; The LED is used for the indicator lamp control; The FM is used to control the buzzer alarm.

Subsequently, the design of the wireless communication interface circuits is also required. The RF905SE module of NewMsgis combined with the upper position computer USB wireless emission module to achieve stable and reliable communication under the same configuration. It uses the highest operating rate of 50 K in 433 MHz band b p s, high efficiency GFSK modulation, strong anti-interference ability, and is especially suitable for identifying control occasions. Functional descriptions of the RF905SE module pins are shown in Table 1:

Table 1. Description of rf905se module pins

Pin	Name	Pipe foot function	Explain
1	VCC	Source	$3.3-3.6$V DC
2	TX_EN	Digital input	TX_EN $= 1$
3	TRX_CE	Digital input	Enabling the chip to launch or receive
4	TRX_CE	Clock out	Enabling the chip to launch or receive
5	PWR_UP	Digital input	Chip on electricity
6	uCLK	Clock out	Enabling the chip to launch or receive
7	CD	Digital input	Carrier detect

According to the data information in Table 1, the actual RF905SE module foot introduction can finally be obtained. With the RF905SE operating voltage of 3.3V and the microcontroller PIC16F884 operating voltage of 5V, the RF905SE pins cannot be directly connected to the microcontroller pins, which need to increase flow limiting resistance that would otherwise burn the RF905SE module. It is worth noting that to simplify the SPI interface of the microcontroller PIC16F884, we connect the serial data output SDO to the MOSI of RF905SE, the serial data input SDI to the MISO of RF905SE, the serial clock SCK to the SCK of RF905SE, and the slave selection signal SS to the CSN of RF905SE. If the microcontroller does not have this SPI function module, an ordinary I/O port can be used to simulate the timing of the SPI interface through the software. When the RF905SE module receives the data, the signal CD is set high once the carrier of the frequency set in the module is detected; the valid address is received,

when the receiving address of the module matches the target address, the signal AM is set high; the valid packet is received, and decoded, the signal DR is set high; when all the valid data is read by the microcontroller, the RF905SE lowers the CD, AM, DR when read by the microcontroller [11].

To complete the communication function must support the interface circuit, the DALI communication interface part includes the transmitting and receiving circuits. The signal transmitted on the DALI bus requires its descent or descent along time between 10 and 100 us. The pressure difference of DALI bus is greater than 9.5 V indicates the high level; the pressure difference is below 6.5 V indicates the low level. The master circuit control system DALI interface circuit is associated with the microcontroller, and has 6 signals between the DALI interface: DALITXEN, DALITX, receiving enabling, receiving data, and receiving analog voltage. When the microcontroller needs to send data to the DALI bus, DALIRXEN = 0, when Q3 is off. The microcontroller gives DALITXEN = 1, DALITX = 0, when the triode Q2 is turned on and Q4 is cut off, thus turning the triode Q1 on, so that the pressure difference between DALIT and DALIR is about 20 V and the DALI bus presents a high level.The microcontroller gives DALITXEN = 0, DALITX = 1, when the triode Q2 is cut off and Q4 turns on, thus the triode Q1, so the pressure difference between DALIT and DALIR is the saturation conduction voltage of the triode Q4, about 0.5V and the DALI bus shows low level as shown in Fig. 3.11. According to the timing diagram analysis of Fig. 3.11, the data sent by the DALI bus is "0011".

To this end, a voltage comparator inside the microprocessor is used. The older negative end input produces a COM1-signal of about 0.2 V for the external resistance partial voltage. When the slave control system provides a high resistance transmission circuit, the pressure drop is generated on R13, R14 and R14, much lower than 0.2 V, so the comparator output is low level When the slave control system provides almost impedance circuit, the high pressure drop on R13 and R 14 (about 0.5 V) is above 0.2 V, so the comparator output is high level, the specific relationship structure is shown in Fig. 2 below:
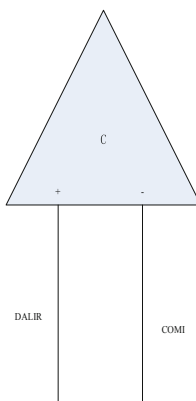
Fig. 2. Output high level diagram

According to the relationship structure in Fig. 2, the actual output is a high-level circuit relationship. When sending "0" through the circuit shown in Fig. 2, the master system microcontroller first controls the bus at high level for 417 us ± 10%, then puts the bus at low level for 417 us ± 10%; when sending 1, the microcontroller controls the DALI bus at low level for 417 us ± 10%, then puts the DALI bus at high level for 417 us ± 10%. High and low level switching is very convenient, and it is easy to design Manchester coding waveforms consistent with the DALI protocol. The master control system sends a complete DALI forward frame, including 1 starting position, 8 address bits, 8 data bits and 2 stops, where the starting position is "1" and a high stop level of 2 cycles, and finally completes the design of the main control circuit of the teaching mode recognition system.

2.2 PCB Plate Design

After completing the master circuit design of the teaching mode recognition system, the analysis and design of PCB board are required. The PCB plate layout and layered structure construction were performed. To judge the standard of the good PCB early design, the layout of this stage is very important. This is because the overall layout often directly determines the direction of the subsequent wiring and the characteristics of the system functions. The PCB board size, height of panel device and forbidden layout area are determined by the interior space of the UAV. The specified size of the plate is 18 mm and 18 mm, and the height of the plate surface device should not exceed 1 cm, so this factor should be taken into account in the selection of the device.

The general principles of PCB circuit board layout are as follows: one is that according to the layout principle of "first large, then small, first main, then auxiliary, first complex and then simple," to the larger or special electronic components, circuit modules with core functions should consider the priority layout. All line length should be as shorter as possible, key signal such as CLK to be the shortest line; analog signal and digital signal; high frequency signal from low frequency signal; high voltage, high current should be completely separated from low voltage, low current; the interval of high frequency elements should be spacious enough. The second is the design should refer to the principle of the layout, prioritithe main devices according to the main signal direction of the circuit board. Third, the same type of plug-in components should be oriented consistently; the same type of polar separation elements of the same type should also be oriented in the same direction; convenient for later welding and testing. Heat elements should be evenly distributed to facilitate the heat dissipation of the overall plate. For the same type of circuit module, try to adopt the layout standard of "symmetry"; the overall arrangement of the components should be convenient for debugging and maintenance, and leave enough room around the elements that need to be debugged. The fourth is to optimize the layout according to the standard of uniform distribution, center of gravity balance, overall coordination, and beautiful layout. According to the principle described above, the design adopts the principle of FPGA as the core, according to the main signal flow direction, SRAM, SDRAM surround, power supply, signal interface edge placement, and the overall beauty and symmetry.

During the layout, Determine the core location of the FPGA, Identity and lock the position in the PCB software tool; Then according to the specific situation of FPGA

round, Under the principle of ensuring that the same class of recognition signals is connected to the same Bank, Determine the location of SRAM and SDRAM around it; Then determine the CF card and USB interface positions with reference to the identification signal flow direction and the plate size, Considering the easy operation and overall beauty of the CF card, Place it symmetrically on the upper side of the plate; Finally, according to the recognition level value of the power supply, Size value of the power chip, power supply mode and heat dissipation to determine the layout of the power supply; Other test feet are left in an open space.Identification stratification of the entire PCB after determining the overall layout of the identification system.In this system, the FPGA has a total of 680 IO feet, encapsulated as BGA, according to the principle of the cruciferous outgoing line, at least six layers of signal outgoing line, combined with the power supply and ground of the whole board, set to a total of 12 recognition layers.

After completion, design the power supply and the identification of the ground line. The design of the high-speed electronic identification system includes all aspects, which needs to be thoughtful about all aspects, and the first consideration is how to reasonably allocate the power supply. The reasonable and effective distribution of power supply will directly affect the stability of the whole circuit system function. Setting up the power supply identification layer mode and the bus mode are two common basic methods in the process of power supply distribution: the bus mode refers to the power supply system composed of the power supply recognition transmission lines, which are determined due to the different power supply voltages of different circuit modules in the circuit system. Under the design of the bus mode, the power supply and the device will produce the impedance during the transmission process, and the function of the transmission line is equivalent to the resistance, which directly leads to the minimum output impedance of the whole power supply system. In the circuit system identification design, the circuit board is required to be layered, and the corresponding power supply or local attributes should be added to the different layers. This whole circuit board frame jointly constructed by the power supply layer and the formation is the power supply layer mode. In this way, the role of the overhole is particularly obvious, which is a bridge between the power supply corresponding to the specific device and the whole system.In the process of return, the signal can always find the shortest circuit path through the perforation, and the minimum output impedance of the corresponding circuit system will also become smaller. The direct effect is to reduce the noise of the power supply. When the whole identification system is performed, different degrees of noise will be generated, so it must be fully considered and corresponding filtering measures to reduce and reduce noise generation.

The power supply recognition filter is a noise reduction filter method often adopted in the design. It is generally realized by placing the filter circuit in the circuit system. The commonly used filtering schemes are L-type filter, T-type filter and T-type filter. Capititors from 1 uF to 10 uF are often used to filter out low frequency noise, while capacitors from 0.01 uF to 0.1 uF are placed at the power input of the active device to filter out high frequency noise. The capacitance should be placed as close as possible to the chip, and the lead is as short and as thick as possible. And for the same level to supply of different digital modules and analog modules, need to be separated through magnetic beads. The identification of the directly adjacent signal layer should not appear

in the design, all the signal layers can be interspersed with the power layer or connected to the formation for spatial isolation. When wiring, the actual wiring of the circuit board should be as thick as possible than the power line, and the power line is thicker than the signal line.Subsequently, in the difference pair walk, the difference pair line width is 8mil, and the differential signal line spacing is less than or equal to between the signal lines.

The distance between the difference and the difference pairs is not less than 50 mil 35, and try to avoid the emergence of the interlayer difference signal, using the difference pairs within the same layer. The line of the difference pair should be as short as possible, the straight line, minimize the number of holes through the line or avoid the gap, the signal line within the difference must remain the same, and the length of the two lines of the difference line signal should be as close as possible. To ensure that the characteristic impedance of the signal line is continuous everywhere along the signal line and maintains a constant, the differential impedance should be controlled at 1002 $\pm$ 5%, to ensure that the impedance matching. Avoid right angle curve, use arc or 45° oblique angle instead. Finally, the design of PCB board cloth is completed, so the design of the overall hardware of the above teaching recognition system is completed.

3   Software Design of Online Teaching Mode Recognition System for Ideological and Political Courses

3.1   Design of Teaching Recognition Function Module Under Hash Algorithm

Generally speaking, the auxiliary function of the identification system involves three links, including collection, processing and transmission, which is also the key to the identification system design. There is a great difference between the computer language and the natural language of the video. How to accurately identify the differences between the two languages is a problem that must be solved when identifying software recognition. Ideological and political feature extraction technology is the basic composition of the speech recognition system. It is mainly responsible for extracting the ideological and political features, providing an accurate language signal to the translator in time, and improving the accurate coefficient of computer translation work. The following are the related processes, as shown in Fig. 3 below:
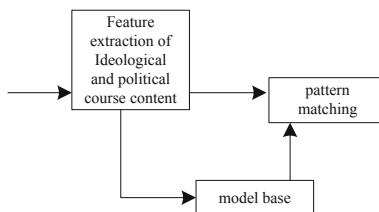
Fig. 3. System teaching identification flow chart

The corresponding software design is continued according to the process in Fig. 3. The speech recognition system should match the corresponding functional modules to

assist teachers and students to translate the language meaning of ideological and political content in a short time, so as to avoid the inconvenience caused by artificial translation and language errors. Therefore, the scope of association identification is need to be determined, as shown in formula. 2:

$$M = \kappa + 0.5a - \frac{1}{2} \tag{2}$$

In Formula 2: M represents the association recognition range, κ represents the recognition control ratio, and a represents the matching range. Through the above calculation, the actual association identification range can finally be obtained. Within this scope, the diversified mode matching technology adopts an intelligent identifier, which automatically recognizes and analyzes after the translator enters the voice, so as to reduce the difficulty of manually translated statements. For example, computer recognition software establishes a matching model, according to the matching translation mode of English words, words, sentences and other structural forms, the final language results can be obtained by executing the program command, and give students help in speech recognition. After this, training on the identification module was performed. The speech recognition system is designed to realize educational informationization. Here, based on the above, the recognized training ratio should be calculated, as shown in formula. 3:

$$L = \frac{1}{Y + 0.25R} + 2\delta \tag{3}$$

In Formula 3: L represents the training proportion of recognition, Y represents the execution time, R represents the recognition speed, and δ represents the actual recognition range. With the above calculation, the actual identified training ratio can finally be obtained. Help the teachers to solve the translation problems encountered in the ideological and political classroom teaching, and deepen the students' understanding of the ideological and political knowledge. After the voice recognition ends, the translator will automatically perform the simulation training operation to create a virtualized voice training platform for the students, which is also a more practical function of the software recognition system. The simulation training technology adopts the man-machine integrated design idea, which combines the translator, the voice identifier and so on to implement the training methods, quickly identify and judge the English voice level, and guide the students to adjust the voice mode. Ideological and political content converter is a necessary operation tool for modern teaching. Teachers and students can accurately understand the deep meanings of various words with the help of the content converter function, which puts forward more requirements for the design of speech recognition system. The author believes that the design of the speech recognition system should consider the specific workflow of the translator, and arrange a scheme that meets the translation software in advance, so as to improve the human-computer language conversion rate, combine the hash algorithm, and control the corresponding recognition conversion vector with the hash algorithm, as shown in Eqs. 4, 5 and 6:

$$G = \sqrt{2d + 1.25} - \gamma + \frac{1}{3} \tag{4}$$

$$T = \sqrt{2d + 1.25} - 2\gamma + \frac{1}{4} \tag{5}$$

$$O = \sqrt{2d + 1.25} - 3\gamma + \frac{1}{5} \qquad (6)$$

In formulas 4, 5 and 6: d represents the recognition conversion vector, and γ represents the control range. With the above calculation, the actual recognition conversion vector can finally be obtained. Set it in the control module of the system, the voice recognition system design and application situation: for the basic recognition module. The speech recognition method is mainly the mode matching method, which conducts matching processing according to different translation requirements to realize the accuracy of ideological and political multi-level vocabulary. First, in the training stage, the user says each word in the vocabulary in order, and puts its feature vector as a template into the template library. The second is in the recognition stage, the feature vector of the input voice compares the similarity with each template in the template library, and the highest similarity is output as the recognition result [12]. In addition, it is the front-end module. Front-end processing refers to the processing of the raw speech before feature extraction, which is the main role of preprocessing operations. Speech recognition system is often affected by external interference and reduces the accuracy of translation [13]. The design of the front section processing module can eliminate some noise and the influence of different speakers, so that the processed signal can more reflect the essential characteristics of speech. The model of speech recognition systems usually consists of two parts: acoustic and language models, corresponding to the calculation of speech-to-syllable probability and byte-to-word probability, respectively. The extraction and selection of acoustic features is an important link in speech recognition. This step is directly related to the overall work efficiency and has a certain influence on the recognition and learning of ideological and political teaching. Complete the design and association of the teaching recognition function module under the hash algorithm.

3.2 Multi-core Identification Database Design Under the Hash Algorithm

After completing the design of the teaching identification function module under the hash algorithm, a multi-core identification database also needs to be created. When designing the software function, the system function code is mainly completed through the use of C, casel ~ case6 indicates no sound detected, retraining, noisy environment, database full, different sound detected, serial number error, RSP_NAMEDIFF and RSP_CMDDIFF indicate two input names respectively, so, according to the actual situation, design and construction, and calculate the identification span coefficient, as shown in formula 7:

$$B = \omega - \sqrt{2m + 1} \qquad (7)$$

In formula 7: B represents the recognition span coefficient, ω represents the feature recognition control ratio, and m represents the time-frequency feature decomposition value. Through the above calculation, the actual identification span coefficient can finally be obtained. Subsequently, in the design of the above system software, the design of the multi-core identification software is performed. The establishment of the database is based on feature decomposition and association dimension feature registration. Then, in the process of teaching, we can identify the existing ideological and political knowledge

of voice, pictures and video, and the automatic detection process of multi-core identification database is to complete the noise reduction processing of ideological and political content signal through the use of time and frequency feature decomposition method. On this basis, through the comprehensive use of time-frequency analysis and extracting related information legitimate characteristics, to further optimize the automatic pronunciation error detection method of multi-core identification database, to improve the error detection ability of pronunciation identification. Combined with the hash algorithm, the feature identification matrix is created and its feature extraction coefficient is calculated as shown in Eqs. 8, 9 and 10:

$$F = \frac{t + 2y}{2} - 3\Re \tag{8}$$

$$S = \frac{t + 4y}{2} - 6\Re \tag{9}$$

$$A = \frac{t + 6y}{2} - 9\Re \tag{10}$$

In formulas 8, 9 and 10: F, S and A represent feature extraction coefficients, t represents detection difference, y represents hash ratio, and $\Re$ represents recognition range. Through the above calculation, the actual feature extraction coefficient can finally be obtained. It is set in the model of the system recognition, the input state parameters are used to represent the length of the recognition input signal, the corresponding difference represents the wavelet coefficient, the output ideological and political change course content signal is characteristic decomposed and registered, the decomposition filter group is represented by AFB, and the soft threshold function of the speech recognition system. It was determined and clearly defined. Combined with the feature extraction coefficients calculated by the hash algorithm, the pronunciation signal reorganization in the process of multi-core identification database is completed by using the wavelet multi-layer reconstruction method. The reconstruction filter group of the system is represented by SFB, and the inverse transformation values of the obtained signal filtering are obtained. After completion, after the error detection and output recombinant pronunciation signal is completed combined with the wavelet multi-layer reconstruction method, the correlation information of the signal is extracted, and the pronunciation signal is converted from the time domain to the frequency domain by the time-frequency analysis method, and the instantaneous frequency of the voice signal is adopted. Finally, the design of the multi-core identification database is completed under the hash algorithm.

4   System Test

4.1   Test Preparation

The ideological and political courses of School A were selected as the main test target. The identification rate of the system was tested, and the system was tested in quiet and noisy environments, each test instruction was tested 10 times, and against specific people in different environments, recording the number of successful identification of the

system by 5%. The sampling frequency was set to 12,500 Hz, and the signal-to-noise ratio during the pronunciation recognition process ranged from −5 dB to 20 dB, which can sample the different recognition ratio of the system by the signal beam. After completing the construction of the above test environment, check whether the corresponding test equipment is in a stable operation state, and there are no external factors affecting the final test results, check correctly, and start the test.

4.2 Test Process and Results

In the test environment built above, test. The specific test procedure is shown in Fig. 4 below:
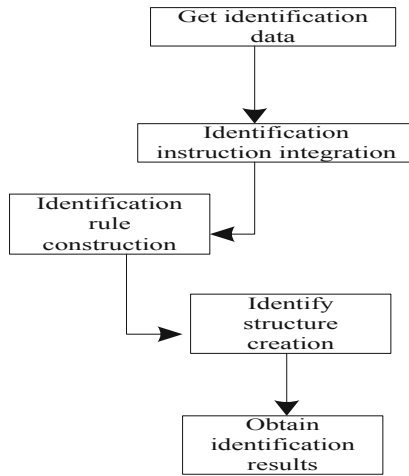
Fig. 4. Test flow chart

According to the test in Fig. 4 above, the corresponding test results can be obtained and comparative analyzed, as shown in Table 2 below:

Table 2. Analysis of test results

Test group	Literature [5] system recognition rate\%	Literature [6] system recognition rate\%	This paper designs the system recognition \%rate
Test group 1	82.15	87.12	92.15
Test group 2	76.34	85.34	90.55
Test group 3	79.15	88.13	94.31
Test group 4	75.12	86.54	93.51
Test group 5	60.25	78.31	90.57

It can be seen from Table 2 that, compared with the systems of literature [5] and literature [6], the recognition rate of the system designed in this paper is higher, up to 94.31%, while the highest recognition rate of the other two methods is only 82.15% and 88.13%, which shows that the recognition effect of the design method is better, with strong stability, high recognition accuracy and high practical application value.

5   Conclusion

In order to accurately identify the online teaching pattern recognition effect of Ideological and political course and improve the quality of Ideological and political teaching, this paper proposes an online teaching pattern recognition system of Ideological and political course based on hash algorithm. In the hardware part of the system, the main control circuit and PCB board of the teaching pattern recognition system are designed; In the software part, the teaching recognition function module adopts hash algorithm, designs the multi-core recognition database under hash algorithm, and then completes the design process of online teaching pattern recognition system of Ideological and political course. The final test results show that the recognition rate of the design system is high, all above 90%, indicating that its recognition effect is good, has strong stability and certain practical significance, in order to provide some help to improve the quality and efficiency of Ideological and political teaching.

Fund Project. 1. Demonstration Project of Grass-roots Teaching Organization (Teaching and Research Office): Teaching and Research Office of Basic Principles of Marxism; No.: (2020JCJS01)

2. University-level undergraduate engineering project: "Introduction to Basic Principles of Marxism" first-class course of social practice; No.: (2020SHSJ03)

References

1. Nie, A : Design of English interactive teaching system based on association rules algorithm. Secur. Commun. Netw. 2021(s1), 1–10 (2021)
2. Liu, J., Yang, B : Design of evaluation system of physical education based on machine learning algorithm and SVM. J. Intell. Fuzzy Syst. (1–3), 1–12 (2020)
3. Yao, F : Design and simulation of integrated education information teaching system based on fuzzy logic. J. Intell. Fuzzy Syst. **37**(4), 1–9 (2019)
4. Yang, J.: Teaching optimization of interior design based on three-dimensional computer-aided simulation. Comput.-Aid. Des. Appl. **18**(S4), 72–83 (2021)
5. Gao, Y., Li, Q.: Pattern recognition algorithm based on container theory. Autom. Technol. Appl. **38**(06), 117–121 (2019)
6. Wang, H., Wang, Q.: Optimization simulation and application of shape pattern recognition based on Legendre polynomial. J. Mater. Metall. **20**(04), 275–281 (2021)
7. Huang, W., Xiao, X., Xu, M : Design and implementation of domain-specific cognitive system based on question similarity algorithm. Cogn. Syst. Res. **57**, 20–24 (2019)
8. Suharyadi, J., Kusnadi, A : Design and development of job recommendation system based on two dominants on psychotest results using KNN algorithm. Int. J. New Media Technol. **5**(2), 116–120 (2019)

9. Wagner, S.P.: Impact of computer-based peer review on college students' performance and perceived self-efficacy in an online graphic design course. J. Appl. Commun. **103**(4), 8–8 (2019)
10. Yang, Y., Zhang, Q., Liu, B., et al.: Case design of linear algebra hybrid teaching model under problem-based learning. Int. J. Inf. Educ. Technol. **9**(9), 618–622 (2019)
11. Hu, C., Chen, C., Zhou, T., et al.: The design and research of a new pharmaceuticals-vending machine based on online medical service. J. Ambient Intell. Hum. Comput. (3), 1–10 (2020)
12. Liu, S., Bai, W., Srivastava, G., Machado, J.A.T.: Property of self-similarity between baseband and modulated signals. Mob. Netw. Appl. **25**(4), 1537–1547 (2019). https://doi.org/10.1007/s11036-019-01358-9
13. Yang, J., Yang, L., Sheng, W.: Fault pattern recognition simulation of distributed optical fiber disturbance sensing system. Comput. Control. **37**(1), 444–447 (2020)

Design of Cross Language Education Resource Sharing Platform Based on Hadoop Framework

Xiaonan Yu[✉] and Li Jiang

Nanchang Institute of Technology, Nanchang 330108, China
`yx202149@163.com`

Abstract. The rapid development of information technology has promoted the modernization of cross language educational resource sharing. Educational resource sharing platform is the platform basis for carrying out distance education activities, which is of great significance to the development of modern distance education. Firstly, the hardware design of cross language education resource sharing platform is carried out, and the MCU controller is designed to play a powerful control function combined with FPGA system. The design is based on resource reconfiguration port and external input port to realize hardware connection. Then, the software design of cross language education resource sharing platform is carried out, the resource sharing algorithm based on Hadoop framework is optimized, and the multi line sharing ability of resource sharing algorithm is enhanced. The data model is designed to realize the design of shared platform. Finally, the platform test verifies that the response time of the platform is less than 4000 ms.

Keywords: Hadoop framework · Cross language · Educational resources · Resource sharing

1 Introduction

The ten year development plan for educational informatization (2011–2020) of the Ministry of education puts forward the guiding ideology of "narrowing the digital divide in basic education and promoting the sharing of high-quality educational resources", and puts forward the action plan of "action for the construction and sharing of high-quality digital educational resources", including the construction of a national public service platform for digital educational resources, the construction of all kinds of high-quality digital educational resources at all levels Establish a mechanism for co construction and sharing of digital education resources [1, 2]. However, in the informatization process of basic education, on the one hand, there is a serious shortage of high-quality educational resources in rural schools, and the contradiction between people's demand for educational resources and the insufficient supply of educational resources is prominent; On the other hand, in the existing resource sharing construction, the software and hardware resources are invested repeatedly, and the resource sharing is poor". The emergence of cloud computing technology provides a feasible solution to the problem of digital education resource sharing. Relying on thousands of cloud servers on the cloud education

platform, it has extremely powerful computing functions, massive network resources, repeated construction, platform operation and maintenance, security and other problems existing in the existing resource sharing construction will be solved, It can save the cost of Digital Education Resource Sharing Construction [3, 4], promote the integration of urban and rural digital education resources and strengthen resource sharing.

Digital educational resources refer to multimedia teaching materials that can be operated under the multimedia computer and network environment after digital processing, such as media materials, question bank and test paper materials, teaching courseware, cases, literature, data catalog index, network courses and other forms of resources. According to the expression of information, digital education resources can be divided into digital slide, digital projection, digital audio, digital video, digital online teaching resources, etc. [5]. Through the investigation of the current situation of digital education resource sharing of basic education in Liaoning Province, it is found that there are the following problems: the distribution of digital education resources in urban and rural areas is unbalanced. Due to the unbalanced economic development in urban and rural areas, urban schools are often better than rural schools in the construction of digital education resources. Due to the lack of communication, schools in the province are often prone to repeated construction of resources in the construction of digital education resources. Due to the limited funds of schools, the construction of resources is often limited to the construction of commonly used teaching resources.

Because there is no unified organization, sharing plan and sharing goal, nor unified interface and effective digital education resource sharing platform, it is impossible to realize the efficient and reasonable sharing and use of resources. In order to better realize the sharing of educational resources, it is necessary to establish a sharing platform with stable performance, unified standards and perfect functions, use standard interfaces to unify the resources of schools and departments, and integrate heterogeneous and dynamic resources [6, 7], so as to establish an efficient and high-quality educational resource sharing mechanism. The digital education resource sharing platform based on Hadoop framework has the characteristics of interconnection, collaboration and sharing and simple operation. It has strong advantages in building a digital learning environment, promoting the popularization and sharing of high-quality education resources and promoting the diversified development of learners. The design of Hadoop platform is based on the principles of compatibility and sharing and openness. The main contents of the design include the overall structure design of the system and the structure design based on the system. Through literature research and teacher-student interviews, determine the demand analysis of cloud teaching and autonomous learning platform.

In view of this characteristic of cloud computing, combined with the current situation of educational resource construction in schools in Liaoning Province and the needs of Educational Resource Sharing Construction among schools [8], from the perspective of regional coordinators, a cross language educational resource sharing platform is constructed based on Hadoop framework, and the hardware part is optimized from three aspects: MCU controller, resource reconstruction port and external input port, The software part constructs the resource sharing algorithm based on gae cloud computing, designs the data model, and finally completes the performance analysis of the design platform through the platform test.

2 Hardware Design of Cross Language Education Resource Sharing Platform

The hardware part of educational resource sharing platform is a complex part. Different functions can be realized by using different functional modules, which requires very rich functional modules. At the same time, this also puts forward two requirements for the control platform: first, there are enough ports; Second, it has rich port types. At present, although the embedded processor has a high degree of integration and rich number of pins, the demand for ports on the education platform is changing. The fixed interface function of the hardware system leads to the limited open resources of the system to the outside, and the use method of the port is also single. In view of this problem and combined with the characteristics of the hardware system of the education platform, this paper puts forward the overall functional requirements of the hardware system of the education resource sharing platform of this subject, as follows:

(1) It has good interactive function and is convenient for students to share educational resources.
(2) Support port resource reconfiguration, and the corresponding relationship between the external output port of the controller and the I/O of the core processor can be reconfigured on site [9, 10], so that the same port can be compatible with the bus interface types of multiple modules.
(3) It has a variety of special external port types, such as SPI, IIC, ADC and motor/steering gear port.
(4) It supports the functions of graphical programming, port configuration and online download of upper computer software.

2.1 Design MCU Controller

The hardware control platform is the bottom and foundation of the whole hardware module. As shown in Figure 1, the whole education control platform is designed based on "MCU + FPGA" as the core control system, including MCU controller, FPGA controller, human-computer interaction part, communication interface part and external input and output port. The hardware platform shall meet the requirements of high modularity, strong reconfigurability and open hardware interface.

The core component of hardware control platform is MCU controller. Combined with the characteristics and demand analysis of educational resource sharing platform, the main factors considered in the selection of MCU controller are: processor running speed, integration, hardware resources, cost and power consumption. The running speed of the processor can improve the response of the system, increase the support of external devices, and quickly complete complex instructions. The consideration of low cost can improve the competitiveness of products, and low power consumption can prolong the service time of batteries.

In this paper, "MCU + FPGA" is used as the core control mode. Generally, the communication between MCU and FPGA can interact with instructions according to the predefined communication protocol. For the port resource reconstruction function of the education resource sharing platform, the MCU can send control instructions to
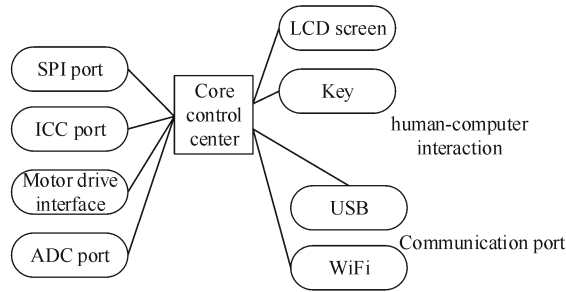
Fig. 1. Bottom layer and foundation of hardware module

the FPGA, and then the FPGA can generate serial port, interrupt, PWM wave and other functions to the external port according to the instructions. However, the implementation of this control method is complex and does not highlight the advantages of "MCU + FPGA" system. MCU itself is embedded with pin resources such as timer, serial port and external interrupt, while FPGA has more abundant wiring resources. The whole chip of FPAG is composed of an array of programmable logic units separated by wiring resources and surrounded by programmable uo units. The logic units arranged in the array are connected through the programmable internal wiring in the wiring channel to realize the logic function, The programmable switch can control the segmented metal wires to connect in any way to form the signal lines required between logic units. In the controller design of this paper, AVR single chip microcomputer ATMega128A is selected as the main control chip of the system. ATMega128A single chip microcomputer is an 8-bit single chip microcomputer with low power consumption and high performance produced by ATMEL company. It is widely used in many fields such as industrial real-time control, communication equipment, instruments and meters. The single cycle instruction execution time and advanced instruction set structure [11, 12] are adopted. Compared with other single chip microcomputer models in this series, it has better performance and higher data transmission rate. It is fully applicable to the design of the hardware control platform of this educational resource sharing platform. Its main characteristics are as follows:

(1) 8-bit microprocessor with high performance and low power consumption. It has a wide working voltage range and strong anti-interference ability of the power supply.
(2) RISC structure design ensures that many instructions are executed in a single clock cycle. The system is under the full static working condition of 16 MHz crystal oscillator, and the processing capacity can reach 16 MIPS.
(3) 8-channel 10 bit successive approximation ADC is integrated internally, and the sampling rate can reach 15 ksps under the highest precision.
(4) Rich IO resources meet system requirements. In addition, the pin resources also include second functions such as timer, counter, serial interface and so on. The I/O interface has a large driving current, and the maximum current can be as high as 40 mA, which can directly drive light-emitting diodes and small relays.
(5) It has 128 K flash storage space and 4 KB ram to meet the basic requirements of the system without additional extended program memory and data storage space.

"MCU + FPGA" is the core control mode. The MCU part was mentioned earlier. Now FPGA and MCU are combined. As a programmable logic device, FPGA provides four kinds of programmable resources: central programmable function unit of the chip, rich programmable IO pin resources, programmable wiring resources and on-chip memory RAM. The rich programmable pin resources and programmable wiring resources meet the functional design requirements of the hardware system of the educational resource sharing platform in this paper. Although CPLD and FPGA have a lot in common, considering the specific application requirements of the system, FPGA has more advantages:

(1)   FPGA is more suitable for timing logic and trigger rich structure [13].
(2)   FPGA is programmed by changing the wiring of internal wiring. Programming under the logic gate has greater advantages in programming flexibility.
(3)   FPGA integration is better and CPLD integration is relatively low.
(4)   Generally, the power consumption of FPGA is lower than that of CPLD.

In conclusion, considering that FPGA is used as the controller of resource reconstruction in this part, the chip model is cyclone series ep1c3t100c8. As a medium-sized product of atlera company, this chip is a low-cost FPGA chip, and rich I/O resources fully meet the system requirements. The system circuit mainly includes crystal oscillator clock, serial configurator circuit and download and debugging interface.

2.2   Design Resource-Based Reconfiguration Port

The main idea of port resource reconfiguration is to connect the hardware resource pin of ATMega128A with the external port through internal rewiring of FPGA, so as to realize the conversion of port resources. FPGA has abundant programmable wiring resources and pin resources to meet the system requirements. In terms of hardware design, there are three types of connecting pins with FPGA: connecting pins between ATMega128A and FPGA, connecting pins between internal modules of the system and FPGA, and connecting pins between reconfigurable ports and FPGA. The pin resources connected between FPGA and ATMega128A include 8 ordinary pins (PA2-PA3, PA6-PA7, PC2-PC3, PC6-PC7), two serial ports (RXDO, TXDO, RXD1, TXD1) and four PWM (OC3A, OC3B, OC1A, OC1B), one external interrupt (PDO (SCL)) The design is divided into default mode and reconfiguration mode according to the use mode of system resources. In the default mode, common pin resources are allocated to the reconfigurable port, and other resources are used by the default internal module; in the reconfiguration mode, serial port, interrupt, PWM and other resources are allocated to the reconfigurable port, and the corresponding internal module using the resource will stop working temporarily. According to this principle For the convenience of users, the system defines the port reconfiguration specification to indicate the resource allocation under different modes, as shown in Table 1:

It is pointed out here that the serial port resources connected to FPGA are ATMega128A, which is connected to FPGA after multiple selection switches. The specific details will be explained in the next section. The INH input pin of CD4052 chip is set low to enable multiple chip selection. The two input control pins a and B are

Table 1. Resource allocation of FPGA in each mode

Serial number	Hardware resource	Default mode		Reconstruction mode
1	common pin	Reconfigurable	Interface	–
2	Uarso serial interface	COM1/COM2/COM3/COM4		Reconfigurable ports COM1/com2
3	Uarts1 serial interface	Wifi module		Reconfigurable port COM31/COM4
4	Pwm0 timer output	–		Reconfigurable port COM1
5	Pwm1 timer output	Motor 1 channel a	Control input	Reconfigurable ports com2
6	Pwm2 timer output	Motor 1 channel Ba	Control input	Reconfigurable ports COM3
7	Pwm3 timer output	Motor 2 channel a	Control input	Reconfigurable ports COM4
8	Intro external interrupt/SCL	Motor 2 channel Ba	Control input	Reconfigurable ports COM1/COM3

controlled by one control signal line. Where P_ CTRL1, P_ CTRL2, P_ CTRL3 is connected and controlled by PE2, PE6 and PE7 pins of ATMega128A respectively. The A and B pins are set high through the pull-up resistance. Y3 channel is gated by default when the system is powered on to ensure the use of the basic functions of the system. Students can change the gating of different paths through programming, so as to realize different functions. As shown in Table 2, the application of each resource in different path selection.

Table 2. Resource allocation of multi-channel selector switch

Serial number	Hardware resource	First path	Second path
1	Uars0 Serial I/O interface	USB module	FPGA resource reconfiguration
2	Uars1 Serial I/O interface	LCD display module	FPGA resource reconfiguration
3	SPI serial I/O interface	Epcs 16 Serial Configurator	SPI dedicated port

In order to realize fast and simple assembly and flexible combination of various modules, the external ports must be standardized and diversified. According to this requirement, the external input and output ports of the system hardware design uniformly use the six wire universal interface of RJ11. The system has a total of 10 external ports. According to the type of ports, they are divided into two categories: dedicated ports and reconfigurable ports.

2.3   Design of External Input Port

The specific connection of each external port is as follows:

(1) SPI port: this port is connected to the multiplexer selection chip. When programming, ATMega128A controls the path selection and connects the port to the SPI function pin of PBO-PB3.
(2) IIC port: pin 5 of the port is connected with FPGA, and the other pins are connected with PD4, PD5 and PD1 (SDA) of ATMega128A respectively. In default mode, pin 5 corresponds to PDO (SCL) pin, which is used as IIC port.
(3) Motor/steering gear ports M1 and M2: two motor drive signals in each port are connected with the output pin of the motor drive chip to control the speed and forward and reverse rotation of the DC motor. PWM2_ 5 , PWM5_ S is directly connected to the timer output pins of pe5 and pb7 of ATMega128A to control the angle of the steering gear.
(4) AD acquisition ports ADC1 and ADC2: connected to pfo-pf7 pin of ATMega128A, providing 8 ad signal acquisition channels for sensor analog signal acquisition. USB is used to standardize the connection and communication between external devices and computers. As a standard external high-speed bus interface, USB is suitable for a variety of devices, such as MP3 players, cameras, high-speed data acquisition devices, etc. Compared with other communication interfaces, the biggest feature of USB interface is easy to use. The system can automatically configure all functions without user participation, and supports hot plug. However, ATMega128A hardware itself does not support USB communication function and requires peripheral circuit conversion. The circuit connection is shown in Fig. 3–17, where a_ Txdo and a_ RXD0 is the Y3 channel pin of the multi-channel selector switch. After path selection, it corresponds to the serial port pins PE3 (TXD0) and PE2 (RXDO).

This platform uses FT232RL as the USB interface conversion chip, which can realize the conversion between Serial uart interface and USB without specific USB firmware programming, and supports the data transmission format of serial interface 7 or 8-bit data and 1 or 2-bit stop bit. FT232RL chip integrates clock circuit, EEPROM, resistance and avcc filtering, which reduces the number of external components, 256 byte receiving buffer and 128 bytes Byte sending buffer can achieve high data throughput, which is very important in daily system data exchange or program download. Considering the power consumption and size of the control system, esp8266 WiFi module designed by Shenzhen Anxin Co., Ltd. is selected as the wireless communication module in the system. The control mode of serial communication reduces the occupation of pin resources, It

supports at instruction set and has strong advantages for subsequent secondary development. Esp8266 is characterized by high on-chip integration, so it can be used with only a few external circuits. As shown in Figure 2, the module utxd and urxd are connected with FPGA controller. In the default mode, FPGA connects the serial port of ATMega128A O The pins are assigned to the module. Esp8266 supports three working modes: AP, station and AP + station. The internal firmware integrates at instruction set and has many rich instructions, which is convenient for the secondary development of functions in the later module function design, so as to realize the wireless communication function between the controller and the computer or mobile device.

3   Software Design of Cross Language Education Resource Sharing Platform

Based on the hardware design, in order to further optimize the performance of the cross language education resource sharing platform in the process of resource sharing, optimize the Hadoop framework resource sharing algorithm, and build a data model to improve the efficiency of cross language education resource sharing.

3.1   Optimize Resource Sharing Algorithm Based on Hadoop Framework

Hadoop is an open source, distributed system infrastructure platform. It is an open source distributed computing and storage project implemented in Java. Hadoop is actually composed of multiple parts. Hadoop MapReduce is the core part, as shown in Figure 2:
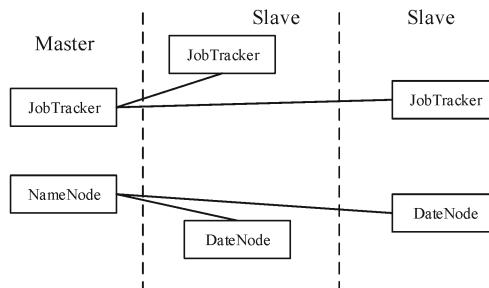
Fig. 2.   Basic structure of Hadoop framework

Compared with traditional distributed systems, Hadoop system has the following advantages: Hadoop can run on large-scale clusters composed of Job Tracker and commercial computers, or on cloud computing servers such as Amazon's elastic computing cloud (EC2); Hadoop enables users to write parallel code quickly and efficiently through the subordinate structure(Slave); The data and backup mechanism in HDFS distributed file system and the task monitoring mechanism in MapReduce ensure the reliability of distributed processing; The scale of Hadoop can be linearly expanded by increasing the number of nodes in the cluster to solve the problem of big data processing [14].

Symbols $P(n)$, $H^2_{k,n}$ and $f(C_n)$ represent the transmission power, channel gain and power required to receive data when transmitting data to user k on subcarrier n, respectively. And the mathematical relationship between them can be expressed as:

$$P_k(n)\frac{f(C_n)}{H^2_{k,n}} \tag{1}$$

It shows the relationship between transmission power and reception power required to share educational resource data with users on a subcarrier. Each subcarrier simultaneously transmits the same data for multiple users in the multicast group, so the transmission power $P(n)$ required on a subcarrier during multicast is related to the channel gain of all users served by the subcarrier. The exact mathematical relationship is as follows:

$$P(n)\frac{f(C_n)}{H^2_n} \tag{2}$$

In formula (2):

$$H_n = \min\{H_{k,n}|pk(n) = 1\} \tag{3}$$

In formula (3), xx is an indicator, which indicates whether subcarrier n is allocated to user k. It reveals that the transmission power required on the subcarrier during multicast depends on the worst channel gain of the users served by the subcarrier. And the detailed definition of indicator $pk(n)$ is as follows:

$$pk(n) = \begin{cases} 1, \ n \to k \\ 0 \end{cases} \tag{4}$$

The optimization objective of the model in this chapter is to minimize the total transmission power required by the system through reasonable and effective bit allocation and subcarrier allocation under the condition of limiting the total rate of the system, and mark the minimum transmission power required by the system as Pmin. Then the optimization model can be mathematically expressed as follows:

$$P_{\min} = \min \sum_{n=1}^{N} P(n) = \min \sum_{n=1}^{N} \frac{f(C_n)}{H^2_n} \tag{5}$$

$$P(n) \geq \frac{f(C_n)pk(n)}{H^2_{k,n}} \tag{6}$$

$$\sum_{k=1}^{k}\sum_{n=1}^{N} C_npk(n) \geq 0 \tag{7}$$

$$pk(n) \geq 0 \tag{8}$$

$$P(n) \in \{0, 1\}, C_n \in B \tag{9}$$

The above formula (5) is the optimization goal of the algorithm to minimize the total transmission power required by the system. The transmission power on the subcarrier must meet the channel gain requirements of all served users; The total rate available to all users of the platform must be greater than the minimum rate requirement of the system; The transmission power on the subcarrier must be non negative: constraint 3 symbol 1 must be taken from positive integers 0 and $pk(n)$, and the number of bits allocated on the subcarrier must be taken from the number of bits set B that the system can provide.

3.2 Design Data Model

Relational database is the main tool in the design and implementation of application system. In the process of traditional database design, the author usually establishes the data model with data table structure. The relational database realized in the design process is the main goal and task of the system.

This paper is an application developed based on Google App Engine development environment. Gae does not support traditional data storage, that is, relational database storage. However, due to habitual considerations, we still use data tables to describe the data model information, and then transplant them into the data storage model recognized by gae through the data storage methods mentioned later. The main data tables involved in the system are shown in Table 3−Table 5.

Table 3. User basic information user data table

Field name	Data type	Remarks
userID	Long	Tag user ID
UserName	String	User name
Password	String	User password
Email	String	Email (Google account)
roleID	Int	User role

Table 4. Role data table

Field name	Data type	Remarks
roleID	Long	ID of the tag role entity
rolename	String	Including students, teachers and adminstrators
roleType	Int	1 (student), 2(teacher), 3(administrator)

After logging in, students can browse all groups under interest groups. On the interest group list page, students can view the name, founder account, founder name, creation time, topic, access times and other information of all interest groups. At the same time, they can filter and query the interest group list by conditions, and the filtering results

Table 5. Data sheet of education resource information

Field name	Data type	Remarks
fileID	Long	ID that marks the uniqueness of the resource
filename	String	Source title
filecontent	Long	Source content
Username	String	Asset publisher user name
useremail	Email	Resource publisher mailbox
upldadDte	Date	Resource upload time
description	String	Resource description
courseID	Long	Course to which the resource belongs
courseID	String	ID marking course uniqueness
courseInfo	String	Course introduction information
postID	String	ID that marks the uniqueness of the post
posttitle	String	Post on title
postcontent	String	Post content

Table 6. Platform operation hardware environment

Equipment	Model	CPU	Memory	Network bndwidth
Controller	HP ProDesk 498G2 MT	Intel(R)Core™i7-4790CPU@3.60GHz	64 GB	100Mbps
Calculation node	HP ProDesk 498G2 MT	Intel(R)Core™i7-4790CPU@3.60GHz	64 GB	100Mbps
Client	MT HP ProDesk 498 G2 MT	HP ProDesk 498G2 MT	64 GB	100Mbps

are still returned in the same form. Students can browse the group or apply to join the group according to their personal interests. They can become members of the group

only after being approved by the group leader. Students can create their own groups, view my group and group information related to me, and delete the groups they create. Share resources in the same group. Group members can upload resources for use by group members and leave messages in the group to communicate with others. Everyone can also delete their own uploaded files, and the team leader can delete all files in the group. Students can invite others to join or leave the group. The teacher's operation of the interest group is consistent with that of the students. The administrator has the function of managing the group, mainly including the addition, deletion, modification and query of group resources and messages, and the deletion of the group.

4 Platform Test

The design of cross language education resource sharing platform based on Hadoop framework studied in this paper integrates and virtualizes the existing hardware infrastructure and constructs a unified resource pool with the help of cloud computing technology. In order to test the practicability of the cross language education resource sharing platform, the public library resource integration and sharing methods pointed out in reference [14] are compared as the original platform, and a comparative test experiment is designed.

4.1 Test Software and Hardware Environment

The educational resource sharing platform built in this paper is based on the cloud platform built by openstack. Due to the limitations of the laboratory environment, the cloud platform in this paper is composed of a control node, a computing node and a client. One PC is used as the openstack control node, one PC is used as the control node, and the other computer is provided as the client. The control node is deployed on a PC. The hardware environment configuration list of this platform is shown in Figure 6:

The software environment for system development and operation in this paper is as follows: (1) The control node takes Ubuntu 14.04.3 server as the underlying operating system, the system version running on the client is windows 10 Ultimate, and the openstack project version is kilo. (2) The running environment of distance education platform, the application layer of this paper, is apache-tomcat-7.0.57 for the server, Oracle11g for the database, and JDK version 1.7.

4.2 Test Results

The testing tool used in this paper is Apache jmeterl42, which is a Java based tool designed and developed by Apache for stress testing. It can simulate the concurrent access of multiple users to the system and record the response time. Although the software can not simulate the real user access, it can control the user request to be sent evenly within a certain time. This chapter tests the response of users accessing the system, sets all requests to be sent within 1 second, sets the number of threads (i.e. the number of simulated users) to 50 to 500, increases 50 each time, and simulates 10 times at the most
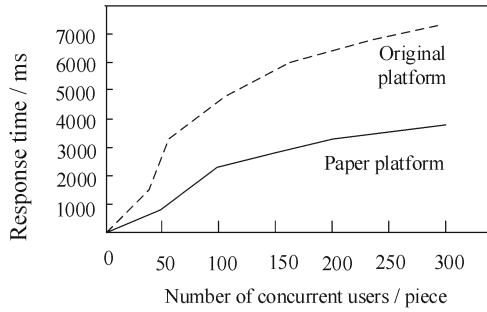
Fig. 3. Comparison of response time

ten sampling points, records the response time comparison each time, and draws the comparison diagram according to the test results, as shown in Fig. 3:

As can be seen from Fig. 3, cloud computing technology adopts virtualization, its own scheduling strategy and load balancing method. Compared with traditional methods, the response time is significantly reduced when the number of concurrent users is large, and has high response speed and service quality.

Use this platform and the original platform for standard resource sharing, then count the completion time of resource sharing, average waiting delay and resource download time, and finally make a comparative analysis. The statistical results are shown in Table 7.

Table 7. Evaluation results of sharing ability

Method	Resourec sharing	Average waiting delay	Resourec download time
	Completion time/s	/s	/s
Paper platform	5.62	1.36	3.66
Original platform	7.23	2.42	5.63

As can be seen from Table 7, when using the cross language education resource sharing platform based on Hadoop framework to share standard resources, the sharing completion time, average waiting delay and resource download time are shorter than those of the original platform, indicating that the sharing ability of the constructed platform is stronger.

5   Conclusion

Based on the research and summary of educational resource sharing platforms at home and abroad and the combined application of cloud computing and distance education, this paper constructs the cloud platform with the help of openstack open source project, and designs and implements each module in the platform. The main work of this paper is as follows:

(1) This paper analyzes the current situation of the existing platform, summarizes the problems of slow response time and long waiting delay of the platform, and designs a cross lingual education resource sharing platform based on cloud computing and Hadoop framework, combined with the advantages of cloud computing technology that can quickly integrate, share and transmit massive multi-source data.

(2) Study the relevant technologies needed to build a platform based on Hadoop framework.

(3) Conduct a detailed demand analysis of the platform and determine the design objectives. Combined with the existing application cases of cloud computing in distance education, this paper designs an implementation scheme of distance education platform based on cloud computing, and analyzes and designs the overall model, hierarchy and components of the platform.

References

1. Schwartz, I.S., Woc-Colburn, L., Mccarty, T.P., Cutrell, J.B., Cortes-Penfield, N.W.: 135. impact of #idjclub, a synchronous twitter journal club, as a novel infectious disease education platform. In: Open Forum Infectious Diseases, vol. 7, no. (Supplement_1), pp. S197–S198 (2020)

2. Zhang, Q., Wang, K., Zhou, S.: Application and practice of vr virtual education platform in improving the quality and ability of college students. IEEE Access (99), 1–1 (2020)

3. Yuan, Y.: To build a solid education platform and give full play to the educational role of student apartment – a case study of sichuan university of arts and science. Int. J. Soc. Sci. Educ. Res. **3**(5), 40–44 (2020)

4. Beecham, N., Unger, C.: Designing the female orgasm: situating the sexual entrepreneur in the online sex-education platform omgyes. Des. Issues **35**(4), 42–51 (2019)

5. Lee, J.Y., Patel, S.J.: Innovating business model in the higher education industry: a platform-based approach in university career services. Ind. High. Educ. **34**(2), 91–99 (2020)

6. Wen, J., Wei, X., He, T., Zhang, S.: Regression analysis on the influencing factors of the acceptance of online education platform among college students. Ingénierie des Syst. D Inf. **25**(5), 595–600 (2020)

7. Yang, Y., Cho, J.H.: Clinical effects of remote glucose monitoring and patient-centered education platform for patients with diabetes. J. Korean Diab. **21**(4), 204–210 (2020)

8. Chen, Z., Xu, M., Hu, Z., Zhang, S., Jumani, A.K.: Multimedia educational system and its improvement using AI model for a higher education platform. J. Multiple-valued Logic Soft Comput. **36**, 25–41 (2021)

9. 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

10. Son, W.: Development of SW education class plan using artificial intelligence education platform focusing on upper grade of elementary school. J. Korean Assoc. Inf. Educ. **24**(5), 453–462 (2020)

11. Zhang, Y., Qin, G., Cheng, L., Marimuthu, K., Kumar, D.S.: Interactive smart educational system using AI for students in the higher education platform. J. Multiple-valued Logic Soft Comput. **36**(1–3), 83–98 (2021)

12. 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

13. Liu, S., Bai, W., Srivastava, G., Machado, J.: Property of self-similarity between baseband and modulated signals. Mob. Netw. Appl. **25**(4), 1537–1547 (2020)
14. Zhang, L.L.: Simulation research on the integration and sharing of public library resources under cloud computing. Comput. Simul. **37**(5), 416–419 (2020)

Design of Interactive Language Education Assistant System Based on Data Classification

Li Jiang[(✉)] and Xiaonan Yu

Nanchang Institute of Technology, Nanchang 330108, China
`jl16023@163.com`

Abstract. Aiming at the poor effect of language education, this paper puts forward the design method of language interactive education assistant system based on data classification, optimizes and improves the hardware structure of the system to ensure the operation quality of the system, further combines the principle of data classification, optimizes the function of the system software and simplifies the operation process of the system, Finally, experiments show that the language interactive education assistant system based on data classification can effectively break the limitations of space and time and improve the quality of language teaching.

Keywords: Data classification · Language interaction · Educational assistance

1 Introduction

With the development of education, computer technology, mobile Internet, and communication technology, computer teaching aid systems based on language interactive education aid platforms have been increasingly used in teaching, and have played an important role in the effective development of teaching activities [1]. With the rapid development of information technology and the continuous innovation of educational information construction, people have higher and higher requirements for teaching auxiliary systems. User needs are gradually developing in the direction of more functions, easier use, and faster efficiency. The role of network-based auxiliary teaching in modern education is becoming more and more obvious. The small remote system for the interactive education auxiliary system makes the communication between teachers and students, and between students and students more convenient and smooth, which is conducive to improving the learning effect [2]. The online auxiliary teaching system allows students to fully understand the teacher before class, preview the content of the course, easily obtain auxiliary materials after class, discuss and answer questions online, and also facilitate teachers to understand the student's dynamics in a timely manner, communicate and adjust the teaching method in a timely manner, So as to improve the quality of teaching and achieve better teaching effect.

Therefore, relevant researchers have done a lot of research on online assisted instruction system and achieved some results. Cao et al. Designed a method based on Net platform. The hardware structure of the system consists of user interface layer, service

selection layer and data management layer Teachers, students and other users enter their own identity information in the user interface layer, log in to the system, and click the corresponding program according to their own application needs. The business selection layer transmits the user selection instructions to the data management layer, and the data management layer selects the corresponding resources according to the user needs and feeds back to the user the interaction of the system is mainly reflected in interactive teaching and information interaction. Interactive teaching is reflected in online teaching between teachers and students; Information interaction is embodied in the information transmission of system information interaction model After testing, the designed system has strong pressure resistance, can respond to the application instructions of a large number of users in real time, and the interactive teaching effect is good, which improves the students' sense of self-efficacy. However, the operation of the system consumes a lot of energy and needs to be further improved. Hu proposed to design an online translation assistant system based on fusion of multilingual interaction. In the design of the system, the goal is to realize online learning. In order for learners to learn in the process of continuous correction, an end-to-end platform is designed. The platform integrates the machine translation server into the user interface most commonly used by professional translators, so that the machine can continuously learn from people's choices and adjust the model according to specific fields or user styles, Save the later editing work of learners and improve the auxiliary effect in the process of online translation However, the application of software in the system design is less, and there are some limitations.

Therefore, this paper designs an interactive language education assistant system based on data classification. The main design route of the system is:

(1) Based on the in-depth investigation of the background and current situation of the teaching assistant system, the hardware structure of the system is optimized.
(2) Combined with the principle of data classification, the structure of the course teaching assistant system is optimized. According to the teaching data of teachers and students' daily learning needs, analyze and design the teaching data management system which is more conducive to teachers' daily teaching work, and optimize the functional structure of the system.
(3) On this basis, the design system is divided into two parts according to the user type: student user data classification teaching information management, teacher data classification teaching information management and administrator user data classification teaching information management center to complete the system design.

2   Language Interactive Education Assistance System

2.1   Hardware Structure of Language Interactive Education Assistance System

Through the in-depth investigation of the background and current situation of the teaching assistant system, this paper summarizes some common basic needs of users, and analyzes the main user roles of the system. For example, in teaching practice, teachers and students play a major role, because language interactive educational assistance is a process of teaching and learning [3]. If a language interactive education assistant system

wants to assist teaching, it must provide certain data, tools, materials, etc. these resources are inseparable from the careful management of the administrator. At the same time, the administrator should also provide guarantee for the safety and normal order of the whole system. Therefore, the role of the administrator is also very important [4]. In order not to block the whole language interactive education auxiliary system, it is also necessary to consider non school personnel, that is, tourists. Their existence can share the excellent resources in the school with other scholars, continuously publicize the campus culture, and receive good feedback and suggestions from more talents and sages. Based on this, the hardware structure of the system is optimized as follows:
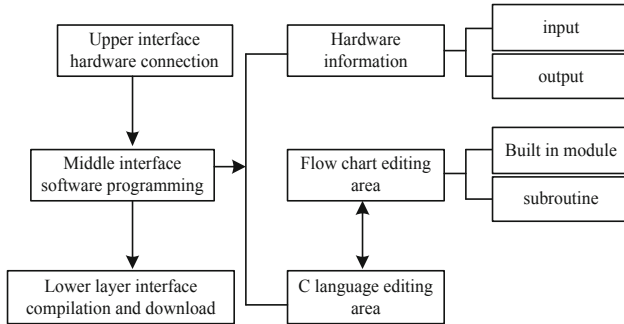
Fig. 1. System hardware configuration

In the system hardware configuration in Fig. 1, in order to realize the effectiveness of the auxiliary system, the hardware connection of the upper interface, the software programming of the intermediate interface and the compiling and downloading module of the lower interface are designed in this architecture, and the data output and input module is designed in the hardware. The PID switch and communication protocol of the network composite tracking sensor are designed in the system output module, as shown in the following Table 1 and Table 2:

Table 1. PID switch setting of composite tracking sensor

Switch pin	Switch condition	Function
2–5	Open	Address value 0H01
2–5	Shut	Address value 0H02
3–4	Open	Find the white line and check that the white line light is on
3–4	Shut	Find the black line and check that the black line light is on

In order to ensure the operation efficiency of the system, it is necessary to further optimize the main control board port and port connection information, and summarize the motherboard interface through a large number of practical search, as shown in the following list (Table 3):

Table 2. Composite tracing sensor communication protocol table

Host command	C sensor return data	Effective value
P	Status values of 9 channels	0–245
PX1– PX9	Monochrome data of corresponding 1–9 channels	0–1352
PX11–PX19	Two color data of corresponding 1–9 channels	0–1352
PX21–PX29	Threshold data of corresponding 1–9 channels	0–1352
PX31–PX39	Current data of corresponding channels 1–9	0–1352

Table 3. Motherboard interface table

Name	Motherboard number	Software identification	CPU port	Connecting parts
Analog interface	A3–A15	P0–P9	FD0–FD9	Gray scale, infrared ranging, light intensity and other analog sensors
Digital interface	A18–A29	F0–F8	FB0–FB9	Digital signal module
Electric drive interface	A28	N0	FB2–FB9	On board dual motor drive output
Liquid crystal display	T6	Y0	–	On board LCD interface

Combine the principle of data classification to optimize the system development process, build a language interactive education auxiliary information management platform, and improve the efficiency of the system. The specific structure is shown in the figure (Fig. 2).
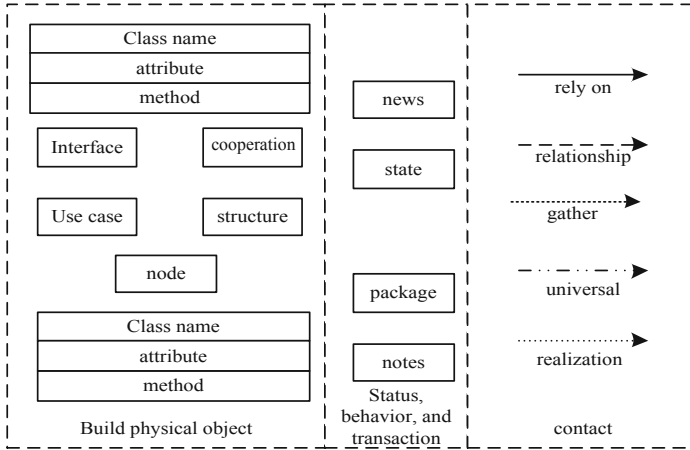
Fig. 2. Auxiliary information management platform for language interactive education

Based on the above method, the system hardware structure can be designed to better guarantee the quality of system operation and prolong the service life of the system [5, 6].

2.2   Optimization of System Software Function Structure

The system designed in this paper supports user types such as teachers and students. Users can query, modify and update their own information [7]. After students register, teachers can make an analysis and statistics on students' learning situation after each
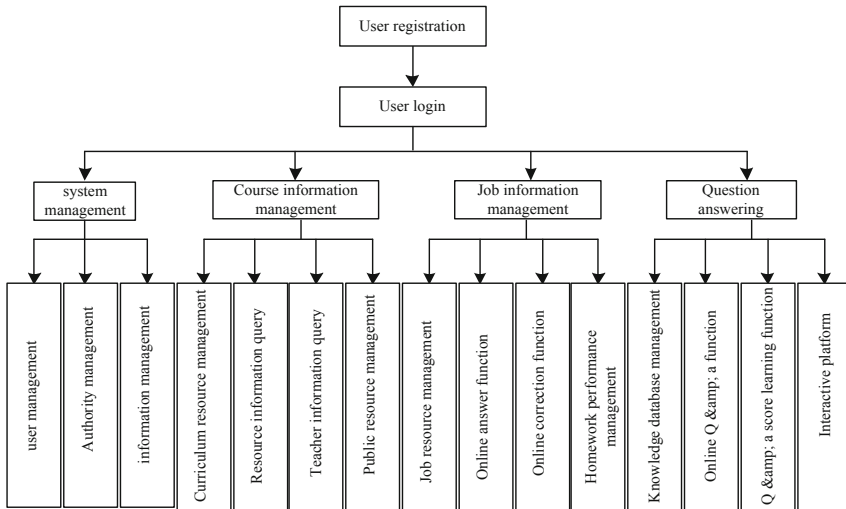
Fig. 3. System function structure diagram

class or after a period of time, timely understand students' learning dynamics, and feed back to students, so that teachers and students can interact and exchange course content in real time, so as to better improve the effect of course teaching [8]. Combined with the principle of data classification, the structure of course teaching assistant system is optimized. According to the work flow of teachers' curriculum teaching data and students' daily learning needs data, analyze and design a teaching data management system that can be more conducive to teachers' daily teaching work, and optimize the system functional structure, as follows (Fig. 3):

After analyzing the entire system of the system, it can be determined that this article will use the MVC architecture and methods, and use the B/S mode for auxiliary system development and design work. After the analysis, the specific architecture is included in this structure diagram as shown in the figure. Three-layer structure of presentation layer, logic layer and data layer (Fig. 4).
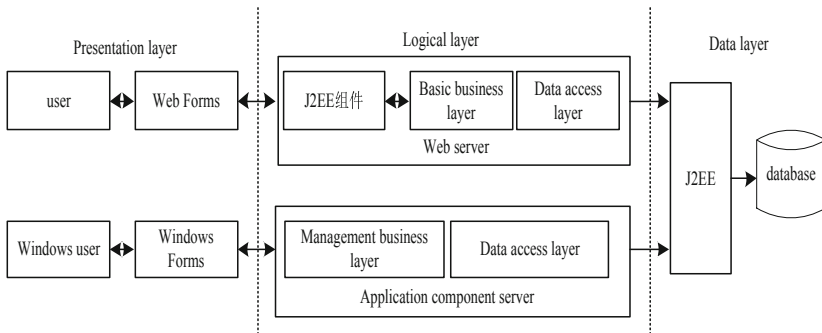
Fig. 4. Software architecture

Through the development and design of this system, it can be used by managers. At the same time, the development system and structure of the three-level architecture are also very clear, and the operation is relatively simple, which can greatly reduce the workload of the software package installation process [9]. In the system presentation layer, the main work is to use this interface to realize the conversion of data information. Commands can be input into the system in the form of machine language. This is very key for the presentation layer. Users can directly use this level. At the same time, the simplicity of the whole interface needs to be considered. The logic layer mainly operates and processes the data layer, issues operation commands through the presentation layer, and then processes the data [10]. After getting the data from the data layer, the data can be fed back to the system to improve the operation quality of the system.

2.3 Realization of Language Interactive Auxiliary Teaching

The designed system is divided into two parts according to the user type: data classification teaching information management for student users, data classification teaching information management for teachers and data classification teaching information management center for administrator users.

The system mainly provides students with six important modules: media learning environment, test question training environment, programming training environment, operation system, learning feedback and user information management. According to the actual language interactive auxiliary requirements, the system provides different user information management functions for users with different user permissions. Student users can only edit their own user information; Teacher users can edit their own user information and view some student user information, such as student number, name, class and other non private information; The system administrator can manage the user information of all users. There are two main management methods: manual editing and importing format files. Based on this, the system data information classification management process is optimized, as shown in the figure below (Fig. 5).
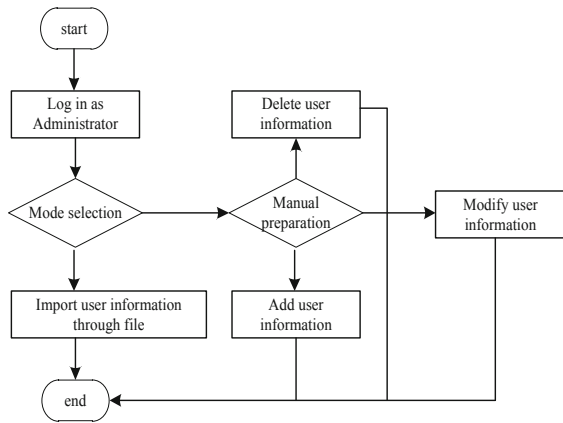
Fig. 5. System data information classification management process

The interactive assistance system provides multiple types of synchronous and asynchronous interactive tools, and combines them with the Steaching process of teachers and the learning process of students to form various "interactive toolbars" suitable for different teaching and learning stages. Facilitate convenient and efficient two-way communication and exchanges between teachers and knowledge transfer, between students and teachers, and between students and knowledge transfer. The main tools of the interactive auxiliary system The main tools used by the interactive auxiliary system are shown in the Table 4.

Table 4. Description of various interactive tools

Type	Tool name	Feature description
synchronization	Hypermedia	The teaching content can be presented to students in various forms, such as text, graphics, animation and video, and connected with hypertext
	E-mail	With the communication means provided on the Internet, teachers can assign tasks and homework through e-mail, and answer one-to-one through e-mail
	Learning Forum	Online information service system can provide various resources, information and contacts to users in need

Use the various teaching tools introduced above to build an interactive auxiliary subsystem to provide corresponding "interactive Toolbar" for different stages of network teaching and learning. The architecture model of interactive auxiliary subsystem is shown in figure (Fig. 6).
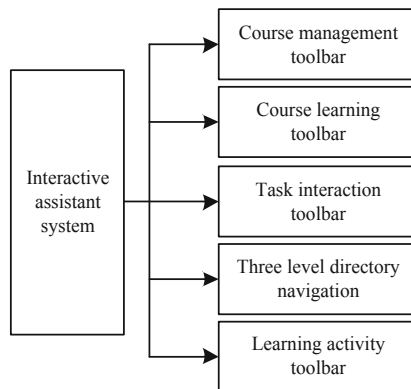
Fig. 6. Block diagram of interactive assistance system

In the process of data classification, a large number of data stored in the database do not meet the classification conditions. In order to enhance the effect of data classification, it is necessary to filter the original data. In order to ensure the horizontal and vertical inconsistency of each evaluation index, it is necessary to classify the data effectively. The classification of data is to consider the needs of users and recommend teaching resources to users who want to use them according to the needs of users. Therefore, this paper classifies the relevant data.

According to the characteristics of the teaching quality evaluation system, the data classification needs to meet the following rules 1: assuming that the college has N total of a teachers, for convenience, it is advisable to set each teacher as s_{kij} students, the number of all evaluation indicators is t, k is the score of the j evaluation index of student

i, and $E_k(j)$ is the mean and variance of the j evaluation index of teacher K.

$$E_k(j) = \frac{1}{N} \sum_{N}^{i=1} s_{kij-t} \tag{1}$$

$$D_k(j) = \frac{1}{N} \sum_{i=1}^{N} \left(s_{kij} - E_k(j)\right)^2 \tag{2}$$

If $D_k(j) = 0$, it means that all students have the same data for M evaluation indicators of teacher k, and the evaluation data will be deleted. E_i is the mean and variance of student i is evaluation of teacher k:

$$E_i = \frac{1}{W} \sum_{W}^{k=1} E_i(k) \tag{3}$$

$$D_i = \frac{1}{WM} \sum_{W}^{k=1} \sum_{M}^{j=1} \left(s_{kij} - E_i\right)^2 \tag{4}$$

If $D_i = 0$, it means that there is no difference in students' scores on all evaluation indicators ot of all teachers, then the data will not have a positive effect on data classification. Therefore, α and β will not be selected. The mean and variance of students' teaching evaluation scores are as follows:

$$\alpha = E_i - E_k(j) \tag{5}$$

$$\beta = ot - D_i \tag{6}$$

Use overall classification accuracy T_n and single classification accuracy F_n to describe the performance of the SVM classifier, which is defined as follows: For the case where the data set sample distribution is relatively balanced, the overall classification accuracy evaluation index is used.

$$A = \frac{T_1 + T_2 + T_3 + \ldots + T_n}{T_1 + T_2 + T_3 + \ldots + T_n + F_1 + F_2 + F_3 + \ldots + F_n} \tag{7}$$

For the unbalanced distribution of data samples, the single classification accuracy evaluation index is used.

$$C = T_i/(T_i + F_i) \tag{8}$$

In the formula, T_i represents the number of correct classifications in the data sample of the type, and F represents the number of incorrect classifications. When describing the performance of the classifier, the overall classification accuracy can objectively reflect the generalization performance of the SVM classifier, and the single classification accuracy can accurately reflect the pros and cons of the SVM classifier. Based on this, this paper divides the students' online behavior data samples into two categories under

the evaluation index with higher overall classification accuracy. Its functions include the formulation of various types of homework questions, the management of homework question banks, the release of homework, online submission, online correction, and the statistical analysis of homework scores and score information query functions. Program function introduction: Multiple choice questions and judgment questions can be automatically corrected and the scores displayed after the homework is submitted. Students can view the homework history and correct answers from the homework record. Teachers can manage homework question types and question banks, make online corrections to some question types, and give scores and comments, as well as statistical analysis and management of results. Work flow: After the students log in to the system, they will complete the homework according to the teacher's assignment.

Students ask questions to the system in the way of natural language, and the intelligent question answering agent gives answers. Intelligent Q & A agent has the ability to learn, accumulate knowledge and improve the ability to answer students' questions. After users ask questions, the system analyzes the keywords, matches them with the answers, and feeds back the results. Both teachers and students can score the answers given by the system, so as to continuously train the intelligent question answering agent of the system.

The steps are as follows (1) The user enters the question in natural language, and the intelligent answering agent obtains the content of the question and displays it in the dialog box. (2) The intelligent answering agent analyzes the question, gets a number of keywords, and filters out the redundant words by merging synonyms, etc., to get the keywords that are actually needed. (3) According to the obtained keywords, query in the knowledge database to get the most consistent answer. (4) Return and display the answer. Based on this, the effective design of the system is realized, and the effect of auxiliary teaching is effectively guaranteed.

3   Analysis of Experimental Results

Any set of system programming software needs supporting software, which is developed using object-oriented V programming software. On the basis of this development software, the supporting software is used for development, and the system uses programming management software to support the software, and the system is programmed by flowchart programming software. And C language programming software is composed of two parts, programming management software is used to establish the system low-level module software and set up the program running environment. The flowchart programming software and C language programming software rely on AVR GCC support, and need to install Winavr GCC at the same time. This software has established environment settings. The default GC version used is: Winavr20070525. The built-in software module is the part executed in the flowchart programming language. This part tests whether its functions can be dragged and used correctly. The test results are shown in the Table 5.

After the development and design of each functional module of the system, it is necessary to fully test the function and performance of the system, so as to prevent abnormalities in the process of system operation. At present, the main test methods

Table 5. Built in software module test table

Type	Can I use it	Operation	C language conversion status
Conditional judgment	Sure	Stable	Accuracy
Count repeat	Sure	Stable	Accuracy
Condition repetition	Sure	Stable	Accuracy
Do repeat	Sure	Stable	Accuracy
Repeat forever	Sure	Stable	Accuracy
Assignment statement	Sure	Stable	Accuracy
Toms delay	Sure	Stable	Accuracy
MS Delay	Sure	Stable	Accuracy
Sensor	Sure	Stable	Accuracy
start-up	Sure	Stable	Accuracy
stop it	Sure	Stable	Accuracy

are black box test and white box test. Black box test is to detect and analyze each functional module of the system to ensure that the functional modules of the system can operate normally. There are some differences between white box test and black box test, which requires designers to have a certain understanding of the structure of the program, and carry out internal detection and analysis. When testing the system function, first of all, it is necessary to test the data interface of the system, determine the relevant parameters and information input by the user in the test process, and also meet the following requirements: (1) divide and test the functional modules of the system, complete the test analysis of the main functional modules, and complete the big data test. (2) When testing the network of the system, users need to log in to the system many times, so as to determine the user's actual operation, which can ensure the normal use of the system. During the specific test, the personnel participating in the test need to adjust according to the parameters, as shown in the Table 6:

Teaching assistant management includes various functions such as examination management, score management, and student attendance management. For the functions of the system, these functions are the main functional modules in the entire system, as shown in the figure (Fig. 7).

The test of the system is not only considered from the function, but also from the performance and other aspects. In particular, different users will participate in the test process and need to be tested and processed according to different user roles. In the process of system test, all functional modules are closely connected with the test process. In addition, the process is easy to understand, the user interface is very friendly and the operation is relatively simple. After the function test, the abnormal performance test is also required. The specific method is that the user inputs the abnormal value to judge whether the system can avoid the occurrence of abnormal conditions during operation. At the same time, the system can prompt that the value entered by the user is illegal. In the test results, it can be determined that there are 160 use case designs selected

Table 6. Functional test cases

Control	Expected results	Is it stable
Software main interface and menu	Operate the main page, including registration, login, use, etc., to determine whether the main interface and menu functions of the software are stable	Stable■ unstable□
Message publishing management	Administrators use this function to publish message data	Stable■ unstable□
Binding effect of official documents	Staff can manage official document data	Stable■ unstable□
Teaching resource management	The system can sort out teaching resources	Stable■ unstable□
Auxiliary teaching	The system can provide courseware, audio and video for auxiliary teaching	Stable■ unstable□
system maintenance	The administrator can back up the database and download the database	Stable■ unstable□
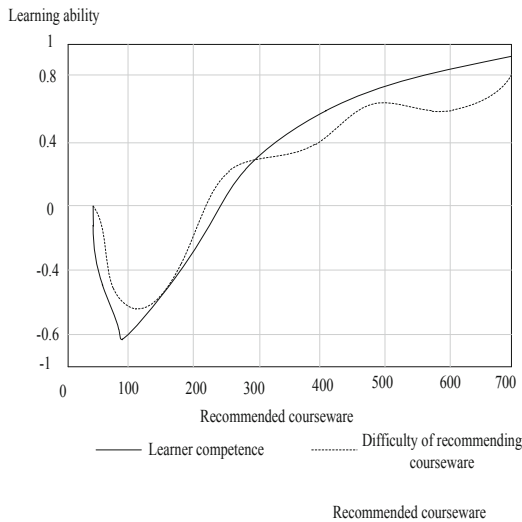
Fig. 7. Evaluation of the recommended effect of courseware

this time, The total number of functional modules involved is 400, and there are 380 normal test items, accounting for 95% of the test proportion; 15 items with light defects, accounting for 3.75% of the test proportion; There are 5 moderate defects, accounting for 1.25° of the test proportion. Although the system runs well, there are still some errors, mainly as follows: users of any identity can view all functions when entering the system. Although users without permission cannot use these functions of the system,

there are still defects. In addition, the description of the interface is inaccurate. These test problems have been corrected after the test results, and all have been changed and improved before deployment.

4 Conclusion

This paper designs a language interactive course assistant system based on data classification. The system design includes hardware and software. In the hardware design, the overall optimization of the system module is optimized, and the connection layer and other modules of the system are designed. In the system software design, the data classification method is used to classify the extracted user demand data to improve the performance of the auxiliary system. The experimental results show that the system designed in this paper has good performance. However, in the current research, more user data and user needs should be considered in the system design, which will be helpful in the future.

References

1. Tassi, A., Gigante, D., Modica, G., et al.: Pixel- vs. object-based landsat 8 data classification in google earth engine using random forest: the case study of Maiella national park. Remote Sens. **13**(12), 2299 (2021)
2. Castrogiovanni, P., Fadda, E., Perboli, G., et al.: Smartphone data classification technique for detecting the usage of public or private transportation modes. IEEE Access, **99**, 1–1. [1] (2020)
3. Li, G., Ouyang, S., Yang, Y.: A study on the construction of a culture pedagogical network learning space – the CASH curriculum idea. Int. J. Emerg. Technol. Learn. (iJET) **14**(17), 73 (2019)
4. Yang, K., Ma, Q.: To improve the quality of distance education and the simulation of effective resources under the big data. Comput. Simul. **34**(4), 212–215, 334 (2017)
5. Omar, S., Hanifah, M.A., Rathakrishnan, L., et al.: The use of kahoot interactive quiz in the teaching of Malaysian nationhood studies course at universiti Utara Malaysia. Solid State Technol. **63**(3), 177 (2020)
6. Kassabolat, A., Kadirsizova, S., Kozybayeva, M., et al.: Future teachers' opinions on preparation and use of interactive materials in teaching. Int. J. Emerg. Technol. Learn. (iJET) **15**(23), 121 (2020)
7. Muoz, L., Villarreal, V., Morales, I., et al.: Developing an interactive environment through the teaching of mathematics with small robots. Sensors **20**(7), 1935 (2020)
8. Sprute, D., Tnnies, K.D., Koenig, M.: A study on different user interfaces for teaching virtual borders to mobile robots. Int. J. Soc. Robot. **11**(3), 1–16 (2019)
9. Boschmans, S.A.: Becoming a pharmacy assistant and technician: education and training. Encyclopedia Pharm. Pract. Clin. Pharm. **15**(02), 15–24 (2019)
10. Liu, S., Fu, W., He, L., et al.: 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

Basketball Posture Recognition Method in Physical Education Based on Machine Vision

Xiaobo Li[✉] and Zhiqiang Wang

Sports Center, Xian Eurasia University, Xian 710000, China
`lixiaobo10034@126.com`

Abstract. The traditional basketball posture recognition method has large error and the recognition effect decreases. A basketball posture recognition method based on machine vision is designed for physical education teaching. Collect basketball posture information; Extract the characteristics of basketball movement in physical education teaching; Basketball posture calibration based on machine vision; Then realize the recognition of basketball posture. By means of comparative experiments, it is verified that the attitude recognition effect of the new method is better and has great popularization value.

Keywords: Machine vision · Physical education · Basketball posture · Identification method

1 Introduction

In the 21st century, the innovation of science and technology has changed with each passing day. Cameras, video recorders and various video acquisition devices have been integrated into all aspects of our life. These devices are widely used in traffic management, medical experiments, shopping mall services and so on [1]. It can be said that our life is closely related to these technologies, and almost everyone can't live without these technology-related devices. Nowadays, the video data we can get has shown explosive growth, which brings some difficulties to our traditional video analysis. In the traditional methods, video analysis and other related processes through simple manpower are no longer applicable in this era. Now we often assist people to do corresponding video analysis by means of computer automatic video classification [2]. Its application scenarios involve video surveillance, human-computer interaction, virtual reality, video retrieval and so on [3]. People have done a lot of research on the content of automatic video recognition through computers, but now some video based recognition done by major companies often requires some special cameras, such as Kinect of Microsoft and somatosensory games developed by various game companies [4, 5]. However, there is little research on the videos taken by mobile phone cameras that are often used in our daily life. Indeed, compared with the video taken by the computer, the video taken by the mobile phone has some "jitter" problems. However, according to the psychology

of ordinary people, people are more willing to accept new products that meet the corresponding functions by transforming the objects they often use, rather than buying expensive equipment [6, 7]. Compared with the former, the equipment used in the latter has great universality and practicability.

In recent years, fitness has attracted more and more attention. Almost everyone pays attention to their own body, and everyone will do some fitness activities every day. Basketball is the most popular among fitness sports [8]. Basketball is a collective sport that throws the ball into the opponent's basket to score and prevents the opponent from getting the ball and scoring under the restriction of specific rules. Compared with other ball games, basketball has many technologies, various tactical forms, strong skills of players, and reflects the characteristics of individual combat and cooperation. In the basketball game, the basketball skill level of the players has an obvious impact on the whole team. If the basketball level of the players is insufficient, the weaknesses of the team will be exposed, and the defense and attack level will be greatly reduced, which is not conducive to the performance of the team in the basketball game. Therefore, it is very necessary to carry out scientific and reasonable basketball training for the players [9, 10]. In the traditional way of basketball training, coaches make training plans according to the training and competition of athletes. This method depends on the coaches' training theory and their own experience, and has a certain subjectivity. In addition, it is difficult to avoid wrong movements and possible injuries to athletes' muscles, soft tissues and bones through scientific observation in the training process, which will affect the normal progress of training and even shorten athletes' sports life. From the perspective of training quality evaluation, the evaluation work is operated manually. Coaches need to calculate the training performance of each athlete with reference to different test standards. This method also has some disadvantages [11]. Firstly, the test of athletes is carried out manually, which needs a lot of time for coaches, the process is complex and the accuracy is poor; Secondly, the test method has limitations. It is difficult to measure some important motion parameters such as acceleration and angular velocity directly, but it is impossible to measure the information such as muscle tension, sprint ability and body balance; Third, coaches lack scientific evaluation methods, and it is difficult to formulate corresponding decision-making schemes according to the test data. Therefore, if the sports parameters of athletes can be collected in real time and accurately, the sports posture of athletes can be analyzed and identified, and the training effect evaluation model can be constructed. Based on this, coaches can reasonably adjust the training scheme and scientifically evaluate the training quality, it is of great significance to improve the competitive ability of sports mobilization and the decision-making ability of coaches.

2 Design of Basketball Posture Recognition Method in Physical Education Teaching Based on Machine Vision

2.1 Collect Basketball Posture Information

At present, there are two main recognition methods of human posture recognition, namely, recognition technology based on image analysis and recognition technology based on inertial sensor. The recognition technology based on image analysis mainly

recognizes human posture by collecting video, image and other information. Therefore, it is necessary to place cameras and other monitoring equipment in the detection environment in advance for data acquisition. The application of image analysis technology in human posture recognition is early and mature, In addition, the technology uses multiple cameras to detect human action posture from multiple angles, and uses neural network algorithm to train and classify image and video data. Although this method can accurately recognize other people's daily actions, it contains a large amount of data, It is difficult to realize real-time monitoring [12, 13]. There are still many deficiencies in the recognition technology based on image analysis, which requires high precision of the equipment, and the equipment is bulky and inconvenient to carry. Video acquisition is easy to produce dead corners, some places are not easy to be observed, the monitoring range is obviously limited, and the large amount of image acquisition data is easy to lead to insufficient storage and can not achieve the purpose of real-time monitoring. The recognition technology based on inertial sensor makes up for the shortcomings of image recognition technology. The development of science and technology drives the improvement of sensor technology. Sensor equipment has become the best method to obtain human posture information with the characteristics of small volume, high precision, flexibility and easy to wear, low environmental requirements, high sensitivity, low energy consumption and good real-time performance, It is widely used in various fields, such as competitive sports, rehabilitation therapy, somatosensory games and so on. Multiple inertial sensor devices are used together to form a body area network, which has been widely used. Therefore, on this basis, this paper designs the data acquisition module by using the way of sensor.

In the data acquisition stage, the attitude information of human body completing different actions is collected by sensors. The sensor node formed by the combination of multiple sensor devices can convert the action information in the process of action completion into electrical signals for uploading, so as to meet the requirements of subsequent logic operation, data storage and communication. According to the requirements of practical application, it is difficult for a single sensor module to meet the work requirements. The information required in human posture recognition is complex and diverse, including physical and physiological information such as acceleration, angular velocity or heart rate. The analysis and processing work needs to be completed inside the node, so the design of the node needs to include multiple sensor modules, It can be used together to meet the work requirements of the system. Generally, a sensor node includes four modules, as shown in Fig. 1 below.

As shown in Fig. 1, it is mainly composed of processor module, power supply, sensor module and communication module. The processor module controls the normal operation of each functional module of the sensor node and carries out the relevant processing of each signal; The sensor module realizes the function of detecting object motion information and realizes the transformation from motion information to electrical signal; The communication module is responsible for signal transmission and transmitting node data to other devices by wireless means; The power supply provides energy for the normal operation of the whole sensor. At present, mobile devices such as mobile phones have also begun to integrate sensor modules, which have the function of wireless communication. They will replace sensor nodes to wear them on key parts of the human body for
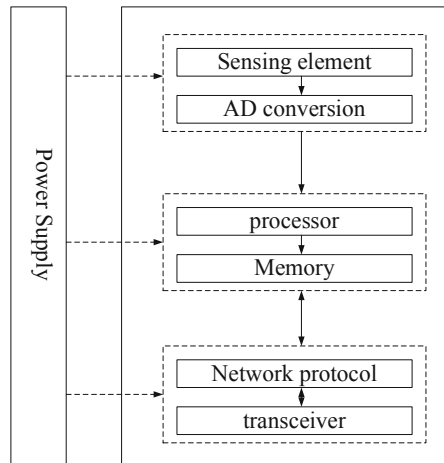
Fig. 1. Sensor node structure

signal acquisition. Compared with sensor nodes, the wearing position of mobile devices is not fixed, which will have an impact on the recognition results of the system. When detecting motion information through sensors, The equipment can be placed in a fixed position to avoid this effect. At present, there are many human posture recognition platforms. Researchers need to build a posture detection platform according to their own research needs.

In data acquisition, multiple sensor nodes are generally used to collect relevant information. In addition to the structure of the node itself, the effective and complete transmission of data is another main problem. At present, according to different data transmission media, data transmission forms are mainly wired and wireless. Wired transmission mode is more stable and reliable, but it is not widely used because of its complex installation and wiring and many restrictions on motion detection; Wireless transmission mode has many advantages in the field of human posture recognition. The commonly used wireless communication technologies include wireless radio frequency identification, Bluetooth, Zig Bee, wireless ultra wideband, etc. wireless transmission mode can reduce the impact of sensors on normal activities, so most systems use this form of data transmission. When designing wireless transmission protocol, we need to consider network architecture, radio technology, communication protocol and energy control. Among the common network topologies, star topology and mesh topology are widely used in practical applications. In the body area network structure, it mainly realizes the aggregation of the data information collected by the nodes attached to the human body, so as to transmit the data to the upper computer for calculation. In the application of body area network, the star topology requires multiple nodes to be directly connected with the receiving node. Its network communication structure is relatively simple and easy to implement, so it is often used. Compared with the star topology, the mesh topology is more complex, but it can use multiple ways to reduce the path loss caused by diffraction, and the data transmission only exists between adjacent nodes, which can keep the node

small transmission energy. In the network protocol setting, we need to set a reasonable network structure according to our own research needs.

2.2 Extraction of Basketball Action Characteristics in Physical Education Teaching

Feature is the abstract representation of a set of data, and the representation of data features can have many forms. At the same time, feature can be equivalent to attribute value, so feature extraction is the process of transforming the original data set into an attribute set that can be used to express these data characteristics. When the amount of input data set is too large and some of the data is redundant, it is necessary to use the feature extraction algorithm to extract the feature attribute set of input data to form the feature vector. As a representative feature set, it simplifies the data set to a certain extent, improves the operation efficiency of the algorithm and reduces the complexity. Generally, the combination of feature extraction and feature selection is used to construct the appropriate feature vector. The selection of the appropriate feature set is very key to the recognition of human posture. After extracting the data set representing each action through data division, analyze the data of unit action to extract the attribute set of reaction characteristics. This process is the feature extraction process [14]. The main methods of feature extraction are to analyze the characteristics of different data and extract the corresponding features. At present, the more common features are time-domain features, frequency-domain features and time-frequency features. Time domain features: there are many time domain features, which are often used in research: mean, variance, extreme value, standard deviation, covariance, correlation coefficient, root mean square, etc. It is also called signal statistical feature. Its calculation method is simple and low complexity. It is often used in the field of human pose recognition. Frequency domain characteristics: in order to obtain the frequency domain characteristics of the signal, the time domain signal is usually transformed into the frequency domain signal, which is generally realized by Fourier transform. The frequency domain signal represents the components of the signal at different frequencies, which is different from the time domain characteristics, and reflects the signal characteristics on the other hand. The commonly used frequency domain features include frequency domain line, spectral energy, etc., and the Fourier transform coefficient $L + L$ and its absolute value are generally used as frequency domain features for reference. Time frequency feature: the time-frequency feature reflects the characteristics of the signal in time domain and frequency domain. This feature is mainly used for feature extraction of unstable signals. In the process of time-frequency feature extraction, wavelet analysis is mainly used to realize the extraction, and the local transformation of frequency and time can be used to realize multi-scale thinning of the signal.

Feature selection is a variable selection method, also known as attribute selection or variable subset selection. It is a process used to select relevant attribute subsets in order to build a classification model. The primary reason why feature selection is proposed is that in the feature set obtained by feature extraction, not all attributes are relevant and useful, and some attribute selection may be redundant. The introduction of irrelevant attributes not only has no effect on the construction of the model, but also makes the constructed model more complex due to the redundancy and irrelevance of data. Therefore,

reasonable feature screening is extremely necessary. There is a big difference between feature selection and feature extraction. The purpose of feature extraction is to extract feature vectors from the original data, and feature selection is to select appropriate feature sets from these feature vectors. There are three main purposes of feature selection: (1) simplify the model and reduce the computational complexity; (2) Shorten training time; (3) Strengthen the promotion to avoid the problem of over fitting. The commonly used feature selection algorithms are generally obtained by combining the evaluation function with sequential forward / backward search, decision tree, best first search and genetic algorithm. Among them, the evaluation algorithm is a function that can reflect the advantages and disadvantages of the selected feature subset, which can be used to solve the correlation between feature and classification, classifier error rate and so on.

Since the concept of machine vision was put forward, many large companies, research institutes and universities have made unremitting efforts in this field. At present, they have made great progress in many fields. Many products of machine vision have been widely used in our society and life, such as human-computer interaction, virtual reality, video retrieval, video surveillance and so on. Most of the major companies do the corresponding research on machine vision to meet their own needs, or to explore the way and clear the obstacles for the future of the company. Therefore, each company has its own focus, and each company also has its own advantages and disadvantages. In various applications of machine vision, gesture recognition based on video is a popular and indispensable direction in the research direction of human-computer interaction. Gestures are closely related to our lives. Almost everyone has to make different gestures every day, such as waving, clapping, boxing and so on. Among the gestures people make, some gestures are complex and some gestures are simple. But every gesture is basically inseparable from the beginning, process and end of the gesture. For the research on the application of gesture recognition based on video, we can roughly make these classifications: one-dimensional gesture recognition, two-dimensional gesture recognition and three-dimensional gesture recognition. The specific application scenarios of gesture recognition of these three classifications are very different. One dimensional gesture recognition is mainly used to recognize some static hand types. It is relatively simple. The gestures he wants to recognize are almost unchanged. For example, you need to recognize a static V hand type, a static fist hand type, and a static number 0 hand type; Compared with one-dimensional gesture recognition, two-dimensional gesture recognition still has no relevant depth information. Its process is mainly detected and recognized by the changing characteristics of a gesture action in the two-dimensional plane. Although this kind of recognition extracts more features than one-dimensional gesture recognition, it is difficult to recognize because it not only recognizes simple hand shape, but a series of continuous actions without corresponding depth information. Because modern people often carry some devices, such as mobile phones and cameras, their cameras have no depth information, but these devices are widely used now, and almost everyone is using these devices. Therefore, two-dimensional gesture recognition still has corresponding research prospects and application scenarios. Compared with two-dimensional gesture recognition, three-dimensional gesture recognition has a layer of depth information, so its recognition efficiency is more accurate than two-dimensional gesture recognition, but the calculation of three-dimensional gesture recognition is much

larger than that of two-dimensional gesture recognition. Therefore, three-dimensional gesture recognition is generally applied to some devices equipped with depth cameras, It can not be used for mobile phones and other equipment, such as home game consoles, some security locks of security companies, places with large personnel mobility in hotel amusement parks, etc.

In this paper, feature extraction is divided into two categories, one is global analysis method, the other is local analysis method. The overall analysis method usually extracts the foreground of the video by means of background subtraction method, optical flow method and difference method to obtain the corresponding binary image, then obtains the overall region of interest through the binary image, and then processes these regions of interest accordingly. The next step is to obtain the contour and edge information of these regions of interest. However, in these information, we may often contain a lot of unnecessary and redundant information, which we call noise. How to filter these noises and obtain the corresponding accurate eigenvalues is often the focus of feature extraction. People have done a lot of research in this area. Extracting video eigenvalues can distinguish the corresponding eigenvalues. This method mainly obtains the contours through the background subtraction method. These contour information often contains corresponding energy information. The second step of this method is to analyze the energy of these contours, in which the analysis mainly analyzes the corresponding motion potential energy, and kinetic energy, Then the corresponding video contour energy map is obtained by analysis, and the motion history map of the contour is obtained by using the motion image. Because the process of obtaining eigenvalues by this method is single perspective and the perspective is often unchanged, it is easy to obtain some features of invariant moments.

2.3   Calibrating Basketball Posture Based on Machine Vision

After the basketball feature extraction above, this paper will further calibrate the basketball posture based on machine vision. In general, in the extracted feature set, only the largest and second largest features are taken as the corresponding real feature values, while for other smaller features, they are deleted on the current interface, so as to ensure that there are two external rectangular features of the image of each frame, Prevent the situation that more than two features are drawn in a connected area. This prevents "system noise". By removing these noises, we can further improve the accuracy of our recognition. The principle of machine vision is shown in Fig. 2 below.

In Fig. 2, P represents the light source, X_1 and X_R represent the left and right cameras, d represents the distance between the left and right cameras, L and R represent the object to be measured, B represents the left and right boundary distance of the object to be measured, D represents the distance between the camera and the object to be measured, and Z represents the distance between the light source and the object to be measured. As shown in Fig. 2, the original data signals collected by the machine vision equipment often have noise signals generated by the influence of the external environment and themselves. These signals are inaccurate. Therefore, it is very necessary to denoise the original data signals. There are many methods to denoise data information, and the signal denoising method implemented in software design is generally called digital filtering. This technology mainly includes classical filter and modern filter. Among them, the
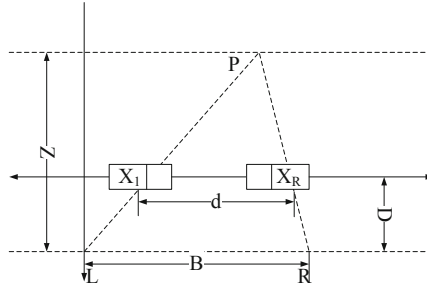
Fig. 2. Principle of machine vision

former first assumes that the useful signal and noise signal in the signal are located in different frequency bands, and the noise signal can be removed simply by placing the signal in the linear system. The common high-pass and low-pass filters are constructed based on the principle of different frequency band distribution of the signal. However, the classical filter has defects, which is no longer applicable to the original signal data with overlapping frequency bands of noise and signal. For this type of signal, the modern filter is generally used. Different from the classical filter principle, the modern filter classifies the useful signal and noise into random signals, and then deduces and estimates the characteristics of the useful signal or noise by using its corresponding statistical characteristics such as autocorrelation function and self power spectrum. The common modern filter algorithms mainly include Kalman filter, Wiener filter and so on.

Normalization, also known as standardization, is a basic step in data information mining and plays a role in simplifying calculation. In the process of human pose recognition, normalization method is also used to process the data. Because in the body area network system, the placement positions of nodes are different, and in the process of movement, nodes will shake with the movement of human body. The change of node position information will cause a certain deviation in the collected data, which may reduce the accuracy of human posture recognition. Data normalization can eliminate the dimensional influence between different data and solve the problem of data comparability. The normalization processing means is to scale the data to a certain range, and transform its expression, so that different data can be comprehensively compared and evaluated. Two methods of data normalization processing will be introduced in detail below. First, linear function transformation. This method maps the original data to the closed interval between 0 and 1 through linear transformation, and can complete the equal scaling function of the original data signal. The specific effects are as follows.

$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

In formula (1), X_{norm} is the result of linear function conversion; X is the original data; $X_{\max}$ is the maximum value of sample data; $X_{\min}$ is the minimum value of sample data. On this basis, the 0-means normalization method is used to process the original data into a normal distribution set with mean value of 0 and variance of 1, as shown below.

$$y = \frac{x - \mu}{\delta} \tag{2}$$

In formula (2), x data set to be processed; y is the result obtained after calculation; μ is the mean value of the original data; δ is the variance of the original data. In order to solve the attitude accurately and reduce the noise interference of the sensor, the angular velocity, acceleration and magnetic field strength are fused in the process of calculating the node attitude, and the quaternion method is used to represent the space attitude, and then the attitude is calibrated. Quaternions are composed of a real number and three imaginary units. The definition formula is as follows:

$$Q = w + x_i + y_j + z_k \tag{3}$$

In formula (3), w, x, y and 4 are real numbers; $i. j$ and k are three imaginary units; Q is a unit quaternion, which is normalized as follows:

$$Q_{norm} = \frac{Q}{\sqrt{w^2 + x^2 + y^2 + z^2}} \tag{4}$$

In formula (4), Q_{norm} represents the normalized quaternion. In this paper, it is applied to describe the rotation of rigid bodies in three-dimensional space, as follows:

$$\hat{Q} = 0.5 \cdot Q \cdot p \tag{5}$$

$$p = 0 + w_{xi} + w_{yj} + w_{zk} \tag{6}$$

In formula (5–6), $\hat{Q}$ is the derivative of quaternion to time; p is the angular velocity. Through visual conversion, the unit quaternion used to describe the conversion of rigid body from one attitude to the next can be obtained. The update formula is as follows:

$$Q_{k+1} = Q_k + \Delta t \cdot \hat{Q}_k \tag{7}$$

$$\hat{Q}_k = 0.5 \cdot Q_k \cdot p_k \tag{8}$$

In formula (7–8), at this time, k is a non negative integer; Q_{k+1} is the unit quaternion of basketball posture at $k+1$ times; Q_k is the unit quaternion of basketball posture at k times; Δt is the time interval between two samples; $\hat{Q}_k$ is the derivative of quaternion to time at time k; p_k is the quaternion at time k. Calibrate the quaternion as follows:

$$x_{k+1} = \phi_k x_k + w_k \tag{9}$$

$$z_k = H_k x_k + w_k \tag{10}$$

In formula (9–10), x_{k+1} is the basketball posture calibration state at time $k+1$; ϕ_k is the state transition matrix; x_k is the basketball posture calibration state at time k; w_k is the noise vector; z_k is the secondary measured value; H_k is the relationship between basketball teaching posture and visual measurement in the ideal state. Through this calibration state, basketball teaching posture can be accurately identified.

```
        ┌─────────────────────┐
        │        Start        │
        └─────────────────────┘
                   │
                   ▼
        ┌──────────────────────────────────┐
        │ Collect basketball posture information │
        └──────────────────────────────────┘
                   │
                   ▼
        ┌──────────────────────────────────┐
        │ Calibration of basketball posture based │◄──────┐
        │        on machine vision         │       │
        └──────────────────────────────────┘       │
                   │                                │
                   ▼                                │
        ┌──────────────────────────────────┐       │
        │      Normalization processing      │       │
        └──────────────────────────────────┘       │
                   │                                │
                   ▼                           N    │
        ◇ Whether it meets the ideal state ◇────────┘
                   │ Y
                   ▼
        ┌─────────────────────┐
        │         End         │
        └─────────────────────┘
```
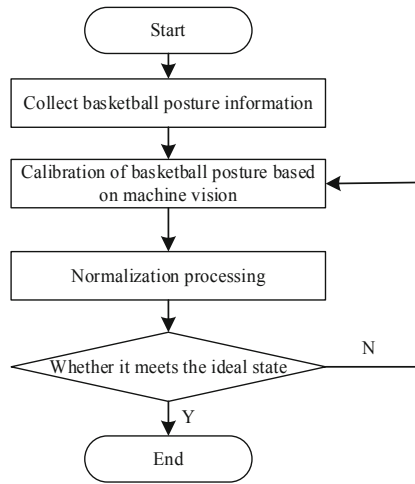
Fig. 3. Basketball posture recognition process

2.4 Recognition of Basketball Posture

The implementation process of basketball posture recognition method based on machine vision is shown in Fig. 3.

In order to realize the recognition of basketball posture, the machine vision equipment is fixed on the back of the basketball player's hand, and the acceleration and angular velocity data of the player's hand in the process of jump shot are collected. The jump shot process is divided into four stages, the shooting posture of each stage is analyzed, and the sound feedback reminder is used to help the athletes correct the shooting posture. However, the jump shot posture of athletes is reflected by the movement posture of their arms and legs. This method only analyzes the posture of hands in the shooting process, and lacks the consideration of the posture of other parts of the body. By collecting the data of athletes' heart rate, oxygen consumption and acceleration, this paper analyzes the physiological characteristics of basketball players in the process of sports, compares the physiological performance of athletes in different forms of competition and training, and proves that wearable devices are very helpful to the quantification of basketball, but there is no specific research on the quantification of basketball; The basketball posture recognition system is constructed by using acceleration machine vision, the lower limb data of basketball players are collected, and the recognition of 8 kinds of actions in basketball is completed. In this paper, only the acceleration data is used as a reference, the characteristics of the constructed data set are single, and the average recognition accuracy is less than 70%. There are many researches on basketball based on machine vision, but they mainly focus on the movement information of some limbs in basketball, such as the hand posture in the shooting process, the lower limb state in the movement process or the physiological characteristics of basketball players. There is little research on the comprehensive analysis of basketball players' upper and lower limb movements. Basketball movement is a complex movement completed by the upper and lower limbs. The recognition of basic basketball movements plays an important role in improving

the skills of basketball players. Therefore, this paper mainly studies the classification of basic movements of upper and lower limbs in basketball, and preliminarily realizes the recognition of basic basketball movements.

3   Experiment and Analysis

In order to verify whether the method designed in this paper has practical effect, this paper carries out experiments on the above methods. The experimental process and results are as follows.

3.1   Experimental Process

The PC of this training and testing program is configured with Intel Core 2 2.70 GHz processor, 2 G memory, Windows XP operating system, and Visual Studio 2005 and OpenCV programming environment. The original resolution of the camera is 1024 × 768, the resolution of the collected video image is adjusted to 512 × 384 by using the interpolation algorithm. The experiment starts at the angle of 0° and ends at the angle of 360°. The sensor node is continuously rotated and the data is sampled every 45°. The experiment is divided into two groups. In order to compare the recognition effect of the method designed in this paper, the first group adopts the traditional attitude recognition method; The second group of the experiment uses the method designed in this paper to recognize the running posture of basketball.

3.2   Experimental Results and Discussion

In the above experimental environment, two methods are used to recognize the basketball action posture in the collected image, and compared with the standard action. The smaller the error between the standard action and the standard action, the better the recognition effect of the corresponding method. The experimental results are shown in Table 1.

Table 1. Experimental results

Rotation angle	Recognition error of traditional attitude recognition methods	The recognition error of the attitude recognition method designed in this paper
0	0.25	0.00
45	0.29	0.03
90	0.32	0.04
135	0.68	0.08
180	0.87	0.12
225	1.38	0.36

(*continued*)

Table 1. (*continued*)

Rotation angle	Recognition error of traditional attitude recognition methods	The recognition error of the attitude recognition method designed in this paper
270	2.72	0.44
315	2.95	0.48
360	3.03	0.51

As shown in Table 1, the traditional attitude recognition method has large recognition error and poor recognition effect; The attitude recognition method designed in this paper has small recognition error and good recognition effect, which is in line with the purpose of this paper.

4 Conclusion

Basketball training is the process of improving competitive ability. It refers to the educational process specially organized to continuously improve and maintain the competitive level of athletes under the guidance of coaches and the participation of athletes. The content of basketball training mainly includes five aspects: Athletes' physical training, technical training, tactical training, psychological training and intelligent training. In the competition, the athletes' technical movement is unstable and do not play well, and they can't keep up with the changes of their opponents and miss people in defense. On the one hand, it shows that the basic skills are not solid, and it also shows that the trainers and coaches don't pay attention to the basic skills practice in peacetime, so it will lead to the athletes' repeated mistakes and omissions due to the basic skills in the competition. Therefore, we should strengthen the basic technical training of athletes. In this paper, the recognition method of basketball action posture in physical education teaching based on machine vision is designed. By collecting basketball action posture information, the basketball action characteristics in physical education teaching are extracted, and the basketball action posture is calibrated by machine vision method, in order to improve the recognition effect, reduce the recognition error, and provide guiding suggestions for the standardization of basketball.

References

1. Wang, S.M., Liao, C.L., Ni, Y.Q.: A machine vision system based on driving recorder for automatic inspection of rail curvature. IEEE Sens. J. **99**, 1–1 (2020)
2. Zhao, C., Ding, H., Cao, G., et al.: A new method for detecting compensation hole parameters of automobile brake master cylinder based on machine vision. J. Adv. Trans. **3**, 1–14 (2021)
3. Sergiyenko, O., Tyrsa, V.: 3D optical machine vision sensors with intelligent data management for robotic swarm navigation improvement. IEEE Sens. J. (99), 1–1 (2020)
4. Xue, B., Wu, Z.: Key technologies of steel plate surface defect detection system based on artificial intelligence machine vision. Wirel. Commun. Mob. Comput. **2021**(6), 1–12 (2021)

5. Qy, A., Jin, F.A., Jian, T.B., et al.: Development of an automatic monitoring system for rice light-trap pests based on machine vision. J. Integr. Agric. **19**(10), 2500–2513 (2020)
6. Liu, S., Wang, S., Liu, X., et al.: Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**(1), 90–102 (2021)
7. Zhang, Y., Hong, G.S., Ye, D., et al.: Powder-bed fusion process monitoring by machine vision with hybrid convolutional neural networks. IEEE Trans. Industr. Inf. **16**(9), 5769–5779 (2020)
8. Chen, Y., Shu, Y., Li, X., et al.: Research on detection algorithm of lithium battery surface defects based on embedded machine vision. J. Intell. Fuzzy Syst. **5**, 1–9 (2021)
9. Ali, M., Atia, M.R.: A lead through approach for programming a welding arm robot using machine vision. Robotica, 1–11 (2021)
10. Chen, J.H., Zhang, J.J., Gao, R.J., et al.: Research on modified algorithms of cylindrical external thread profile based on machine vision. Measure. Sci. Rev. **20**(1), 15–21 (2020)
11. Liu, S., Liu, D., Muhammad, K., Ding, W.: Effective template update mechanism in visual tracking with background clutter. Neurocomputing **458**, 615–625 (2021)
12. Li, M., Jia, J., Lu, X., et al.: A method of surface defect detection of irregular industrial products based on machine vision. Wirel. Commun. Mob. Comput. **1**, 1–10 (2021)
13. Liu, W.: Simulation of human body local feature points recognition based on machine learning. Comput. Simul. **38**(6), 387–390,395 (2021)
14. Wang, S., Liu, X., Liu, S., et al. Human short-long term cognitive memory mechanism for visual monitoring in IoT-assisted smart cities. IEEE Internet of Things J. **9**(10), 7128–7139 (2021). https://doi.org/10.1109/JIOT.2021.3077600

Image Recognition Method of Educational Scene Based on Machine Learning

Yingjian Kang[(✉)] and Lei Ma

Beijing Polytechnic, Beijing 100016, China
kangyingjian343@163.com

Abstract. The traditional image recognition method mainly relies on the similarity expansion calculation of the prominent features of the image to realize the image recognition. This method not only reduces the recognition accuracy of the image, but also makes the recognition efficiency of the image low due to the complex calculation process. In response to the above problems, this research designed an image recognition method for educational scenes based on machine learning. After performing normalization, denoising, and enhancement preprocessing on the educational scene image, the HOG, SIFT, and Haar features in the image are extracted. Then use the convolutional neural network model in machine learning technology to complete the recognition of educational scene images. Experimental results show that the effective recognition rate of this method is higher than 92%, and compared with traditional methods, the recognition efficiency of this method is significantly improved.

Keywords: Machine learning · Convolutional neural network · Educational scene image · Image recognition · Image enhancement · Feature extraction

1 Introduction

With the explosive growth of the number of digital images, the image processing method of using manual methods to manage and classify digital images has been difficult to meet people's needs because of its low efficiency. People urgently need a method that can automatically process and manage digital images, and computer vision task was born under this background [1].

Computer vision includes many disciplines, such as machine learning, image processing and pattern recognition. Through the powerful processing ability of the computer, it can analyze and process the obtained video images, and detect, recognize and analyze the targets in the video scene. Through the analysis and data mining of a large number of video data, the content information in video is analyzed, so as to replace some human work.

In the field of computer vision, how to use computers to realize the automatic recognition and management of digital images is an urgent problem to be solved. At present, the main development direction is to use machine learning methods to train computers

to have their own "thinking" so that they can automatically "recognize" and "understand" images. Indoor scene is more complex, which makes the research of indoor scene recognition more difficult.

The general process of scene image recognition includes feature selection and extraction and classifier learning. Among them, feature selection is the key step in the whole recognition task. The more abstract the feature, the richer its structural information and semantic information, and the stronger its expression ability. Therefore, in the scene recognition task, the feature of the target image is extracted first, and then input into the classifier for recognition. In previous studies, scene recognition methods can be divided into the following two types according to the nature of the extracted features: one is the scene recognition method based on the underlying features, and the other is the scene recognition method based on the middle and high-level semantics [2]. The above methods belong to scene recognition methods based on semantic features. In the process of feature extraction, the selection of features is artificially set by researchers, which is not only time-consuming and laborious, but also expensive. What features can accurately express scene information and achieve ideal recognition effect is still a difficult problem in the research.

In the early stage of scene recognition research, scholars often choose features such as color and texture to describe the scene. With the development of technology, related scholars have proposed the use of SIFT, HOG, GIST, CENTRIST and other features to recognize scenes. When faced with indoor scenes with complex structures, it is often impossible to accurately describe the most important features of the scene. Wang et al. used image segmentation algorithms to identify high-speed railway scenes. Sun et al. used the different importance of the information in the scene image, introduced the privilege information and attention mechanism, and realized the recognition of the special scene. However, when the above technology was applied to indoor scenes, the expected recognition result could not be obtained. Li et al. fused RGB images from multiple perspectives to realize indoor scene understanding and complete indoor scene recognition. This method had a large amount of calculation in practical application, and the efficiency of scene image recognition was low.

With the continuous development of machine learning, experts and scholars have proposed a large number of machine learning models. These models are used in different fields of machine learning according to different characteristics to solve different problems [6]. Applying machine learning to image recognition can improve the effect and accuracy of image recognition with the help of good learning.

Education scene image refers to the video image information of education and teaching collected by AR technology, which can reflect the teaching status and other information. By analyzing the education scene image, the improvement in the teaching environment can be analyzed. However, the existing image recognition methods are seldom used in educational scene images. Therefore, according to the above research background, aiming at the problem of low image recognition efficiency existing in traditional methods, this paper designs a new image recognition method for educational scenes based on the convolutional neural network model in machine learning technology.

2   Methodological Research

2.1   Image Preprocessing of Educational Scenes

The acquisition of digital images is the process of imaging objects, which refers to the process of obtaining image data in the scene through a specific imaging device and reconstructing the obtained data [7]. In this paper, the Baslel ACE series acA2500-149c industrial camera with a high-definition webcam is used to capture video in educational scenes to provide better depth images. Among them, the specific parameter information of the Basler camera is shown in Table 1.

Table 1. Specific parameters of Basler cameras

Parameter item	Specific value
Horizontal/vertical resolution	2590 × 1942px
Horizontal/vertical pixel size	2.2um × 2.2um
Frame rate	14fps
Black and white/color	Color
Interface	GigE
Pixel depth	12bits
Video output format	Mono 8, Bayer GB 1 2, YUV 4: 2: 2

Image normalization plays a very important role in processing digital images. It can eliminate the requirements of illumination and pose size for image processing. Generally, the pictures obtained by the camera contain a large part of background information, which will adversely interfere with the later image processing. In order to improve the algorithm performance of face detection and subsequent clothing segmentation in the classroom scene, the image needs to be processed as follows:

1) Image scale normalization. According to the image size obtained by the camera, in order to improve the speed of image calculation and processing and without affecting the loss of original image information, the image size is scaled to 640 × 480 pixels. In the face classifier training process, in order to segment the face region from the original image, we use the calibration software to calibrate and cut the face region of the original image, with a size of 20 × 20 pixels.
2) Noise reduction. For two-dimensional images, there are some noise points in both gray image obtained directly and gray image converted from color image. In this paper, considering the performance indicators of denoising effect and computational efficiency, median filtering is used to reduce the noise of the collected image.

The mathematical expression of median filtering for two-dimensional images is [8]:

$$y_{ij} = Mid_A\{f_{ij}\} \tag{1}$$

Among them, A represents the template formed by the neighborhood pixels selected during filtering; f_{ij} represents the pixel value at (i, j) in the two-dimensional image. Neighborhood median filtering can effectively filter out the isolated noise in the image, it is very effective for salt and pepper noise, and at the same time it can preserve the details of the edge and other information in the image to the greatest extent.

3) Image enhancement. High-pass filter can enhance image edge information and repair blurred details in the image. In this study, histogram equalization is used for image enhancement. For an image with a pixel area of A_0, it is assumed that the gray value of the image pixel is divided into N effective gray levels. If the number of pixels in each gray level of the image is A_0/N, then this image is called a histogram equalized image. Assuming that the gray level of the original image is g, the gray level histogram of the image before transformation is denoted as H_1, the gray level after transformation is denoted as G, the gray level of the transformed image is denoted as H_2, according to the principle of histogram equalization, it can be known that the transformed image has A_0/N pixels at each effective gray level. So there are the following expressions:

$$G = F(g)$$

$$\sum_{i=0}^{g} H_1(i) = \sum_{i=0}^{G} H_2(i) = G \times A_0/N \tag{2}$$

On this basis, we can get:

$$G = F(g) = \frac{N}{A_0} \sum_{i=0}^{g} H_1(i) = \frac{N}{A_0} A(g) \tag{3}$$

$A(g)$ represents the cumulative function of pixel distribution, and the calculation formula is:

$$A(g) = A(g-1) + hist[g], g = 0, 1, \ldots, 255 \tag{4}$$

After preprocessing the education scene image according to the above process, the feature extraction of the education scene image is performed.

2.2   Image Feature Extraction of Educational Scenes

There are many features to be recognized in educational scene images. Based on HOG feature, SIFT feature and Haar feature, this study identifies scenes and people in the image.

The histogram can retain the edge and contour information of the image to the greatest extent, and is insensitive to light intensity and small offsets. Since the edge contour of the object is represented by the gradient distribution and the edge direction at the same time, the image can be divided into the same unit cell when extracting the HOG feature,

and the gradient of all pixels in the image can be calculated. The gradient of the pixel in the image is [9]:

$$G_x(x, y) = H(x + 1, y) - H(x - 1, y)$$
$$G_y(x, y) = H(x, y + 1) - H(x, y - 1) \tag{5}$$

Among them, $G_x(x, y)$ represents the horizontal gradient of the image pixel (x, y); $G_y(x, y)$ represents the vertical gradient of the image pixel (x, y); $H(x, y)$ represents the pixel value of the image pixel (x, y). Then the gradient magnitude and direction at (x, y) are:

$$G(x, y) = \sqrt{G_x^2(x, y) + G_y^2(x, y)} \tag{6}$$

$$\alpha(x, y) = \tan^{-1}\left(\frac{G_y(x, y)}{G_x(x, y)}\right) \tag{7}$$

Then count the gradient histograms of each cell one by one, then cascade all the cells, and finally generate the HOG descriptor. The calculation of the HOG feature is small, and it can effectively represent the shape and contour of the target in the image within a specific range, and it is not sensitive to changes in illumination.

SIFT features include two steps: key point detection and key point feature description [10]. Among them, key point detection includes scale space extreme value detection and
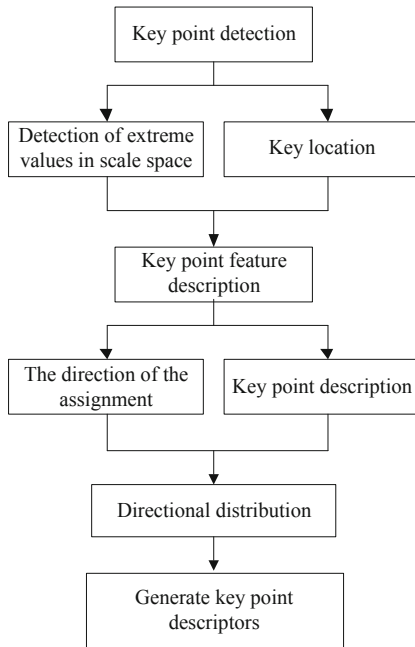
Fig. 1. Schematic diagram of SIFT feature extraction process

key point positioning, and key point feature description includes direction assignment and key point description. The process of extracting SIFT features is shown in Fig. 1.

For the character feature extraction in the educational scene image, this paper uses Haar-like features to represent the character features in the educational scene. Figure 2 is a schematic diagram of the representation of Haar-like features.
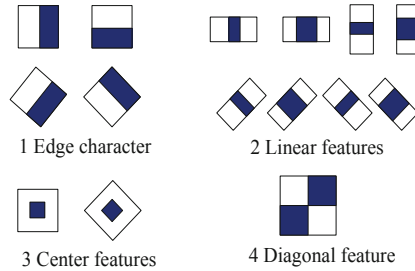
Fig. 2. Representation of Haar-like features

Haar feature is a performance feature that reflects the change of image gray level information. It is sensitive to some simple graphic structures with relatively large gray level changes, such as line segments and edge images. At the same time, it describes a structure with a specific trend [11]. By translating and telescoping the Haar feature template to change its position information and adjusting the size of the template, many rectangular area features can be obtained. Among them, the value of the rectangular feature is called the feature value. Denote the integral image of image I as I', where point (x, y) is represented as any pixel in image I, which is defined as follows:

$$I'(u, v) = \int_{x=0}^{u} \int_{y=0}^{u} I(x, y)dxdy \tag{8}$$

According to the definition of integral image, the integral value of any region can be obtained. When calculating the integral value of a rectangular area in the image, it can be calculated only by using the integral values of the current point and the three points in front of the current point. After extracting the features of education scene image, the convolution neural network in machine learning is used to realize education scene image recognition.

2.3 Using Convolutional Neural Networks to Recognize Educational Scene Images

Convolutional neural network is developed on the basis of artificial neural network. Convolutional neural network simulates the machine learning algorithm of biological visual nervous system, which consists of input layer, convolutional layer, pooling layer, full connection layer and output layer. Among them, the input layer can input 2d image data, the convolution layer and pooling layer are used to extract features, the full connection layer is used to integrate the features transferred from the convolution layer and pooling

layer, and the output layer can judge the relevant images by category to achieve the goal of image recognition. The most important part of convolutional neural network is the convolutional layer and the pooling layer [12].

The output layer of the convolutional neural network is designed as a support vector machine classifier, the features are extracted through the convolutional neural network, and then the support vector machine classifier is used for classification to achieve a better image recognition effect [13]. The convolutional layer is a related operation used to extract feature surfaces in a convolutional neural network. Many abstract feature surfaces can be obtained through convolution operations. At the beginning, the input is an $M \times N$ pixel input image, and then the convolution process is performed according to the specific convolution and $P \times T$ convolution. If the sliding step size of the sliding window is O, then the size of the feature image obtained by the convolution operation is O Calculated as follows:

$$S = [(M - P)/O + 1] \times [(N - T)/O + 1] \tag{9}$$

The receptive field of the convolutional layer is to only connect the local area of the upper layer to the neurons of the next layer through the convolution kernel. It can be regarded as the first process of the image. The connection of the local area of the receptive field can significantly reduce the parameters. It can be assumed that the size of the output feature map is $R \times S$, then the number of fully connected parameters is:

$$N = M \times N \times R \times S \tag{10}$$

The number of parameters connected locally is:

$$N = P \times T \times R \times S \tag{11}$$

It can be seen from the comparison of the formulas that the parameter reduction is reduced from the multiple of the original image size to the multiple of the convolution kernel size. Subsequently, through weight sharing, all output neurons share the same weight, then the number of parameters is directly reduced to the number of parameters of convolution kernel, and the number of parameters shared through weight is:

$$N = P \times T \tag{12}$$

Pooling layer is the relevant operation of convolutional neural network to sample the feature surface. Through the sampling operation of pooling layer, the feature surface with smaller scale and fewer parameters can be obtained. Figure 3 is the schematic diagram of the two pooling methods selected in this paper.

The pooling layer immediately follows the convolutional layer and can play the role of secondary feature extraction. Full connection using the full connection method will cause more parameters, then the calculation of the entire network will become larger and it is prone to overfitting, which can be handled by the method of Dropout abstention. The waiver formula is:

$$r = m\left(a\left(\sum_t \omega_t v_t\right)\right) \tag{13}$$

3	0	2	1
0	1	1	0
0	1	6	1
2	1	0	2

3	2
2	6

(A) Maximum pooling

3	0	2	1
0	1	1	0
0	1	6	1
2	1	0	2

1	1
1	2

(B) Average pooling
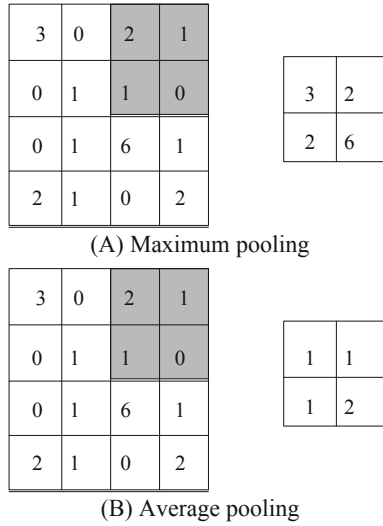
Fig. 3. Schematic diagram of convolutional neural network pooling operation

Regarding the waiver formula, consider a as the activation function, and then m as the waiver function. The value calculated by the activation function a represents the value of the neuron, and then the abstention function m represents whether the neuron is acting as an input. ω is the connection weight between the pooling layer and the fully connected layer in the convolutional neural network; v is the output weight of the convolutional layer.

In convolutional neural networks, both the S function and the R function are widely used, and the two functions have their own excellent performance. For the S function, the function is closer to the expression of the neuron and has good smoothness, but it has the disadvantages of gradient disappearance and slow convergence speed; for the R function, the function can have sparseness and faster convergence speed, which just compensates for S The lack of function, so combining the advantages of the two functions, a new activation function SR function is designed. The function formula is as follows:

$$f(x) = \max\left(\frac{1}{1 + e^{-x}} \left(x + \frac{1}{2} \right) \right) \tag{14}$$

The training samples of convolutional neural network are composed of education scene images with typical characteristics, and the convolutional neural network established above is used for image recognition of the training sample set. According to the recognition results of the sample set, the parameters of the convolutional neural network are continuously adjusted to minimize the recognition error of the convolutional neural network. The specific training and testing process of convolutional neural network is shown in Fig. 4.

The parameters of neural network with minimum identification error are taken as the final identification parameters. According to the above contents, the recognition results of education scene images are output by convolution neural network, and the content
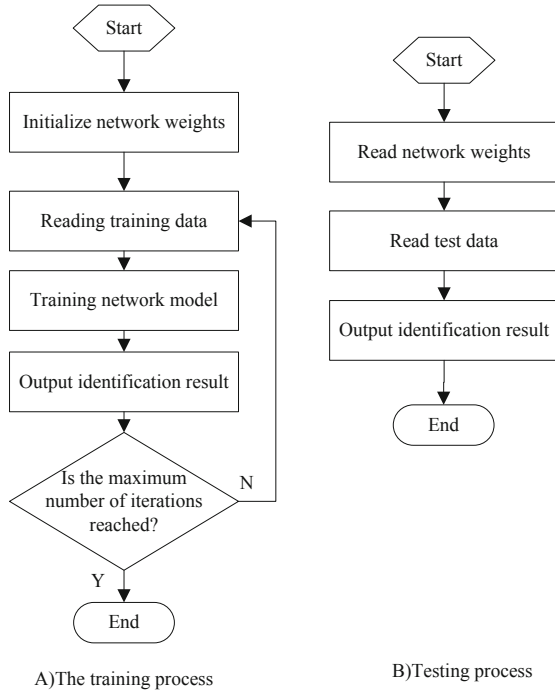
Fig. 4. Flow chart of training process and testing process of convolutional neural network

research of education scene image recognition method based on machine learning is completed.

3   Experimental Research

In order to verify the effectiveness of the above machine-learning-based educational scene image recognition method, the following experiments are designed. In order to avoid the uniformity of experimental results, the traditional recognition method based on image segmentation algorithm is compared with the image recognition method based on feature fusion.

3.1   Experiment Content

In the experiment, the recognition method designed in this paper was taken as the experimental group, the recognition method based on image segmentation algorithm was taken as the comparison group 1, and the image recognition method based on feature fusion was taken as the comparison group 2.

From the monitoring of different types of educational places in a university, 5000 educational scene images are taken as experimental data set. All images to be recognized have the same resolution and size to avoid interference to the experimental results. Of these, 2,700 were used for machine learning training and 2,300 for comparative testing.

Due to the large difference in the number of sample data, the training and testing of the model will be affected, and the scenario category with a large number of data is prone to overfitting, while the scenario category with a small number of data has little influence on the training process. Therefore, data enhancement operations such as magnification rotation, horizontal flip and vertical flip were selected for the scene image in the experiment.

The data set is divided into training set, validation set and test set according to the ratio of 0.6, 0.2 and 0.2. The training set is used to train the network model, the validation set is used to adjust the parameters of the network model, and the test set is used to test the recognition performance of the network model. Perform unified preprocessing on the experimental images, and annotate the images manually.

On three experimental platforms with exactly the same configuration, three image recognition methods of the experimental group and the comparison group were used to recognize the images. The recognition results of the 3 groups of recognition methods are compared with the real image annotations, and the correct recognition numbers of the recognition methods under different recognition numbers are counted, and the effective recognition rate of the recognition methods is calculated. It takes time to record the recognition of 3 groups of recognition methods at the same time, so as to compare the recognition efficiency of the methods. Comprehensive analysis of the two index data, evaluation of the performance of the identification method.

3.2 Experimental Results

Figure 5 shows the comparison of the effective recognition rates of the three groups of methods.
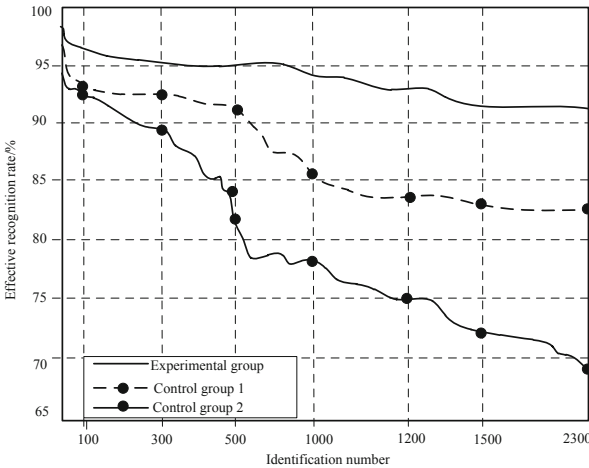
Fig. 5. Comparison of effective recognition rate of methods

Table 2 shows the identification time of the three groups of methods. The longer the time, the lower the identification efficiency.

Table 2. Recognition time-consuming comparison/s

Identification number	Test group	Comparison group 1	Comparison group 2
100	1.62	1.78	1.85
500	1.71	1.99	2.13
1000	1.84	2.07	2.54
1500	1.95	2.36	2.91
2300	2.16	2.63	3.52

According to the analysis of Fig. 5, with the increase of the number of samples to be identified, the effective recognition rate of comparison group 2 first decreased, followed by the effective recognition rate of comparison group 1. However, the effective recognition rate of control group 1 will not be reduced after it is reduced to 83%. The effective recognition rate of the experimental group method has been maintained at more than 92%.

According to the data in Table 2, under the condition of high effective recognition rate, the identification time of the experimental group method is the shortest and the identification efficiency is the highest.

To sum up, the education scene image recognition method based on machine learning designed in this paper has the advantages of high recognition accuracy and high recognition efficiency.

4   Conclusion

Image recognition is the research focus of machine learning and the foundation of machine vision. This paper studies the image recognition method of educational scenes based on machine learning. Correlative experiments prove that this method effectively improves the recognition effect of pictures. However, the amount of data in this experiment is relatively small. Since the larger the amount of data in the image library, the more essential the characteristics of things can be learned. Therefore, follow-up studies will be applied to larger-scale image recognition tasks.

References

1. Zhu, X., Fu, Y.: Development of Image–based classroom attendance system. Hubei Agric. Sci. **58**(03), 111–115 (2019)
2. Fang, G., Hu, Q., Fang, S., et al.: Face image quality evaluation in video stream and its application in classroom attendance system. Comput. Appl. Softw. **35**(10), 140–146+251 (2018)
3. Wang, Y., Zhu, L., Yu, Z., et al.: Segmentation and recognition algorithm for high-speed railway scene. Acta Optica Sinica **39**(06), 119–126 (2019)
4. Sun, N., Wang, L., Liu, J., et al.: Scene recognition based on privilege information and attention mechanism. J. Zhengzhou Univ. (Eng. Sci.) **42**(01), 42–49 (2021)

5. Li, X., Zhang, B., Sun, F., et al.: Indoor scene understanding by fusing multi-view RGB-D image frames. J. Comput. Res. Dev. **57**(06), 1218–1226 (2020)

6. Cai, X.: Automatic recognition simulation of non-orthogonal building images in complex scenes. Comput. Simul. **36**(10), 339–343 (2019)

7. Zhang, K., Hou, J.: A new image recognition method in improved convolution neural network. Sci. Technol. Eng. **20**(01), 252–257 (2020)

8. Liu, S., Wang, S., Liu, X., et al.: Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**(1), 90–102 (2021)

9. Ma, Y., Ma, H., Wang, Y.: An image recognition method based on CD-WGAN. J. Nat. Sci. Heilongjiang Univ. **38**(03), 348–354 (2021)

10. Xia, C., Meng, Q.: Fuzzy edge recognition of ship monitoring video image based on machine learning. Ship Sci. Technol. **42**(18), 85–87 (2020)

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. He, X., Yang, F., Chen, Z., et al.: The recognition of student classroom behavior based on human skeleton and deep learning. Mod. Educ. Technol. **30**(11), 105–112 (2020)

13. Liu, S., Liu, D., Muhammad, K., Ding, W.: Effective template update mechanism in visual tracking with background clutter. Neurocomputing **458**, 615–625 (2021)

Human Motion Recognition Method in Physical Education Based on Wearable Perception

Pengli Liu[(✉)] and Zhiqiang Wang

Sports Center, Xian Eurasia University, Xian 710000, China
`liupengli7563@163.com`

Abstract. The traditional human behavior recognition technology takes a long time to recognize human behavior. Aiming at the shortcomings of traditional methods, this paper designs a new human action recognition method in physical education based on efficient wearable sensor network. Firstly, the human action recognition characteristics of physical education teaching are extracted. Secondly, the human action recognition model of physical education teaching is constructed based on wearable perception, so as to realize human action recognition. The results show that the shortest action recognition time of physical education teaching people designed in this paper is 0.169 s. The designed physical education teaching human action recognition method has good recognition effect and certain application value.

Keywords: Wearable perception · Physical education · Human body · Action recognition

1 Introduction

As an advanced information processing tool, computer has provided information services for mankind since its birth. In recent years, with the continuous improvement of hardware manufacturing and software development technology, all kinds of miniaturized devices with communication, perception and computing capabilities have become popular [1]. The services provided by computers have also changed from the original proprietary and centralized computing services to more flexible services around people's needs. Computing systems and human users are more closely combined through natural interaction [2–4]. In this kind of computer system, an important supporting technology is the perception and recognition technology of human behavior, so that it can provide services according to human behavior [5]. On the basis of human behavior recognition, a series of typical applications can be developed. In academic circles, a lot of research has also been carried out on the development of such a new application system [6]. To realize human behavior recognition, we must solve many problems, including human behavior perception, behavior modeling and recognition. In perception, we should not only capture the data related to human behavior, but also provide a good user experience [7]. In modeling and recognition, we should fully consider the complexity of human behavior

and meet the application requirements in performance [8]. Human behavior recognition technology has attracted extensive attention in academia and industry because of its broad application prospects and rich research problems. Under this background, this paper makes a systematic and in-depth research on human behavior recognition in daily life.

Despite the support of rich perceptual means and computing resources, it is still a challenge to realize the recognition of human behavior [9]. Firstly, from the perspective of behavior perception, what type of data can best reflect the characteristics of behavior and is most conducive to subsequent recognition is still an unsolved problem. Secondly, from the perspective of perception methods, how to reduce the interference to people's daily life as much as possible on the basis of ensuring that the perceived data can be effectively used for behavior recognition Protecting people's privacy and improving the wearing experience [10] is still a very challenging problem; Third, from the perspective of human behavior, people's daily behavior is complex, the implementation of behavior is arbitrary, and there is complex interaction between people. How to model complex behaviors and identify them accurately is a very difficult problem; Finally, in some application scenarios, there are performance requirements represented by real-time performance of the recognition system. How to give consideration to the accuracy and performance of recognition at the same time needs to be deeply studied. Therefore, how to realize real-time human behavior recognition technology for complex behavior based on wearable sensor network is a problem worthy of research. The research on Chinese elementary mathematical knowledge extraction based on CRF algorithm proposed in document [11] introduces the traditional CRF process of named entity recognition. Then, an improved algorithm CRF++ for conditional field model is proposed. Aiming at the low recognition rate of named entities based on traditional machine learning methods, a post-processing method of entity recognition with automatic dictionary generation is proposed. After identifying mathematical entities, a pruning strategy combining Viterbi algorithm and rules is proposed to improve the recognition rate of basic mathematical entities. The distribution and application of the main additional error in the fractal coding method proposed in reference [12]. By extracting the original additional error value, a new fast fractal coding method is proposed. Then, using the extracted main additional error values, we analyze the distribution of these values. We found that different distributions of values represent different parts of the image. Finally, we analyze the experimental results and find some properties of these values. Experimental results also show the effectiveness of this method. And the existing human motion recognition methods have not been applied to physical education teaching, which cannot ensure the recognition accuracy and recognition performance on the basis of ensuring that the perceptual data is effectively applied to sports motion recognition.

In order to solve the above problems, this paper systematically studies the human behavior recognition technology in physical education teaching based on wearable sensor network. Aiming at the complexity of human behavior, the problem of behavior recognition in the case of single person complex behavior execution is deeply studied. Various data related to human behavior are obtained through various sensing means contained in wearable sensor networks, and the corresponding human behavior is recognized

through behavior model based on behavior recognition algorithm. The recognition algorithm based on EP and ESP is used to accurately segment the boundary between adjacent behaviors. The recognition of multi - person interaction behavior is studied from the perspective of group From the perspective of behavior recognition system performance, the problem of real-time behavior recognition is studied. From the perspective of perception means, the low-cost maintenance free human behavior perception technology based on passive wearable devices is studied.

2   Design of Human Motion Recognition Method in Physical Education Teaching Based on Wearable Perception

2.1   Extracting Human Motion Recognition Features in Physical Education Teaching

Recognition is one of the most frequent activities of human beings and other living bodies in their life. It is a basic ability that human beings must have to know, understand and adapt to the environment. In the above understanding, not as a special mode of observation methods and concluded that the provisions of the observation can be through biological capacity, such as eye view, hear, touch to get the feeling in the form of observations, can also is to use technical means, such as instrument is to obtain the data in the form of observations, likewise, recognition process can be a brain thinking, can also be algorithm is carried out. Based on the above basic understanding, a formal modeling of a recognition system to complete the recognition task is presented as shown in (1) below.

$$RS = S \circ R \tag{1}$$

In Formula (1), S stands for perception and R stands for recognition and identification process. In general, behaviors can be named according to people's subjective wishes, such as drinking water, eating, watching TV, etc. According to the above basic understanding of human behavior, when people perform specific behaviors, they will produce relatively significant external performance, which is defined as (2) below.

$$E = RA \longrightarrow O \tag{2}$$

In Formula (2), A represents the behavior set and O represents the external performance set. At this point, the framework designed is shown in Fig. 1 below.

As can be seen from Fig. 1, after obtaining various data related to human behavior through various sensing means contained in wearable sensor networks, these data need to go through a series of processing to identify the human behavior information contained therein. Although different types of perceptual data need different processing methods, the basic system framework is basically the same when considering the whole human behavior recognition system. Behavior recognition is often regarded as a special classification problem. A behavior recognition system conforms to the architecture of a typica pattern recognition system. The wearable sensor network at the bottom of the system is responsible for sensing all kinds of data related to human behavior, and transmitting the sensed data to the processing node through data communication

In formulas (3)–(6), T represents time, x_t represents frequency domain, and the interquartile distance is used to describe the dispersion degree of acceleration data. The calculation method is to arrange the acceleration data from small to large, and calculate the difference between the third quartile and the first quartile. The mean crossing times is used to describe the fluctuation degree of acceleration data with time, and its calculation method is to count the times of acceleration data crossing the mean.

In addition to the above common features, there are also some features that can be used, such as maximum value, minimum value, histogram, etc. the spatial domain features mainly include the correlation coefficient of acceleration data of multiple axes. These correlation coefficients describe the degree of correlation between accelerations in different body parts and directions. Given two random variables representing accelerations of different axes, X and y, the correlation coefficient (x, y) between them can be calculated by formula (7).

$$correlation(X, Y) = \frac{\text{cov}(X, Y)}{\delta x \delta y} \tag{7}$$

In formula (7), $\text{cov}(X, Y)$ represents covariance and δx, δy represents standard deviation. Since the frequency domain characteristics are periodic, the calculation formula is as follows (8).

$$Energy(X) = \frac{\sum_{i=1}^{T} |F_i|^2}{T} \tag{8}$$

In formula (8), $|F_i|$ represents the amplitude after Fourier change. When processing environmental and physiological index data, the common features of sensor data including numerical types such as illumination, temperature, humidity and pulse also include simple time-domain features and frequency-domain features. The range of features used is basically covered by the previous section. Because the number of environmental and physiological sensors involved in a wearable sensor network is generally one sensor, the use of spatial domain features is rare. For speech signal, in addition to ordinary numerical calculation, speech recognition is also done to help mark the category of behavior.

2.2 Construction of Human Motion Recognition Model in Physical Education Based on Wearable Perception

Wearable sensor network, due to its rich types of sensors, can comprehensively sense all kinds of data related to human behavior, including acceleration, temperature, sound, location, object use, etc. at the same time, it has the advantages of small privacy invasion, small volume, light weight, small living interference, simple maintenance, etc., It has become a very advantageous technology of behavior perception. In a wearable sensor network composed of active sensor devices, people wear various types of sensor nodes, which are equipped with various types of sensors including acceleration, temperature, sound, light and so on. After the sensor node obtains the sensor reading, it sends the data to the sink node through wired, wireless and other communication methods. The sink node transmits the received data to the processing node for subsequent behavior identification. The advantage of active sensor equipment is that it has strong sensor

ability and can sense various types of data. However, one of the obvious limitations is that active sensor devices need battery power to maintain their normal work. Therefore, a human motion recognition model for physical education teaching is constructed based on wearable perception.

The "quasi" mentioned above is the basic requirement to be met by a behavior recognition system, and its basic meaning is accurate behavior recognition. More specifically, when evaluating the functional indicators of a behavior recognition system, we try to carry out the evaluation with the help of confusion matrix. The confusion matrix is a square matrix. The rows in the matrix represent the recognition results. Based on the confusion matrix, many performance indicators described below can be calculated. Firstly, for the overall recognition accuracy, there are the following metrics. Recognition accuracy is one of the most commonly used performance indicators in behavior recognition. The definition of recognition accuracy is that the behavior instances correctly recognized account for the total number of all behavior instances. The recognition accuracy of the algorithm can be easily calculated on the basis of the confusion matrix. In addition to counting the recognition accuracy of the algorithm on the basis of examples, for the behavior recognition system whose input is continuous data flow, another common method of statistical accuracy is the time slice accuracy. Given a period of time series data, first slice it with a fixed length time slice, The essence of time slice accuracy is to calculate the proportion of correctly classified time slices.

For the identification of single person complex behavior, the following problem description is given: Based on the behavior data obtained by the perception platform, firstly segment the data to obtain the end behavior example. Considering the dynamics of behavior execution length, how to ensure the accuracy of data segmentation is the problem to be solved in this paper. Then, assuming that the system considers m behaviors, the problem to be solved in single person behavior recognition is to construct the recognition algorithm RA so that the recognized behavior category is equal to the corresponding real behavior category, considering various complex situations that human behaviors may include sequence, crossover and parallel execution in real life, The corresponding human behavior may be a single behavior executed sequentially, or a complex behavior composed of multiple basic behaviors executed alternately or in parallel. Due to the diversity of people's complex behaviors in daily life, it is impossible to show all kinds of complex execution of behaviors in daily life in training data. Therefore, it is necessary to design an algorithm framework that can identify complex human behaviors only by modeling and training basic behaviors.

Firstly, solve the problem of data source, complete the construction of single person behavior perception platform and obtain behavior data. On the obtained continuous sensor data stream, the data is segmented by using the 1-s sliding window segmentation technology to obtain the sequence of data units. After using classical methods to extract features from data units, the feature vector sequence is obtained. Based on the feature vector sequence, the sliding window model is used to segment and obtain behavior examples. Finally, the recognition algorithm based on EP and ESP, including sequential behavior model and complex behavior model, is used for recognition. In order to accurately segment the boundary between adjacent behaviors, the recognition results are used as feedback information to fine adjust the previous segmentation points.

In this process, how to model and identify sequential and complex behaviors is the biggest difficulty. To solve this difficulty, firstly, the problem of modeling and identifying the basic behavior of sequential execution is solved. Because of its strong ability to distinguish between categories and easy to understand, the classification method based on EP is more suitable for the use scenario proposed in this paper as a basic behavior modeling and recognition method. Considering the sequence characteristics of basic actions during behavior execution, in order to realize more accurate modeling of behavior, it is necessary to model the sequence of basic action elements in the behavior model. Therefore, based on the traditional EP method, the concept of ESP is proposed and used for behavior modeling. The complexity of single person behavior requires that the behavior model can model complex behaviors including sequential, cross and parallel execution. However, due to the diversity of complex behaviors, it is impossible to collect corresponding training examples for each complex behavior and model and identify them separately. Therefore, based on the basic behavior model of sequential execution, a training free complex behavior model acquisition method based on behavior length accumulation and optimistic score estimation is proposed. It realizes the goal of single person complex behavior recognition under diversified execution conditions without a large number of special training data.

After obtaining the perceived behavior data, the corresponding features are extracted from different types of data for recognition. One second of data is taken as a basic unit for feature extraction. Specifically, the extracted features include the following categories.

The features extracted from three-dimensional acceleration data include its mean value, variance, energy, entropy and correlation coefficient. Where the average value is the average value of acceleration over a period of time. Variance is used to characterize the stability of acceleration data. Energy characterizes the periodicity of acceleration data. The method of calculating energy is to calculate the square sum of the amplitudes of each component of the acceleration data after Fourier transform. Di is used to help distinguish different behaviors when the acceleration values of different behaviors have similar energy. The calculation method of Di is to calculate the average information Di of each component of acceleration data after Fourier transform. The correlation coefficient of acceleration is calculated by selecting the pairing between each axis of each acceleration sensor and calculating its correlation. The correlation coefficient is used to describe the correlation between the readings of axes in different directions of different acceleration sensors. In addition to the acceleration data, the average values of temperature, humidity and light intensity sensor readings are calculated as characteristics. For the micro RFID sensor worn by both hands, the article corresponding to the perceived tag is used as the characteristic value. When the tag is not perceived or the perceived tag reading error occurs, it is represented by null. For location data, the perceived room of the person is used as the characteristic value.

In general, a total of 75 features are extracted from the sensor data, of which 72 features are numerical, while the remaining three features (enumerative) can only be selected from several fixed values. For continuous sensor data flow, a 1-s sliding window is used to extract features, and every 1-s sensor reading is extracted into a 75 dimensional feature vector.

For numerical features, when the corresponding eigenvalues are extracted from the original data of the sensor, they need to be discretized for subsequent recognition. The discretization method based on the direct value is adopted. This method uses the category information to calculate the direct value of different discretized intervals, and selects the interval division method with the smallest direct value as the discretization method for continuous eigenvalues By using this method, the characteristics of 72 numerical types are discretized into 1046 disjoint intervals.

Firstly, the framework of single person complex behavior recognition algorithm based on EP is given. Given a new behavior observation sequence and assuming its time, the goal of the recognition system is to mark the observation value at each time with the correct category of one or more behaviors. Assuming that the current time is t, for each possible behavior a, first use a sliding motion with length L The window intercepts a section from the observed value sequence as an instance to wait for recognition, which is the average length of the behavior obtained from the training data. After obtaining an instance, calculate the possibility that the behavior category corresponding to the instance is a. after calculating the possibility for all possible behaviors, select the behavior with the highest possibility as the behavior corresponding to the current instance. Process After the current instance is completed, slide the sliding window forward and continue to process the next instance in a similar way. In order to accurately divide the boundary between two adjacent behaviors, a boundary detection and adjustment algorithm is proposed to realize the accurate segmentation of behavior boundary in the recognition process. The above processing steps will be executed repeatedly until the end of the input observation sequence.

In the process of recognition, the possibility of an instance corresponding to a certain behavior is estimated by calculating different scores. The method of calculating these scores is the core of this algorithm. In the algorithm, three different types of scores are proposed. The recognition model based on this is shown in (9) and (10).

$$E = \frac{\sum_{i=1}^{T} |F_i|^2}{correlation(X, Y)} \tag{9}$$

$$R = \frac{E}{correlation(X, Y)} \tag{10}$$

Given an instance of an observed value to be identified, the sequence of observed eigenvectors contains an instance of the behavior SA, followed by an instance of the behavior SA, starting at t time. During the recognition, the length Ls A of behavior SA is firstly used to intercept A section of observation value instance. In order to achieve accurate recognition, coverage score is introduced to measure the proportion of observation values unrelated to the current assumed behavior in A section of data.

2.3   Realize Human Motion Recognition

Combined with the sliding window based recognition algorithm and boundary detection algorithm proposed above, assuming that a sliding window contains a complete example of the current behavior, this combined algorithm can very effectively and accurately

identify all behaviors in the data set to be identified. But in daily behavior, even if it is the same behavior, the time used by people to perform it may be different every time. As mentioned above, this difference in time length, if not effectively solved, will have a serious impact on the performance of the whole recognition system. Therefore, when designing the final behavior recognition algorithm, we need to further optimize this problem.

The search space of sequential pattern mining can be expressed as a tree structure. In this tree, each node represents a sequence pattern, and the sequence pattern represented by each node is the prefix of the sequence pattern represented by its child nodes. It is to mine new patterns by continuously expanding the sequential patterns in the current tree. Given a sequence pattern, there are two ways to extend it. Add a new item to the last element E of S. sequence extension adds a new element containing only one item after the last element E of S. For element extension, the items in each element are arranged in dictionary order. Only when a new element is arranged after all existing items, the item is considered to be added to the element. For example, for elements, e can be extended with item C, but e is not considered to be extended with a. The reason for this is to avoid duplicate sequences in the search process.

In the process of mining, esp miner traverses the search space through depth first search. For the support of sequences, it is obvious that the support of a sequence pattern is always less than or equal to the support of its subsequences. ESP miner uses this feature to preliminarily prune the search space. Starting from an empty sequence $< >$, esp miner continues to add new elements to the current sequence for expansion until the support of the new sequence pattern is lower than the minimum support threshold.

As a preliminary exploration of human behavior recognition based on wearable sensor networks, firstly, the method of human behavior perception and recognition using active wearable sensors with rich perception types and intuitive perception data is studied. From the problem of behavior recognition, this paper first selects the recognition of single person complex behavior as the research content. The problem of single person complex behavior recognition based on active wearable sensor networks is discussed in detail.

First, data acquisition. In a typical active wearable sensor network, the data collected by each sensor node is often sent to the sink node through a single hop wireless network, and then transmitted to the processing node for processing. When considering the problem of single person complex behavior recognition, the focus is on how to identify, and the problem of data acquisition can learn from the existing structure of typical wearable sensor networks. Therefore, on this issue, we do not make too much in-depth research on the problem of data acquisition.

Second, data segmentation. In the traditional behavior recognition system, sliding window is a widely used data segmentation technology. However, when considering practical application scenarios, it is found that it is difficult to accurately locate the boundary between two adjacent behaviors by using a simple sliding window model. Inaccurate data segmentation will lead to incomplete or mixed behavior data contained in a behavior instance, which will affect the effect of subsequent recognition. Therefore, when considering the problem of single person complex behavior recognition based on active wearable sensor networks, how to accurately segment the data is the key problem.

Third, feature extraction. In the field of behavior recognition, the commonly used sensor types based on active sensor devices include acceleration sensor, sound sensor and RFID sensor. The features used in the existing work include various time-domain and frequency-domain features of acceleration sensor data, various time-domain and frequency-domain features of sound sensor, and tag information of RFID sensor. In general, the sensors currently used and the features extracted according to the sensors basically cover various types of data that can be used under the conditions of the prior art. However, because human behavior recognition is a new topic, there is no final conclusion on what features extracted by each sensor can best reflect the behavior information contained in the data. The quality of the extracted features directly affects the accuracy of subsequent behavior recognition, so this is a problem worthy of in-depth exploration.

Firstly, the overall framework of espar algorithm is given. The input of the algorithm is the continuous observation data stream obtained from the human wearable sensor network, and becomes the sequence of feature vectors according to the preprocessing. The operation of espar algorithm is divided into two stages: model training and behavior recognition. In the training phase, the basic behavior data set executed in the order of marked behavior categories is used to train the model and mine the corresponding esp. In the recognition stage, given a sequence s of eigenvectors, s is segmented by using a sliding window with length LX to obtain an instance of behavior, where LX is the average length of behavior a that may correspond to the assumed current instance. After obtaining the instance, the recognition algorithm is used to identify the behavior corresponding to the instance. Corresponding to the data of two adjacent behaviors, the boundary detection algorithm is used to adjust their boundaries and adjust the length of behaviors, so as to achieve accurate behavior data segmentation and recognition. This process can be regarded as a feedback process. The purpose of this process is to correct the inaccuracy of the simple sliding window based method in behavior instance segmentation. The above identification process will be executed circularly until the whole input data is identified.

Based on the above definition, the complete process of espar algorithm is given. Firstly, for each possible behavior a, a sliding window is used to intercept a segment from the eigenvalue sequence as the behavior instance to be identified. The size of the sliding window is la. After obtaining the instance, calculate the possibility of the behavior using the specified method. Specify the most likely behavior as the behavior category for this instance. In order to correct the inaccurate behavior boundary caused by using a simple sliding window to segment data, a boundary detection algorithm is used to adjust the behavior boundary. The algorithm continues to cycle through the above process until the whole input is marked with a behavior category. The length of behavior has a great influence on the accuracy of recognition. Therefore, for ESP model, a fine adjustment method for sliding window size is also proposed.

3 Experiment

In order to verify the recognition effect of the human motion recognition method designed in this paper, it is compared with the research on Chinese elementary mathe-matical knowledge extraction based on CRF algorithm proposed in literature [11], and experiments are carried out.

3.1 Experimental Preparation

The performance of the proposed recognition method is verified by experiments. This experiment aims to measure the performance of the proposed method through recognition accuracy, verify the impact of different system parameter selection on the proposed method through model and parameter analysis methods, and highlight the effect of the proposed method through comparative experiments. The selected sports behavior list is shown in Table 1 below.

Table 1. List of sports behaviors

Number	Behavior
1	Walk
2	Run
3	High jump
4	Long jump
5	Basketball
6	Football
7	Skipping rope

As shown in Table 1, the sports behavior at this time has complex randomness, and subsequent identification method detection experiments can be carried out.

3.2 Experimental Results and Discussion

The action recognition of physical education teachers designed in this paper and the research on Chinese elementary mathematics knowledge extraction based on CRF algorithm proposed in literature [11] are used to detect the time spent in the above behavior. The detection results are shown in Table 2 below.

Table 2. Experimental results

Number	The identification method designed in this paper is time-consuming/s	The recognition method proposed in reference [11] is time-consuming/s
1	0.456	1.459
2	0.674	2.164
3	0.169	1.942
4	0.176	1.692
5	0.369	2.061

(*continued*)

Table 2. (*continued*)

Number	The identification method designed in this paper is time-consuming/s	The recognition method proposed in reference [11] is time-consuming/s
6	0.582	1.169
7	0.364	1.698

It can be seen from table 2 that the shortest action recognition time of physical education teaching people designed in this paper is 0.169 s. The human motion recognition method designed in this paper has good recognition effect, short time consumption, good recognition effect and certain application value.

4   Conclusion

This paper summarizes the fundamental problems to be solved in human behavior recognition. The basic structure of a behavior recognition system is proposed and the problem is subdivided. By reviewing the existing related work, it is found that the existing work has made some progress in behavior data acquisition technology, data segmentation technology, feature extraction and behavior recognition algorithm. These existing technologies have important guiding significance for the work. However, when it comes to the behavior recognition technology based on wearable sensor network and meeting the principle of "how fast and save", it is found that the discussion of human behavior in the existing work is still in the initial exploration stage, and there is no in-depth research on the complexity of human behavior, the interaction between multi-person behaviors and the real-time requirements of behavior recognition Systematic analysis and research.

Aiming at the recognition of single person complex behavior, this paper deeply discusses the sequential, cross and parallel execution of human behavior in daily life, puts forward a pattern matching algorithm with strong ability to distinguish between categories based on EP, and puts forward the basic behavior model of sequential execution, Through behavior length accumulation and optimistic score estimation, the complex behavior recognition algorithm framework including sequential, cross and parallel execution can be identified without additional training. At the same time, this paper further discusses the sequence of basic actions in people's daily behavior, puts forward a new pattern esp after serialization expansion for EP, and applies it to the above algorithm framework to effectively improve the recognition accuracy of the algorithm. A human behavior perception platform based on active wearable sensor network is designed and implemented. On this basis, a single person complex behavior recognition prototype system is implemented and applied to verify the above single person complex behavior recognition methods.

References

1. Gao, Y., Ma, G.: Human motion recognition based on multimodal characteristics of learning quality in football scene. Math. Probl. Eng. **2021**(7), 1–8 (2021)

2. Gao, Z., Wang, P., Wang, H., Xu, M., Li, W.: A review of dynamic maps for 3D human motion recognition using ConvNets and its improvement. Neural Process. Lett. **52**(2), 1501–1515 (2020). https://doi.org/10.1007/s11063-020-10320-w

3. Yan, H., Zhang, Y., Wang, Y., et al.: WiAct: a passive wifi-based human activity recognition system. IEEE Sens. J. **20**(1), 296–305 (2019)

4. Li, Y., Miao, Q., Tian, K., et al.: Large-scale gesture recognition with a fusion of RGB-D data based on optical flow and the C3D model. Pattern Recogn. Lett. **119**, 187–194 (2019)

5. Lou, Y., Wang, R., Mai, J., et al.: IMU-based gait phase recognition for stroke survivors. Robotica **37**, 2195–2208 (2019)

6. Wang, Z., Fang, Y., Li, G., et al.: Facilitate sEMG-based human-machine interaction through channel optimization. Int. J. Humanoid Rob. **16**(04), 797–809 (2019)

7. Zhao, R., Ma, X., Liu, X., et al.: Continuous human motion recognition using micro-doppler signatures in the scenario with micro motion interference. IEEE Sens. J. **21**(4), 5022–5034 (2020)

8. Zhao, R., Ma, X., Liu, X., et al.: An end-to-end network for continuous human motion recognition via radar radios. IEEE Sens. J. **21**(5), 6487–6496 (2020)

9. Yang, J.: Study of human motion recognition algorithm based on multichannel 3D convolutional neural network. Complexity **2021**(6), 1–12 (2021)

10. Huang, R., Sun, M.: Network algorithm real-time depth image 3D human recognition for augmented reality. J. Real-Time Image Proc. **18**(2), 307–319 (2020). https://doi.org/10.1007/s11554-020-01045-z

11. Liu, S., He, T., Dai, J.: A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese. Mob. Netw. Appl. **26**(5), 1891–1903 (2021). https://doi.org/10.1007/s11036-020-01725-x

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. 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)

14. Liu, W.: Simulation of human body local feature points recognition based on machine learning. Comput. Simul. **38**(06), 387–390+395 (2021)

15. De-kun, J., Memon, F.H.: Design of mobile intelligent evaluation algorithm in physical education teaching. Mob. Netw. Appl. **27**, 527–534 (2021). https://doi.org/10.1007/s11036-021-01818-1

16. Liu, F.: Era of big data is based on the study of physical education teaching mode in MOOC. J. Phys: Conf. Ser. **1744**(3), 032008 (2021). (7pp)

17. Chang, J., Li, Y., Song, H., et al.: Assessment of validity of children's movement skill quotient (CMSQ) based on the physical education classroom environment. Biomed. Res. Int. **2020**(1), 1–11 (2020)

Acquiring the Cultural Characteristics of English Online Courses Based on Wavelet Transformation

Ying Lin[1,2](✉), Megawati Soekarno[1], and Yangbo Wu[3,4]

[1] Faculty of Psychology and Education, University Malaysia Sabah, 88000 Sabah, Malaysia
Linying0813@126.com
[2] School of Foreign Languages, Xinyu University, Xinyu 338000, China
[3] Faculty of Computing and Informatics, University Malaysia Sabah, 88000 Sabah, Malaysia
[4] School of Mathematics and Computer, Xinyu University, Xinyu 338000, China

Abstract. In recent years, China's online English curriculum has received extensive attention, not only because it is more in line with the educational development background of the new era, but also because this teaching mode has stronger flexibility, coupled with the assistance of information technology, which further improves the extraction effect of English curriculum cultural character. Therefore, the research on the extraction method of English online curriculum cultural character based on wavelet transform. Set the theoretical objectives and construction principles of cultural character of English online course, build a multi-level feature extraction model under wavelet transform, and extract the cultural character of the course by wavelet phase transform extraction method. The final test analysis shows that under the same test environment, compared with the traditional Gabor correlation extraction test group, the extraction test group of wavelet transform finally obtains the higher accuracy of cultural character extraction of English online course, which shows that it has better effect on cultural character extraction of English online course and has practical application significance.

Keywords: Wavelet transformation · English online course · Cultural characteristics · Characteristics acquiring · Acquiring method · Transformation structure

1 Introduction

English curriculum has always been in a key position in China's educational structure. The objectives and values of education are mainly reflected and implemented through different types of curriculum. Therefore, the renewal and reform of curriculum is the core content of educational innovation and development. At present, the English curriculum reform all over the world is in full swing. Profound changes have taken place in curriculum concept, curriculum objectives, curriculum content, curriculum implementation and curriculum evaluation. The curriculum research paradigm is gradually breaking through

the constraints of the traditional research framework and starting to make an in-depth exploration of curriculum theory from a broader perspective [1].

In addition, with the advent of the era of globalization, international economic cooperation and trade exchanges are becoming closer and more frequent, and a multicultural global village is taking shape [2]. In this case, relevant foreign scholars began to pay more attention to and examine the significance and value of English education. Foreign researchers attach importance to the research results of the character of English online culture courses through the method of observation and research, and believe that the purpose of English education is no longer limited to being able to speak a few foreign languages and translate, but to enable students to better survive and develop in a multicultural global village [3]. Looking back at the development of English Curriculum in China, we can find that in the past few decades, although the pace of reform has never stopped, there are few real major breakthroughs. The main methods used are genetic algorithm, neural network and other methods, resulting in the overall effect is not very ideal. The narrow horizon of English curriculum research is the main reason [4]. The narrow research horizon will inevitably not provide a broader development space for the development of English curriculum. Therefore, breaking the limitations of traditional research fields, starting with linguistics, culturology, sociology, anthropology and other disciplines, and adopting interdisciplinary research paradigm to carry out in-depth research on English curriculum is bound to become the development trend of English curriculum research in China. Nowadays, with the continuous development of China's information technology and network technology, it also expands the space and scope for the further development of China's English teaching, and further adjusts the overall English teaching mode by using relevant innovative computing methods. For example, English online teaching, multimedia network practice, flipped classroom, etc. while innovating the teaching mode, it also needs to be based on the promotion of cultural character, this paper focuses on a comprehensive and profound analysis of the cultural nature and cultural function of English online courses. In view of the inseparable relationship between language and culture, curriculum and culture, it is considered that English online curriculum itself has significant cultural character. Based on the interpretation of the connotation of cultural character of English online courses and the in-depth investigation of its current situation, this paper reveals the phenomenon of "loss of cultural character" in the research and development of English courses in China, and puts forward the trinity of "language culture curriculum" English curriculum theory construction framework. From the dimensions of curriculum theory research and English curriculum construction, This paper probes into the cultural reconstruction of English Curriculum in China.

Wavelet transform is a new transform analysis method. It inherits and develops the idea of localization of short-time Fourier transform. At the same time, it overcomes the shortcomings that the window size does not change with frequency. It can provide a "time-frequency" window that changes with frequency. It is an ideal tool for signal time-frequency analysis and processing. The main feature is that it can fully highlight the characteristics of some aspects of the problem through transformation, analyze the localization of time and space frequency, and gradually refine the signal through expansion and translation operation, so as to finally achieve the time subdivision result at high

frequency. In recent years, it has been widely used in the cultural character extraction method of English online courses, and achieved relatively good results [5]. Therefore, the cultural character extraction method of English online course based on wavelet transform is analyzed and studied. This paper discusses from different directions, constructs a multi-level feature extraction model under the wavelet transform, realizes the extraction of cultural character in the course by the wavelet phase transform extraction method, explains the close relationship between English course and cultural character extraction, and enhances the actual teaching quality level of English online course.

2 Acquiring the Cultural Characteristics of English Online Courses Based on Wavelet Transformation

2.1 The Theoretical Goal Setting of Cultural Characteristics in English Online Courses

The formation of the cultural character of English online courses has a solid theoretical foundation. It is mainly manifested in the following aspects: as a language course, English online courses are deeply influenced by linguistics; as a subject, English courses are influenced by course related theories; English online courses as a language teaching activity are deeply affected by social and cultural theories. The following is mainly based on linguistics, curriculum theory and social cultural theory, to explain the theoretical basis of the cultural character of the English curriculum [6]. Therefore, it is necessary to set the basic goal of extracting cultural characteristics in English online courses. In fact, language is the unique spiritual wealth of mankind. It is inseparable from life. In the whole process of human life, all the material and spiritual civilization of mankind are naturally connected with language. Any healthy and normal person has the ability to use language, and all human activities are inseparable from language. Language is a social product, which develops with the development of society, changes with social changes, differentiates with the division of society, and unifies with the unity of society. It is also the most important social phenomenon of mankind, and sociality is one of the essential characteristics of language. This attribute of language is closely related to human sociality. It can be said that people and society have created language. Therefore, the target setting range for the acquiring of cultural characteristics in English courses should be set to a wider range, and the actual cultural characteristics acquiring attributes should be established.

The establishment of the cultural characteristics acquiring attribute of English online courses makes it obtain a real cultural role. However, curriculum can not be directly equated with culture, because curriculum, as a part of culture, can not represent the whole culture. Obviously, through the analysis of the culture in the English online course, we can know that what can really become the course is only a small part of the culture. In real society, due to the lack of reasonable analysis of the cultural characteristics of English online courses, a considerable part of culture can not be included in the courses. Only according to the rationality of educational theory, can we really start from the rational construction of standardized curriculum in education and teaching, take English online curriculum as a cultural existence, restore the cultural characteristics of English

online curriculum itself, and establish the cultural noumenon nature of English online curriculum.

At the same time, it should be noted that a certain degree of autonomy is essential in the setting of cultural characteristics acquiring goals. From a cultural perspective, the tool-oriented English online course is a course that lacks vitality, and it is a course that has lost the foundation of survival [7]. Therefore, its goal can only be to cultivate people who become tools through indoctrination and training. The autonomy of the cultural character of the English online course shows that the English course is an independent cultural pattern. It is neither a copied "other culture", nor a vassal or puppet of other cultures, nor a passive cultural expression, but a self-conscious and self-contained cultural state. Looking back on the history of the development of English online courses, it is not difficult to find that there has never been an independent English course with independent cultural style [8]. The instrumental role of English online courses makes its own cultural qualities inherently obscured. On the surface, English online courses also seem to have the performance of processing and selecting culture, but this selection and processing is only formal, not the slightest. Therefore, cultural subjectivity is crucial to the reconstruction of English online courses [9].

The value of cultural characteristics autonomy of English online course is mainly reflected in its consideration of human subjectivity. Human subjectivity is not formed naturally, nor is it the result of self evolution, but needs education to play an important role [10]. In the era of carrying forward people's subjectivity and emphasizing people's dominant position, it has become a very important proposition to endow the English curriculum to cultivate students' individual subjects with the cultural quality of autonomy. It is one of the aims of contemporary education to cultivate learners' initiative and subjectivity and make learners truly independent individuals. The relationship is shown in Fig. 1 below:
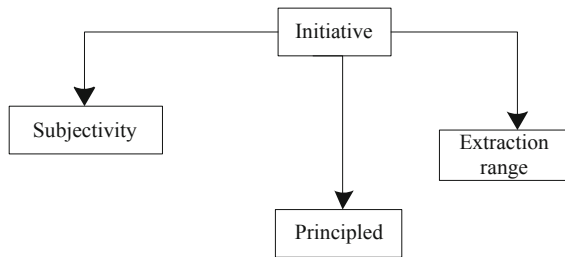
Fig. 1. Value relationship of character autonomy

Figure 1 helps to understand the above-mentioned structural relationship. At the same time, it is also the intended meaning of the independent cultural character of English online courses. The autonomy of the cultural character of English online courses is also manifested in its criticism and innovation of the culture it loads. In other words, in the process of reflecting and inheriting the external objective culture of English online courses, it is necessary to adopt a critical and dialectical point of view, instead of adopting an attitude of full acceptance or complete identification. It is necessary to critically reform

these external cultures. This is not only the intended meaning of cultural heritage itself, but also the mission of English online courses. At the same time, we must realize that if we want to use these external cultural materials to criticize, reconstruct and innovate the ideas of English online courses, without the autonomy of the cultural characteristics of the English courses as the foundation, it is difficult to achieve the expected acquiring goals. Therefore, when setting goals, we must ensure that they are diverse and stable.

2.2   Setting the Principles of Cultural Character Construction

The scientific and reasonable construction principle is an important guarantee to ensure that the cultural characteristics of the English curriculum can be demonstrated. To truly highlight the cultural attributes of the English curriculum in theory and practice, it must follow the integration, contextualization, equality, development and principles of practicality.

One is the principle of integration. For a long time, in our English online courses, the teaching content is relatively abstract and has no immediate effect on language acquisition. Therefore, the training involves little or no teaching of cultural knowledge, and education about culture is only English courses. The appendages in the Chinese language are independent of online English teaching and are in a dispensable situation. The principle of integration means that within the setting of English online courses, the selection and implementation methods, as well as the evaluation methods and content, etc. are organically integrated in the field of language, and cultural knowledge and language knowledge are integrated into the same. In a process, the current situation of the separation of the two is changed, so that the students' cultural awareness and language knowledge can be developed simultaneously. The fundamental purpose of foreign language online courses is to improve learners' language skills and cultivate learners' cross-cultural communication skills. Therefore, as a foreign language course, it must integrate different cultures and use a variety of comprehensive practical activities to finally achieve it. Unique educating function. Integration, in itself, is an important basic principle in the process of constructing the cultural characteristics of English online courses.

The principle of integration is also reflected in the integration of English teaching content and the acquiring of cultural characteristics as well as integration of cultural knowledge and language knowledge in English teaching. It has been fully demonstrated before that cultural knowledge and language knowledge are the key to truly mastering a language. If there is bias or neglect at any end, only by closely connecting culture and language can we really give full play to the practical function of cultural characteristics acquiring. Integration is also reflected in the integration between English online courses and other disciplines. As an English online course for foreign language learning, it can organically integrate the curriculum content of English and other disciplines according to the specific situation of different students and the reality of students in different grades, break the barrier between English and other disciplines, and establish a diversified cultural characteristics acquiring goal.

The second is the principle of contextualization. The so-called situational principle is a curriculum implementation environment that is close to students' daily life, resonates with students and conforms to students' interests in the construction of English online

curriculum cultural characteristics, that is, in the process of English online teaching and learning, pay attention to the intake of situational content, actively create a life-oriented learning environment and lead out the corresponding cultural characteristics. Guide students to connect what they have learned with the reality of life so that students can deeply understand the cultural connotation of the language they have learned. For English online courses, focusing on situational creation can enable students to constantly adjust and stimulate students' excitement points, attract students' attention and keep students in an excited and active state in changeable teaching activities. In this way, the teaching effect will get twice the result with half the effort.

The third is the principle of equality. Culture is shared by all mankind. To a certain extent, the cultures of different countries in the world share commonality, but there are also significant differences. Differences are caused by different social conditions and other factors, and are by no means caused by the advantages and disadvantages of different cultures. Therefore, when facing the impact of different cultures in various countries in the world, and when comparing and analyzing different cultures The characteristics of the culture should be fully valued and respected, and there should be no discrimination or ridicule of any culture, and prejudice against certain cultures.

English online courses not only have cultural attributes themselves, but are also an important way for students to understand world culture. Therefore, in the process of constructing the cultural characteristics of English courses, the principle of equality must be adhered to. The principle of equality refers to the ability to treat different cultural expressions with an equal attitude when dealing with different cultures in the world, to understand the differences among the cultures of the world, and to maintain an inclusive attitude to these differences. Specific to the cultural characteristics of English online courses, the principle of equality mainly includes the following aspects. The specific structural relationship is shown in Fig. 2 below:
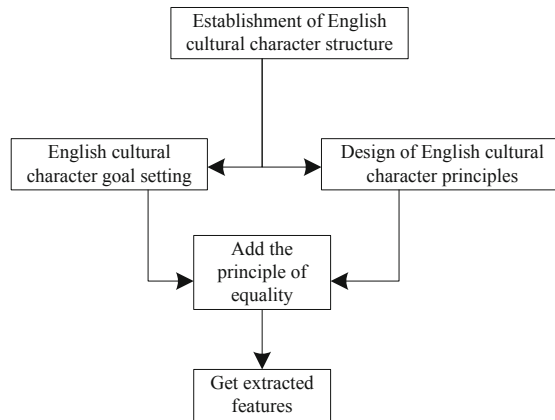
Fig. 2. Structure diagram of equality principle

According to Fig. 2, under the guidance of the principle of equality, we should respect cultural character. In the context of cross-cultural communication, rationality

and practicality are equal, and there is no distinction between high and low. Therefore, the subject of cross-cultural communication must treat the communication object and the cultural environment of the communication object with a respectful attitude, which is conducive to the gradual unification of the cultural character of using a foreign language.

The fourth is the principle of development. This development determines the influence scope of the cultural characteristics of English online courses, and English online courses are also developing. In addition, because the cultural character subject of English online course is developing continuously, we must adhere to the concept of development in the cultural character construction of English online course. The principle of development includes two aspects: one is the development of culture itself. For a long time, whether we regard culture as the material product or spiritual product of human society, we are used to understanding the connotation of culture from a static point of view and understanding culture as a noun. Of course, this explanation has its own reasons, but looking at the history of human cultural development, we can find that the cultural character itself is not invariable. At the same time, cultural character and its types also develop and change with the development of society. The second is diversified development. We should build a development system from three aspects: establishing scientific cultural teaching values, building a cultural education system with humanistic characteristics and implementing a reasonable cultural evaluation mechanism, expand English culture teaching and improve the cultural consciousness of English curriculum learning.

Fifth, the principle of practicality. This means that in the process of constructing the cultural characteristics of English online courses, we must participate in the practical activities of English language, gradually improve the expression ability of English language in the continuous process of English expression, and test the effectiveness of English learning in combination with the cultural characteristics of language. If we lack English practice, we will lose the motivation of English learning and development, there will be no improvement in comprehensive language literacy, and various mistakes will be made in the process of language use. Therefore, principle of practicality is the fundamental principle that must be adhered to in the process of constructing the cultural character of English online course. Its implementation are directly related to the implementation of other principles, and it also governs other principles to a certain extent. Only the language ability that can stand the test of practice can be called the real language ability.

2.3 Construction of Multi-level Characteristics Acquiring Model Under Wavelet Transformation

The diversity of cultural characteristics of English online courses is mainly reflected in its artistic character, cultural heritage character, communication character. The specific structure distribution is shown in Fig. 3 below:
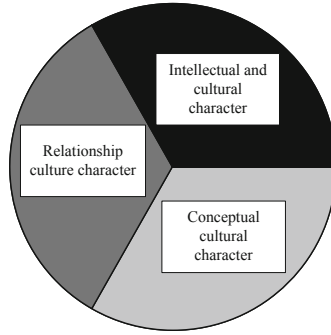
Fig. 3. Proportion of diversified value of English cultural character

According to the proportion analysis in Fig. 3, we can understand the importance of cultural character applied in teaching. Therefore, on the basis of the above information, the actual value of the proportion needs to be extracted first, as shown in Formula 1 below:

$$K = (2\phi + \frac{a}{3}) - 3\Re \tag{1}$$

In formula 1: K represents the actual value of the acquiring ratio, ϕ represents the acquiring range, a represents the variation coefficient, and $\Re$ represents the acquiring error. Through the above calculation, the actual value of the actual extraction proportion can be obtained. Take it as the corresponding proportion range, set it in the overall structure of extraction, and set the structure on the initial wavelet change basic extraction model. In order to improve the actual effect of extraction, an extraction wavelet transform matrix can be created in the model, and then calculate the relevant extraction transform trend value, as shown in formulas 2, 3 and 4 below:

$$M = \sqrt{3R + 1.25\chi} - \frac{t + 2e}{2} \tag{2}$$

$$B = \sqrt{6R + 1.25\chi} - \frac{t + 4e}{2} \tag{3}$$

$$D = \sqrt{9R + 1.25\chi} - \frac{t + 6e}{2} \tag{4}$$

In formulas 2, 3, and 4: M, B, D represent the acquiring transformation trend value, R represents the wavelet acquiring ratio, χ represents the extension transformation coefficient, t represents the target acquiring range value, and e represents the allowable limit error value. The above coefficients are set according to the calculation scale parameters. Through the above calculation, the actual extraction trend value can be obtained. Set it in the initial extraction model, summarize and integrate the set corresponding extraction principles, and finally form a more complete multi-level extraction structure.

In the context of quality education, teachers should keep up with the times and learn new educational ideas, such as flipped classrooms and micro-classes in order to improve

students' English cultural character. In addition, use flexible and effective evaluation strategies in the practice of English teaching, pay attention to the growth and progress of students, strive to increase their confidence, stimulate their curiosity, increase their interest in learning English, tap their potential, and truly improve English online teaching quality to achieve the goal of the new curriculum reform. Take the new acquiring target as the core, use the diversity of the model to process the existing acquiring tasks, use the multi-level principle to extract the structure, and add the wavelet transform processing structure to extract the characteristics of the cultural character of the English online course, complete the construction of multi-level feature acquiring model under wavelet transform.

On this basis, in order to further improve students' cultural character by using the model, teachers can also carry out a variety of English class activities, such as quiz game and English story telling competition. When these activities are held, some small prizes can be set to reward active participants and activity winners, so as to attract students' attention and stimulate students' interest in learning. These activities not only attract English loving students to actively participate, but also mobilize the interest of students with learning difficulties in English and give full play to their other strengths. In English teaching, English teachers should consciously introduce some short cultural stories to students, and prompt students to think in some key places and deliberately cultivate students' ideological and moral character, so that students will improve their cultural character unconsciously.

2.4 Wavelet Phase Transformation Acquiring Method to Achieve the Acquiring of Curriculum Cultural Characteristics

Before acquiring, it is necessary to align the saliency and make the corresponding analysis. The extracted feature vector calculated by the direct Gabysbdy wavelet transform method can effectively identify the image, and has rotation and scale invariance, but it has the defect of high dimension. Feature acquiring algorithms usually hope to not only extract features with significant identification, but also make the calculated extracted feature vector have low dimension. According to this idea, a feature acquiring method based on Gabysbdy wavelet transform is proposed in this paper.

Assuming that $PS = \{0, ..., P - 1\}$ is used to denote the acquiring set, and $QS = \{0, ..., Q - 1\}$ is used to denote the direction set, there are a total of $P \times Q$ types of acquiring and direction combinations. If CS is used to denote the combined set, it can be expressed as shown in formula 5:

$$S = (\omega + \frac{E}{2} - 4A) - \frac{1}{V} \tag{5}$$

In formula 5: S represents the combination set, ω represents the verification range, E represents the set condition of the wavelet transform, $\aleph$ represents the actual value acquiring, and V represents the application transformation acquiring range.

Through the above calculation, the actual combination set can be finally obtained. If the first four cultural character features are selected from ES, and the maximum energy level value is calculated, the cultural character features are extracted from the direction

corresponding to the level value, and the image identification is constructed based on these features, which may only be partially, it can be extracted in layers. For the acquiring of cultural character in English online teaching, it is also necessary to consider the phase of its acquiring, and the wavelet phase transform acquiring method is used to finally realize the acquiring of the cultural character of the course. According to the above analysis and the acquisition of data information, the calculation of the wavelet phase transformation ratio is carried out, as shown in the following formula 6:

$$G = \kappa + \frac{\sqrt{F}}{3} - 1.25\lambda \tag{6}$$

In formula 6, G represents the wavelet phase transformation ratio, κ represents the wavelet fixed value, F represents the actual range of the phase, and λ represents the phase processing acquiring coefficient. Through the above calculations, the actual wavelet phase transformation ratio can be finally obtained, which is set in the acquiring model to extract the cultural character of the English online course. From different angles, the wavelet phase transformation acquiring method is used to achieve multiple acquiring of English online course cultural character.

3 Method Test

This test is mainly to verify the accuracy of cultural character extraction of English online course under wavelet transform. The test is divided into two groups. One group is the traditional Gabor correlation extraction method, which is set as the traditional Gabor correlation extraction test group; The other group is the wavelet change culture character extraction method designed in this paper, which is set as the extraction test group of wavelet transform. The two methods are carried out at the same time, and the results are compared and analyzed.

3.1 Test Preparation

Select 6 classes of school A as the main goals of this test, set and calculate related indicators, as shown in Table 1 below:

Table 1. Experimental parameter table

Test index	Fixed index	Dynamic change index
Fusion ratio	0.5214	0.6741
Average score	89.21	86.31–92.15
Class size	65	65
Highest score	98.62	89–98.62
Minimum score	52.12	50.09–83.52

According to the data information in Table 1 above, the relevant test conditions can be obtained finally. After completion, set the wavelet transform range, as shown in Table 2 below:

Table 2. Wavelet transformation parameters

Range index	Standard range	Adjustment range
Extract the associated value	0.5214	0.116
Orientation equalization coefficient	16.37	20.92
Bias function	−0.3941	−0.1574
Dynamic imbalance value	4.4	4.12

According to the data information in Table 2 above, the actual wavelet transform range can be set finally. After completion, check whether the relevant information of the test is correct and in an error free state.

3.2   Test Process and Result Analysis

Start the test, arrange the English online courses of six classes evenly in the course structure, test their teaching effect at the beginning of each month, record the relevant results, and set the period as six months. Finally, verify the students' course completion rate and cultural character, extract the students' English course cultural character, and match the extraction results of different methods with the actual character results, to propose the verification results of character extraction accuracy. The details are shown in Table 3 below:

Table 3. Analysis of experimental results

Number of experiments	Traditional Gabor correlation extraction test group character extraction accuracy	Wavelet transform extraction test group character extraction accuracy
10	90.2	80.3
20	90.5	80.1
30	90.6	80.6
40	90.9	80.4
50	90.4	80.5
60	91.1	80.7

(*continued*)

Table 3. (*continued*)

Number of experiments	Traditional Gabor correlation extraction test group character extraction accuracy	Wavelet transform extraction test group character extraction accuracy
70	90.5	80.2
80	90.3	79.6
90	90.8	80.1
100	90.7	80.2

According to the data information in Table 3, we can finally draw the following conclusions: under the same test environment, compared with the traditional Gabor correlation extraction test group, the wavelet transform extraction test group finally obtains relatively high character extraction accuracy, which shows that it has better effect on the cultural character extraction of English online courses and has practical application value.

4   Conclusion

The cultivation of English cultural character needs a long process, which is essentially the cultivation of a good habit. It requires the continuous efforts and support of students and teachers in the process of learning. Therefore, this paper studies the cultural character extraction method of English online course based on wavelet transform, and verifies the performance of the method from both theoretical and experimental aspects. This method has a high extraction accuracy when extracting the cultural character of English online course, which is always maintained at more than 90%. Therefore, it shows that the character extraction method in this study can meet the needs of use. In the future research work, we should further improve the accuracy of cultural character extraction and improve the application value of the method.

References

1. Li, W., Xu, W., Zhang, T.: Improvement of threshold denoising method based on wavelet transform. Comput. Simul. **38**(06), 348–351+356 (2021)
2. Zhang, J., Liu, W., Zhang, Z.: Study on characteristics location of pantograph-catenary contact force signal based on wavelet transform. IEEE Trans. Instrum. Meas. **68**(2), 402–411 (2019)
3. Xie, G., Qian, Y., Yang, H.: Forecasting container throughput based on wavelet transforms within a decomposition-ensemble methodology: a case study of China. Marit. Policy Manag. **46**(12), 178–200 (2019)
4. 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
5. Too, J., Abdullah, A.R., Saad, N.M.: Classification of hand movements based on discrete wavelet transform and enhanced feature acquiring. Int. J. Adv. Comput. Sci. Appl. **10**(6), 83–89 (2019)

6. Meng, Y.: Research and analysis of intelligent English learning system based on improved neural network. J. Intell. Fuzzy Syst. **39**(2), 1–11 (2020)
7. Liu, S., Bai, W., Srivastava, G., Machado, J.A.T.: Property of self-similarity between baseband and modulated signals. Mob. Netw. Appl. **25**(4), 1537–1547 (2019). https://doi.org/10.1007/s11036-019-01358-9
8. Liu, S., Chen, X., Li, Y., Cheng, X.: Micro-distortion detection of lidar scanning signals based on geometric analysis. Symmetry **11**(5), 14–19 (2019)
9. Xu, X., Zhang, Y., Tang, M., et al.: Emotion recognition based on double tree complex wavelet transform and machine learning in Internet of Things. IEEE Access **33**(9), 142–147 (2019)
10. Li, H.J., Peng, M.: Online course learning outcome evaluation method based on big data analysis. Int. J. Contin. Eng. Educ. Life-Long Learn. **29**(4), 349–361 (2019)

Sentiment Analysis of Opinions over Time Toward Saudi Women's Sports

Norah J. Almateg$^{(\boxtimes)}$, Sarah M. BinQasim, Jawaher N. Alshahrani,
Ahad Y. Marghalani, and Zahyah H. Alharbi

Management Information Systems Department, College of Business Administration,
King Saud University, Riyadh, Saudi Arabia
{442920287,442920306,442920303,442920320}@student.ksu.edu.sa,
zalharbi@ksu.edu.sa

Abstract. The examining and promoting women's health is essential to the general well-being of society. As sports activities are important for maintaining physical, psychological, and social health, female participation in sports is one of the significant factors in achieving the goals of Saudi Arabia's Vision 2030. In this paper, sentiment analysis was used to compare the society opinions pre and post permit-ting Saudi women's sports, between 2017 and 2021. To identify the sentiment of a given tweet, a lexicon for the Saudi dialect was developed. In total, 12,000 tweets were collected and prepared. After data preparation, the tweets were reduced to 1,999 across all selected hashtags for this the initial study. We used four different hashtags related to Saudi women's sports, namely, (#Officially_female_sports_in_schools) represented as a Pre-Hashtag, whereas (#Women_Sport) and (#Female_Sport) as Pre and Post Hashtags, and (#Tahani _Alqahtani) as Post-Hashtag. The data in each hashtag were classified as positive, negative, or neutral. To build the sentiment classifier model, A Support Vector Machine (SVM) classifier was applied. The highest average accuracy was for the Pre-Hashtag with a score of 91%, followed by the Pre and Post Hashtag with a score of 85%. Finally, the Post Hashtag has the lowest score of 72%. The results show that 81% of the sample are positive. Accordingly, women have been becoming more motivated to engage in sports participation, as well as Saudi society is being more encouraging.

Keywords: Women's sport · Sentiment analysis · Natural Language Processing (NLP)

1 Introduction

Women's sports history started back in the nineteenth century, particularly when the second Olympic games took place in 1900 and female athletes participated for the first time [1]. Women's sport has always been recognized in some countries, but in countries such as Saudi Arabia, women's sport has been a controversial subject for many years. In

recent years, Saudi Arabia has been witnessing the rise of female athletes, and women are increasingly becoming open about participating in sports compared to previous years.

To analyze opinions in society toward women's sport, sentiment analysis was used in this research. Sentiment analysis involves using Natural Language Processing (NLP) techniques, computational linguistics, and text analytics to identify and extract subjective information from the source materials, aiming to determine the attitude of a speaker or a writer toward a certain topic or incident [2].

The Saudi community has seen an increase in the use of social media platforms such as Twitter [3]. Twitter is a powerful tool for disseminating information and an excellent source of opinionated text about a wide range of topics: politics, buisness, economic, and social. As a result, the NLP research community has become interested in studying this rich language resource. To the best of the author's knowledge, no previous study has used sentiment analysis to analyze Twitter users' opinions about women's sports in Saudi Arabia. Therefore, this initial study seeks to compare the society opinions pre and post permitting Saudi women's sports.

This paper is organized as follows: Sect. 2 covers the history of women sports in Saudi Arabia; Sect. 3 discusses related studies; Sect. 4 describes the study's methodology; Sect. 4.3 addresses the sentiment analysis process; Sect. 5 presents the results of the analysis; and Sect. 6 includes the conclusion and the future work.

2   History of Saudi Women's Sports

In 2013, Saudi Arabia's first dedicated sports center for females was opened in Al-Khobar city [4]. That same year, females in private schools were officially allowed to engage in sports, while in public schools, it was not permitted until 2017 [5]. The decision to allow physical education in 2017 was made by Saudi Arabia's Ministry of Education (MoE) in order to fulfill the goals of the Saudi Vision 2030 and promote healthy practices in society. Women in 2018 were also allowed to attend events at sports stadiums [6]. As part of Saudi Arabia's push to advocate for more female participation in sports and develop a more inclusive sports environment, the country in 2017 also appointed Princess Reema Bint Bandar, the first women in the kingdom to take on such an important role as the head of the Saudi Federation for Community Sports (SFCS) [7]. In her role as Vice President for Development and Planning at the Saudi Arabian General Sports Authority, the Princess has supported female participation in sports as well as contributing to the development of an inclusive sports environment for women. She has also worked on legitimizing women's gyms and focused heavily on encouraging women to use the streets and public parks to exercise. As she stated, "I've been telling women they don't need permission to exercise in public; they don't need permission to activate their own sports programs. And more and more, they are doing it. The choices that women have today are greater than yesterday and every day they will grow more" [8].

3 Related Work

In this section, previous research articles are summarized to give readers a general overview of women's sports in Saudi Arabia. Following this, studies that have used sentiment analysis in Arabic are examined.

3.1 Women's Sports in Saudi Arabia

Alruwaili in [9] studied the relationships between sport, gender, education, region, and religion in order to provide recommendations on how to improve gender inclusivity in sport in Saudi Arabia. The study focused on answering four questions. First, what are the dominant attitudes towards women's participation in sport in Saudi Arabia? Second, what are the key social, cultural, and civic issues that affect women's participation in sport in Saudi Arabia? Third, how do different interpretations of Islam influence attitudes towards women's sport in Saudi Arabia? Fourth, how are the ideas about women's participation in sport in Saudi Arabia changing? The researcher used a mixed-methods approach. In particular, an exploratory survey was undertaken via a questionnaire that was distributed to 890 individuals (444 responded: 196 women and 248 men). In addition, thousands of tweets from the micro-blogging site Twitter were examined with a representative sample of 96 selected for discussion. A thematic analysis of both the interview and Twitter data was performed. Alruwaili concluded that there is support for women to participate in sport and physical activity, most frequently on health grounds. This support was not explicitly constrained by male authority or Islamic teachings, although religiously-motivated reasoning was apparent in a proportion of the sample. More specifically, support was relatively high in the data across education levels, city of origin, and gender for women to participate in sport and physical activity. However, one restriction on this participation that most respondents across all data-gathering methods agreed upon was that women's participation in sport should be in accordance with the teachings of Islam, sex-segregated, and occur in a private setting.

Another study on women's sports was undertaken by Fakehy, Alfadhi, and Alotaibi [10]. The study focused on identifying the factors affecting the attitude of undergraduate female students in Saudi Arabia toward sports. The relationship between physical fitness, social experience, formal competition, physiological experience, and sports attitude of female students was examined. The researchers used questionnaire data to validate their hypotheses, which were as follows: H1) Physical fitness has a relationship with psychological experience; H2) Physical fitness has a relationship with sports attitude. 645 female students responded to the survey. After analyzing the data, the authors found that physical fitness, social experience, and formal competition positively influenced physiological experience, which in turn had a positive effect on the attitudes of female students toward sports.

Al-Haramlah, Merza, and Albakerin [11] investigated the level of physical activity among Saudi women and explored differences in terms of factors such as place of residence, age, weight, educational level, profession, and marital status. This study used an interventional approach to support efforts that encourage physical activity in Saudi women. A pilot sample of 80 females was interviewed to secure the validity and reliability of the preliminary instrument. The study was driven by two main questions. First, what

are the attitudes of Saudi women toward practicing physical activities? Second, are there any significant statistical differences in Saudi women's attitudes towards practicing physical activities that are based on the provinces of residence, age group, weight, educational level, nature of profession, or marital status? The researchers concluded that Saudi women in general are interested in practicing physical activities if opportunities and suitable facilities are available; they positively support the idea of providing women's sports clubs, as well as the idea of incorporating physical activities within girl schools. The findings also revealed that there are statistical differences in Saudi women's attitudes toward engaging in physical activity that are attributable to their provinces of residence, age, weight, educational level, and the nature of the profession.

Sayyd, Zainuddin, Zulaika, and Altowerqi in [12] discussed the current state of physical activity and sports activities in the Kingdom of Saudi Arabia (KSA) compared to other countries in the world. The authors also examined the issues that hinder the success and development of physical and sports activities development in KSA. The researchers conducted their review of the literature using ScienceDirect, Springer, the Journal of Health Sciences, and Google Scholar databases. A total of 58 articles were included in their research. The study revealed that the main factors hindering the success of physical and sports activities in the KSA's universities among males were lack of energy, motivation, self-confidence, and time, while among females, the factors were lack of social support and resources. The researchers concluded that there was no motivating environment that encouraged sports participation. They also found that there is limited action and initiative in terms of sports activities and sports participation in KSA compared to many other countries worldwide.

The research undertaken by Al-shahrani [13] aimed to determine the extent of sports practices among women in Saudi society by identifying the associated motives and obstacles to participation. Moreover, the researcher sought to develop suggestions and solutions to advance the culture of sports practice among women in Saudi society. The study focused on answering three questions. First, what are the motives for practicing sports for Saudi women in society? Second, what are the obstacles to women's sports in Saudi society? Third, what are the proposals and planning indicators for activating sports for women in Saudi society? The researcher conducted a questionnaire with 432 Saudi women in Riyadh. The study was based on a descriptive approach. Al-shahrani concluded that Saudi women are generally eager to maintain their fitness and increase their motivation for sports practice. The high cost of using fitness centers was one of the obstacles identified for women's sport participation in society. In addition, the absence of school curricula for sport reduces Saudi women's awareness of the importance of practicing sport.

Notably, several methods were used in the studies mentioned in this literature review, but none has applied sentiment analysis. Therefore, we found it worthwhile to apply sentiment analysis to analyze Arabic tweets in order to study Saudi society opinions pre and post permitting women's sports.

3.2 Arabic Sentiment Analysis

Al-Twairesh, Al-Khalifa, Al-Salman and Al-Ohali in [3] aimed to define a methodology that could be used in collecting and constructing a large dataset of Arabic tweets. Their

research was motivated by the lack of sufficient resources that allow the application of Arabic sentiment analysis. The extracted dataset contained almost 2.2 million tweets and was used to generate an Arabic corpus of tweets. The researchers cleaned and preprocessed the collected dataset by first filtering retweets, URLs, and mentions, and then by removing Arabic letters using normalization and tokenization. After manual annotation, the corpus was reduced to 17,573 tweets. Four labels were used: positive, negative, neutral, and mixed. The corpus included tweets written in Modern Standard Arabic (MSA) and the Saudi dialect. Three annotators were recruited to resolve the conflicts in annotation through majority voting. A list of guidelines for annotation was defined. After completing the annotation, a questionnaire was developed to evaluate the methodology. The results showed that two annotators stated that the guidelines were clear. The annotators were asked if the annotation of tweets was clear; all three annotators chose sometimes. The annotators were also asked which label was the hardest to determine; two annotators said mixed and one annotator said neutral.

Another study on Arabic sentiment analysis undertaken by Alqmase, Al Muhtaseb, and Raabaan [14] aimed to build a classification model using sentiment analysis by formulating sports Arabic text into fanatic and anti-fanatic contexts. Anti-fanatic text was defined as text that helps to decrease sports-fanaticism, while fanatic text increased sports-fanaticism. This was achieved by formalizing the social text into 21 fanatic and anti-fanatic contexts using proposed indicators. Fanatic indicators were aggression, agitation, hatred, and passion, whereas anti-fanatic indicators were adaptation, knowledge, respect, affections, and intimacy. Then, the authors developed a fanatic lexicon with 1,780 terms. After that, 919,000 domain-specific tweets were collected and labeled. To build a classification model, machine learning algorithms were applied. As a result, the best-built classifier achieved 91% accuracy. The proposed classification model can help governments to measure the impact of their efforts to reduce sports-fanaticism.

Ali in [15] aimed to conduct a comprehensive emotion mining and sentiment analysis task during the pandemic by collecting Arabic tweets related to online learning. The author extracted data using Twitter APIs, where the collected tweets were associated with seven different hashtags related to COVID-19. Then, the data were prepared for intensive preprocessing. This included removing hashtags, URLs, identifying emoticons, user mentions, and extra spaces. Also, punctuation was replaced with a single space. Spelling correction was applied to prepare the dataset for stemming. In addition, specific letters were normalized and stop words were removed. All emojis and emoticons provided by Twitter were kept and considered as a part of the texts. The most frequent emojis were defined, after which every emotion was replaced with its typical weight using the NRC lexicon. The different emotion annotations for a target term were consolidated by selecting the emoji with the highest weight. Two different datasets were used for the experiment. The datasets were collected between 20 September 2020 and 15 October 2020, and the total number of records across both datasets was 10,487. Finally, different classification algorithms were applied, including Naïve Bayes (NB), Multinomial NB (MNB), K-Nearest Neighbor (KNN), Logistic Regression (LR), and Support Vector Machine (SVM). The results showed that the proposed model performed well in analyzing people's perceptions about the coronavirus, achieving an accuracy of 89.6% using SVM classification. As for emotion analysis, anger was found to dominate the

tweets, followed by the fears surrounding the first attempt to engage in distance learning. This was mostly due to the lack of face-to-face communication, network system failure, ambiguity.

4 Methodology

This section presents the research methodology, which consisted of three phases: data collection, data preparation, and sentiment analysis, as shown in Fig. 1.
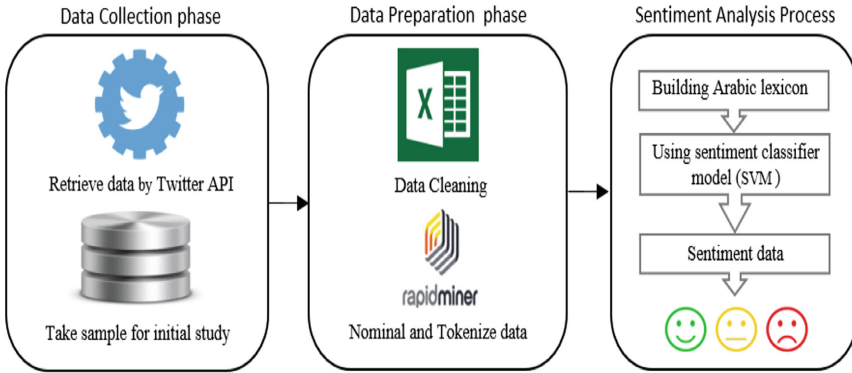
Fig. 1. The applied Arabic sentiment analysis methodology.

4.1 Data Collection

Data were retrieved from Twitter using the Python Tweepy library and an API key. 12,000 Arabic tweets were extracted using four different hashtags. For each hashtag,

Table 1. Hashtags used to form dataset

Dataset ID	Hashtag	Type	Date	Size
S01	#رسميا_رياضه_البنات_بالمدارس Officially_female_sports_in_schools	Pre	July 2017	670
S02	#رياضة_البنات #Female_Sport #الرياضة_النسائية #Women_Sport	Pre/ Post	2017 − 2021	660
S03	#تهاني_القحطاني #Tahani _Alqahtani	Post	30 July 2021	669

tweets were collected over a different period in order to compare the society opinions pre and post permitting Saudi women's sports, as well as changes in sentiment over time. The dataset size and period for each hashtag were identified as shown in Table 1.

4.2 Data Preparation

Since the text in tweets is known to be noisy, it should be cleaned and pre-processed before being analyzed. We started by removing the noisy tweets that contained ads, coupons, political materials, and unrelated tweets. We also excluded tweets that were posted on a date outside the covered period specified in this study. To accomplish the preparation and preprocessing tasks, we used a variety of tools, which are detailed in this section.

Microsoft Excel

We first cleaned the collected data using Microsoft Excel formulas. This involved excluding retweets and tweets containing URLs or media, as well as removing user mentions (@user) from the tweets along with hashtags (#), emojis, and punctuation. Then, we normalized several Arabic letters that have different shapes, including (أ, ة), by converting them into a single shape. For example, the different forms of "alif" (إ, آ, أ) were converted into (ا), and the letter "ta'a"(ة) was converted to (ه).

RapidMiner

The second step involved using the RapidMiner tool. RapidMiner is a software platform developed by the company that integrates machine learning, data and text mining, predictive analytics, and business analytics into a single platform. It is a sophisticated offering with over 1,500 drag-and-drop operators, allowing for the most efficient and quick data mining operations [16].

For our work, we utilized the "Nominal to Text" operator to convert all nominal attributes to string attributes, after which we used text processing techniques. "Tokenization" was applied, which breaks a document's text into a sequence of tokens. "Filter stopwords" was also used to filter Arabic stopwords from a document by removing every token which matched a stop word in the built-in stop word list. The stemming process was excluded due to its poor outcomes, which works to reduce the inflectional forms of each word into a common base word or root word or stem word. The authors in [17] demonstrated that using the stemming operator produces unsatisfactory results in the Arabic language.

After data preparation, the quantity of tweets across all four hashtags was reduced from 12,000 to 7,500. Then, a limited sample was randomly constructed for the initial study, consisting of 1,999 tweets covering all hashtags. It is worth referring to [18], which proved that the performance of a classification algorithm does not depend on the size of the datasets used.

4.3 Sentiment Analysis Process

Sentiment analysis is a technique of recognizing and extracting subjective information from source materials by Using NLP, computational linguistics, and text analytics. It

involves calculating people's perceptions, attitudes, and emotions regarding entities, persons, situations, events, themes, and their characteristics [19]. This study used a lexicon-based approach to extract sentiments from tweets. An Arabic lexicon was required to label the data before applying the classifier algorithm.

Lexicon-Based Classification

Lexican-based classification is a method of assigning labels to documents based on the number of words in two contrary lexicons, such as negative and positive sentiment [20]. Lexicon-based approaches require the manual collection of opinion words, and so they have been criticized for requiring too much human effort. Furthermore, the Arabic language is dialectically rich and its diverse structural properties in the various dialects need to be fully captured in order to derive maximum benefit from Arabic sentiment analysis, particulary for less formal channels such as social media [21].

It was challenging to find a lexicon that included Saudi dialects related to sports for this study. Two prior studies have constructed Arabic lexicons with the Saudi dialect [22, 23]. Accordingly, a lexicon was built for this study. First, we collected data from Twitter, particularly from the sport context, focusing on the Saudi dialect. Then, we tokenized these data. In turn, we classified the most important words into positive and negative. In the classification process, all authors participated, and then their results were compared to assure the polarity of the classification. Lastly, duplicated words were removed.

As a result of this process, 4,765 were the total number of words in the dictionary. 2,701 words were positive and 2,062 were negative. The dictionary is available on GitHub [24] for other researchers to use in future work.

Experimental Setup

The classification process was performed using the Python programming language (specifically, Jupyter notebook in Anaconda navigator). First, the built lexicon was applied on the dataset to assign the polarity of each word in the tweets to either positive or negative. Each tweet was then labeled based on its maximum polarity by counting the number of positive and negative words in the tweet. For example, if the tweet contained three positive and one negative word, it would be classified as a positive tweet (and vice versa); if the number of the negative words is greater than the positive words, the tweet would be classified as positive. Moreover, if the number of positive and negative words is equal, the tweet is classified as neutral. An illustration is shown in Table 2.

Once the dataset is labeled, Support Vector Machine (SVM) was applied to build the sentiment classifier model. SVM is a collection of supervised learning algorithms used for classification, regression, and outlier detection [25]. We used SVM due to its effective performance in Arabic sentiment analysis, according to [26].

We applied the train-test split() function to split that dataset into 70% training data (to train the classifier) and 30% test data. Moreover, to ensure accuracy and prevent overfitting, 10-fold cross-validation was performed on the SVM. This process was undertaken on all three datasets mentioned in Table 2.

Table 2. Sample of tweets labeling

Dataset ID	Hashtag	Pos count	Neg count	Label
S01	رياضه البنات دس السم العسل	2	2	Neutral
	Translated to: Female' sport put poison in honey			
S02	العقل السليم الجسد السليم	2	0	Positive
	Translated to: A healthy mind a healthy body			
S03	خسرت نفسها قبل ان تخسر المباراة	1	2	Negative
	Translated to: She lost herself before losing the game.			

5   Results and Discussion

To evaluate the performance of SVM methods for each dataset, we used different eval-uation metrics: Accuracy = (TP + TN)/(TP + FP + TN + FN), Recall = TP/(TP + FN), and Precision = TP/(TP + FP).

In these three equations, TP (true positive) is the number of correct positive predic-tions, TN (true negative) is the number of correct negative predictions, FP (false positive) is the number of incorrect positive predictions, and FN (false negative) is the number of incorrect negative predictions [27].

Table 3 shows the results for the SVM model for each dataset. As indicated, SVM demonstrated superior performance in 3-class classification, since it is able to differentiate between sentiment polarity.

Table 3. Results of SVM model

Classifier	Dataset ID	Accuracy	Recall	Precision
SVM	S01	0.91	0.90	0.90
	S02	0.85	0.85	0.85
	S03	0.72	0.71	0.71

In the next subsection, further results are discussed for each dataset, where each dataset represents a certain hashtag during a specific period, as shown in Table 1.

5.1   Pre-hashtag (S01)

To evaluate and confirm the sentiment analysis for society toward Saudi women's sports at the beginning of the decision of permitting sports in female schools. 670 tweets were randomly selected from our dataset as a balanced sample from the hashtag (رسميا_رياضه_البنات_بالمدارس#, which is translated as (#Offi-cially_female_sports_in_schools). This was a trending hashtag on Twitter in July 2017

when the Ministry of Education (MoE) in Saudi Arabia announced that sports classes would soon be introduced officially in female schools.

Table 4 shows that 90% of the sample was satisfied with the introduction of sport and physical activity in female schools, with 604 positive tweets. By contrast, only 10% of the tweets showed negative or neutral sentiments.

Table 4. Comparison of positive, negative, and neutral tweets for pre-hashtag

Dataset ID	Size	Positive	Negative	Neutral
S01	670	604	36	30

According to the results above, it is confirmed that there was a need for female sports classes in schools, referring to [9], which proved that there is support for women to participate in sport and physical activity. However, most of the tweets agreed that women's participation in sport should be undertaken in accordance with the teachings of Islam, sex-segregated, and occur in private settings. On the other hand, the authors in [13] proved that women are interested in maintaining their fitness and increasing their motivation for sports practice. The main obstacle that women face was the high cost of participating in the gym, with the absence of school curricula for sport having been overcome by the decision of the MoE.

5.2 Pre and Post Hashtags (S02)

In this hashtag, 660 tweets were randomly selected from the dataset as a balanced sample in order to evaluate and confirm the sentiment analysis for the pre and post hashtag (الرياضة النسائية# - رياضة البنات#), which translates to (#Female_Sport #Women_Sport). To compare societal opinions over time, two separate periods were selected (2017 and 2021). The results show that 85% of the sample was classified as positive, where some were supporters of women participating in sports, while others were female athletes already participating in sport.

Table 5. Comparison of positive, negative, and neutral tweets for pre and post hashtags

Dataset ID	Size	Positive	Negative	Neutral
S02	660	575	53	32

According to the results in Table 5, women in general are interested in sports participation in the case of the availability of facilities, a suitable environment, and necessary supplies and materials.

As mentioned in [12], that there is no motive environment that encourages sports participation, Therefore, it is noteworthy that in this study, the results suggest otherwise; in particular, comparing 2017 with 2021, women are becoming more motivated to engage

in sport, and the environment is increasingly becoming encouraging. Referring to [10], the study proved that physical fitness, social experience, and formal competition had a positive effect on physiological experience, which in turn positively influenced the attitudes of female students towards sports.

5.3 Post-hashtag (S03)

To evaluate and confirm the sentiment analysis for the post hashtag (تهاني_القحطاني#), which translates to (#Tahani _Alqahtani), 669 tweets randomly were selected from the dataset as a balanced sample.

72% of the results were supporters of Tahani's participation in the Olympics, while the rest criticized her. The results in Table 6 show unrealistic responses arising from the high level of rejection of Tahani's participation owing to religious and political factors, which results in getting out of context in a way that is not related to sport.

Table 6. Comparison of positive, negative, and neutral tweets post-hashtags

Dataset ID	Size	Positive	Negative	Neutral
S03	669	442	118	109

Another study conducted in 2020 [12], which mentioned that females lack social support and resources, is proved by the results above. Tahani Al-Qahtani, a Saudi Women who participated in the Tokyo Olympics in 2021, received different points of view on her participation: some were supporters, while others attacked her due to religious and political affiliations and motivations.

Additionally, some tweets mentioned that Saudi women's activities must be improved. For example, one tweet stated: "There is very little action in sports activities in the KSA compared to other countries in the world." The supportive tweets showed that women are inspired to participate if they have the opportunity and a suitable place, which was mentioned in [11].

6   Conclusion and Future Work

The aim of this study was to compare the society opinions pre and post permitting Saudi women's sports. We collected a dataset of 12,000 tweets that were prepared, consisting of text related to Saudi women's sports from Twitter. The tweets reduced to 7,500 across all four hashtags after data preparation. For the initial study, a limited sample was randomly constructed consisting of 1,999 tweets covering all hashtags. For the analysis process, we performed sentiment analysis using the SVM classifier, which showed a good performance. The highest accuracy was for the Pre-Hashtag with a score of 0.91, followed by the Pre and Post Hashtag with a score of 0.85. Finally, the Post Hashtag has the lowest score of 0.72.

Our results, which compared people's opinions toward women's sports from previous years with nowadays (2017 vs. 2021), in addition to the progress between these years.

The pre opinions dataset showed that people previously wanted women to engage in sport, as well as participate in physical activity during education; therefore, most of the opinions agreed with this decision, suggesting that this would be helpful and healthy for society. In contrast, as of 2021, citizens and residents across Saudi society have already engaged in many kinds of sport. They are also in agreement, support, and are proud of this engagement. This trend is reflected in the result of the third dataset, which is related to a Saudi woman, Tahani Alqahtani, who participated in the Olympics 2021. However, the outcome of the second dataset, concerning pre and post opinions sentiment in 2017 vs. 2021, showed that the societal interest has moved toward greater women's engagement in sport. Moreover, the available data suggest that sport is becoming a major aspect of Saudi women's lifestyles.

This paper is an initial study that analyzes opinions in Saudi society toward women's sports. Future work will continue to consider more hashtags that are recent, as well as comparing them with older dates, to find better results. Another recommendation is to use more general hashtags that target the society more. Also, include thematic analysis to be able to provide recommendations that improve women's sports in Saudi Arabia.

References

1. Vadhera, N.: Historical sketch of women's participation in sports: an overview. Int. J. Yogic Hum. Mov. Sports Sci. **3**(2), 417–422 (2018)
2. Tejwani, R.: Sentiment analysis: a survey. arXiv, pp. 1–3 (2014)
3. Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., Al-Ohali, Y.: AraSenTi-Tweet: a corpus for Arabic sentiment analysis of Saudi tweets. In: 3rd International Conference on Arabic Computational Linguistics, ACLing, Dubai, UAE, pp. 63–72 (2017)
4. Saudi Arabia opens first sports centre for women. https://gulfnews.com/world/gulf/saudi/saudi-arabia-opens-first-sports-centre-for-women-1.1192220. Accessed 21 Dec 2021
5. Saudi Arabia approves physical education program in girls' schools. https://www.arabnews.com/node/1127811/saudi-arabia. Accessed 27 Dec 2021
6. Saudi Arabia to let women into sports stadiums. https://edition.cnn.com/2017/10/29/middleeast/saudi-arabia-women-sports-arenas/index.html. Accessed 27 Dec 2021
7. Determined and Inspiring: 50 Saudi Women in Sports. https://www.abouther.com/node/15196/people/leading-ladies/determined-and-inspiring-50-saudi-women-sports. Accessed 27 Dec 2021
8. Ladies Leading Team, Princess Reema on the Future of Sports and Women in Saudi Arabia, About Her. https://www.abouther.com/node/9071/people/leading-ladies/princess-reema-future-sports-and-women-saudi-arabia. Accessed 16 Oct 2021
9. Alruwaili, M.: Females and sport in Saudi Arabia: An analysis of the relationship between sport, region, education, gender, and religion. Ph.D., eTheses, Faculty of Health Sciences and Sport, University of Stirling, Scotland, UK (2020)
10. Fakehy, M., AlFadhil, A., Hassan, Y., AlOtaibi, H.: Empowering Saudi women for sports as a response to KSA Vision 2030: attitudes of undergraduate female students toward practicing sports. J. Sports Psychol. **30**(2), 53–66 (2021)
11. AlHaramlah, A., Merza, H., Albakr, F.: Saudi women's attitudes toward physical activity: an interventional approach to improving future health. Asian Soc. Sci. **12**(7), 95 (2016)
12. Sayyd, S., Zainuddin, Z., Diyana, Z., Altowerqi, Z.: Sports Activities for undergraduate students in Saudi Arabia universities: a systematic literature review. Int. J. Hum. Mov. Sports Sci. **8**(1), 1–16 (2020)

13. Al-Shahrani, H.: The reality of women's sport in Saudi society. Int. J. Hum. Mov. Sports Sci. **8**(6), 525–535 (2020)

14. Alqmase, M., Al-Muhtaseb, H., Rabaan, H.: Sports-fanaticism formalism for sentiment analysis in Arabic text. Soc. Netw. Anal. Min. **11**(1), 1–24 (2021). https://doi.org/10.1007/s13278-021-00757-9

15. Ali, M.: Arabic sentiment analysis about online learning to mitigate covid-19. J. Intell. Syst. **30**(1), 524–540 (2021)

16. RapidMiner Documentation Operators. https://docs.rapidminer.com/latest/studio/operators/. Accessed 15 Nov 2021

17. Alami, N., Meknassi, M., Ouatik, S. A., Ennahnahi, N.: Impact of stemming on Arabic text summarization. In: 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), Tangier, Morocco, pp. 338–343 (2016)

18. Althnian, A., et al.: Impact of dataset size on classification performance: an empirical evaluation in the medical domain. Appl. Sci. **11**(2), 796 (2021)

19. Tripathi, P., Vishwakarma, S.K., Lala, A.: sentiment analysis of English tweets using rapid miner. In: 2015 International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, India, pp. 668–672 (2015)

20. Eisenstein, J.: Unsupervised learning for lexicon-based classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1 (2017)

21. Assiri, A., Emam, A., Aldossari, H.: Arabic sentiment analysis: a survey. Int. J. Adv. Comput. Sci. Appl. **6**(12), 75–85 (2015)

22. Almujaiwel, S.: GitHub - salmujaiwel/Covid-19_1M_Saudi_Tweets: 1M Saudi Tweets on Covid-19. GitHub. https://github.com/salmujaiwel/Covid-19_1M_Saudi_Tweets. Accessed 24 July 2020

23. Alrumayyan, N., Bawazeer, S., AlJurayyad, R., Al-Razgan, M.: Analyzing user behaviors: a study of tips in foursquare. In: Alenezi, M., Qureshi, B. (eds.) 5th International Symposium on Data Mining Applications. AISC, vol. 753, pp. 153–168. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78753-4_12

24. Github, Women Sport Arabic Tweets Dataset Dictionary (2022). https://github.com/jewelshahrani/Women-Sport-Arabic-Tweets-Dataset-Dictionary

25. Scikit learn, Support Vector Machines (2011). https://scikit-learn.org/stable/modules/svm.html

26. Alabbas, H., Al-Khateeb, M., Mansour, A., Epiphaniou, G., Frommholz, I.: Classification of colloquial Arabic tweets in real-time to detect high-risk floods. In: 2017 International Conference on Social Media, Wearable And Web Analytics, Social Media 2017, pp. 1–8 (2017)

27. Windasari, I.P., Uzzi, F.N., Satoto, K.I.: Sentiment analysis on Twitter posts: an analysis of positive or negative opinion on GoJek. In: 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), pp. 266–269 (2017)

"I Show You How I Solved It!"

Empowering Novice University Students to Learn Programming and Mathematics Through Self-produced Videos to Potentially Teach to Their Peers

Terry Inglese[1], Lukas E. Fässler[2], and Patrik Christen[3(✉)]

[1] Institute for Information Systems, FHNW University of Applied Sciences and Arts
Northwestern Switzerland, 4002 Basel, Switzerland
[2] Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland
[3] Institute for Information Systems, FHNW University of Applied Sciences and Arts
Northwestern Switzerland, 4600 Olten, Switzerland
`patrik.christen@fhnw.ch`

Abstract. A relevant concern of Java programming and mathematics instructors is that first-year college students usually have difficulties in grasping the abstract concepts of these two disciplines. Consequently, a meaningful part of students may fail to pass their core exams in BIT, *Business Information Technology*. To overcome this issue, two maths and programming instructors and a researcher in psychology of education, implemented the so-called *exploratory video-based instructional intervention*, through which BIT students were asked to explain specific Java programming concepts and to describe how to solve three maths exercises using self-produced videos. From a diagnostic perspective, the instructors were able: a) to recognise the correctness of the syntactic knowledge, the conceptual knowledge and the strategic knowledge of Java programming and b) to distinguish whether students were correctly applying the foundations of mathematics, which are essential skills for becoming a good programmer. The results of this experimental study showed that first-year students appreciated the production of these self-explaining videos, resulting in mastering complex abstract concepts in mathematics and in programming.

Keywords: Programming · Mathematics · Self-explaining · Teaching · Student video · Transfer learning

1 Introduction

Programming and mathematics are two essential disciplines for learners who study *Business Information Technology* (BIT) at the FHNW University of Applied Sciences and Arts, based in Northwestern Switzerland. Furthermore, being a BIT student also means becoming a future qualified expert, able to professionally communicate and work professionally with diversified clients and specialists within the two fields of Business and Information Technology. However, in the past years, as a trend in our BIT programme, approximately half of

the first-year students do not pass their programming exam. One of the problems we encountered, as instructors, was the difficulty for many students to relate new abstract concepts to existing knowledge. One of the characteristics of our students are their diversified academic and professional backgrounds. Not all of them have a strong foundation in mathematics and almost many of them never experienced programming before enrolling in the study program.

The focus of this paper is to show how instructors (two programming/mathematics instructors – the second and the third authors of this paper – together with an instructor of psychology of learning, didactics and business communication) introduced in two courses (programming and mathematics) the so-called *exploratory video-based instructional intervention*, through which students were asked: a) for the programming course, to explain a specific programming concept, and b) for the mathematics course, to describe how to solve three maths exercises, using – in both courses – self-produced videos.

The instructional aim of this *exploratory video-based instructional intervention* (which was for students an optional task to earn additional learning points, added to their final grade) was threefold: first, to promote the choice of using multimedia in self-explaining (and explaining to other peers) specific concepts and procedures that are rather abstract and theoretical. Second, we, as instructors, wanted to analyse in which ways the multimedia modality could promote a shift – *from* a concrete to an abstract transfer of concepts – in student learning. We used the *video-transfer-task* methodology, through which students were asked to explain (in business mathematics) linear algebra exercises and (in programming) some Java programming elements. Moreover, this instructional intervention provided a basis for understanding how we can better design our instruction, using multimedia videos as an additional learning output. In fact, from a diagnostic perspective these videos can help us recognise the correctness of the syntactic knowledge, the conceptual knowledge, and the strategic knowledge of Java programming and understand whether students are grasping the foundations of mathematics and programming. Thirdly, we believe that promoting the communication techniques of the so-called *procedural discourse* could be beneficial, considered that in the future, these BIT students will need to communicate and to collaborate with diversified clients and business professionals using information technology.

2 Theoretical Insights from Computing Education Research

Within the so-defined *computing education research*, based on the thoroughly and insightful literature review on student misconceptions and mistakes in introductory programming courses, Qian and Lehman [25] provided an insightful explanation about the types of difficulties students are usually encountering while learning the foundations of programming. As early as the 80ies, scholars in education and cognitive psychology started to focus on understanding the quality student's thinking while learning programming, including the analysis of misconceptions,

their "alternative conceptions", preconceptions and others. Specifically, in computer science education and programming, series of pivotal studies [1, 2, 16–19, 26] were conducted to comprehend students' incomplete understanding when learning computer programming. These studies are still very insightful nowadays, while today's instructors are searching for the right balance between designing the pedagogical teaching strategies and structuring the programming teaching content.

One aspect is the possibility to offer to novice learners more *concrete-abstract transfer* instructional opportunities to novice learners [17]. In fact, Qian and Lehman [25] (p. 1) underscored that "many sources of students' difficulties have connections with students' prior knowledge." The most common difficulties were found in *encountering unfamiliar syntax, natural language, maths knowledge, inaccurate mental models, lack of strategies, programming environments and teachers' knowledge and instruction.* More specifically, scholars including Qian and Lehman [25] recognised three main types of students' difficulties, based on general types of programming knowledge: first, *syntactic knowledge* (referred to knowledge of language features, facts and rules); second, *conceptual knowledge* (meaning how programming constructs and principles work and internal computer mechanisms) and third, *strategic knowledge*, especially the application of syntactic and conceptual knowledge of programming to solve novel problems, such as tracing and explaining code [25] (p. 16; see also [2]).

Specifically, for Qian and Lehman [25] (ibid.) *syntactic knowledge* referred to frequent errors, namely: the mismatched parentheses, brackets and quotation marks; using irresolvable symbols, because failing to declare a variable before using it, missing semicolons, using illegal start of Java expressions, which happens, because of unfamiliarity with the Java expressions; or the mistakenly use of the assignment operator (=) instead of the comparison operator (==). These mistakes are easily detectable and straightforward to address. The *conceptual knowledge* denotes the errors in the basic mechanics of programming languages, which can also lead to students' misconceptions that are related to student's mental models of code execution and computer systems (examples are the concept of variable, variable scope, conditionals, the looping construct, program execution, and object-oriented programming concepts and principles). The *strategic knowledge* is defined as the "conditional knowledge in cognitive psychology" [25] (p. 6), which is denoting the expert level of knowledge on planning, writing, debugging programs for finding solutions to new problems by applying syntactic and conceptual knowledge.

The authors emphasised that "students' difficulties in strategic knowledge are highly correlated to their difficulties in syntactic knowledge and conceptual knowledge" [25] (p. 6), caused by the inexact mental model of reference. Therefore, it is not enough to know only syntactics and semantics is not enough to be a good programmer; being able to understand the problem to solve and how to decompose it, is also paramount. In fact, "When a novice programmer debugs a program, he or she typically reads and traces code in a local manner – line by line – without a holistic view about programming ... (in) debugging

novices usually is not fixing the error but rather comprehending the program and locating the error" [25] (p. 7). Furthermore, Qian and Lehman [25] (ibid.) reported on the factors that contribute to these difficulties, such as: task complexity and cognitive load, natural language, existing maths knowledge, flawed mental models, inadequate patterns and strategies, environmental factors, and teachers' instruction and knowledge. For example: novices, which are learning to program, might be not familiar with all the requested programming languages' syntax, forgetting some elements, such as parenthesis, braces, semicolons, etc., therefore, *task complexity and cognitive load* might be a common error condition. In addition, the own *natural language* knowledge might hinder the use of the specific programming language, as the terms do not match.

Another related aspect is the level of proficiency in English, which can contribute to the success in learning to program (an example is the use of *then* in the *if-then-else* programming construct). The *existing maths knowledge* is an indispensable precondition of being able to program. If the levels of prior maths knowledge are low, this situation could become a source of misconceptions and of potential errors. Other related factors are *flawed mental models*, which could cause incomplete and inaccurate execution of coding and the tracing of the code. Consequently, *inadequate patterns and strategies for solving problems* might develop, due to the fragmentary mental model instead of a well-organised structure. In such cases, patterns and strategies for solving programming problems might be applied, failing "to reason at an abstract level when comprehending, writing and debugging code ... the programming knowledge gained by novice programmers usually is not organised into meaningful patterns and the connections between pieces of knowledge are not well established." [25] (p. 9). The *environmental factors* are language features related to the programming environment, which are supporting the debugging activities. Finally, the *teacher's instructional and knowledge* is referring to a type of instructor that *teaches rules rather than reasons*, promoting the memorisation of syntactic knowledge, rather than understanding and reason, thereby – unintentionally – fostering the development of incorrect mental models, which are then difficult to change.

Qian and Lehman [25] underlined that the dissemination of new instructional approaches and tools to overcome these issues have been limitedly researched and studied. In fact, they warmly advised going beyond simply documenting and addressing students' difficulties and incorporate the *conceptual change theories* and also consider the pedagogical content knowledge (PCK). They refer to two theoretical perspectives within the *learning science* and the *scientific disciplines*, which apply this scientific knowledge in everyday life. These two theoretical perspectives are: *knowledge as theory and knowledge as elements* [24] (p. 351).

Historically, the *knowledge as theory* has been the predominant approach, and it refers to a student's knowledge accurately represented as a coherent unified framework of a *theory-liked knowledge*, involving a student's interpretation of subordinate models and ideas. It refers to the Piaget's concepts of assimilation and accommodation [24]. "If a learner's current conception is functional and if the learner can solve problems within the existing conceptual schema, then the

learner does not feel a need to change the current conception" [24] (p. 352). The learner needs to be dissatisfied with this initial conception, to abandon it and to accept a conceptual change, and often novices are not equipped to do so.

Knowledge as elements, on the other hand, refers to an ecology of quasi-independent elements, "where a combinatorial complexity of the system constraints students' interpretations of a phenomenon ... student's understanding in terms of collections of multiple quasi-independent elements" [7] (pp. 352–354) This definition was proposed by diSessa [7], "the knowledge structures of novices consist primarily of unstructured collections of many simple elements that he calls p-prims (phenomenological primitives) ... developed through a sense-of-mechanism that reflects our interactions with the physical world ... do not have a status of a theory ... are generated from a learner's experiences, observations, and abstractions of phenomena... have more exploratory power than conceptual framework theory." [7] (p. 355).

Summarising, through the *knowledge as elements*, composed by facts, narratives, concepts, and mental models at various stages of development, novices connect and activate these knowledge elements, based on the relevance of the situation. During the conceptual change process, which resembles a *piece-by-piece* progression of knowledge building, these elements find their way to learning and to making sense, compared to the *theory-liked knowledge* process. Therefore, these two models of knowledge define different ways of designing curricula, with the focus on helping students organise and reorganise their learning process.

3   Practical Insights from Cognitive Psychology

The *knowledge as elements* theory has some elements in common with the *generative learning* theory. In fact, according to Fiorella and Mayer [10], there are eight ways to promote generative learning and conceptual change in learning, which means when learners are actively making sense of the information to be learned. These are: learning by *summarising*, by *mapping*, by *drawing*, by *imagining*, by *self-testing*, by *self-explaining*, by *teaching to others*, and finally by *enacting*. The self-teaching technique, using videos to explain concepts in mathematics and in programming (as in our cases), has been scientifically proven to be very effective in previous research cases (see [9,12]).

Additionally, based on his four decades of research in the establishing his science of instruction, Mayer [16] (p. 121; see also [1]) defined *meaningful learning* the process by which the learner combines the new learning material with their *schema* (the already existing knowledge), through the process of assimilation. As early as the early 1980s, he asked himself whether concrete models – *advance organiser* – for novice programmers could support the meaningful learning of computer programming, while promoting more understanding than memorisation, because "the payoff for understanding comes not in direct application to the newly learned material, but rather in the transfer to new situations." [16] (p. 123). The same conclusions were reached by Soloway [26] (p. 852), when he differentiated the difficulties, novices had between *syntax* and *semantics* of programming. Based on the recent scientific contributions of Fiorella and

Mayer [11] and by Bétrancourt and Benetos [3], important criteria for designing instructional videos were considered and shared within the instructional design research community. An instructional video "is intended to help people learn targeted material ... it is a form of multimedia instruction" [11] (p. 465), including visual material (video) and verbal material (voice and/or onscreen text). Beneficial is the *segmenting*, being able to break down the multimedia presentation in meaningful segments, to provide learners with a control of their own learning process. Hindering aspects of instructional videos are *faces on screens*, *adding practice without feedback* and *inserting pauses* [11] (p. 465).

Soloway [26] stated that a program has two audiences: *the computer*, because the instructions in a program turn the computer into a mechanism that dictates how a problem can be solved and *the human reader*. Therefore, the programmer needs to have an explanation as to why the program solves the given problem, "learning to program amounts to learning how to construct mechanisms and how to construct explanations. In teaching programming and problem solving in general a key objective is to develop useful methods of abstraction. A hallmark of expertise is the ability to view a current problem in terms of old problems, so that solution strategies can be transferred from the old situation to the current situation." [26] (p. 853). Programming is a *design discipline*, which is producing an artefact that performs desired functions. Being an artefact leads to the concept of *mechanism*, which specifies a chain of actions that promotes some desired effects. In these processes *change* is the norm and not the exception; therefore, programmers need to provide the evidence of *how* and *why* the artefact was designed in a certain way, to ensure that the next programmer can effectively modify the artefact, if needed. This complexity also includes being able *to explain*, because programming processes are basically *mechanism and explanations*. Nevertheless, syntax and semantics are not enough; on the contrary, being able to break problems down into sub problems, and knowing their interconnections, this is also important for mastering problem-solving-programming tasks.

This exploratory video-based instructional intervention provided a basis for how we can better design our instruction, using these multimedia videos, as an additional learning output. In fact, from a diagnostic perspective, through these videos help us: a) to recognise the correctness of the syntactic knowledge, the conceptual knowledge, and the strategic knowledge of Java programming and to understand whether students are grasping the foundations of mathematics; b) to understand how students elaborate their contents from a *procedural discourse* communication, and how to teach this skill. Another aspect that we wanted to understand is the level of the so-defined *procedural discourse*, which the students intuitively used to create and to comment their videos.

Procedural discourse means *the how to do-to perform definite communication's procedures* [8, 22]. According to Farkas [8] (p. 42) *procedural discourse* relates to written and/or spoken discourse guiding people in the performance of a certain task. *Procedural discourse* is more than a logical and structured information. Based on two recognised perspectives: a) human problem solving in the context of systems theory [23] and b) rhetoric and the source of credibility [6],

procedural discourse stems from the purposeful human behaviour, intended as the "*telling someone who is in one set of circumstances how to transition to another set.*" [8] (pp. 42–43). *Procedural discourse* implies four steps: 1. a desire state, which is the goal of the user; 2. the prerequisite state, as a condition for moving towards the goal; 3. the interim state, referring to the milestones and sub-goals to reach through our actions; and finally, 4. the unwanted states, which the user wants to avoid. This type of communication, *rhetorical* by nature, is procedure-based, and guided through tasks to accomplish. By analysing the way how students explained their programming and maths videos, we learn how to integrate this *procedural discourse* component, as a communication *added value skill* to be develop.

In the next section, we will provide more context about the two courses: its instructional design and the methodology used, students' results and how they approached the *exploratory video-based instructional intervention*.

4   The Two Courses: Programming and Business Mathematics 2

4.1   Participants

The two courses *Programming* and *Business Mathematics 2*, were held in the second semester of 2019. Two different classes were enrolled in the two courses. 17 students participated in the programming course, which covers the basics of Java programming, including object-oriented programming. In the mathematics course, 18 students were enrolled, learning the basics of linear algebra with an emphasis on computational approaches.

4.2   Method

Based on the instructor's motivation to drastically reduce the gap between students who are successful in programming and maths exams and those who fail, and based on the fact that programming is strongly related to maths, the instructors of this paper designed and applied the *exploratory video-based instructional intervention*, offering students the option of producing self-video-based explanations of programming and maths concepts, together with a flipped classroom intervention, using the *E.Tutorial*, which will be explained later. The foremost instructional aim was to understand how students can represent maths and programming concepts and exercises, using multimedia explanations and whether they would gradually undergo cognitive change through these video-based interventions *from a concrete to an abstract transfer* of programming and maths concepts [4, 13, 14].

In the programming class, a 4-step-model, a blended learning concept, called *E.Tutorial* and developed by the ETH Zurich, was used. First, students read short text documents of the most important definitions, concepts, and tasks (referred to the *SEE* part). Second, guided by an electronic tutorial, the concepts of the *SEE* part were applied (referring to the *TRY* part). Third, these

concepts were then applied individually to a problem-based task, transferring of the knowledge (the so-called *DO* part). Fourth, individual solutions of the *DO* part were presented and discussed in a face-to-face meeting with the instructor (the *EXPLAIN* part). Some topics were extended with traditional lecturing. It can be sustained that the exam's outcomes improved, compared to the previous years, with a failing rate of 30% and an average grade of 4.25 (maximum grade 6). Although a rather traditional teaching approach was used in the maths class, approximately half of the in-class time was invested to solve maths problems. The failing rate was 0% and an average grade of 5.5 (maximum grade 6) was achieved.

To keep track of their own learning experience and to engage students in this process, we asked them to produce the *exploratory video-based instructional interventions*. In the programming class, they explained a concept of Java programming and how to program it. In the mathematics class, they were asked to create three videos explaining how to solve three exercises. The creation of the videos was rewarded with additional learning points.

5   Results

5.1   The Programming Course

Using a post-questionnaire, students expressed their opinions about the entire course. The *E.Tutorial* was very much appreciated, because of the freedom it provided to study alone, at their own pace, introducing one Java concept at the time to promote understanding of specific concepts and methods. Correspondingly, self-learning was perceived correspondingly as a challenge for other learners as they appreciated the teaching option of experiencing a live-coding explanation on a screen in the classroom. In terms of exam performance, the programming and maths instructor appreciated the fact that the exam's outcomes improved, compared to the previous years, reaching an average grade of 4.25 (maximum grade 6). Certain exam questions were easily accessible, because students could have solved these even applying rote learning. However, one question was designed as a *transfer knowledge* one. Students received a description in a non-programming terminology of what a program should do, and they had to implement its functionality in the form of a method. They were aware of the exam format.

From the video production perspective, no student complained about producing a video with their own tools. On the contrary, 16 out of 17 students took the chance to earn learning points by creating a video, where they needed to explain a particular Java programming element and, if possible, to relate the abstract concept to a familiar and concrete object. The duration of these videos ranged from 30 s to 5 min. To analyse the videos, we used the three-part framework of Bétrancourt and Benetos [3]: the *representational approach*, the *cognitive approach*, and the *instructional approach* (see also Carliner's physical, cognitive, and affective information design framework [5]).

From the *cognitive information design approach*, the instructor analysed the presented content for correctness: 12 out of 16 were correct. For only five out of

16 there was a relation to concrete objects; and finally, only in three cases (out of 16) abstraction was mastered.

From the *instructional approach*, the programming videos showed the following features: two out of 16 introduced a definition; 13 out of 16 provided an explanation; 12 out of 16 introduced an example; two out of 16 used highlighting features, such as arrows, to let the audience follow the content presented within the videos; four out of 16 used rhetorical questions, like: *what is a variable*?

From the *representational approach*, six out of 16 addressed themselves to the audience, either with using *I*, *you* or *we*; 11 out of 16 used background music; students represented themselves, using animated-anonymous personas provided by the software they used; or they represented students in a classroom at the blackboard, personas with names, personas in stressful situations, like being a procrastinator and needing to deliver the programming homework, or creating a funny Christmas story, where a student needed to solve a programming problem, before being able to finally buy a gift for his girlfriend. Some students added slides with explanations from educational material, without commenting or elaborating on these contents. Some used their own words in explaining concepts. Some video-animations were completely disconnected, as if these presented their own mental schema, completely full of useless elements, disordered, superficial and with no context or relations to the instructional content.

Summarising, we were not fully satisfied with the video-based contents produced by the students. These were merely descriptive contributions, somewhat superficial, and not highly relevant to the course. Nevertheless, only the one contribution really surprised us. In fact, one out of 16 created a serious, but also intriguing and humorous video, featuring three students, becoming themselves examples of Java code and its explanation (Fig. 1). As mentioned previously,
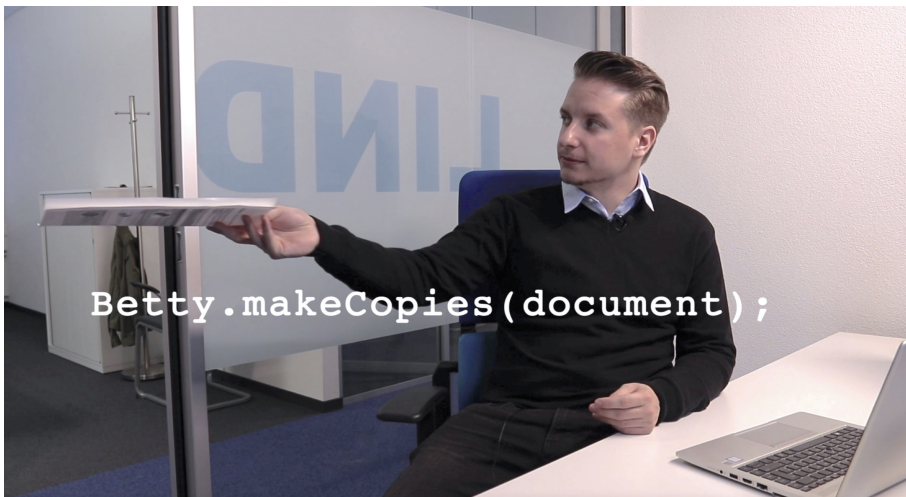

Fig. 1. A screenshot of an example video explaining abstract programming concepts on the basis of a concrete story and concrete objects.

according to Fiorella and Mayer [10], there are eight ways to promote *generative learning*, which means when learners are actively making sense of the to be learned information. These eight ways are: learning by *summarising*, by *mapping*, by *drawing*, by *imagining*, by *self-testing*, by *self-explaining*, by *teaching to others*, and finally by *enacting*. In this one case we considered this one video as a combination of *self-testing*, *self-explaining*, *teaching to others*, *enacting*, through a sort of theatrical representation, through which students became Java coding and Java programming concepts *actors and actresses*. The students, who engaged themselves in such a theatrical representation filmed on a video, were good students. One student started with no programming knowledge but was intrinsically motivated and matched with a student with considerable programming knowledge. Already in the class, they immediately selected the topic and started to sketch a story, based on clear instructional contents to communicate and to personally embed, representing Java objects, like people working in an office, as Fiorella and Mayer would define as "*embodied cognition*" [11] (p. 466).

This choice embodies how to transform abstract Java codes and Java concepts into concrete *actors and actresses*, finding relationships between Java codes and Java concepts. For example, in programming it is possible to copy an object; consequently, these students produced a copy of objects, presenting an assistant in an office represented by themselves. Additionally, they chose the topic: *objects and classes*, which is the most abstract and fundamental topic in the Java object-oriented programming language. These students applied a humorous approach to represents themselves as Java codes and Java concepts, covering from very basic to more advanced Java programming topics. Through the enactment, they tried to transform *an abstract programming concept to a concrete one, and promoting the transfer: from a concrete to an abstract*, through a serious storytelling theatrical representation [15]. As instructors we asked ourselves, if showing this excellent video example would help novice students to understand these programming concepts.

5.2 The Business Mathematics 2 Course

The final grades were showing that all students successfully passed the exams. 16 out of 18 students took the chance to earn learning points in creating three videos, where they needed to explain one exercise per video. Also, in this case, no student complained to produce a video with their own tools.

A total of 48 (16×3) videos were produced by the students, with approximately a total of 167 min distributed over 16 students, ranging from a maximum of 21 min to a minimum of three minutes for each student explaining the three exercises. The instructor analysed the videos considering the three categories: *syntactic knowledge*, *conceptual knowledge* and *strategic knowledge* [25]. In terms of these three categories, all the 48 videos were correct. But a *no audio explanation in own words hypothesis* was suggested. In fact, a total of three student videos were produced without audio explanation. The instructor was not able to understand whether the procedures shown in these videos were correct or not.

In the good videos, students showed how to work with clear, orderly procedural writing, combined with clear explanations. For example, one student copied from the book all the elements to solve the exercises, giving context to the exercises before solving these and defining the work to be done, and afterwards explaining in their own words the beginning of the problems to be solved. Another provided solutions with their own words; someone else added their own insights and their own context to arrive at the correct result. Other students were able to link the solutions to the formulas and theorems on a formula sheet provided by the instructor during the exam. Some students enhanced the explanation with own content (summarised on cards), which was not explained in class, other used predefined screens and one student also corrected himself during the explanation, reaching the correct solution even if the exercise was challenging. Some spoke very freely and confidently. Others suggested hints on how to solve problems, using alternative strategies.

The less good videos reflected the behaviour of certain students in class. For example, distracted students produced videos without context or explanation, not mentioning the type of exercise he/she was dealing with, showing a lack of involvement, adding music to the video explanation instead of thoroughly explaining the exercise. When music instead of oral explanation was used, it became a difficult task for the instructor to verify whether the video-content was clearly understood or not.

It is interesting to note that the maths explanations required an abstract-to-concrete shift, since abstract concepts had to be defined first and then elaborated through concrete cases, e.g., calculations with concrete values. This shift was only recognisable if students used their own words to explain what they did and if they provided context or their own insight.

According to the *adagio* of the media philosopher Marshall McLuhan "The medium is the message" [20,21], he provocatively proposed to focus on the medium and not just the content. The medium impacts the content (the message) to be delivered, also becoming also the message itself. The videos were very insightful from an instructional point of view, because students were able to produce their videos without any technical difficulty. No one showed their own face, but they worked with their own voice. This way of creating the videos mirrored their own behaviour and personality features as students in classroom.

6   Limitations and Discussion

We are aware that our exploratory studies present several limitations. First, we had two experimental groups for both courses: Students were exposed to the same amount of content. Organising a control group was beyond the scope of the present study. Second, the two classes (*Programming* and *Business Mathematics 2*) were not the same ones. For a next educational experiment, we plan to work with a class that covers mathematics and programming instructional contents, to understand how students may combine the two disciplines. Third, the production of the videos was an optional and non-mandatory task, designed to help students with greater difficulties.

Through these videos, we learnt that the *exploratory video-based instructional intervention* could greatly improve the understanding of where students are in their learning processes. We do believe that the *conceptual change theories*, especially the *knowledge-per-element* could promote a beneficial aspect in designing learning activities within the curriculum. For a next round of instructional experiments, a suggestion is using a content analysis methodology, considering the following analysis lenses: a) the three levels of explanation, according to *the representational approach*, *the cognitive approach*, and *the instructional approach*, developed by Bétrancourt and Benetos [3], b) the logical and rhetorical construction of procedural discourse, designed by Farkas [8] and finally c) the Morain and Swarts [22] research questions, which are: *With which convectional forms of instructional discourse, do students use sound, text, still images and moving images? How do these modal forms of content help the viewers to access, understand, stay engaged with the instructional message? How do these usages of content differ in these videos that other students can rate good, average, poor? If these good videos were shown to weaker students, would they represent helpful learning tools for them? If yes, why, and how? How could we measure this additional video-enhanced understanding (the shift from concrete to abstract)?*

Additionally, we plan the creation of a multimedia accessible collection of all students' self-produced videos to be stored on the Moodle platform. The instructor will provide "an assessment grid", together with a vote system. During the last lesson of the course, the class will be looking at all the videos, voting these and give students the chance to promote more discussion around programming and maths concepts.

Summarising, through the *knowledge-as-elements*, composed by facts, narratives, concepts, and mental models at various stages of development, novices connect and activate these elements, based on the relevance of the learning situation. During these conceptual change processes, which may be perceived as a piece-by-piece progression of knowledge creation, first-year students could find their way to enhance their programming and mathematics understandings and skills, improved through the procedural discourse communication skill, which represent another vital competence for becoming a competent BIT professional. As mentioned, for Fiorella and Mayer [11], there are eight ways to promote *generative learning*. Through the *exploratory video-based instructional intervention*, students experienced four of the eight strategies (*summarising, self-testing, teaching to others, and enacting*). Providing students with more opportunities to learn by using more media that is spiced with creativity could make programming and abstract mathematics concepts more accessible to novice students.

References

1. Adelson, B.: Problem solving and the development of abstract categories in programming languages. Memory Cognit. **9**(4), 422–433 (1981). https://doi.org/10.3758/BF03197568

2. Bayman, P., Mayer, R.E.: Using conceptual models to teach basic computer programming. J. Educ. Psychol. **80**(3), 291–298 (1988). https://doi.org/10.1037/0022-0663.80.3.291

3. Bétrancourt, M., Benetos, K.: Why and when does instructional video facilitate learning? A commentary to the special issue "developments and trends in learning with instructional video". Comput. Hum. Behav. **89**, 471–475 (2018). https://doi.org/10.1016/j.chb.2018.08.035

4. Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., Krathwohl, D.R.: Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain. David McKay Company, New York (1956)

5. Carliner, S.: Physical, cognitive, and affective: a three-part framework for information design. Tech. Commun. **47**(4), 561–576 (2000). https://www.jstor.org/stable/43748975

6. Coney, M.B., Chatfield, C.S.: Rethinking the author-reader relationship in computer documentation. ACM SIGDOC Asterisk J. Comput. Documentation **20**(2), 23–29 (1996). https://doi.org/10.1145/381815.381826

7. diSessa, A.A.: Toward an epistemology of physics. Cognit. Instr. **10**(2/3), 105–225 (1993). https://www.jstor.org/stable/3233725

8. Farkas, D.K.: The logical and rhetorical construction of procedural discourse. Tech. Commun. **46**(1), 42–54 (1999). https://www.jstor.org/stable/43088601

9. Fiorella, L., Mayer, R.E.: Role of expectations and explanations in learning by teaching. Contemp. Educ. Psychol. **39**(2), 75–85 (2014). https://doi.org/10.1016/j.cedpsych.2014.01.001

10. Fiorella, L., Mayer, R.E.: Eight ways to promote generative learning. Educ. Psychol. Rev. **28**(4), 717–741 (2016). https://doi.org/10.1007/s10648-015-9348-9

11. Fiorella, L., Mayer, R.E.: What works and doesn't work with instructional video. Comput. Hum. Behav. **89**, 465–470 (2018). https://doi.org/10.1016/j.chb.2018.07.015

12. Hoogerheide, V., Deijkers, L., Loyens, S.M.M., Heijltjes, A., van Gog, T.: Gaining from explaining: learning improves from explaining to fictitious others on video, not from writing to them. Contemp. Educ. Psychol. **44**, 95–106 (2016). https://doi.org/10.1016/j.cedpsych.2016.02.005

13. Krathwohl, D.R.: A revision of bloom's taxonomy: an overview. Theory Pract. **41**(4), 212–218 (2002). https://doi.org/10.1207/s15430421tip4104_2

14. Krathwohl, D.R., Anderson, L.W.: A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Longman, New York (2009)

15. Lugmayr, A., Sutinen, E., Suhonen, J., Sedano, C.I., Hlavacs, H., Montero, C.S.: Serious storytelling-a first definition and review. Multimed. Tools Appl. **76**(14), 15707–15733 (2017). https://doi.org/10.1007/s11042-016-3865-5

16. Mayer, R.E.: The psychology of how novices learn computer programming. ACM Comput. Surv. (CSUR) **13**(1), 121–141 (1981). https://doi.org/10.1145/356835.356841

17. Mayer, R.E.: Models for understanding. Rev. Educ. Res. **59**(1), 43–64 (1989). https://doi.org/10.3102/00346543059001043

18. Mayer, R.E.: How to be a Successful Student: 20 Study Habits Based on the Science of learning. Routledge, New York Oxon (2018). https://www.routledge.com/How-to-Be-a-Successful-Student-20-Study-Habits-Based-on-the-Science-of/Mayer/p/book/9781138319851

19. Mayer, R.E., Fay, A.L.: A chain of cognitive changes with learning to program in logo. J. Educ. Psychol. **79**(3), 269–279 (1987). https://doi.org/10.1037/0022-0663.79.3.269

20. McLuhan, M.: Understanding Media. Routledge, London (1964)

21. McLuhan, M., Fiore, Q.: The Medium is the Massage: an Inventory of Effects. Penguin Press, London (1967)

22. Morain, M., Swarts, J.: YouTutorial: a framework for assessing instructional online video. Tech. Commun. Q. **21**(1), 6–24 (2012). https://doi.org/10.1080/10572252.2012.626690

23. Newell, A.: Reasoning, problem solving, and decision processes: the problem space as a fundamental category. In: The Soar Papers (vol. 1): Research on Integrated Intelligence, pp. 55–80. MIT Press, Cambridge, MA, USA (1993). https://dl.acm.org/doi/abs/10.5555/162580.162585

24. Özdemir, G., Clark, D.B.: An overview of conceptual change theories. Eurasia J. Math. Sci. Technol. Educ. **3**(4), 351–361 (2007). https://doi.org/10.12973/ejmste/75414

25. Qian, Y., Lehman, J.: Students' misconceptions and other difficulties in introductory programming: a literature review. ACM Trans. Comput. Educ. (TOCE) **18**(1), 1–24 (2017). https://doi.org/10.1145/3077618

26. Soloway, E.: Learning to program = learning to construct mechanisms and explanations. Commun. ACM **29**(9), 850–858 (1986). https://doi.org/10.1145/6592.6594

Design of Enterprise Economic Dynamic Management System Based on Spark Technology

Lu Zhang and Yipin Yan[✉]

Faculty of Management, Chongqing College of Architecture and Technology, Chongqing 400000, China
zl13452873019@163.com

Abstract. Aiming at the problem that the currently used dynamic management system based on Hadoop and B/S architecture is affected by the slow data mining rate, resulting in low efficiency of data dynamic management, a design of enterprise economic dynamic management system based on Spark technology is proposed. Deploy the physical architecture of the Spark-based economic dynamic management system, and build a data warehouse in this architecture to facilitate users to quickly view data in real time. The B/S (browser/server) model is adopted to design data collection modules, business service modules and performance modules to meet the needs of big data analysis and decision-making. When using Spark technology to dynamically adjust the difference data in the database, the rule base needs to be updated in time to convert the automatic conversion system to a detection system. Use the optimization algorithm of Spark Join operator to optimize the entire connection operation, filter out the project data without specific categories in the bank flow data, reduce the data entering the shuffle stage, and design a dynamic management process. It can be seen from the test results that the system has a maximum management efficiency of 92% in a safe environment. In a non-interference environment, the highest management efficiency is 0.95, which has an efficient management effect.

Keywords: Spark technology · Enterprise economy · Dynamic management · Data warehouse · Spark Join operator

1 Introduction

At present, the enterprise finance department does not have a data management and analysis platform. It mainly relies on general office software such as ERP software and UFIDA financial management system to process data. However, general office software such as ERP has problems such as huge system, multiple functions, information redundancy and poor information sharing. In terms of function, such software is mainly used to input data and store data, Instead of collecting and analyzing data, the company's leaders cannot view the enterprise's financial data and understand the current company's

operation through the visual information management platform in a timely, accurate and real-time manner. In addition, such software usually has high cost, complex system upgrade, difficult system maintenance and slow data migration. These problems greatly affect the enterprise's office efficiency. Therefore, the dynamic management of enterprise economy is necessary. The previously proposed enterprise economic dynamic management system based on Hadoop uses clusters to complete high-speed operation and storage of data. At the same time, it is transparent to developers and supports agile development. It is mainly composed of HDFS and MapReduce. HDFS not only has high fault tolerance, but also has low hardware requirements and high throughput, You can also access the data in the file system as a stream. MapReduce implements task fragment processing, distributes fragments to multiple nodes through map, and then synthesizes data sets through reduce and loads them into the data warehouse, which truly realizes parallel computing [1]; The economic dynamic management system based on B/S architecture and data warehouse technology are used to realize data query and analysis, so as to ensure that the management can view the analysis results in real time and accurately understand the overall operation of the company [2]. However, the above two systems read the disk and file system more frequently, which makes the data mining speed slower. Therefore, the design of enterprise economic dynamic management system based on spark technology is proposed.

2 System Hardware Structure Design

According to the demand analysis of the economic dynamic system and the characteristics of the big data platform, the physical architecture of the Spark-based economic dynamic management system is shown in Fig. 1.

As can be seen from Fig. 1, in the whole process of designing the system, designers and developers do not need to care about the specific physical environment, but only need to know how to collect the data source. The economic dynamic management system is mainly composed of big data platform, data analysis management platform and MySQL database [3]. The big data platform runs on multiple servers in parallel, which is used to complete the standardized and structured processing of customer transaction behavior, details and other log information, conduct data analysis, and then store the results in the database, so that users can easily query and analyze the results. The data source of the big data platform comes from its own HDFS distributed file system. HDFS is used to store persistent log information, that is, all log information is stored in HDFS, while the corresponding analysis result database only stores the information of the day, which not only ensures the data mining of long-term accumulated big data, but also reduces I/O operations, So as to improve the overall efficiency of the system. The data analysis management platform provides external web services to facilitate users to query the analysis results more clearly and intuitively [4].

2.1 Data Warehouse Design

The traditional data warehouse is combined with the Spark big data platform to expand the real-time analysis function of big data to meet the original business needs. Figure 2 shows the architecture design of the system data warehouse.
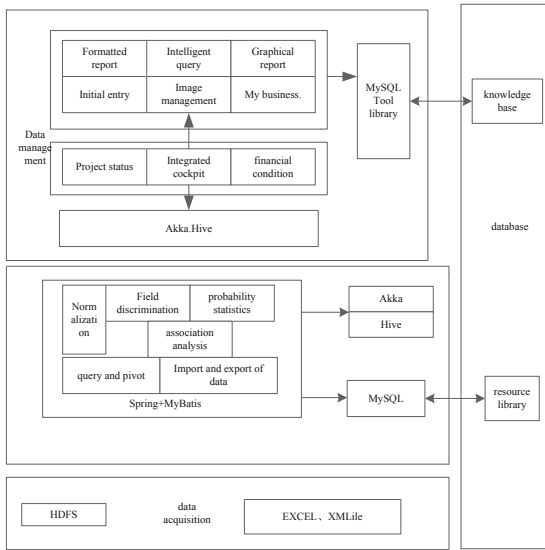
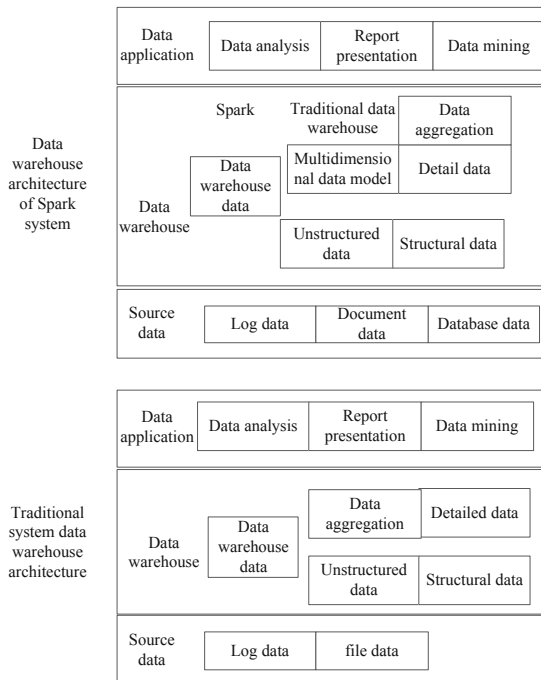**Fig. 1.** Schematic diagram of the hardware structure of the Spark-based system

Fig. 2. System data warehouse architecture

As can be seen from Fig. 2, this architecture increases the number of types in the source data layer compared with the traditional data warehouse and expands the data source. In the data warehouse layer, spark is integrated with the traditional data warehouse layer, in which spark is responsible for processing real-time, batch and unstructured data; Other unstructured data is processed by traditional data warehouse. At the same time, the connection between the data source layer and the data warehouse layer has also been changed accordingly. If it is integrated into the big data platform, it needs to be processed by ETL (extraction, conversion and loading) [5]. In the data warehouse layer, identify the data information mode, fractal compress the image and video information [6], extract the initial additional error value, abstract the distribution of error value, save database memory and realize effective data compression [7]. In the design of data warehouse of the whole system, the design of database occupies a key position, and the conceptual structure design of database is the core of the whole database design. E-R diagram is the basic method of conceptual model design. E represents entity and R represents relational model. It is composed of entity, relationship and attribute. If the system function module design gives the specific function module and behavior implementation of the developer, the data structure of the system provides the data storage mode and the correlation between data [8]. Although the system uses HDFS to store unstructured data such as logs, real-time data and data analysis result files, the final normalized analysis results will still be stored in the relational database for users to view the data in real time and quickly.

2.2 Data Acquisition Module

The economic dynamic management system adopts the B/S (browser/server) model. The B end is responsible for generating various requests to the S end, and the S end responds to related requests to complete data collection, which is sorted and stored in a database or file [9]. The data collection layer specifically corresponds to the initial input and business function modules of the system. The initial input mainly enters some basic company information such as organizational framework, employee files, bank files, etc. and some data directly related to the project at the initial stage of the project, such as the owner Files, supplier files, etc., can be entered in the form of direct entry through the browser form input box, or by importing the corresponding existing Excel file. Manual entry ensures the real-time performance of system data, while file import provides convenience for initial batch entry of data [10]. The business module divides different business scenarios according to the different positions of the employees, and enters real-time dynamic business data. Because the business data entered is complex and diverse, there are also many types of entry forms provided [11]. The direct input and file import are the same as those described above. At the same time, in order to meet the needs of big data analysis, it also provides a way to read files from the distributed file system and import batch data from the database. The composition of data acquisition module is shown in Fig. 3.
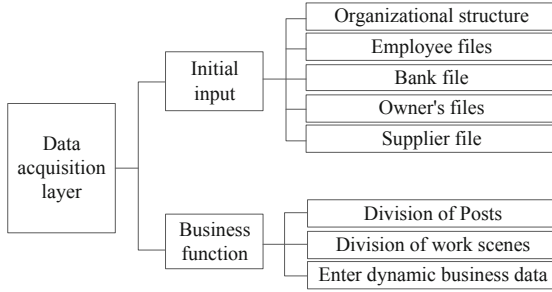
Fig. 3. Composition of data acquisition module

2.3 Business Service Module

The business service module is located on the S side, receives requests from the B side, and uses Java Bean, Servlet, SparkStreaming, and SparkSQL technologies to complete the corresponding processing of each request. The business service layer mainly corresponds to the three business analysis modules of the system's project status, integrated cockpit and capital status [12]. Among them, the project status and funding status belong to the analysis of the data in the traditional data warehouse to get the specific situation of a certain project, and the detailed analysis of the funding status of a certain project in the time dimension; while the integrated cockpit mainly combines large-scale analysis. The data platform analyzes the unstructured data from the file system or the batch structured data in the database to obtain the macro operation of the entire company, such as the geographical distribution of projects in the spatial dimension and the contract time of projects in the time dimension. The system business service module is shown in Fig. 4.
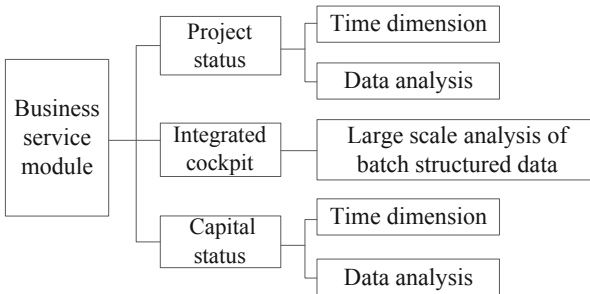
Fig. 4. Composition of business service module

2.4 Performance Module

The presentation module is mainly located at the B end. The data analysis results and data query results are dynamically refreshed and displayed on the page through the web page. It mainly applies HTML and CSS for page layout, jqueryeasyui for page

beautification, jQuery for page logic processing, and Ajax for asynchronous submission and refresh. The core of the presentation layer is the data query module, which mainly includes intelligent query, graphical report, formatted report and image query. Of course, it provides operation interfaces for project status, comprehensive cockpit, fund status, initial entry and my business. The system displays various data processing results through the beautiful and easy-to-use operation interface to meet the final analysis and decision-making requirements.

2.5 Real-Time Data Processing Module

Facing the new challenge of fast real-time streaming data calculation and fast batch calculation, the system needs to find a non-complicated implementation solution instead of combining frameworks for various computing scenarios to ensure concise data in the system Flow direction, so as to better ensure the high responsiveness of the system. Based on the distributed framework Spark and the distributed message queue Kafka as the main data transmission medium, a fast data processing module with two major functions of real-time computing and offline computing is constructed, as shown in Fig. 3.
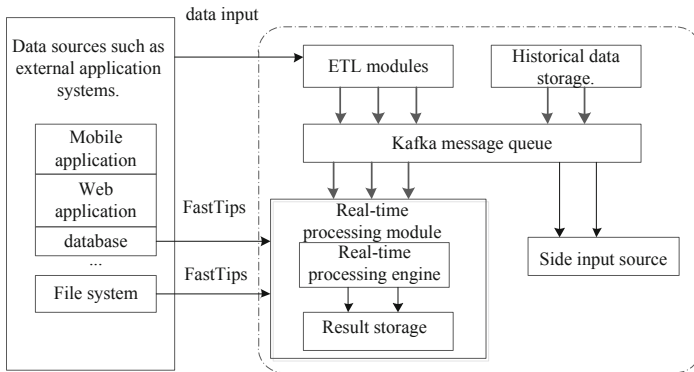
Fig. 5. Schematic diagram of the structure of the real-time data processing module

As can be seen from Fig. 5, the system mainly includes ETL module and real-time processing module. Kafka message queue can be regarded as an internal data transmission bus, which is connected between the two large modules to buffer data and distribute consumption data, so that the system will not go down due to excessive load when a large amount of data is generated by the data source, Kafka is also connected to historical data storage external disk devices and side input data sources.

(1) ETL module obtains the data to be processed by establishing a connection with external data sources. It can adapt to a variety of heterogeneous data sources and data in different formats, and support offline data and stream data, such as mysql, HDFS, Kafka, network socket, etc. one or more data sources can be selected for connection according to business scenarios. The received original data is cleaned, converted and sent to Kafka message queue.

(2) Kafka message queue is a distributed high-performance message queue system. Data receiving and distribution is to receive the converted data from ETL module through publish and subscribe (producer consumer) mode, classify these data according to business topics, facilitate applications to realize business functions, and also play the role of data buffer. In addition, it is also connected with a historical data storage device and a side input data source, which can be extended according to the actual needs of users.

(3) The purpose of storing historical data is to facilitate the tracing of historical data, to be able to analyze the complete and large-scale historical data in the past, and to output the specified range of historical data incrementally to the real-time processing module, thereby enabling data "replay" And the "recording and broadcasting" function, generally use a distributed storage system like HDFS.

(4) The side input data source is to save some metadata transferred from external application systems, because generally metadata changes and updates are not too frequent, these metadata are sometimes indispensable when processing data, so by establishing a side Save these metadata by inputting the data source, which saves the bandwidth of data transmission, and only updates when it needs to be updated. Redis or HDFS can be used as the storage device.

(5) The real-time processing module pulls data through the Kafka message queue for calculation in the manner of consumers, and provides calculation methods such as statistical analysis, complex aggregation, and machine learning. Users can implement their own calculation logic according to business needs. The calculated results are stored in the local file system or database, and provide an interface for external applications to quickly query and view the results in real time.

3   System Software Part Design

3.1   Enterprise Economic Data Mining Based on Spark

Spark is a fast and universal cluster computing platform for large-scale data processing. It provides services in the form of Web Service. Users can access the service by calling the relevant Web Service interface, or perform two operations on the WebService interface provided by the system. Development to provide richer services.

Emphasis is placed on the analysis of dynamically changing business economic data, and the detection of dynamic change data is regarded as a data analysis process. This process requires mining and analysis of massive data in order to obtain dynamic changing business economic data under the normal mode. When the mining process reaches a state of dynamic data changes, the automatic conversion system needs to be converted to a detection system. The rule base needs to be updated in time during this process, as shown below:

Let $R = \{r_1, r_2, \cdots r_m\}$ be a collection of projects, and $F = \{f_1, f_2, \cdots f_n\}$ represent a collection of database projects. Each transaction T has a separate project subset $F \subseteq R$, and has a unique identifier. The association rule is the logical implication formula of the formal term $A \Rightarrow B$, and $A \subset F, B \subset F, A \cap B = \emptyset$.

The support of the logical implication formula $A \Rightarrow B$ in the rule base includes the percentages of both A and B transactions. The condition concept is $P = (B/A)$, which can be expressed as:

$$\sup\ part(A \Rightarrow B) = P(A \cup B)$$
$$confidence(A \Rightarrow B) = P(B/A) \tag{1}$$

Confidence is responsible for measuring the accuracy of relevant rules, and support is responsible for measuring the accuracy of relevant relationship matrix. Through confidence and support, we can evaluate how representative the rule is in the whole mining process. Obviously, the greater the support, the more tense the relationship rules.

When using Spark technology for dynamic data mining, it is necessary to combine the characteristics of dynamic data in large differential database. The specific implementation process is as follows:

Step 1: Obtain interest

Through the query of related data, the mining target and the reference set are collected into the related database. According to the above association rules, the percentage of all things is regarded as the expected confidence level, and comparative analysis is performed. The obtained interest level W is:

$$W = \frac{confidence(A \Rightarrow B)}{\sup part(A \Rightarrow B)} \tag{2}$$

The interest degree W in formula (2) can include the correlation degree of all logical implication formula $A \Rightarrow B$.

Step 2: Set the minimum limit matrix

At the level of rough calculation, the mining target is set as the minimum limit matrix, and the corresponding matrix values are all stored in the database, where the attribute value is a single value.

Step 3: Calculate the matrix support

There are different degrees of support between different matrices, the calculation method is:

$$confidence\{A, B\} = W \times confidence(A) \times confidence(B) \tag{3}$$

The support degree of the minimum threshold is obtained according to formula (3), and the minimum database is formed through the expenditure degree.

Step 4: check the matrix relationship

Check the matrix relationship between common databases, eliminate the data inconsistent with the actual relationship, and form a topology.

Step 5: form a new topology

The topological relations obtained in the above contents are generalized to form a new topological structure, so as to realize the mining of enterprise economic dynamic management data.

3.2   Join Operation Based on Spark Join Operator

Data connection is a key step, so the optimization algorithm of Spark Join operator is used to optimize the entire connection operation. The specific analysis process is as follows:

Step 1: Use Kafka to receive real-time messages from external systems in real time; Copy real-time bank log data to HDFS file system;

Step 2: Use the hdfsutii toolkit to read the data in the HDFS file system; Use the data generator to generate simulation data to verify the reliability of the program;

Step 3: Convert data to RDD; Call rddoperatorutils class to group RDD and aggregation operations; Call bfjoin to perform RDD connection operation;

Step 4: Store the result set in MySQL; Read the data in MySQL and display the analysis results.

Bank flow data has the characteristics of complex payment types, which makes it impossible to accurately count many subtle types of payments. At this time, they can be classified into the same type of data. Generally, users pay more attention to certain data, such as a certain item in a certain period of time. The total income, total expenditure, the proportion of certain types of funds in the total expenditure/income of the project and other macro information, etc., therefore, six types of data are divided.

In the specific processing, when the RddBFJoin is processed, the BF Join algorithm designed in the previous section is used. In this business, before the connection of rddl and rdd2 is executed, the item data that does not have a specific category in the bank flow data can be filtered out. Reduce the data entering the Shuffle stage.

3.3   Dynamic Management Process Design

Spark provides a very direct way for users to submit compiled Spark applications, spark-submit scripts. Users only need to package the application and its dependencies together, and then use the script to submit to the cluster, and set the application deployment mode, program entry, parameters and other information to complete the submission, and start the application task on the cluster. Enter the cluster management interface, the first is the list of data types to be created, including data type name, brief description, jar package name, jar package location, creation time, etc. The list is sorted by the data type creation time. Buttons for creating, updating, and deleting data types are provided at the top of the list, and pagination components are provided at the bottom of the list. When creating a data type, click the Create button to pop up a window, enter the name of the data type (check the name), click the upload button, select the corresponding jar package, and upload, or you can briefly describe the data type.

Dynamic management process design, as shown in Fig. 6.
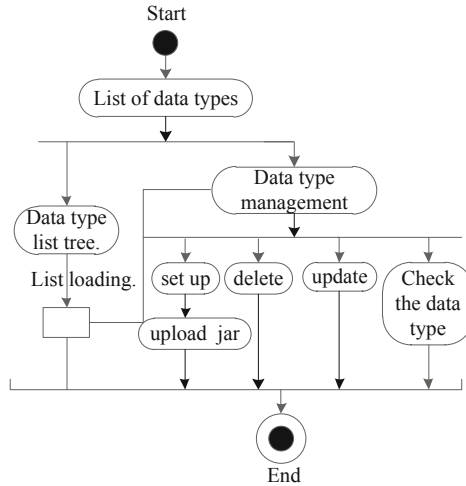
Fig. 6. Dynamic management process

When updating the data type, select the data type to be updated, and click the update button to pop up a window. The window type is the same as the pop-up window for creating the data type, but the data type information already exists in the window. You can modify the information to be updated. If you need to modify the corresponding jar package, re upload the jar package, and the original jar package will be deleted automatically. When deleting a data type, select the data type to delete and click the delete button to delete it. After deleting the data type, the corresponding data type information in the data type information table in the database will be deleted, and the corresponding data type jar file in the file system will be deleted automatically.

4   System Test

This test is derived from the ATL data processing analysis system project test report. The purpose is to summarize the test phase of the enterprise economic dynamic management system based on Spark technology and analyze the test results, evaluate whether the system meets the requirements, and at the same time, discover the system as much as possible Bugs and defects to ensure product quality.

4.1   Test Data

The data used in this experiment are nearly 300W bank journal data of a company. The following information is extracted. The specific data set format is shown in Table 1.

Table 1. Bank flow data field information

Field description	Type
Abstract	String
Borrow	Float
Loan	Float
Balance	Float
Bank account	Int
Transaction type	String
Transaction hour	String

Based on the data in Table 1, the system test and analysis are carried out.

4.2   Test Results and Analysis

Use the Hadoop-based enterprise economic dynamic management system, the B/S-based economic dynamic management system, and the Spark technology-based enterprise economic dynamic management system to compare and analyze the management efficiency in a safe environment and a disturbed environment. The result is shown below.

Safety Environment

In a safe environment, the management efficiency of the three systems is compared and analyzed, and the results are shown in Fig. 7.
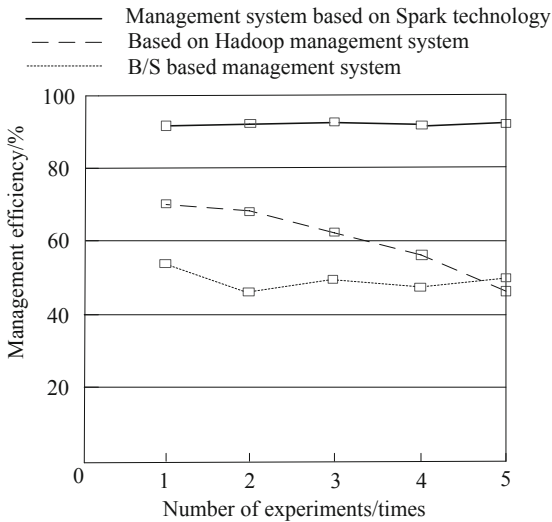
Fig. 7. Comparative analysis of the management efficiency of the three systems in a safe environment

As can be seen from Fig. 7, the management efficiency of the management system based on Spark Technology under the security environment is always maintained at more than 90%; The management efficiency of Hadoop based management system is 46%–70%; The management efficiency of the management system based on B/S is 46%–56%. It can be seen that the management efficiency of the management system based on spark technology proposed in this paper is higher than that of the other two systems.

Interference Environment

In order to verify the actual operation performance of the system designed in this paper, the management efficiency of the three systems under different concurrency is analyzed, and the results are shown in Fig. 8.
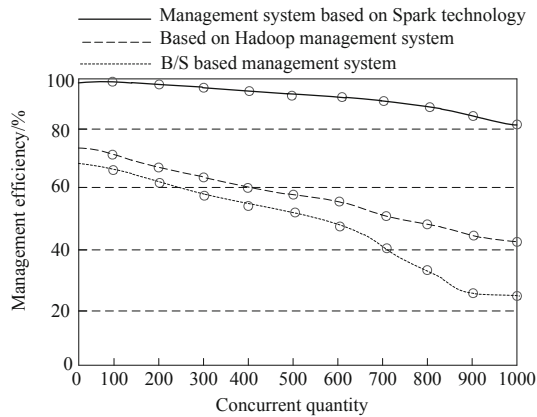
Fig. 8. Comparison and analysis of management efficiency of three systems under interference environment

As can be seen from Fig. 8, when the concurrency of the three systems increases, the management efficiency gradually decreases. The management efficiency of the management system based on Hadoop fluctuates between 42% and 74%, and the management efficiency is low; The management efficiency of the management system based on B/s fluctuates between 25% and 68%, and decreases greatly when the concurrency is 500–900, so the system is relatively unstable; Although the management efficiency of the management system based on spark technology proposed in this paper has been declining, it can always be maintained at more than 80%; It shows that the management system proposed in this paper can maintain 80% management efficiency under the condition of 1000 concurrency.

5   Conclusion

Based on the new management and analysis requirements of enterprises for financial data, an economic dynamic management system integrating multi-source data collection, statistics, analysis and management is constructed. The hierarchical design idea is used in the construction of the whole system, the mature spark architecture is used to build

the server side of the system, and the spark big data technology is used to complete the basic development of the system.

This paper mainly designs and develops the economic dynamic management system based on spark technology, studies the impact of data connection operation in spark on data processing and analysis in the economic dynamic management system, and optimizes the large table equivalent connection algorithm in spark. However, at present, the system has only completed independent development, and cannot interact with external systems. At the same time, some functions have not been migrated from traditional platform to large number platform. In addition, there are still many deficiencies in the optimization of data connection algorithm. Here, we put forward our prospects for the future:

Although the proposed optimization algorithm BF join improves the connection performance by effectively filtering the data that does not meet the connection conditions, the pre filtering of RDD partitions may lead to data skewing of data with roughly the same size due to different matching degrees of filtering conditions, resulting in the decline of connection performance. Therefore, the optimization of data skewing connection algorithm needs to be further studied.

References

1. Tang, X., Song, Z.: Evolutionary game analysis of collaborative consumption behavior of participants in sharing economy. Enterp. Econ. **461**(01), 66–72 (2019)
2. Wang, S., Xue, X., Ge, Y., et al.: Geo-Economic strategies assessment based on computational experiment: taking the customs clearance time adjustment of China-Indonesia and China-Vietnam as an example. Econ. Geogr. **39**(02), 12–21+63 (2019)
3. Li, G.: The benefit distribution of tourism enterprise strategic alliance under active cooperation mode. Enterp. Econ. **03**, 145–152 (2020)
4. Xu, W., Guo, S., Wang, L., et al.: Dynamic control system of material flow schedule for building material equipment manufacturing enterprise oriented to production task. Comput. Integr. Manuf. Syst. **25**(03), 105–118 (2019)
5. Zhao, X., Zhou, Z., Wu, Y., et al.: Construction of village resource management system based on perspective of poverty alleviation development. Mod. Electron. Tech. **43**(10), 103–107 (2020)
6. Liu, S., Bai, W., Liu, G., et al.: Parallel fractal compression method for big video data. Complexity **2018**, 2016976 (2018)
7. 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
8. Su, J., Li, Y.: Practice teaching reform of agricultural and Forestry Economics and Management undergraduate major based on rural revitalization strategy. For. Econ. **41**(04), 94–98 (2019)
9. Li, J.: Theoretical design and exploration of modern economic management system in public hospitals. Friends Account. **21**, 2–8 (2020)
10. Liu Y.: Study on the particularity and innovation path of village collective Economic Management in China. Agric. Econ. (04), 40–42 (2020)
11. Zhao, C., Sun, J.: Try to discuss computer information technology and economic management optimization integration. Economics **3**(4), 85–86 (2020)
12. Yu, S.: Analysis on internal control method of enterprise economic management risk. Economics **4**(1), 21–23 (2021)

Design of Enterprise Financial and Economic Data Accurate Classification Management System Based on Random Forest

Junlin Li[1](✉) and Haonan Chu[2]

[1] Beijing Union University, Beijing 100101, China
lijunlll1512@163.com

[2] School of Labor Relations and Human Resource, China University of Labor Relations, Beijing 100048, China

Abstract. The conventional system has the problems of low recall rate, accuracy rate and high false positive rate of financial data classification. Therefore, an accurate classification management system of enterprise financial and economic data based on random forest is proposed. In the hardware, the front end, middle layer, server end and enterprise financial and economic data display end are used to form the overall architecture of the system, optimize the data memory of the server end, and transform the serial communication circuit of the development board; In the software design, abnormal financial data are filtered, a decision tree is established for each sample data, the utility function value is learned through the membership of the decision tree, and the optimal classification category is selected for the data through the random forest classifier. The experimental results show that the designed system improves the recall and accuracy of data classification, reduces the false positive rate, and the financial data classification results are more accurate and reliable.

Keywords: Random forest · Financial data · Management system · Data classification

1 Introduction

Reasonable classification is the basis for in-depth mining and analysis of financial data. However, due to the large scale and weak regularity of enterprise financial and economic data, the problem of data imbalance has become the key problem of financial data mining. Unbalanced data leads to the decline of the accuracy of financial data classification and the actual effect of data. Therefore, it is necessary to accurately classify financial data [1]. The research on the accurate classification management system of financial data is conducive to improve the classification management and information analysis ability of financial data, and is of great significance to the safe operation of enterprises [2].

With the rapid development of big data and information technology, foreign data classification management systems have achieved good development. Through the feature extraction and fusion clustering processing of financial data, the internal association

rule feature information of financial data is extracted, and then the automatic classification and identification of financial data are carried out according to the distribution of feature information, so as to realize the adaptive fusion processing of financial data, It can improve the business process management ability of financial data [3]. The domestic data classification management system has also made great progress. The software is designed in the embedded environment, the multiple regression analysis method is used to analyze the statistical characteristics of financial data, the retrieval structure model of financial database is constructed, and the statistical analysis method is used for automatic statistics of financial data to realize the optimal classification of financial data [4].

Liu et al. Proposed to design a text classification method of distributed financial data based on multi neural network fusion. This method designs a multi-element neural network path including word embedding layer, convolution layer, bidirectional gating cycle unit layer, attention mechanism layer and softmax layer; On this basis, the demand effect resource classification strategy is adopted to complete the mapping transformation from the demand of qualitative science and technology resources to the solution of quantitative resource service effect, and then to the output of qualitative science and technology resources. It focuses on solving the significant long-distance dependence characteristics of distributed text, obtaining effect knowledge quickly and accurately, and improving the analysis effectiveness of financial data. Liu et al. Proposed to design a financial data classification method based on width learning system. This method extracts the deep features of financial data through simple structure to speed up the classification speed The input data is constructed by using the time series of voxel mean of the region of interest in the data, the shallow and deep features of the financial data are extracted respectively, mapped into feature nodes and enhancement nodes of width learning, and the model framework is constructed. The connection weight of the classification model is calculated by inverse ridge regression to classify the financial data and greatly reduce the training time.

However, the data classification of conventional systems mostly presupposes the uniform distribution of the number of samples in different categories, resulting in insufficient data classification. In response to this problem, combined with existing research theories, a random forest-based accurate classification management system for corporate financial and economic data is proposed.. Random forest is to artificially synthesize new minority samples to reduce the imbalance of data categories, and perform linear difference between neighboring minority samples to synthesize new minority samples.

2   Design of an Accurate Classification Management System for Enterprise Financial and Economic Data Based on Random Forest

2.1   Hardware Design of Accurate Classification Management System for Enterprise Financial and Economic Data

The Overall System Architecture Design
The overall architecture of the system is composed of four parts: front end, middle layer,

server end and enterprise financial and economic data display end. The server side adopts micro server, uses database to transmit information, provides business interface through middle tier architecture, and connects the front end and server side. The front end and server side exchange information with the middle tier through interfaces respectively, uses the front end to control enterprise financial and economic data, and outputs enterprise financial and economic data at the display end. The overall architecture of the system is shown in Fig. 1:
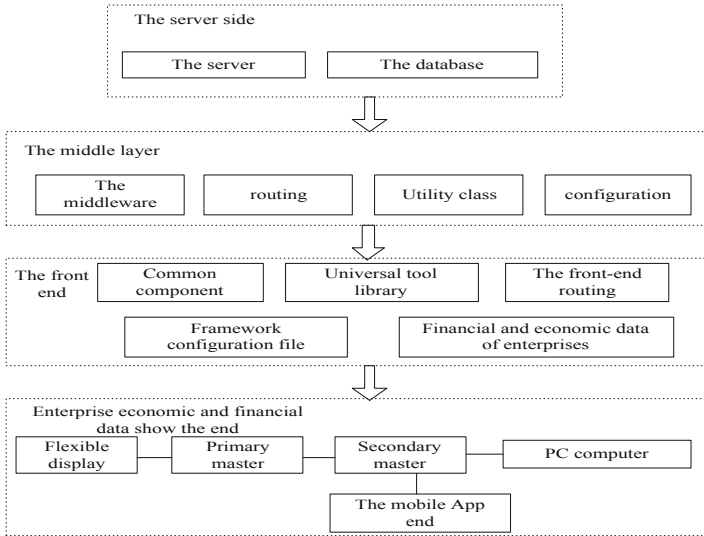
Fig. 1. The overall architecture of the accurate classification management system for corporate financial and economic data

The front-end adopts a component-based development method, using chart customization, file segmentation and uploading and other components to form reusable universal components, abstract processing of element event-related libraries to form a universal tool library, and the front-end routing layout components adopt a skeleton structure. All business components are embedded in the layout components to ensure that the business can be reused. According to different business functions, switch different business components, select front-end routing configuration, front-end environment configuration, packaging tool configuration, multi-national language configuration, and form a framework configuration file, Corporate financial and economic data includes static resources, image resources, etc. When classifying corporate financial and economic data, each business component independently references label resources internally. The middle-tier architecture uses the middleware organizational framework to add user authentication, log, and exception handling middleware to the middle-tier to check the login status requested by the interface, record the corporate financial and economic data classification program behavior, and handle the wrong program behavior in a timely manner. The interface request is mapped to the tool module through routing, the encapsulated business component is used to forward the request, and the configuration of the

framework environment is formed through the server address, deployment environment, and external device address [5]. The server adopts a B/S structure, including multiple controllers and filters, as well as multiple business entities and database entities, to check and process different interface requests and business logic. The enterprise financial and economic data display terminal adopts the secondary micro-control method. The single-chip microcomputer and the development board are respectively used as the primary and secondary master controllers. The development board is regarded as the corporate financial and economic data motherboard, equipped with Wi-Fi modules, The power conversion module, etc., convert the communication protocol to realize the communication between the primary control node and the PC end and the APP end. The PC end uses serial communication and the APP end uses wireless communication. At this point, the overall system architecture design is completed.

Optimize the Storage of Corporate Financial and Economic Data
Optimize the enterprise financial and economic data memory on the server side of the system, and expand the storage capacity of the memory for enterprise financial and economic data. Firstly, the input and output interfaces of the memory are optimized. The optimized interfaces are shown in Fig. 2:
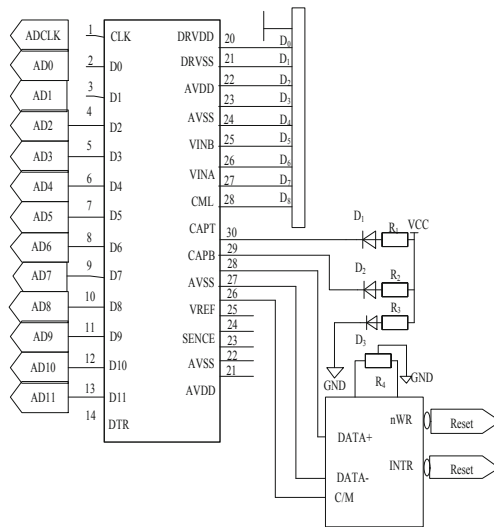
Fig. 2. Schematic diagram of enterprise financial and economic data storage interface

LDPC decoder, Q hard decision message memory, buffer buffer, decoding result memory and cyclic shift coefficient memory are selected as the configuration devices of server-side memory. Firstly, the enterprise financial and economic data meeting the security level is written into the buffer, the buffer is full, the data is read out, it is input into the Q hard decision message memory, the corresponding column address of the network information is read, the corresponding cyclic shift value is taken out from the address, the information is cyclically shifted left by the same digit according to the value, and the left

shift results are transmitted to the LDPC decoder in turn, Update the information node repeatedly, obtain the latest decoding message, output the decoding results of enterprise financial and economic data, and finally store the decoded information in the memory [6]. Define the interface form of memory input and output, as shown in Table 1:

Table 1. Memory input and output interface

Port	Input/Output	Bit width	Explain
DRVDD	Enter	1	Input from outside, system clock
AVDD	Enter	2	Information input enable terminal, input the quantized data to be decoded
VINA	Enter	1	Reset terminal, ensure low level is valid
CAPT	Enter	1	Information input terminal, to ensure high level effective
AVSS	Enter	1	Bit rate selection input
SENCE	Output	8	Decode data output terminal to ensure high level is effective
AVDD	Output	2	Output enable indicator

The 0.5 code rate represents the low level of the code rate selection input, and the 0.8 code rate represents the high level. For the output enable indicator, the parallel output mode is adopted, the decoding result is set to 1 bit, and 8 decoding results of information are output at one time, so as to increase the size of system memory and improve the maximum processing number of user access requests. So far, the optimization of enterprise financial and economic data memory has been completed.

Optimize the Main Board Structure of Corporate Financial and Economic Data
Optimize the main board structure of the display end of the financial and economic data of the enterprise, and transform the serial communication circuit of the financial data development board to reduce the interference of data collection, update, and communication. The development board uses EEPROM chip as the core chip, and the main board uses STM32F103ZET6 signal microprocessor, equipped with 512 KB Flash memory, provides high-density code instructions, efficiently stores business tag data, and integrates comparators and timers in the peripheral modules of the microprocessor., Power supply, etc., the external power supply adopts AMS1117 power conversion core, and provides 3.3 V DC voltage for the main board through the positive and negative interface. Configure the network access point inside the development board, use the 433 MHz wireless communication module, and use the TCP communication protocol to communicate with the APP and the first-level main controller, set the wireless communication module parameters, and maintain the APP-side communication channel and the first-level main controller The communication channels are the same, and the 433 MHz frequency band is selected [7]. Use serial communication to connect the corporate financial economic

data mainboard and PC upper computer, and the optimized serial communication circuit is shown in Fig. 3:
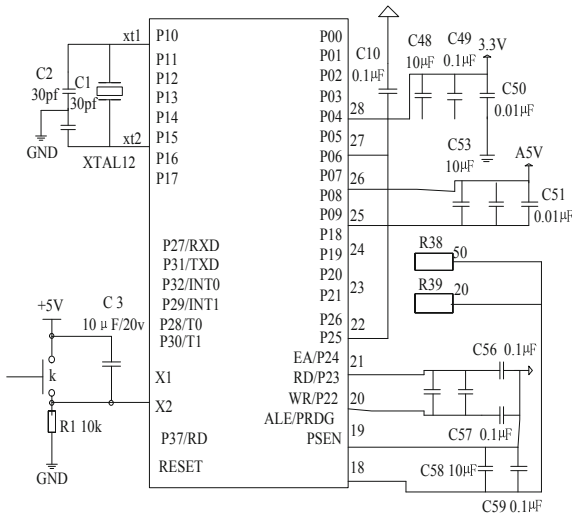
Fig. 3. The serial communication circuit of the main board of enterprise financial economic data

The serial communication circuit of the development board is divided into amplification circuit and filter circuit. AD620 differential amplification chip is selected as the amplification circuit, and the amplification factor of the chip is 7 times. Set the zero point of the output end of the amplifier, ground the 6 pins, and connect with the polar capacitor and ceramic capacitor to reduce the signal noise. AD8676 amplifier is selected as the filter circuit to connect the 6 pins of the amplification circuit, The circuit is disassembled into low-pass filter and high pass filter, and the voltage following method is used to realize the main stage amplification and band-pass filter of the signal. The optimized serial communication circuit includes data collector signal processing circuit, power supply circuit and serial port circuit. The financial data development board adopts 3.3 V power supply, configures the power monitoring reset chip for the power supply circuit, integrates a certain capacity of serial memory, and connects the reset signal to the reset pin of the reset chip, so that the power supply circuit has the function of power down protection. In order to keep the output signals of each data collector consistent, a differential processing chip is configured for the signal processing circuit of the collector, the original signal is input into the differential processing chip, and the three differential electrical signals are converted into three single ended electrical signals, in which the two signals of x-axis and y-axis are processed into single ended signals, the signal of z-axis is processed into zero clearing signals, and then the three signals are inverted, Ensure that the signal matches the voltage of the differential processing chip pin. After the differential processing chip outputs three logic electrical signals, it transmits the signals to the serial port circuit, configures a voltage stabilizing chip for the serial port circuit, performs pulse counting processing on the three logic electrical signals, installs

a static register and a clock backup register at the serial port, and uses the two registers to obtain voltage from the pins of the voltage stabilizing chip to provide reliable power supply for the serial port circuit, Parallel capacitors are connected to the output of the core controller to filter the three signals, so as to improve the transient response and stability of the signal. So far, the optimization of the structure of the main board of financial data has been completed, and the hardware design of the accurate classification management system of enterprise financial and economic data has been realized.

2.2 Software Design of Precise Classification Management System for Enterprise Financial and Economic Data

Preprocessing Enterprise Financial and Economic Data
Through abnormal identification, find the source of weak data, and correct the abnormal financial and economic data of the enterprise. Obtain the financial and economic data files of the system enterprise, simplify and merge the data, and keep the data format consistent. Traverse all data files, treat each sample data as a two-tuple, including sample observations and sample labels, and check data integrity. Through time-effect correlation, test the data relevance of different time series, consider the existence of small-scale statistical samples in the system, and use statistic Z for testing. The statistic calculation formulas for normal samples and test samples are:

$$Z = \frac{z_1 - z_2}{\sqrt{\frac{\sum u_1^2 + \sum u_2^3}{W_1 + W_2 - 2} \times \frac{W_1 + W_2}{W_1 W_2}}} \tag{1}$$

Among them, z_1 and z_2 are the mean value of the data, u_1, u_2 is the standard deviation of the data, and W_1, W_2 is the data capacity. The larger the Z value, the greater the degree of difference between the normal sample and the test sample. Through the completeness and correlation test, the abnormality of the financial and economic data of the enterprise can be obtained [8].

Identify the sample data that is significantly different from the normal sample as the abnormal data in the financial and economic data of the enterprise. Using a clustering algorithm based on the density of clustered data groups, clustering analysis of emergency repair data, for the multi-dimensional big data of the power distribution system, the normal sample is taken as the core point, the array to be tested is $V = \{v_0, v_1, \cdots, v_n\}$, the core point is v_0, and the inspection data is $v_1, \cdots, v_n$, n is the number of arrays, and the formula for the distance between the normal sample and the test sample is:

$$U(v_j, v_0) = \sqrt{(s_j - s_0)^2 + (w_j - w_0)^2} \tag{2}$$

Among them, $U(v_j, v_0)$ is the core distance of the two sets of data, s_j is the observation value of the inspection data, $j = 1, 2, \cdots, n$ and s_0 are the observation values of the core points, and w_j and w_0 are the types of the inspection data and the core points respectively. Set the ideal core distance S and the reachable distance T, when $U(v_j, v_0)$ is less than the ideal core distance, it is judged that the test sample is in the same cluster

as the normal sample, and when $U(v_j, v_0)$ is greater than the ideal core distance, it is judged that the two sets of samples are of different clusters. Regarding the core distance and the reachable distance as the neighborhood and density of the array, respectively, they are used as the measurement standard for the clustering of the sample data, and the clustering constraint conditions are obtained. The expression is:

$$\begin{cases} S(v_j, v_0) = \{U(v_j, v_0) \leq S\} \\ S(v_j, v_0) \geq T \end{cases} \tag{3}$$

where $S(v_j)$ represents the data in the cluster with v_0 as the cluster center. After the enterprise financial and economic data is classified into clusters, calculate the statistics $Z(v_j, v_0)$ of each test data and the normal sample, sort the statistics in the same cluster, and use the sample data with the largest $Z(v_j, v_0)$ value as the abnormal data, and filter the abnormal data to obtain the revised data set. At this point, the preprocessing of corporate financial and economic data has been completed.

Accurate Classification of Corporate Financial and Economic Data Based on Random Forest

For the preprocessed corporate financial and economic data, a random forest classifier composed of multiple decision trees is used to accurately classify data categories. The classification results of the random forest classifier include quick ratio, current ratio, accounts receivable turnover ratio, market-sales ratio, price-to-book ratio, cost-to-interest ratio, total asset growth rate, asset-liability ratio, shareholder equity ratio, equity ratio, etc., Establish a decision tree for each financial and economic sample data, and select the best classification result for each decision tree among all the classification results. Analyze the membership degree of each decision tree to the data category, set the membership degree of the i decision tree to the b data category as B_{ib}, and the replacement accuracy of the membership degree B_{ib} as $C(B_i)$, calculate the node purity $d(B_{ib})$ of the membership degree B_{ib}, and the formula is:

$$d(B_{ib}) = D(B_{ib}) log \frac{E}{e_i} \tag{4}$$

where $D(B_{ib})$ is the ratio of the sum of membership degrees of B_{ib} and i decision tree to all data categories, E is the number of categories of the random forest classifier, $b \in [1, E]$, e_i are the personalized parameters of the i decision tree, and the actual sample data Semantic correlation [9]. Calculate the membership reliability c_i of the i decision tree, the formula is:

$$c_i = \frac{C(B_{ib})}{A} d(B_{ib}) \tag{5}$$

where A is the total number of financial and economic sample data, and $i \in [1, A]$. The larger the value of c_i, the higher the credibility of the membership B_{ib} of the i decision tree to the b data category. According to the credibility c_i, the decision trees of all sample data are arranged in descending order. When the system when adding new enterprise financial and economic data, the membership reliability of each decision tree is dynamically adjusted, and the new decision tree is ranked in descending order.

Arrange the results in descending order of the decision tree, match the weights of the weight set in turn, set the weight set to $\{k_1, k_2, \cdots, k_A\}$, and assign a weight value k_i to the membership B_{ib} of the i decision tree. Use the learning utility function K to evaluate the sample set, make each enterprise financial and economic data learn a utility function, assign a utility score for each data, and keep the constraint conditions of all the effective scores consistent. The expression of the learning utility function Q is:

$$Q = m \sum_{i=1}^{A} \frac{C(B_{ib})}{A} M(B_{ib}) d(B_{ib}) k_i \tag{6}$$

Among them, m is the utility score, and $M(B_{ib})$ is the priority of membership B_{ib} [10]. The Q value is closely related to the accuracy of the membership degree of the sample data. Set the high and low thresholds of the function Q. When Q is greater than the high threshold, the decision tree's membership of the data category must be accurate. When Q is less than the low threshold, the decision is determined the membership of the tree to the data category must be inaccurate. If Q is less than the high threshold and greater than the low threshold, the membership of the decision tree to the data category is judged to be a fuzzy number. When the learning utility function value of the membership degree of the decision tree is inaccurate or fuzzy value, the membership degree of the decision tree to the data category is re-selected, and the above process is repeated until the Q value of the membership degree is greater than the high threshold. Count the accurate membership degrees of each decision tree to all categories of the random forest classifier, and select the classification category with the highest membership degree as the classification result of the financial data. So far, the accurate classification of enterprise financial and economic data based on random forest is completed, the system software design is completed, and the hardware design and software design are combined to realize the accurate classification and management system design of enterprise financial and economic data based on random forest.

3 Experiment and Analysis

The designed system is compared with two conventional enterprise financial and economic data accurate classification management systems to compare the recall rate, false positive rate and accuracy of the three systems.

3.1 Experimental Data

The experiment is built on the Matlab platform. In order to verify the effectiveness of the system, the data set uses the quarterly financial statements of all listed companies from 2015 to 2020, including income statements, cash flow statements, and balance sheets. The statistical information sampling scale of financial data is 1000 MBit, the data set contains abnormal samples of abnormal financial conditions of enterprises, and normal samples of normal financial conditions. The random forest classifier starts from six different dimensions of capital structure, cash flow, operating capacity, development capacity, debt clearing capacity, and profitability. The classification categories of all financial and economic data included are shown in Table 2:

Table 2. Classification categories of financial and economic data

Type	Primary coverage
Capital structure	Asset liability ratio, long-term applicability of assets, dynamic liabilities, shareholders' equity, proportion of fixed assets and current assets
Cash flow	Cash recovery rate of total assets, sales revenue, cash proportion of operating income, and capital expenditure
Service power	Total asset turnover rate, fixed asset turnover rate, current asset turnover rate, accounts payable turnover rate, accounts receivable turnover rate, inventory goods turnover rate
Development ability	Growth rate of total assets, growth rate of net assets, net amount of cash held by enterprise operation, growth rate of net profit, total profit, growth rate of operating profit, growth rate of income from single share
Debt clearing capacity	Current cash-liability ratio, net operating current cash, net operating current cash, profit before amortization, depreciation and tax, property right ratio, overspeed moving ratio, current ratio
Profitability	Cost profit margin of capital, operating profit margin, asset impairment loss, total operating cost, net profit, net interest rate of assets, return on assets, return on equity

3.2   Analysis of Experimental Results

The three systems divide the categories of each sample data, and manage enterprise financial and economic data sets. Compare the data classification behavior recall rate R of the three systems to measure the proportion of normal samples that are correctly classified. The formula for calculating the R value is:

$$R = \frac{x_1}{x_1 + x_2} \qquad (7)$$

Among them, x_1 is the number of normal financial data classified correctly, and x_2 is the number of normal financial data classified incorrectly. The comparison result of the recall rate experiment is shown in Fig. 4:
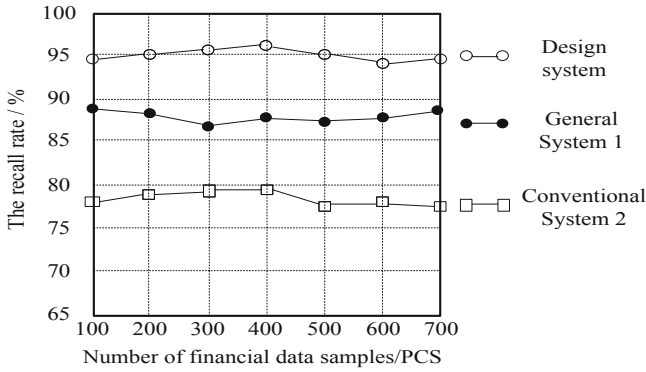
Fig. 4. Comparison results of the financial data classification recall rate experiment

It can be seen from the above figure that the average recall rate of data classification of the design system is 95.0%, the average recall rate of data classification of conventional system 1 is 86.8%, and the average recall rate of data classification of conventional system 2 is 78.9%. The recall rate of the design system has increased by 8.2% and 16.1% respectively, which improves the proportion of normal samples correctly classified.

Compare the misreport rate ξ of data classification behaviors of the three systems to measure the misclassification ratio of abnormal samples. The formula for calculating the value of ξ is:

$$\xi = \frac{x_3}{x_3 + x_4} \tag{8}$$

where x_3 is the number of abnormal financial data classified incorrectly, and x_4 is the number of abnormal financial data classified correctly. The experimental comparison result of false alarm rate is shown in Fig. 5:
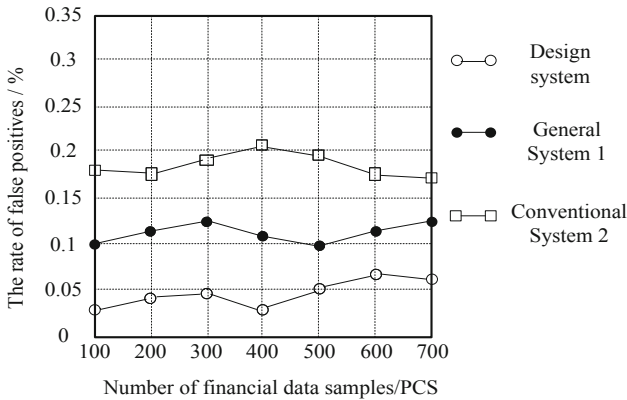
Fig. 5. Experimental comparison results of the false alarm rate of financial data classification

It can be seen from the above figure that the average false alarm rate of the design system is 0.05%, and the average false alarm rates of the other two systems are 0.11% and 0.19% respectively. The false alarm rates of the data classification behavior of the design system are reduced by 0.06% and 0.14% respectively, reducing the misclassification proportion of abnormal samples.

Compare the accuracy rate β of the data classification behavior of the three systems, and measure the proportion of the normal sample and the abnormal sample that are correctly classified. The calculation formula of the β value is:

$$\beta = \frac{x_1 + x_4}{x_1 + x_2 + x_3 + x_4} \tag{9}$$

The comparison result of the accuracy rate experiment is shown in Fig. 6:
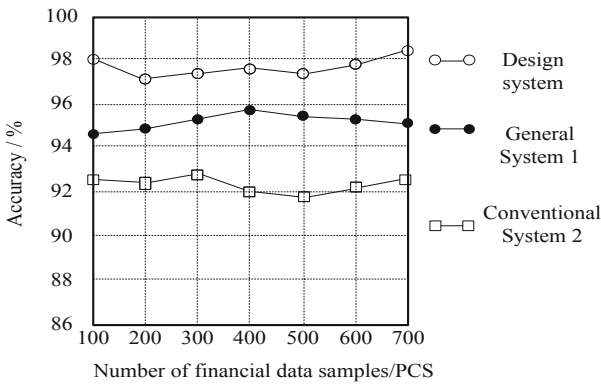
Fig. 6. Experimental comparison results of the accuracy of financial data classification

It can be seen from the above figure that the average accuracy of the design system is 97.8%, the average accuracy of the other two systems are 95.1% and 92.3% respectively, and the accuracy of data classification behavior of the design system is increased by 2.7% and 5.5% respectively, which improves the proportion of correct classification of normal data and abnormal data.

4   Conclusion

In order to improve the classification effect of financial data, this paper designs an accurate classification management system of enterprise financial and economic data. Classify the financial and economic data of enterprises, and the classification results of financial data are more accurate and reliable. In the hardware, the front end, middle layer, server end and enterprise financial and economic data display end are used to form the overall architecture of the system, optimize the data memory of the server end, and transform the serial communication circuit of the development board; In the software design, abnormal financial data are filtered, a decision tree is established for each sample

data, the utility function value is learned through the membership of the decision tree, and the optimal classification category is selected for the data through the random forest classifier. The experimental results show that the designed system improves the recall and accuracy of data classification, reduces the false positive rate, and the financial data classification results are more accurate and reliable. However, there are still some deficiencies in this design system. In the future research, the characteristic parameters of financial data will be filled in the thing characteristic table. Through the intelligent filling of characteristic parameters, the data of classification management will be more accurate.

References

1. Cui, B., Gao, J., Tong, Y., et al.: Progress and trend in novel data management system. J. Softw. **30**(1), 164–193 (2019)
2. Zhao, Z., Shen, Z.: An interactive analysis framework for multivariate heterogeneous graph data management system. Data Anal. Knowl. Discov. **3**(10), 37–46 (2019)
3. Ma, L., Wang, J., Chen, H.: Business data security of system wide information management based on content mining. J. Comput. Appl. **39**(2), 488–493 (2019)
4. Zang, H., Zhao, Q., Li, G., et al.: Design and implementation of data acquisition and management system of agronomic trait for maize. J. South. Agric. **50**(11), 2606–2613 (2019)
5. Li, S., Li, Z., He, Y., et al.: Design and implementation of information management system for ballastless track monitoring data. Railway Stand. Des. **63**(9), 28–33 (2019)
6. Zhu, F., Guo, J.-F., Cao, L.: Hierarchical recognition of data multi label features based on classification rule mining. Comput. Simul. **38**(4), 310–314 (2021)
7. Li, W., Zhao, F.: Application of data statistics management system in hospital performance management. Bull. Sci. Technol. **35**(2), 178–182 (2019)
8. Li, T., Qiu, W., Liu, Y.: Development and application of refined management system of tunnel geological information based on data drive. Tunnel Constr. **39**(1), 68–74 (2019)
9. Liu, S., Bai, W., Srivastava, G., Machado, J.A.T.: Property of self-similarity between baseband and modulated signals. Mob. Netw. Appl. **25**(4), 1537–1547 (2019). https://doi.org/10.1007/s11036-019-01358-9
10. 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)

Marketing Risk Assessment Method of Industrial and Commercial Enterprises Based on Convolutional Neural Network

Yang Li[1(✉)] and Shuang Wang[2]

[1] Guangzhou Huali College, Guangzhou 511325, China
lyang424210@163.com
[2] College of Humanities and Information, Changchun University of Technology, Changchun 130122, China

Abstract. As a part of the whole enterprise system function, marketing management always runs through the beginning and end of the enterprise operation process. Its running status determines whether the whole enterprise system operates effectively. With the increasingly complex and changeable internal and external business environment, enterprises are facing more and more uncertain factors and more marketing risk events. Facing the increasingly prominent marketing risk problems, enterprises lack enough attention and effective preventive measures and means, which makes many enterprises fall into the dilemma of loss or even bankruptcy. Therefore, it is of great practical significance to strengthen the research on enterprise marketing risk management in complex environment. Therefore, a new marketing risk assessment method for industrial and commercial enterprises is designed based on convolutional neural network. Firstly, the marketing risk management objectives of industrial and commercial enterprises are analyzed. Secondly, the risk index system is constructed based on convolutional neural network, so as to realize the marketing risk assessment of industrial and commercial enterprises. The results show that the evaluation effect of the designed marketing risk assessment method is good, It has certain application value.

Keywords: Convolutional neural network · Industrial and commercial enterprises · Marketing · Risk assessment

1 Introduction

With the continuous deepening of China's economic system reform, the gradual development and improvement of the market economic system and the gradual integration of the domestic market with the international market [1], the institutional environment and business environment of Chinese enterprises have also undergone extensive and profound changes, such as the increasingly fierce market competition, the acceleration of the pace of industrial structure adjustment, the fast-paced change of market demand, the continuous emergence of new technologies and products [2–4], the change of public

consumption habits, etc. All these make the internal and external environment of the enterprise more complex and uncertain. This complex and uncertain business environment not only brings unlimited market opportunities to enterprises, but also brings huge risks [5]. The first is that the marketing management of enterprises will face great risks and challenges. Because, as a part of the whole enterprise system function, marketing always runs through the beginning and end of the enterprise operation process, and is contained in the enterprise functions such as strategy, brand, quality, advertising planning, public relations image, channel, after-sales service and so on [6]. The function of marketing management not only highlights the importance of its position, but also shows the risk and complexity of marketing management. Looking back on the development process of Chinese enterprises in recent ten years, behind the abnormal marketing phenomena such as price war, advertising war, brand extension war and diversification war that once prevailed in the market, how many once brilliant enterprises fell down overnight and even withdrew from the historical stage [7]. The painful market outcome tells us that the marketing management of Chinese enterprises generally lacks the awareness of risk prevention. This not only directly caused the confusion of market competition and irreparable huge losses to enterprises, but also had an adverse impact on the development of the whole national economy.

Marketing risk is a common phenomenon in market economy, and it is the possibility of causing damage to enterprises randomly due to the influence of various factors in enterprise marketing activities. Enterprises in market economy, as market subjects [8–10], should participate in market activities independently in the fierce market competition, fully capture market opportunities and avoid market risks. However, due to various reasons, many enterprises are often unable to effectively evaluate, analyze and even make decisions when facing marketing risks. For example, the enterprise risk avoidance management organization is not perfect, the operation procedures are not standardized, the implementation of the system is not strict, the management of loopholes. Marketing personnel lack the ability to analyze and identify risks, lack of awareness of risk prevention, enterprise marketing risk prevention measures ineffective, marketing behavior does not adapt to the rapidly changing market demand. How to avoid marketing risk has become an important factor for the development of enterprises. Chinese enterprises should first have a deeper understanding of marketing theory, and then establish a set of advanced theoretical system of marketing risk analysis, prediction and prevention, and also need a set of standardized, practical and efficient technical means and methods. Gradually establish the corresponding close to the market, flexible operation of the marketing system, so as to avoid marketing risks. At present, Chinese enterprises are generally in a complicated marketing environment with high risks and many opportunities. However, enterprises lack sufficient understanding of the increasingly serious risk problems, and lack of effective preventive measures and means, resulting in many enterprises facing losses and even bankruptcy.

Therefore, it is necessary to deeply reflect on the marketing management links of enterprises under the new environment and further analyze the marketing management mechanism of enterprises. Based on this, this paper designs a new marketing risk assessment method for industrial and commercial enterprises. This paper analyzes the marketing risk management objectives of industrial and commercial enterprises, constructs the

risk index system based on convolutional neural network, and realizes the marketing risk assessment of industrial and commercial enterprises. The experimental results show that the designed marketing risk assessment method has good evaluation effect and certain application value.

2 Design of Marketing Risk Assessment Method for Industrial and Commercial Enterprises Based on Convolutional Neural Network

2.1 Analyze the Marketing Risk Management Objectives of Industrial and Commercial Enterprises

Risk is a kind of uncertainty, which is manifested in that it may bring losses to the people facing risks and huge benefits. Marketing risk means that in the process of enterprise marketing, due to the influence of various unpredictable uncertain factors, the actual income of enterprise marketing deviates from the expected income, so that there is the opportunity or possibility to suffer losses and obtain additional income. The meaning of marketing risk is similar to that of risk. The difference is that he emphasizes that the subject of risk is the participants and competitors of marketing activities. Its loss is the punishment suffered due to violation of market laws or their own mistakes, mainly refers to the reduction or loss of economic interests. Most of its risks arise from marketing activities or related aspects, The risk condition is the uncertain event caused by the market behavior or marketing event of the business entity.

Marketing risk is a kind of complex risk form, which has many contents and forms. According to the different degree of risk, marketing risk can be divided into several levels. The first level is fatal marketing risk, which refers to the risk of large loss and serious consequences. The direct consequences of such risks often threaten the survival of marketing subjects and lead to heavy losses, which means that they can not recover or suffer bankruptcy for a time. For example, major changes have taken place in the market situation, and the enterprise fails to fully predict, resulting in major market behavior mistakes. This makes the frustrated enterprise products have no market and no further development, so that the enterprise falls into serious difficulties. The second level is the general marketing risk, which is all kinds of risks with moderate loss and obvious consequences but does not pose a fatal threat. The direct consequences of such risks have certain sequelae. If part of the enterprise's accounts receivable cannot be recovered on time, the enterprise still has a large amount of bad debt losses after efforts and invalid legal procedures, and the enterprise suffers economic losses, which makes it difficult to turnover working capital. The third level is the slight enterprise risk, which makes the loss small, the consequences are not obvious, and does not have an important impact on the marketing activities of the marketing subject. Generally, such risks do not affect the overall situation and only cause local and minor harm to the marketing subject. For example, the enterprise caused some product losses during transportation due to packaging problems, resulting in events affecting customer distribution and customer claims, which had an adverse impact on the enterprise. The division of these three levels of risk is not absolute. Under certain conditions, general marketing risk and slight

marketing risk will be transformed into special fatal risk, especially after a certain period of accumulation, there will be qualitative changes. If the accounts receivable cannot be recovered for a long time and are illegally occupied by other enterprises for a long time, the consequences will be disastrous for the enterprise, and the general risk and minor risk will be transformed into fatal risk. Therefore, the understanding, analysis, prediction and control of marketing risk is mainly aimed at fatal marketing risk and general marketing risk, because this is the main aspect of the contradiction and the main task of marketing risk management. At the same time, it is also necessary to pay attention to minor marketing risk and prevent it.

In economic activities, in addition to the irregular changes in the natural and social environment, the complexity of marketing activities, and the continuous changes in consumer demand and power, but also due to the limitations of the experience and ability of marketing subjects, they do not fully understand and grasp the generation, development and consequences of risks, or fail to take timely and effective measures to prevent them, And cause all kinds of losses. The most common is the marketing staff's own mistakes, that is, the risks and losses caused by the marketing staff's sense of responsibility and work quality problems. Under what circumstances should marketers take the initiative to bear part of the risk losses and avoid some risk losses, which risks must be borne by marketers, which can and should be avoided, and whether the avoidance is due to cost considerations or strategic and strategic considerations, correct decisions and arrangements should be made, otherwise improper bearing or avoidance will lead to new risk losses, The actual loss will be much greater than expected, that is, a risk will lead to a new risk, and a small risk will lead to a large risk. For example, if the enterprise plans to sell at a lower price and make profits, that is, it will first make some financial sacrifices to defeat its competitors. However, due to improper deployment and implementation, the enterprise has suffered significant economic losses and not only failed to occupy the expected market. However, they were caught by their opponents, denounced as unfair competition and sued in court. The enterprise was not prepared enough to lose the lawsuit, and new economic losses occurred. In the communication between enterprises and other business entities or competitors in the market, if they are careless, they will be used by the other party, forming the risk of their own loss, which is common in the complex market economy.

After the marketing loss of an enterprise, a problem that marketing managers are very concerned about is the impact of the loss event on the profitability of the enterprise. Generally speaking, an enterprise will have a minimum reward string. It is not only the standard to judge whether a marketing activity is feasible, but also the standard to formulate the marketing risk management plan. Marketing risk managers must control the loss within a certain range, in which the profitability of the enterprise will not be lower than the minimum rate of return. The stability of income is very important for enterprises, because it can help enterprises establish a good image of normal development and enhance the investment confidence of investors. For most investors, an enterprise with stable income is more attractive than an enterprise with high income and high risk. Stable income means the normal development of enterprise marketing; In order to achieve the goal of stable income, enterprises must increase risk management expenditure.

The production and operation of the enterprise is like "sailing against the current, if you don't advance, you will fall back". With the increasing competition in modern society, enterprises can firmly attract customers only by constantly introducing newer and higher quality products. Only by constantly exploring new markets can enterprises occupy a leading position in the market. If the enterprise stagnates and lingers on the original performance, the competitors will ruthlessly take away its customers and squeeze it out of the market through strength expansion. Therefore, the enterprise must continue to develop in order to obtain permanent survival. However, the existence of risk has become a potential resistance to the development of enterprises, because the losses caused by risk accidents will have a great impact on the development of enterprises. In order to achieve the development goal, the marketing risk manager must establish a high-quality marketing risk management plan to deal with all kinds of loss results timely and effectively, so that the enterprise can survive the loss. It can quickly obtain compensation and create good conditions for the continuous development of the enterprise. As stated in the pre loss goal, the enterprise can timely and effectively deal with the losses caused by marketing risk accidents and reduce the adverse impact of losses, which can reduce the impact on the national economy and protect the interests of personnel and economic organizations related to the enterprise, which is conducive to the enterprise to bear social responsibility and establish a good social image.

There is a certain connection between all these pre loss goals and post loss goals. For example, in order to achieve the safety factor goal, the marketing risk should be transferred before loss, so as to reduce the loss and make up for the loss to a certain extent, so as to reduce the impact on enterprise marketing. However, it is difficult to achieve all pre loss and post loss goals at the same time. Because there are various conflicts between pre loss objectives and post loss objectives, the realization of any post loss objective requires a certain amount of capital investment. Moreover, with the improvement of the post loss target level, the amount of funds required is also rising, which obviously conflicts with the economic target in the pre loss target. In addition, the safety factor goal in the pre loss goal also conflicts with the economic goal. In order to obtain greater security, marketing risk managers need to make more use of some high-cost marketing risk treatment technologies and increase marketing managers' measures to control risks and risk guarantee, so as to reduce losses and obtain timely and sufficient economic compensation after losses, which will inevitably lead to a sharp rise in marketing risk management costs. Marketing risk managers should properly handle the conflict between objectives, take the overall objective of the enterprise as the commander, widely solicit the opinions of relevant departments, and formulate a marketing risk management objective suitable for the specific situation of the enterprise.

2.2   Risk Index System is Constructed Based on Convolutional Neural Network

Convolutional neural network is generated on the basis of modern neuroscience, biology, psychology and other scientific research achievements, reflecting the basic characteristics of biological nervous system, is a kind of abstraction, simplification and simulation of biological nervous system. The convolutional neural network is composed of many parallel interconnected same neuron models, and the signal processing of the network is realized by the interaction between neurons. The neuron model and structure of a

convolutional neural network describe how a network transforms its input vector into output vector [11]. This transformation process is a computational process from a mathematical point of view, that is to say, the essence of convolutional neural network reflects a functional relationship between network input and its output. By selecting different model structures and activation functions, different convolutional neural networks can be formed to obtain different input/output relations, and achieve different design objectives and complete different tasks. Therefore, before using convolutional neural network to solve practical application problems, it is necessary to first master the model structure and characteristics of convolutional neural network as well as the calculation of its output vector. Therefore, it is necessary to design the neuron model when constructing the risk indicator system, as shown in Fig. 1.
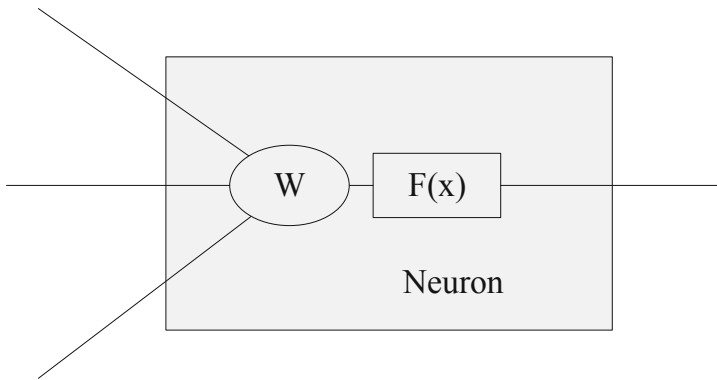
Fig. 1. Schematic diagram of neuron model

It can be seen from Fig. 1 that the model is generally a multi input and output nonlinear element. The neuron output is not only affected by the input signal, but also affected by other factors inside the neuron. Therefore, in the modeling of artificial neuron, an additional input signal is often added, which is called deviation, sometimes called threshold. At this time, it is necessary to calculate the column vectors of different matrices, As shown in formulas (1) and (2).

$$W = [w_1 w_2 ... w_r] \tag{1}$$

$$P = \left[p_1 p_2 ... p_r \right]^r \tag{2}$$

In formulas (1) and (2), $w_1 w_2 ... w_r$ and $p_1 p_2 ... p_r$ represent the array vector set, which need to be integrated to build the neuron model and output the vector, as shown in formula (3).

$$A = f(W \times P + b) \tag{3}$$

In formula (3), B represents the deviation. Activation function is the core of a neuron and network. In addition to the network structure, the problem-solving ability and efficacy of the network depend largely on the activation function adopted by the network.

At this time, the relationship between the activation function and neurons is shown in formula (4).

$$A = \begin{cases} 1 \ W \times P + b \geq 0 \\ 0 \ W \times P + b \leq 0 \end{cases} \tag{4}$$

At this time, the neuron relationship meets the risk assessment criteria, that is, the constructed index is shown in formula (5).

$$f = \frac{1}{1 + \exp[-(n + b)]} \tag{5}$$

In formula (5), n represents the number of indicators. Using this index, we can accurately carry out risk assessment and reduce the difficulty of risk assessment.

2.3 Realize Marketing Risk Assessment of Industrial and Commercial Enterprises

Convolutional neural network has a very outstanding performance in the image classification data set. Due to the large normal image information, a large number of weights need to be set if the fully connected neural network is used, and the neural network evaluation model needs to be built to realize the marketing risk assessment of industrial and commercial enterprises, as shown in Fig. 2.
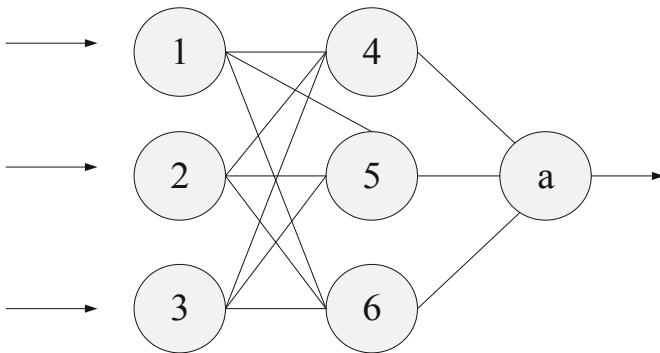
Fig. 2. Neural network evaluation model

According to the neural network evaluation model in Fig. 2, marketing risk evaluation can be further carried out. Due to the complexity of the marketing system, the indicators reflecting the marketing risk of enterprises are also quite complex. Therefore, we should follow certain principles to select early warning indicators. It is not only the goal of risk early warning management, but also can comprehensively and truly reflect the operation effect and existing problems of the early warning system, and really play the role of early warning. This is especially reflected in the definition and calculation method of indicators.

Sensitivity of indicators to system changes. That is, the indicators are required to not only quickly reflect the subtle changes of the system, but also advance the actual fluctuations, sensitively predict and reflect the generation and development of enterprise marketing activities, and timely reflect the real state of enterprise marketing, which is pioneering or universal.

The more indicators are not the better. The main, basic and representative comprehensive indicators should be selected as quantitative calculation indicators to facilitate horizontal and vertical comparison. At the same time, the quantification of indicators and the difficulty and reliability of data acquisition should be considered; The evaluation model and method shall be easy to master, the required data shall be easy to statistics, and the existing statistical data shall be used as much as possible. Relevance requires that each index included in the index system can correctly reflect the content and status of marketing risk from a certain angle, so as to make the index and index become an organic whole connected with each other; Independence means that the overlapping area between indicators should be minimized when designing indicators. At the same time, the relationship between indicators is not contradictory, but dialectical and unified.

Independence makes the interrelated indicators have their own characteristics, and relevance makes the independent indicators become a whole and serve the marketing risk early warning system. In the long run, the internal and external environment of the enterprise is developing and changing. Specifically, marketing risk early warning management is not only a goal, but also a process. This determines that the index system should have dynamic characteristics. That is, with the development of enterprises, the index system should be amended correspondingly, and the contents of new monitoring should be supplemented to ensure the advanced nature of the risk early warning monitoring system. Meanwhile, the contents of the index system should not be changed frequently, and should be relatively stable in a certain period.

In addition to quantitative indicators, some indicators can not be measured by accurate data, but can only be described and evaluated qualitatively. However, they are very important for marketing risk early warning management, so they must be used in order to comprehensively reflect the current situation and trend of enterprise marketing. The establishment of the rating index system should conform to the objective facts and correctly reflect the real appearance of the rating object. The index system and calculation method shall not be biased towards any party of the rating object or rating subject. Rating agencies and rating personnel must be fair, objective and based on facts, and must not arbitrarily change the index items, calculation methods and evaluation standards according to personal preferences. Credit rating must comply with relevant national policies, laws and regulations. The index system should reflect the guidance of national macro policies. Some economic benefit indicators and risk supervision indicators should reflect the specified requirements if the government stipulates standard values. In the determined evaluation indicators, it is mainly divided into objective quantitative indicators and subjective qualitative indicators.

Enterprise external marketing risk is determined according to four major indexes: market risk, customer risk, supplier risk and competitor risk. The range of index range can be determined according to the general situation of Chinese enterprises, relevant data and the results of marketing questionnaire survey. In the specific evaluation, we should

adjust these indicators according to the marketing situation of enterprises, and set the weight of specific enterprises. Some objective indicators that are difficult to obtain can be regarded as qualitative indicators to evaluate. Therefore, the establishment of indicators is for external risks, and according to the competition risk, customer risk, supplier risk, competitor risk four modules. The determination of index range can be based on the current general situation of Chinese enterprises, relevant data and the results of marketing questionnaire survey as the demonstration of setting index range. In the specific evaluation, these indicators should be adjusted according to the marketing situation of enterprises, and the weight of specific enterprises should be set. Some objective indexes that are difficult to obtain can be evaluated as subjective fuzzy indexes.

3 Case Analysis

3.1 Overview

An industrial and commercial enterprise is selected for marketing risk assessment. It has been proved theoretically that the BP neural network with only one hidden layer using sigmoid function can approach any rational function. Considering that the scale, training time and complexity of the system are reduced as much as possible, a three-layer BP network structure is adopted in this example, that is, a hidden layer is adopted. According to the basic principle of neuron selection and common empirical formula, combined with the specific situation, the number of neurons in the hidden layer is determined through trial calculation. The output layer establishes the neural network model according to the classification of marketing risk level. At this time, the initial data of risk indicators are shown in Table 1.

Table 1. Initial data of risk indicators

Indicators	x1	x2	x3
1	0.55	0.27	0.95
2	0.47	0.22	1.0
3	0.43	0.13	0.75
4	0.40	0.18	0.11
5	0.42	0.10	0.06
6	0.45	0.15	0.26
7	0.44	0.12	0.76
8	0.37	0.09	0.87
9	0.40	0.04	0.88
10	0.60	0.09	0.74

Table 1. (*continued*)

Indicators	x1	x2	x3
11	0.60	0.14	0.69
12	0.75	0.19	0.92
13	0.65	0.36	0.90
14	0.52	0.25	0.95
15	0.63	0.20	0.83

It can be seen from Table 1 that there are differences among various risk indicators at this time, which meets the subsequent evaluation needs, and the risk evaluation test can be carried out.

3.2 Application Effect Analysis

The enterprise marketing risk assessment method designed in this paper and the traditional enterprise marketing risk assessment method are used for risk assessment respectively. The application effect is shown in Table 2.

Table 2. Application effect

Index	Accurate strength	This paper designs a risk assessment method to assess the intensity	The assessment intensity of traditional risk assessment methods
1	3	3	3
2	2	2	1
3	1	1	1
4	1	1	1
5	2	2	2
6	3	3	3
7	2	2	2
8	1	1	1
9	3	3	2
10	2	2	1
11	1	1	1
12	3	3	2

(*continued*)

Table 2. (*continued*)

Index	Accurate strength	This paper designs a risk assessment method to assess the intensity	The assessment intensity of traditional risk assessment methods
13	2	2	1
14	1	1	2
15	2	2	2

Among them, 3 indicates high accuracy intensity of marketing risk grade, 2 indicates medium accuracy intensity of marketing risk grade, and 1 indicates low accuracy intensity of marketing risk grade.

It can be seen from Table 2 that the marketing risk level evaluated by the method designed in this paper is more accurate, which proves that the evaluation effect of the designed method is good and has certain application value.

4   Conclusion

By designing the marketing risk evaluation index system, the marketing risk evaluation model based on BP neural network is designed according to the theoretical knowledge of neural network. Among them, the marketing risk model based on neural network established in this paper has the function of self-adaptation and self-learning. Therefore, it has practical and extensive application in the monitoring and prevention of enterprise marketing risk. The evaluation result of this method is good and has certain application value.

Marketing risk management is a complex system engineering, limited by knowledge and theoretical foundation, the study of this paper is only a preliminary theoretical framework, and only provides a research perspective and method. The requirements and operation methods in practical application still need to be further improved in practice. In addition, there are still many deficiencies in this paper, such as the effectiveness of the enterprise marketing index system and the relevance and independence of each index, whether the scope of grade division meets the requirements of practice, the established index system and model has not been effective practice test, etc., all these need to be further studied and discussed.

References

1. Song, Y., Wu, R.: The impact of financial enterprises' excessive financialization risk assessment for risk control based on data mining and machine learning. Comput. Econ. **10**(6), 1–23 (2021)

2. Huang, X., Sun, J., Zhao, X.: Credit risk assessment of supply chain financing with a grey correlation model: an empirical study on china's home appliance industry. Complexity **2021**(2), 1–12 (2021)
3. Korshunov, G.I., Kabanov, E.I., Cehlár, M.: Occupational risk management in a mining enterprise with the aid of an improved matrix method for risk assessment. Acta Montanistica Slovaca **25**(3), 289–301 (2020)
4. Liu, J., Yin, Y., Yan, S.: Research on clean energy power generation-energy storage-energy using virtual enterprise risk assessment based on fuzzy analytic hierarchy process in China. J. Clean. Prod. **236**(1), 1–14 (2019)
5. da Silva Etges, A.P.B., et al.: Proposition of a shared and value-oriented work structure for hospital-based health technology assessment and enterprise risk management processes. Int. J. Technol. Assess. Health Care **28**(4), 1–9 (2019)
6. Xivry, G., Quesnel, M., Vanberg, P.O., et al.: Focal plane wavefront sensing using machine learning: performance of convolutional neural networks compared to fundamental limits. Mon. Not. R. Astron. Soc. **505**(4), 5702–5713 (2021)
7. Mukherjee, P., Mazumdar, C.: "Security Concern" as a metric for enterprise business processes. IEEE Syst. J. **13**(4), 4015–4026 (2019)
8. Saeidi, P., Saeidi, S.P., Sofian, S., et al.: The impact of enterprise risk management on competitive advantage by moderating role of information technology. Comput. Stand. Interfaces **63**(5), 67–82 (2019)
9. Zhou, C., Wang, D.X.: A risk assessment algorithm for college student entrepreneurship based on big data analysis. Complexity **2021**(5), 1–12 (2021)
10. Deng, H., Zhang, A.: Fuzzy hierarchy analytic method of enterprise supply chain financial risk. J. Intell. Fuzzy Syst. **1**, 1–10 (2021)
11. Yang, S.L., Sun, J., Yan, Z., et al.: Research on image interframe compensation based on deep convolutional neural network. Comput. Simul. **37**(1), 452–455 (2020)

Evaluation Method of Enterprise Economic Management Model Effectiveness Based on Deep Data Mining

Yipin Yan and Lu Zhang[(✉)]

Faculty of Management, Chongqing College of Architecture and Technology,
Chongqing 400000, China
`pengqian123678@163.com`

Abstract. Traditional methods have some problems in evaluating the effectiveness of enterprise economic management model, such as low evaluation accuracy and long time-consuming. An evaluation method of enterprise economic management model effectiveness based on deep data mining is proposed. According to the screening principle of the effectiveness evaluation indicators of enterprise economic management mode, screen the effectiveness evaluation indicators of enterprise economic management mode, refer to the characteristic attributes of candidate indicators, analyze the dimension of candidate indicators, quantify the evaluation indicators, and calculate the index weight by using deep data mining technology to realize the effectiveness evaluation of enterprise economic management mode. The example analysis results show that the evaluation method effectively improves the accuracy of the effectiveness evaluation of enterprise economic management model.

Keywords: Deep data mining · Economic management mode · Effectiveness evaluation · Weight determination · Evaluation system · Index screening

1 Introduction

With the analysis of social and economic development and reform process, the ultimate pursuit of enterprises in the development process is to maximize profits. After the long-term practice of modern enterprise economic management, the standardized development of enterprise economic management model has a long way to go [1]. Enterprise economic management is to realize the value of enterprise resources and the actual management of the enterprise economy. At the same time, in the process of economic development, enterprises conduct scientific and reasonable management, organization, planning and monitoring of economic activities, and realize the sustainable development of the enterprise economy by improving the economic benefits of the enterprise [2]. Therefore, the process of developing economic management is mainly divided into two aspects: First, according to the company's own characteristics and the development

law of production and operation, through a reasonable method, the company's product pricing and a reasonable salary system for employees are set to ensure.

In domestic research, Liu Zhenhua et al. [3] considered that the environmental pollution caused by heavy polluting enterprises in the process of production and operation has become one of the important factors restricting the healthy development of China's economy and society, and built a performance evaluation system for heavy polluting enterprises to identify the existing problems of enterprises and provide reference for the formulation of improvement measures. Firstly, based on the sustainability theory, the performance evaluation index system of heavy polluting enterprises is established from the perspective of economy, society and environment. Then, the performance index weighting model of heavy polluting enterprises is constructed based on COWA operator. Finally, The performance evaluation model of heavy pollution enterprises is constructed based on the cloud model. In addition, the feasibility of the performance evaluation system is verified by case analysis; Zhang Yulan et al. [4] established an investment efficiency evaluation system for manufacturing enterprises from the perspective of technological innovation, and used DEA model to evaluate the investment efficiency of 158 Manufacturing Listed Companies in Beijing, Tianjin and Hebei from 2016 to 2018.The study found that the average investment efficiency of listed companies in the Beijing-Tianjin-Hebei manufacturing industry is between 0.75 and 0.79, and the investment efficiency is low. From the perspective of manufacturing industry segments, the investment efficiency of companies related to the petrochemical industry is the lowest, and from the perspective of property rights. The investment efficiency of state-owned enterprises is above 0.8, which is always higher than that of private enterprises. From the perspective of regional distribution, the investment efficiency of listed manufacturing companies in Beijing is the highest, followed by Hebei Province, and Tianjin is the lowest. In addition, the Malmquist index model is used for investment efficiency. The dynamic change value was measured and found that the average total factor productivity change was 0.944, indicating that the overall investment efficiency of the Beijing-Tianjin-Hebei manufacturing industry has declined. Based on this, suggestions for improving the investment efficiency of manufacturing enterprises are put forward from the perspectives of both the enterprise and the government.

The process of enterprise economic management mode reform is reflected through the change of economic management level, and to comprehensively strengthen enterprise economic management, we should start with reforming the operation mechanism of economic management mode and improve enterprise economic operation and financial management system [5]. Thus, it is of great practical significance to change the extensive economic management mode of enterprises in the past, improve enterprise operation efficiency, reduce costs, realize fine management, mobilize the enthusiasm of enterprise staff, and return to public welfare.

Based on the above research background, this paper uses deep data mining technology to design an evaluation method of enterprise economic management mode, so as to improve the economic benefits of enterprises.

2   Design of Effectiveness Evaluation Method of Enterprise Economic Management Model

2.1   Screening Evaluation Indicators of Enterprise Economic Management Mode

Since this article is only for the evaluation of the economic management model of the enterprise, not the overall work development of the enterprise, it is necessary to screen and analyze the evaluation indicators of the economic management mode of the enterprise. The following principles must be followed when selecting the indicators: 1. Integrity; 2. Relevance; 3. Statistics; 4. Independence. "Integrity" emphasizes that the indicator must be aimed at the entire enterprise rather than some employees, "relevance" emphasizes that the indicator is closely related to the economic management of the enterprise, and "statistics" emphasizes that the indicator can be quantified and easily obtained and cannot be a rigid absolute value., "Independence" emphasizes that the indicator is not interfered by subjective factors of statisticians and objective factors that are not the enterprise itself. The indicators that meet the above conditions will be included as candidate indicators into the final screening range [6].

The focus of enterprise financial management is emphasized from the six dimensions of budget management, balance and risk management, asset operation, cost management, revenue and expenditure structure and development capacity, plus a total of 25 indicators of the depreciation life of fixed assets. The specific screening results are as follows (Table 1):

Table 1. Evaluation and screening indicators of enterprise economic management mode

Indicator name	Concrete content
Budget management	Budget revenue implementation rate. Budget expenditure implementation rate. Implementation rate of special financial appropriation
Balance and risk management	Balance rate of business income and expenditure. Asset liability ratio. Current ratio
Asset operation	Total asset turnover. Days sales outstanding. Inventory turnover
Cost control	Single product revenue of production department. Single product expenditure. Production cost rate
Revenue and expenditure structure	Personnel expenditure ratio. Public expenditure ratio. Cost management rate
Development capacity indicators	Personnel expenditure ratio. Public expenditure ratio. Cost management rate
Development capacity	Growth rate of total assets, net assets and fixed assets

2.2 Analysis of Candidate Index Dimensions for Enterprise Economic Management Model Evaluation

According to the tasks and objectives of the enterprise economic management model evaluation system, that is, taking economic management as the entry point, further strengthen the construction of enterprises in improving the level and capacity of economic management, and ensure that enterprises give full play to the important work of economic management characteristics and advantages. At the same time, refer to the candidates The characteristic attributes of the indicators and their role in the evaluation of enterprise economic management models have finally determined eight dimensions [7], namely: budget management, financial capital management, cost management, production expense management, production efficiency, and production quality, Comprehensive satisfaction, special services.

Budget management: these indicators are mainly based on the requirements of modern enterprise management. The enterprise needs to gradually establish a comprehensive budget management system. According to the long-term plan and operation objectives formulated by the enterprise, the scientific and standardized budget method should be used to reasonably arrange the funds required for various business work of the enterprise, and all fund arrangements should be supervised, accounted, evaluated, rewarded and punished in the whole process.

Financial fund management: these indicators are mainly aimed at the management of enterprise financial fund use, financial fund risk and financial burden. They are the main aspects of economic management. They are the comprehensive control and evaluation of enterprise economic operation and business development, which can effectively reduce financial fund risk and reduce enterprise economic loss, it is of great significance to improve business operation efficiency.

Cost management: Cost management is of great significance for companies to effectively use the company's production materials, reduce or even eliminate waste, accurately measure business consumption, and rationally purchase equipment. Enterprise managers can strengthen cost management, timely grasp the specific situation of cost changes, analyze the reasons, strengthen rectification, and summarize the key nodes of cost control, so as to continuously improve operational efficiency.

Production cost management: Although enterprises are under tremendous pressure for survival and development, they try to improve economic efficiency, but this is contrary to the public welfare attributes of enterprises. Therefore, enterprises must put the management of production costs to a certain level, continuously adjust the income structure of the enterprise, and ensure the sustainable development of the enterprise.

Production efficiency: the overload work of enterprise staff in China has become a norm, and the construction mode of "no holiday enterprise" has been popularized throughout the country. However, there is still a huge gap with the huge service demand of the people, which requires enterprises to strengthen management, continuously improve production efficiency, narrow the gap as much as possible and meet the service demand.

Production quality: production quality is not only the basis for the survival of enterprises, but also the core competitiveness of enterprises in the fierce market competition. The consequences of low production quality are very serious, which will directly endanger the economy and life of the people. Therefore, every enterprise regards production

quality as the "lifeline" of the enterprise, and strictly manages it to ensure the safety of staff.

Comprehensive satisfaction: comprehensive satisfaction mainly refers to the enterprise's satisfaction with the expectations of employees in many aspects. Satisfaction is mainly aimed at production service level and attitude, service quality and safety, production cost burden and so on. Employee satisfaction mainly focuses on the working environment and atmosphere, employee treatment and welfare, comprehensive strength of the enterprise and logistics support. The comprehensive satisfaction index can enable enterprise managers to grasp the enterprise management status in time, constantly rectify various problems, ensure the stability of employees, and ensure the orderly development of various work.

2.3   Determine the Weight of the Evaluation Index of the Enterprise Economic Management Model

The evaluation of enterprise economic management model is a complex process. The indicators involved are qualitative indicators, which are difficult to quantify. Due to different views on evaluation problems, different categories of people will have different ideas when determining the index weight [8]. The given index weight is usually in the real number interval. Therefore, it is suggested to use deep data mining technology to calculate interval number to evaluate the effectiveness of enterprise economic management model.

In the decision-making process, the enterprise economic management information based on deep data mining technology has the advantages of convenient operation, simple, intuitive and easy to understand. Therefore, deep data mining technology is widely used in the field of economic management model evaluation of various enterprises. Assuming that $\tilde{a} = [a_1, a_2] = \{a_1 \leq x \leq a_2, a_1, a_2 \in R\}$ represents a closed interval obtained by data mining technology, if $\tilde{a}$ satisfies $\tilde{a} = \{x|0 \leq a_1 \leq x \leq a_2\}$, then $\tilde{a}$ is a positive interval number. Let $\tilde{a} = [a_1, a_2]$, $\tilde{b} = [b_1, b_2]$, $k \geq 0$, then we can get the two interval number arithmetic rules, namely:

$$
\begin{aligned}
&\text{Number multiplication}: k\tilde{a} = [ka_1, ka_2] \\
&\text{addition}: \tilde{a} + \tilde{b} = [a_1 + b_1, a_2 + b_2] \\
&\text{subtraction}: \tilde{a} - \tilde{b} = [a_1 - b_2, a_2 - b_1] \\
&\text{Number multiplication}: k\tilde{a} = [ka_1, ka_2] \\
&\text{multiplication}: \tilde{a} \cdot \tilde{b} = [\min\{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}, \\
&\qquad\qquad \max\{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}]
\end{aligned}
\tag{1}
$$

In order to make up for the insufficiency of the in-depth data mining technology in the evaluation index weight determination process, the method of replacing the point value is used to calculate the validity evaluation index weight of the enterprise economic management model, and then the original data and results are calculated. Assuming that $\widetilde{A} = \left(\tilde{a}_{ij}\right)_{n \times n}$ represents a judgment matrix for evaluating the effectiveness of an economic management model, and the expression of $\tilde{a}_{ij}$ is $\tilde{a}_{ij} = \left[a_{ij}^{L}, a_{ij}^{U}\right]$, mark $A^{L} = \left(a_{ij}^{L}\right)_{n \times n}, A^{U} = \left(a_{ij}^{U}\right)_{n \times n}, \tilde{A} = [A^{L}, A^{U}]$, and the evaluation index vector $\tilde{x}$ can be

expressed as $\tilde{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$. Assuming that $\tilde{A} = [A^{\mathrm{L}}, A^{\mathrm{U}}]$ is given, then the steps of using in-depth data mining technology to calculate the effectiveness evaluation index weight of the enterprise economic management model are as follows:

Step 1: use the deep data mining technology to calculate the economic management mode eigenvectors x^L and x^U corresponding to the maximum eigenvalues of A^{L} and A^{U} [9];

Step 2: calculate α and β according to the maximum eigenvalues of A^{L} and A^{U}, and the formula is:

$$\alpha = \left[\sum_{j=1}^{n} \frac{1}{\sum_{i=1}^{n} a_{ij}^{\mathrm{U}}} \right]^{\frac{1}{2}} \tag{2}$$

$$\beta = \left[\sum_{j=1}^{n} \frac{1}{\sum_{i=1}^{n} a_{ij}^{\mathrm{L}}} \right]^{\frac{1}{2}} \tag{3}$$

In formula (2) and formula (3), α and β both represent the optimal vector values of different indicators, and n represent the number of index optimization.

Step3: Calculate the weight vector $\tilde{\omega} = [\alpha x^{\mathrm{L}}, \beta x^{\mathrm{U}}]$ of the effectiveness evaluation index of the enterprise economic management model.

In the process of collecting opinions from enterprise employees and economic managers, according to the calculation and evaluation method of index weight based on deep data mining technology, statistical management is adopted to form a unified standard for the effectiveness evaluation of enterprise economic management model, and the interval number weight vector of primary evaluation index in the effectiveness evaluation system of enterprise economic management model can be calculated [10], For the same reason, the weight of each secondary evaluation index can be obtained, as shown in Table 2.

Table 2. Validity evaluation index weight of enterprise economic management model

Indicator name	Weights	Indicator name	Weights
Budget revenue implementation rate	[0.20, 0.24]	Consumption of sanitary consumables	[0.36, 0.40]
Budget expenditure execution rate	[0.25, 0.29]	Consumables ratio	[0.10, 0.14]
Implementation rate of special fiscal appropriations	[0.14, 0.18]	Production settlement rate	[0.51, 0.55]
Assets and liabilities	[0.27, 0.31]	Single line cost	[0.52, 0.56]
Current ratio	[0.18, 0.23]	Daily production quantity	[0.22, 0.26]

(*continued*)

Table 2. (*continued*)

Indicator name	Weights	Indicator name	Weights
Accounts receivable turnover days	[0.28, 0.32]	Sales volume per day	[0.44, 0.48]
Depreciation period of fixed assets	[0.20, 0.25]	Utilization rate of production equipment	[0.29, 0.34]
Inventory turnover	[0.14, 0.18]	Rate of qualified products	[0.38, 0.43]

Due to the complexity of the effectiveness evaluation of enterprise economic management model, it is difficult to quantify the qualitative evaluation indicators. The weight of the effectiveness evaluation indicators of enterprise economic management model is calculated by using deep data mining technology, and the weight of the effectiveness evaluation indicators of enterprise economic management model is determined.

2.4   Establish the Effectiveness Evaluation System of Enterprise Economic Management Model

According to the weight of the effectiveness evaluation index of enterprise economic management mode, the effectiveness evaluation system of enterprise economic management mode is established, as shown in Table 3.

Table 3. Effectiveness evaluation system of enterprise economic management model

Target layer	Criterion layer	Index layer
Effectiveness Evaluation System of Enterprise Economic Management Model	Budget management	Budget revenue implementation rate
		Budget expenditure execution rate
		Implementation rate of special fiscal appropriations
	Financial Fund Management	Assets and liabilities
		Current ratio
		Accounts receivable turnover days
	Cost management	Depreciation period of fixed assets

(*continued*)

Table 3. (*continued*)

Target layer	Criterion layer	Index layer
		Inventory turnover
		Consumption of sanitary consumables
	Production cost management	Consumables ratio
		Production settlement rate
		Single line cost
	Productivity	Daily production quantity
		Sales volume per day
		Utilization rate of production equipment
	Production quality	Rate of qualified products
		Staff quality

Through the establishment of the effectiveness evaluation system of enterprise economic management mode, the effectiveness evaluation of enterprise economic management mode is realized.

3   Case Analysis

3.1   Sample Source

The case analysis takes the economic management data of an enterprise in 2018 and 2019 as the research object, and the specific samples are shown in Table 4.

Table 4. Experimental samples

First level indicator	Secondary indicators	2018	2019
Budget management	Budget revenue implementation rate	94%	97%
	Budget expenditure implementation rate	96%	97%

Table 4. (*continued*)

First level indicator	Secondary indicators	2018	2019
	Implementation rate of special financial appropriation	99%	99%
Financial fund management	Asset liability ratio	62%	66%
	Current ratio	56%	61%
	Days sales outstanding	53 days	55 days
Cost control	Depreciation life of fixed assets	-	-
	Inventory turnover	262 times	357 times
	Consumption of sanitary consumables	19 yuan	18 yuan
Production cost management	Consumable ratio	12.4%	12.2%
	Production settlement rate	76.32%	76.41%
	Cost of single pipeline	9016.5 yuan	9220.5 yuan
Production efficiency	Daily production quantity	803900	867400
	Daily sales quantity	25518	26039
	Utilization rate of production equipment	91.7%	91.7%
Production quality	Product qualification rate	99.9%	99.9%
	Staff quality	high	high

3.2   Evaluation Method

The evaluation indicators of this study are quantitative indicators, which are convenient for calculating the relative ratio score. However, due to the existence of proportion, value and negative value, in order to organically combine the evaluation results and facilitate comparative analysis, the index results need to be normalized. The processing principles and sequence are as follows:

Step 1: direct use of quantitative indicators;

Step 2: divide each indicator by the sum of each indicator in two years;

Step 3: according to the accounting system, the depreciation life of fixed assets of enterprises has different types and different types have unified provisions. Therefore, this item is not counted and replaced by "0";

Step 4: the higher the value of some indicators, the higher the priority (high priority), and the lower the value, the higher the priority (low priority). For high-quality indicators, use the results directly, and for low-quality indicators, use "1" minus the results of "(2) above;

Step 5: multiply the processing score of each evaluation index for two years by the weight of each index to obtain the final comparison score.

Although the index weights in this study total 100 points, the calculation of relative scores is used in specific applications rather than the calculation of absolute values, so the index system is suitable for comparative analysis.

3.3   Evaluation Results

The evaluation results of the effectiveness of the enterprise's economic management model in 2018 and 2019 are shown in Table 5.

Table 5. Effectiveness evaluation results of economic management model

Primary index	Secondary index	2018	2019
Budget management	Budget revenue implementation rate	2.8637	2.9541
	Budget expenditure implementation rate	1.1430	1.1548
	Implementation rate of special financial appropriation	3.2756	3.2756
Financial fund management	Asset liability ratio	2.7133	2.5511
	Current ratio	1.5424	1.6776
	Days sales outstanding	1.1707	0.1289
Cost control	Depreciation life of fixed assets	0.0000	0.0000
	Inventory turnover	0.9302	1.2680
	Consumption of sanitary consumables	1.5436	1.6248
Production cost management	Consumable ratio	1.1368	1.1554
	Production settlement rate	1.7188	1.7204
	Cost of single pipeline	1.4249	1.3935
Production efficiency	Daily production quantity	1.4069	1.4589
	Daily sales quantity	1.2260	1.3229
	Utilization rate of production equipment	3.3733	3.3733
Production quality	Product qualification rate	0.9800	1.9800
	Staff quality	1.9438	1.9438

It can be seen that the evaluation score of the effectiveness of the enterprise's economic management model in 2018 was 47.0339 and that in 2019 was 50.0017. Overall, the economic management in 2019 has improved and improved compared with that in 2018.

In order to further verify the effectiveness of the proposed method, experiments have compared the accuracy of this method, literature [5] method and literature [6] method in evaluating the effectiveness of enterprise economic management model. The results are shown in Fig. 1:
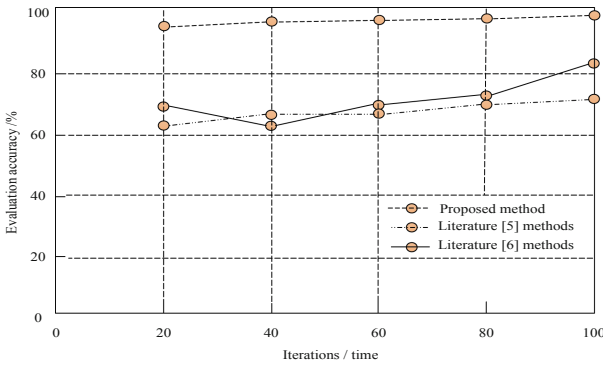
Fig. 1. Accuracy analysis of effect evaluation of different methods

By analyzing the experimental curve in Fig. 1, it can be seen that there are some differences in the accuracy of evaluating the effectiveness of enterprise economic management mode by using the methods of this paper, literature [5] and literature [6]. Among them, the accuracy of using this method to evaluate the sample data is always higher than 90%, and the accuracy of literature [5] method and literature [6] method to evaluate the effectiveness of enterprise economic management model is always lower than this method, which verifies that this method is feasible.

3.4 Result Analysis

Budget management mainly assesses the budget implementation rate. On the one hand, the budget implementation rate reflects the scientificity and operability of budget preparation, and it also reflects the changes in the budget implementation process. The company prepares a budget every year. Because the local permanent population is relatively stable, the income and expenditures of the company have not changed much over the years, and it adopts a relatively safe way of budgeting revenue. In terms of budget expenditures, considering that the company borrowed money for plant construction in 2015, there were bank loans of more than 100 million yuan, and some of the loans were due. Therefore, active expenditure budgeting was adopted when preparing the budget. Financial special appropriation has always been a concern of enterprises. In 2018 and 2019, the government special appropriation was about 7 million yuan, which is a drop in the bucket for enterprises with production income of 438 million. In terms of special financial appropriations, the company generally prepares budgets for special financial appropriations based on the annual budgets of the higher-level departments, so the implementation rate is 100%, but it is necessary to pay attention to the practical problem of the government's special appropriations being too low.

The balance rate of business revenue and expenditure was −1.63% in 2019 and −1.57% in 2018, indicating that when the sales revenue of the enterprise in 2019 increased compared with that in 2018, the range of production expenditure was greater than that in 2018. The two-year data are negative, indicating that the enterprise has no balance of business revenue and expenditure in the past two years, and also indicating that the

operation cost of the enterprise is high. This can also be reflected from the data of asset liability ratio. The asset liability ratio was 63% in 2018 and 67% in 2019, which has reached the warning value. We need to pay close attention to the economic operation. Current ratio indicates the ability of current assets to be realized and used to repay current liabilities before the maturity of short-term liabilities. The main factors affecting the current ratio include production cost receivables and product turnover speed. Based on the unit nature of the enterprise, appropriate liabilities can be considered, but it is better to keep the current ratio at about 200%. Although the current ratio of the enterprise has improved in the past two years, it is still low on the whole, even lower than 1, indicating that the enterprise is short of cash flow and has prominent debt problems. The turnover days of accounts receivable in the two years are 54 days and 56 days respectively. There is no significant difference between the data, but it also reflects that the realization time of accounts receivable in 2019 is longer than that in 2018.

In 2019, the company's inventory turnover rate increased significantly compared with 2018, indicating that there have been major improvements in cost management such as consumables and production. Basically "increasing revenue and reducing expenditure". The company introduced a consumable management system, and many production procedures adopted an outsourcing model, which greatly reduced The cost of consumables. The production revenue cost rate in 2019 has increased compared to 2018, indicating that the sales revenue is greater than the production expenditure, and the cost of the company's staff is higher. It is necessary for managers to further strengthen cost management in the production process or develop new high value-added products project.

4   Conclusion

This paper proposes a research on the effectiveness evaluation method of enterprise economic management model based on in-depth data mining. The results of the case analysis show that the evaluation method can find the dilemma faced by the enterprise economic management model. However, there are still many shortcomings in the research of this article. In future research, while enterprises continue to strengthen their own economic management, government departments must carefully calculate production income and operating costs, and the establishment of price charging standards must be scientific and reasonable, and reflect the staff. Labor value; the production settlement must be timely and in full, and the enterprise must not be held responsible for insufficient production expenditure funds; financial compensation must be made for the difference between the income and expenditure of the enterprise, and the funds must be in place in time to prevent the enterprise from breaking the capital chain. At the same time, enterprises must strengthen their own standardized management to ensure strict implementation of price and charging standards, careful calculation and use of special fiscal funds. At the same time, they must conscientiously do a good job in cost control, reduce losses, and eliminate waste, so that the enterprise can embark on sustainable development.

References

1. Liu, Y., Zhang, Y., Yang, L.: Design and application of enterprise environmental performance evaluation index system. Commun. Finance Account. **20**(5), 28–30 (2021)

2. Cai, S., Fang, H.: The establishment of demonstration enterprises evaluation system in service-oriented manufacturing. Modernization Manag. **39**(05), 70–76 (2019)

3. Liu, Z., Zhang, C.: Research on performance evaluation of heavily polluted enterprises from the perspective of sustainability: based on COWA operator and cloud model. Sci. Technol. Manag. Res. **39**(01), 235–241 (2019)

4. Yulan, Z., Siting, J., Shuang, N., et al.: Research on investment efficiency evaluation of listed manufacturing companies in Beijing, Tianjin and Hebei – from the perspective of technological innovation. Friends of Accounting **12**(18), 14–19 (2020)

5. Zhou, Y., Zuo, Y.: A study of the construction of performance evaluation system of enterprise's patent operation. Shandong Soc. Sci. **281**(01), 147–153 (2019)

6. Huijin, C., Hao, D., Youcai, M.: Supply chain risk evaluation and application of export-oriented enterprises based on improved SCOR. Enterp. Econ. **39**(1), 80–89 (2020)

7. Sun, Y., Mou, L.: Performance management of private environmental protection enterprises under the background of ecological civilization. Appl. Anal. Fuzzy Synth. Eval. Method **14**(2), 3–7 (2020)

8. 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

9. Liu, S., Chen, X., Li, Y., Cheng, X.: Micro-distortion detection of lidar scanning signals based on geometric analysis. Symmetry **11**(15), 1471 (2019)

10. Liu, S., Bai, W., Srivastava, G., Machado, J.A.T.: Property of self-similarity between baseband and modulated signals. Mob. Netw. Appl. **25**(4), 1537–1547 (2019). https://doi.org/10.1007/s11036-019-01358-9

Machine Learning

Research on the Whole Process Quality Control Method of Water Conservancy and Hydropower Construction Based on BP Neural Network

Mingdong Yu[1](✉) and Qian He[1,2]

[1] Department of Civil and Hydraulic Engineering Institute, Xichang University,
Xichang 615013, China
`xccyymd@126.com`

[2] Department of Civil EngineeringFaculty of Engineering, Technology and Built Environment,
UCSI University, 56000 Kuala Lumpur, Malaysia

Abstract. The traditional quality control method of the whole construction process is lack of process quality evaluation, which leads to the low overall quality score of the project after the completion of the project construction. A whole process quality control method of water conservancy and hydropower construction based on BP neural network is designed. First, set up water conservancy and hydropower construction site monitoring, and design data acquisition equipment for the whole construction process according to different construction positions; According to the acquisition equipment set above, collect the water conservancy and hydropower construction data at different locations and use them as the control data; On this basis, the framework structure of BP neural network is established, the control network level is determined, the back propagation function is determined, and the collected data of the whole process of water conservancy and hydropower construction are trained to complete the whole process quality control of water conservancy and hydropower construction. The example analysis results show that the whole process quality control of the project using the design method can effectively improve the construction quality.

Keywords: BP neural network · Water conservancy and hydropower projects · Whole process quality control · Data acquisition

1 Introduction

The construction of small and medium-sized water conservancy projects is generally small in scale, simple in structure, few key technical problems, and the construction is not difficult, so the quality problems become particularly prominent. Quality is the lifeline of small and medium-sized water conservancy projects, but often due to the lag in planning and design, the quality of the preliminary planning work is not high; the project volume is small, the unit price is low, and there are many local contradictions; the quality control system is not perfect, the evaluation mechanism is not perfect, and

the management is not in place [1, 2]; The quality awareness is weak. The front-line construction workers generally have low quality and high mobility, which cause the project quality to fail to meet the standards, and some even have serious consequences such as quality defects and quality accidents. These problems The control of construction quality brings great difficulties [3, 4]. With the advent of the Twelfth Five-Year Plan, Daxing water conservancy has become a key task for the country to protect people's livelihood and promote development, especially the rapid development of a large number of small and medium-sized water conservancy projects that serve the "agriculture, rural areas, and farmers" such as small and medium-sized river management, drinking water safety, and water-saving irrigation.. Therefore, strict quality control is both a technical task and a political task. It is necessary to combine quality control and quality evaluation to promote efficient, comprehensive and in-depth implementation of quality control through quality evaluation to ensure that the quality of the project meets the design standards and specifications., To enable small and medium-sized water conservancy projects to give full play to economic and social benefits in accordance with quality and quantity, and to ensure the safety of people's lives.

Therefore, this paper designs the whole process quality control method of water conservancy and hydropower construction based on BP neural network. The main technical route of this method research is as follows:

(1) Set up water conservancy and hydropower construction site monitoring, and design data acquisition equipment for the whole construction process according to different construction positions;
(2) According to the above set acquisition equipment, collect the water conservancy and hydropower construction data at different locations as the control data;
(3) On this basis, the framework structure of BP neural network is established, the control network level is determined, the back propagation function is determined, and the collected data of the whole process of water conservancy and hydropower construction are trained to complete the whole process quality control of water conservancy and hydropower construction.

2   Research on the Whole Process Quality Control Method of Water Conservancy and Hydropower Construction Based on BP Neural Network

2.1   Layout of Water Conservancy and Hydropower Construction Site Monitoring

The construction site based on hydropower project is usually in a remote place, and there is usually no suitable office place around it. Therefore, the office place is often located at a certain distance from the construction site. It is usually troublesome to check the construction situation on the construction site. Setting up surveillance cameras to remotely monitor the construction situation is a good solution. Cameras shall be arranged at the construction site to transmit the working conditions of the construction site in real time in the form of video to the office away from the construction site through wired network.

Cameras shall be set at the construction sites of various concealed works to control and manage the construction of concealed works Video data shall be stored according to the needs of the project for post observation. In the field monitoring of this paper, wireless network transmission technology is generally used for monitoring, and wireless video monitoring system is used to monitor the whole construction site and key positions in the project construction, so as to comprehensively and timely understand the construction dynamics, respond quickly to unexpected situations, eliminate potential safety and quality hazards in time, and provide analysis basis for existing problems, So as to further improve the safety management and quality management of engineering construction, and greatly improve the modern, scientific and standardized level of engineering construction site management and construction technology. The video monitoring system consists of three parts:

1) remote part (monitoring point): mainly composed of camera and video server. The camera collects video information and transmits it to the video server, which converts, encodes and compresses the video signal [5, 6].
2) Data transmission part: responsible for transmitting the remote video signal to the monitoring center, and transmitting the control commands of the monitoring center to the remote. There are two transmission methods: wireless and wired. Wireless transmission includes analog microwave and digital wireless networks, and wired transmission includes cables or optical cables.
3) Monitoring center: All remote monitoring signals are gathered here for information browsing, storage, conversion and interactive control operations. It is mainly composed of video server, computer, monitoring terminal and other equipment [7, 8]. Due to the temporary nature of project construction, the deployment of wired or optical cables in the data transmission link part is costly and has a long period. However, wireless transmission is in line with the needs of project construction due to its flexible system construction and short construction period.

In the process of the deployment of no online webcams, in addition to the deployment of wireless webcams on the construction site of the dam, for hydropower plants, spillway tunnels and other projects, the cameras are added or moved at any time according to the actual project.

After the installation is completed, in the surveillance video storage and subsequent use, in addition to real-time monitoring video screens of the construction site, it also provides the current monthly construction data video. The camera data that needs to be retained such as key processes and concealed projects are stored in the monitoring center On the server of, for those who need to inquire about the recording.

2.2   Data Acquisition Equipment in the Whole Process of Design and Construction

In the process of quality control, data collection in the construction process is also an important link. For the signal acquisition in the construction process, the secondary voltage and current signals in the experimental process of the acquisition point are generally used. The method adopted in this paper is to obtain the voltage and current signals at the secondary side of the acquisition point at the same time, and then calculate

the dynamic resistance of the acquisition point according to Ohm's law. The advantage of this method is that the measurement is direct and high measurement accuracy can be obtained. However, due to the large secondary side current, a large range current sensor needs to be used. This chapter will introduce the coil sensor and other data acquisition equipment used for high current sensing in detail.

Since there are many factors that affect the quality of water conservancy and hydropower construction, it is necessary to carry out various combinations of these factors for testing to fully reflect the quality changes in various situations. In order to combine various influencing factors and reduce experimental workload, the designed data acquisition equipment is shown in the figure below:
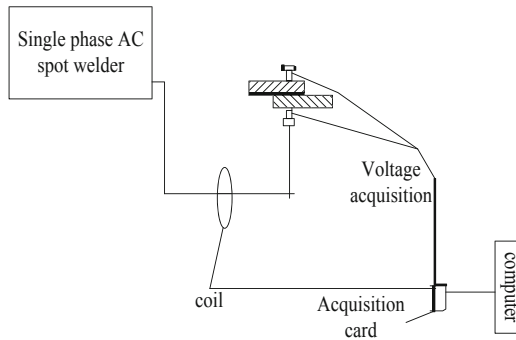

Fig. 1. Schematic diagram of the composition of data acquisition equipment

The main components of data acquisition are shown in Fig. 1. The system consists of a single-phase AC spot welder, secondary side current air-core coil current sensor, inter-electrode voltage sampling wire, A/D signal acquisition card, PC, and sampling software.

The air-core coil current sensor can output the differential waveform of the secondary side current and convert it into a voltage waveform that meets the amplitude requirements of the A/D signal acquisition card. This signal is sent to the A/D signal acquisition card together with the voltage signal between the electrodes. The A/D signal acquisition card in the experiment can perform 16-channel 12-bit analog-to-digital conversion, which can meet the accuracy and rate requirements of the measurement.

The traditional transformer with iron core has a very narrow frequency band, and the secondary signal waveform is seriously distorted when the magnetic saturation is used as relay protection, the reaction speed lags behind and is easy to cause relay misoperation. Air core coil current sensor can overcome a series of problems brought by traditional electromagnetic transformer. Its characteristics are simple structure, no direct circuit connection with the measured current, no iron core, no hysteresis effect and no magnetic saturation. The coil is mainly composed of annular hollow coil. As shown in Fig. 2.
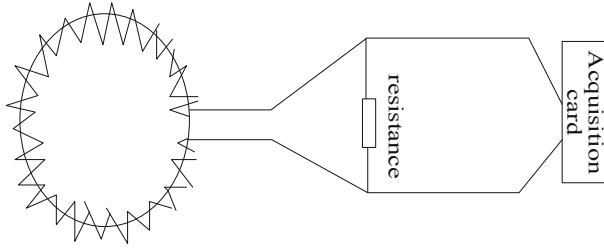
Fig. 2. Schematic diagram of coil structure

Connect a suitable resistor in parallel with the output terminal of the sensor coil to obtain a suitable output voltage amplitude. In order to obtain the original waveform of the welding current, it is necessary to integrate the waveform on the sensing coil and calibrate the integrated amplitude. The parameters of the sensor coil used in this article are as follows: the number of turns is 1200, the wire diameter is 0.4 mm, the size of the flexible skeleton is 3×25 mm, and the output resistance is 100 Ω/0.25 W. This test needs to collect two analog signals: the differential waveform of the secondary current and the voltage between the electrodes. The differential waveform of the spot welding current is an alternating signal with a maximum amplitude of about 4 V. The voltage between the electrodes is also an alternating signal, the maximum amplitude is about 3 V. Therefore, select the 0 channel and 1 channel of the A/D signal sampling card to perform spot welding current differential waveform and inter-electrode voltage sampling respectively, and the range is set to $\pm$ 5 V. Start AD1674 for conversion. When A/D current channel conversion is completed, it can be shorted to PC bus interrupt request signal RQ5 through jumper SX1 1–2. A/D occupies 4 consecutive port addresses: 310 H – 313 H. The A/D input range is selected by the 4-digit DIP switch SW2. The A/D input signal is accessed by the XS1 25D socket. Pins are No. 1 and No. 14, which are voltage and current signals respectively, and the input voltage range is $\pm$ 5 V. This completes the design of the data acquisition equipment.

2.3   Design a Quality Evaluation Model Based on BP Neural Network

BP neural network is also called multilayer feed forward neural network. The network has input layer, output layer and one or more hidden layers. The network is composed of several layers of neurons. It is a multilayer perceptron structure.

Figure 3 shows a very typical BP neural network model with input, output and hidden layers. The figure shows the topology of a BP network with only one hidden layer, and there is no feedback connection between neurons.. The non-linear function is the activation function generally used in the hidden layer, and the S-type function is the most commonly used. The output layer of the network can use linear functions as activation functions and non-linear functions as activation functions [9, 10], which depends on the mapping relationship between the input and output of the specific situation.

In the learning process of the BP neural network, the working signal is propagated forward, and the error signal is propagated back, so it is repeated to train the neural network.
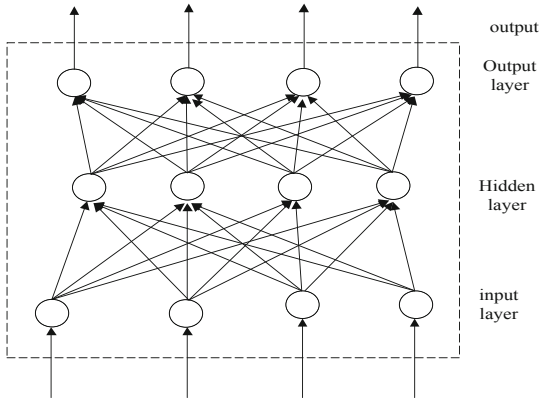
Fig. 3. Frame structure of BP neural network

Table 1. Meaning of letters

Letter	Meaning	Letter	Meaning
P	Network input	s_2	Number of neurons in output layer
r	Number of neurons in input layer	f_2	Output layer activation function
s_1	Number of hidden layer neurons	S	Network output
f_1	Hidden layer activation function	T	Target vector

Table 1 shows the letters and their meanings, which will be used in the following description.

The sequential propagation of working signal along input layer, hidden layer and output layer is called forward propagation. In the process of signal transmission, the weight of neural network is fixed, and the state of neurons in any layer only determines the state of neurons in the next layer. When the output layer does not get the expected output, this process will turn into the back propagation process of error signal.

The output of the i neuron in the hidden layer is:

$$\begin{cases} a_{1i} = f_1\left(\sum_{j=1}^{T} w_{1ij}a_{1i} + b_{2k}\right) \\ i = 1, 2, \cdots, s_1 \end{cases} \tag{1}$$

The output of the k neuron in the output layer is:

$$\begin{cases} a_{2k} = f_2\left(\sum_{j=1}^{s_1} w_{1kj}a_{1i} + b_{2k}\right) \\ k = 1, 2, \cdots, s_2 \end{cases} \tag{2}$$

The error of the function is defined as:

$$E(W, B) + \frac{1}{2}\sum_{K=1}^{S_2} (t_k - a_{2k})^2 \tag{3}$$

Back propagation of error signal: the error signal is the difference between the actual output and the expected output of the neural network. The back propagation of error means that the signal is transmitted from the output layer to the input layer through the hidden layer. In the back propagation process, the weight of the network is adjusted by error feedback.

The weight of the output layer changes from the i input to the k output as follows:

$$\begin{cases} \Delta w_{2ki} = -\eta \frac{\partial E}{\partial a_{2k}} \frac{\partial a_{2k}}{\partial a_{2ki}} \\ \qquad = \eta(t_k - a_{2k})f_2'a_{1i} \\ \qquad = \eta\delta_{ki}a_{1i} \\ \delta_{ki} \quad = (t_k - a_{2k})f_2' \\ \qquad = e_k f_2' e_k f_2' \\ e_k \quad = t_k - a_{2k} \end{cases} \tag{4}$$

The same can be obtained:

$$\begin{cases} \Delta b_{2k} = -\eta \frac{\partial E}{\partial b_{2ki}} \\ \qquad = -\eta \frac{\partial E}{\partial a_{2k}} \frac{\partial a_{2k}}{\partial a_{2ki}} \\ \qquad = \eta(t_k - a_{2k})f_2' \\ \qquad = \eta\delta_{ki} \end{cases} \tag{5}$$

The weight of the hidden layer changes. From the j input to the i output weight, the amount of change is:

$$\begin{cases} \Delta w_{1ki} = -\eta \frac{\partial E}{\partial w_{1ij}} \\ \qquad = -\eta \frac{\partial E}{\partial a_{2k}} \frac{\partial a_{2k}}{\partial a_{1i}} \frac{\partial a_{1i}}{\partial w_{1ij}} \\ \qquad = \eta \sum_{k=1}^{S_2}(t_k - a_{2k})f_2'w_{ki}f_1'P_j \\ \qquad = \eta\delta_{ij}P_j \\ \delta_{ij} \quad = ef_1', e_i \sum_{k=1}^{S_2}\delta_{ij}w_{2ki} \\ \Delta b_{1i} \quad = \eta\delta_{ij} \end{cases} \tag{6}$$

When dealing with problems such as non-linear classification, it is achieved by adjusting the scale of the BP neural network (the number of input nodes, the number of output nodes, the number of hidden layers and the number of hidden layer nodes) and the connection weights in the network. The BP neural network can Approximate any nonlinear function with arbitrary precision.

The determination of the number of layers of the network, the number of neurons in the hidden layer, the selection of the transfer function, the training method and the selection of parameters are the main contents of the network structure design. The reliability of the evaluation results is directly affected by the designed network performance.

(1) Determine the number of layers of the network

A neural network must have at least one input layer and one output layer, which is the premise. The selection and design of a reasonable hidden layer will directly determine

the performance of the neural network. The complexity of the problem to be solved determines the number of hidden layers. Research shows that if you want to strengthen the solution to more complex problems, you can achieve it by increasing the number of hidden layers.

(2) Number of neurons in the hidden layer

In BP neural network, the reasonable setting of hidden layer neurons is very important, which directly determines the realization of the function of the network model. The number of neurons should be appropriate, not too much or too little. Too much will make the learning time of the network longer. If it is too short, it is difficult to deal with complex problems. Repeated comparison is the main method to determine the number of hidden layers, because there is no effective and scientific method to determine the number of neurons.

(3) Training method and parameter selection

For dealing with different problems, BP neural network has many training methods, as well as how to select the training function and its parameters.

First, we must establish a sound and reasonable high-rise building construction site safety evaluation index system and neural network evaluation model. Then carry out the application of safety evaluation and ensure the accuracy of the mathematical model of safety evaluation according to the actual situation of the high-rise building construction site. The steps for the safety evaluation of high-rise building construction site based on neural network are as follows:

1) Determine the number of hidden layers, the number of nodes in the input layer, output layer and hidden layer;
2) Establish a safety evaluation index system;
3) Collect appropriate learning samples and determine a reasonable neural network training method;
4) Choose a suitable transfer function;
5) Training the BP neural network model as a knowledge base for enterprise safety evaluation;
6) Use the trained BP neural network to obtain the evaluation result by using the learned knowledge through the MATLAB simulation technology;
7) The evaluation results obtained can be used as training samples of BP neural network, and can enrich and strengthen the performance of enterprise safety evaluation knowledge base.

3 Case Analysis

In the experiment, firstly, the basic situation of the experimental project is analyzed, the experimental scheme is set according to the situation, the water conservancy project construction data acquisition system is designed in detail, and the quality of water conservancy project facilities is detected according to different modules set in the system.

Finally, the experimental results are analyzed to reflect the effectiveness of the proposed method through the analysis of experimental data.

3.1 Project Overview

The experiment selects a large-scale water conservancy and hydropower project in the middle reaches of the river where a hydropower station is located, which mainly focuses on power generation and has comprehensive utilization benefits such as flood control and sediment blocking. The dam controlled drainage area of the hydropower station is 68512 km^2, accounting for 88.5% of the whole drainage area. The power station has 6 installed units, single unit capacity of 550 MW (maximum capacity of 600 MW), total installed capacity of 3300 mw, guaranteed output of 926 mw, and multi-year average annual power generation of 14.58 billion kw. H.

The hydropower station selected in the experiment is composed of river blocking dam, water diversion and power generation structure, flood discharge tunnel, emptying tunnel, Niri river water diversion project and so on. The river retaining dam is composed of gravel soil straight core rockfill dam and 3 spillways with a width of 12 m; The crest elevation of gravel soil straight core rockfill dam is 856 m. The dam crest is 14 m wide, the dam axis is 573.5 m long, and the maximum dam height is 186 m. The water diversion and power generation structure is composed of bank tower water intake, 6 pressurized water diversion tunnels, underground powerhouse, underground main transformer chamber, underground Tailrace Gate Chamber and 2 pressureless tunnels. Niri river diversion project consists of head hydroproject and diversion tunnel.

The normal storage level of the power station reservoir is 850.00 m, the operating limit water level during the flood season is 841.00 m, the dead water level is 790.00 m, the drawdown depth is 60 m, and the total storage capacity is 5.390 billion m^3, of which the flood regulation storage capacity is 1.056 billion m^3 and the regulation storage capacity is 3.882 billion m^3, which is an incomplete year. Regulate the reservoir. The maximum Kui water height in front of the dam is 173 m, and the main stream backwater reaches the Shimian County Town Crossing River Bridge. The backwater length is 72 km and the reservoir area is 84.14 km^2. The river course of the reservoir is 72 km, involving 3 counties including Hanyuan, Ganluo and Shimian.

3.2 Experimental Scheme

The experimental selection of hydropower stations also used remote video monitoring, electronic document management, and the use of three-dimensional models to express the appearance of the dam body for management. However, the various subsystems are not well integrated, and an integration needs to be established. This chapter mainly takes the GPS real-time monitoring system established in this mode as an example to build a unified platform for similar large-scale projects in the future.

In the Pubugou Dam construction process, if conventional inspection methods that rely on manual on-site control of rolling parameters (such as rolling speed and number of passes) and manual digging test pit sampling are still used to control construction quality, it is different from large-scale mechanized construction. It is also difficult to meet the corresponding construction quality requirements. Therefore, it is necessary to study

a dam filling and rolling construction quality monitoring system with the characteristics of real-time, continuity, automation, and high precision, to adapt to the experimental selection of the river where the hydropower station is located. The experiment selects the construction quality of the main dam of the hydropower station. Management requirements. Accumulate experience for the construction of gravel soil dams similar to core walls in the future.

3.3 Experimental Data Acquisition

The experiment developed by the research institute selects the river where the hydropower station is located. The experiment selects the GPS real-time monitoring system for the construction quality of the dam of the hydropower station. With the GPS system, wireless network communication system and computer hardware as the core, it has developed a filling and grinding system for the dam. The software platform for real-time, high-precision, continuous, and automatic monitoring of the construction quality of the compaction project, realizes the highly automatic operation and operation monitoring of the rolling compaction construction, and can grasp the location and operation execution status of the rolling construction vehicle in real time, The system can record the trajectory, speed and direction of each vehicle, and calculate the thickness and number of passes of the paving layer, providing a basis for accurate construction and inspection judgments.

The GPS real-time monitoring system for dam filling construction quality adopts C/S (client/server) architecture to realize mutual communication and service between mobile remote end and monitoring center. In addition to the traditional keyboard and mouse operation, the mobile terminal is equipped with touch screen operation to realize man-machine dialogue. The communication adopts GPRS/CDMA wireless network to realize network data transmission through public Internet network. The acquisition and setting of GPS data are realized through RS232 serial port. The server side of the system uses broadband static IP address to access the Internet. The system operation principle is shown in Fig. 4 below:
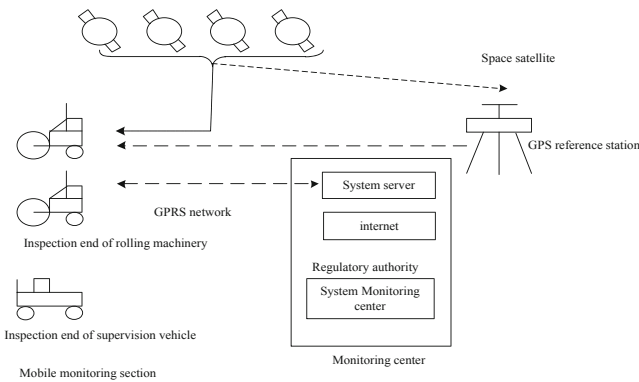

Fig. 4. GPS real-time monitoring system for dam construction and rolling

The GPS real-time monitoring system for dam filling construction quality is mainly composed of three parts (as shown in Fig. 1):

Reference station part, mobile monitoring part and monitoring center. The mobile monitoring part is composed of various mobile monitoring terminals, including all rolling machinery and engineering supervision vehicles. It navigates the operators, requires the operators to operate according to the specifications, reports to the operators, provides parameters and information to the server, and receives server instructions. The benchmark station is mainly used to set the engineering measurement parameters and provide differential information. Considering the long-term site stability, simplicity of setting and diversity of environmental conditions, a separate part is established. The system server provides data channel and data storage, and provides services to supervisors and management departments.

The GPS mobile monitoring terminal installed on the quality supervision vehicle is mainly composed of GPS and industrial computers. According to the needs of project monitoring and at the same time, in order to further improve the accuracy of GPS observation, the system is specially equipped with quality supervision work vehicles for construction supervision units. The main task is to carry out GPS mobile observation. The main observation item is to measure the thickness of the paving soil after the soil is paved, and to provide data for calculating the compaction rate; save the data in the industrial computer in the car and save it in real time The observation results are sent back to the monitoring center, and the monitoring center receives the positioning data of the GPS mobile monitoring terminal sent by the mobile monitoring terminal (including the data of controlling the number of rolling passes of the rolling machine and the thickness of the rockfill layer before rolling, and the data of the rolling range, etc.) At the same time, it can request the monitoring center to receive the results of its further analysis, and guide the on-site construction from the perspective of quality management.

The GPS mobile monitoring terminal installed on the rolling machinery is similar to the GPS mobile monitoring point installed on the quality supervision vehicle. The hardware configuration is also composed of GPS and industrial computers. The observation items are mainly the rolling passes and driving of the rolling machinery. The speed, driving direction and the final rolling thickness of the paving soil are stored in the industrial computer in the car and displayed on the touch screen, and the observation results are sent back to the monitoring center in real time. The computer's touch screen also reflects the rolling situation in real time, and guides the construction work of the rolling truck operator.

3.4   Results and Analysis

Through the investigation and investigation of the bidding documents and the early stage of the construction site of the project, based on the communication with the project management personnel, this paper comprehensively analyzes the preparation control measures before the construction of the project, classifies and arranges the collected data and documents, and lists the specific indicators of the construction quality prediction of the project, The research results of traditional water conservancy and hydropower construction whole process quality control method and water conservancy and hydropower

construction whole process quality control method based on BP neural network are shown in Table 2:

Table 2. Data of construction quality prediction indicators for high-rise buildings

Influence factor		①	②
People	Technical proficiency of construction personnel U11	0.12	0.10
	Comprehensive quality of management personnel U12	0.41	0.50
	Training proportion of construction personnel U13	0.73	0.76
Material Science	Concrete U21	0.34	0.50
	A steel bar U22	0.35	0.50
	Template U23	0.08	0.10
Mechanics	Working condition inspection of mechanical equipment U31	0.65	0.63
	Repair and maintenance of machinery U32	0.52	0.37
	Allocation of construction machinery and equipment U33	0.09	0.10
Method	Construction technical scheme U41	0.87	0.90
	Construction organization design U42	0.41	0.37
	New technologies and methods U43	0.76	0.90
Environment	Construction environment U51	0.69	0.67
	Construction quality management environment U52	0.57	0.63

In Table 2, ① represents the traditional method, ② represents the quality control method designed in this paper based on the BO neural network, and the data obtained from the quality prediction in the BP neural network model. Enter the evaluation results obtained in the above table into the software. After the traditional method is calculated, the score obtained under a certain weight is 70 points, that is, the construction quality of the proposed building predicted according to the prior control measures is qualified. If the quality target of this project is not reached, it indicates that there are problems with the preventive measures for these factors that affect the construction quality in advance, so strict improvement is needed.

Using the BP network model to change the data of each factor, the evaluation score obtained is 83.7 points. Under the evaluation of this score, the predicted value can reach the quality target of the project. When changes are made to the factors of concrete, steel bars, and on-site construction environment, the predicted value can meet our quality requirements, indicating that there are problems with the control measures in these aspects during the pre-construction preparation stage. Therefore, the quality control method of the whole process of water conservancy and hydropower construction based on BP neural network is better than the traditional method of quality control of the whole process of water conservancy and hydropower construction.

The source of concrete and reinforcement in the project is purchased by the construction party. When selecting the source of material supplier, it is easy to buy some

materials with poor quality because the information of material market and manufacturer is not comprehensive enough. Therefore, it is suggested to use the bid inviter to purchase important materials, The quality requirements of engineering materials can be met through the review of manufacturers and the assessment of field investigation and research. In addition, the construction environment control measures have not been done well. We should pay attention to the improvement of the plane and working environment conditions of the construction site, plan the roads on the site, stack materials reasonably, and have good flood control and drainage capacity, construction lighting, safety protection facilities, etc., so as to ensure the construction quality of the project. This result can be used as an important reference for the construction unit to control the construction quality of the project in advance.

4   Conclusion

Combined with the specific examples of water conservancy and hydropower construction projects, starting from the factors affecting the construction quality, this paper analyzes the specific measures of construction quality prior control, uses the BP neural network model to predict the water conservancy and hydropower construction quality level, and proves the objectivity and superiority of BP neural network in water conservancy and hydropower construction quality prior control.

References

1. Yang, N., Dai, Z.: Estimation and simulation of indoor convective heat transfer in green ecological high-rise buildings. Comput. Simul. **38**(3), 442–446 (2021)
2. Wang, Y., Zhang, X., Lu, L., et al.: Estimation of crop coefficient and evapotranspiration of summer maize by path analysis combined with BP neural network. Trans. Chin. Soc. Agric. Eng. **36**(07), 109–116 (2020)
3. Liu, Z., Zhang, J., Deng, F.: Monitoring and identification of state of opening or closing isolation switch based on BP neural network. Power Syst. Prot. Control **48**(05), 134–140 (2020)
4. Wang, L., Li, X., Cao, B., et al.: Prediction of leakage current on insulator surface of transmission line based on BP neural network. High Voltage Apparatus **56**(02), 69–76 (2020)
5. Zhang, C., He, C., Liu, X., et al.: Magnetic barkhausen noise technology for surface hardness evaluation in steel shaft based on BP neural network. J. Exp. Mech. **35**(01), 1–8 (2020)
6. Cui, S., Li, S., Wang, X.: Joint de-noising method of seismic data via BP neural network and SVD algorithm. J. Electr. Meas. Instrum. **34**(02), 12–19 (2020)
7. Li, G., Liu, Z., Jin, G., et al.: Ultra short-term power load forecasting based on randomly distributive embedded framework and BP neural network. Power System Technology **44**(02), 437–445 (2020)
8. 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
9. Huang, S., Chen, G., Wu, C., et al.: Construction of evaluation index model of scientific research project database based on improved AHP-BP neural network. Inf. Sci. **38**(01), 140–146 (2020)
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

Intelligent Recommendation Method of Sports Tourism Route Based on Cyclic Neural Network

Xiangyu Xu[✉] and Zhiqiang Wang

Sports Center, Xian Eurasia University, Xian 710000, China
`xuxiangyu1456@163.com`

Abstract. Due to the low matching degree between scenic spot characteristics and tourists' interests, the accuracy of route recommendation is low. Therefore, an intelligent recommendation method of sports tourism route based on cyclic neural network is designed. On the premise of determining the recommendation target of sports tourism route, the characteristics of sports tourism attractions and routes and tourists' interest are extracted. After clustering, the recommendation list is collaborative filtered from the perspective of tourists. Finally, the circular neural network is used to optimize the recommended route. The test results show that the MAE of the design method opinion results is basically within 0.1, which has high accuracy.

Keywords: Recurrent neural network · Sports tourism route · Intelligent recommendation · Clustering processing · Collaborative filtering

1 Introduction

The World Tourism Organization and the United Nations Statistical Commission have made a scientific explanation of "tourism": people leave the environment of daily life for leisure, business or other purposes, go to certain places and stay there, but do not continue activities for more than one year. Most people think that travel is just for entertainment, vacation, and relaxation of their lives. In fact, it has other purposes, including: leisure, visiting relatives and friends, business, professional visits, health care, religion/pilgrimage, etc. [1–3]. Tourism has become an indispensable part of people's lives. It brings the baptism of both body and mind to busy people. At the same time, it allows us to feel the beauty of nature, feel the customs of exotic folk customs, and feel the joy of traveling together with relatives and friends. In recent years, my country's travel industry has developed rapidly [4, 5]. According to the "Communiqué on the Statistical Survey of National Travel Agencies in the Fourth Quarter of 2013" published by the National Tourism Administration on February 21, 2014, my country's tourism industry revenue reached 223.3 billion during the November long holiday last year, a year-on-year comparison of 2012. The Mid-Autumn Festival and National Day holidays increased by 6.1%. In the fourth quarter of 2013, national travel agencies and domestic tourism organizations had 35,737,500 person-times, 114,798,100 person-days, received 40,040,400

person-times, and 92,446,900 person-days. Numerous data show that the tourism industry has become one of the industries with the strongest development momentum and the largest scale. On the contrary, the problems that people encounter during the tourism process are also endless. The most important ones are the planning of tourist routes and the navigation of tourist attractions. And the real-time conditions in the scenic area, etc. [6, 7]. In the past, the solution to these problems was to use paper maps and scenic spots announcements, etc., which had strong limitations and could not meet the needs of tourists. With the development of science and technology, electronic technology positioning and navigation has brought fresh blood to the tourism industry, and they influence each other and develop in coordination. The navigation equipment we use now generally has a GPS antenna placed inside it. In addition, there are 24 global satellites in the sky above our earth. Generally, at least 3 of them can be received at any time. The GPS antenna receives the data information transmitted by the satellites and combines the stored electronic map to determine the position coordinates for positioning., And then display a route for tourists' reference. The positioning accuracy deviation of the positioning equipment on the market today does not exceed 3 to 5 m. There are many types of positioning systems. Among them, the most commonly used one is similar to car GPS, and its terminals can be smart phones, Pads, and so on.

In the traditional tourism industry, the tour guide tools used by tourists are basically paper maps, which can be used to tour the scenic spot by manually drawing the route from the starting point to the destination. The main disadvantages of manually drawn tourism route map are: first, it has unreasonable use and low efficiency in terms of time and money. Because tourists must first determine a path from the starting point to the target point, and then piece together the maps containing these scenic spots into a set of paper maps covering scenic spots for navigation and play; Second, it can not give tourists real-time scenic spot change information, which will bring a lot of inconvenience to tourists' playing process.

With the continuous development of computer technology, electronic navigation technology has become more and more mature. More and more tourists are used to querying their travel routes through positioning and navigation in the process of playing. However, most mobile terminal navigation software on the market only provides a route based on a single tourist preference, for example, a shortest time route, a shortest distance route and so on. These methods often can not provide tourists with a route that is really suitable for their current actual situation, and usually do not provide alternative routes [8, 9]. In order to solve these problems, a circuitous route system is proposed, which provides another route when tourists deviate from the predetermined route. However, this type of system requires the tourist guidance system to recalculate and generate routes, rather than automatically generate alternative routes. Another disadvantage of the currently available navigation system is that when tourists choose a route to play, the situation in the scenic spot may change. For example, the performance of a scenic spot is about to begin. It is wise to change the route at this time.

Based on this, this paper designs an intelligent recommendation method of sports tourism route based on cyclic neural network. Based on the route of sports scenic spots and the preference characteristics of tourists, the optimization is realized, the most qualified tourism route is obtained, and it is used as the recommended route. Finally, the

effectiveness of the design method is verified by experimental tests. Through this study, in order to provide a valuable reference for the development of tourism industry.

2   Recommended Goals for Sports Tourism Routes

The recommendation of tourist route information is highly subjective and plays a decisive role in the overall evaluation. The recommended information cannot be absolute. It can only provide tourists with a tourist route that is more satisfactory than the current situation and other helpful tourist routes. Information [10]. Therefore, most of the recommended content of this article is subjective. Let us study the subjective and objective evaluation criteria used in similar topics below:

(1)   Subjective evaluation criteria

   The subjective evaluation standard is that tourists compare the given route information and then give feedback. Obtain the value of the information from the feedback information, and then make corresponding changes, forming an iterative and continuous improvement of the recommended information. The value of information is a subjective evaluation standard, and the positive feedback given by most tourists is the basis.

(2)   Objective evaluation criteria

   The objective evaluation criterion is to study the best path method, take the factors in the current scenic spot into consideration, and combine the path generated by the best path method. This path is the best path. After that, we reasonably compare our recommendation information with the path information, and finally get an evaluation. However, the difficulty of this standard is how to study the best path method. There are different opinions on this method, and there is no standard answer. Therefore, objective evaluation is difficult to achieve in similar subjects.

3   Feature Extraction

3.1   Extraction of Sports Tourist Attractions and Route Features

This paper does not consider the self characteristics of tourists, but focuses on the domain characteristics of scenic spots and routes in the field of sports tourism, and then extracts the feature points of tourists' interest from the domain characteristics. Therefore, before the extraction of tourists' interest features, we must extract the interest feature points of scenic spots and routes in the field of sports tourism. The characteristics of sports tourist attractions are extracted from the existing data of sports tourist attractions. When analyzing the content of the crawled original web page, keyword matching and information extraction are carried out according to the attribute description of sports tourist attractions on the web page. Finally, it is concluded that the interest characteristics of sports tourist attractions are divided into provinces, districts and counties, categories, stars Development time and ticket price. According to the classification of China's sports tourism

resources, sports tourism attractions are divided into eight categories: geographical landscape, water scenery, biological landscape, ruins, astronomical and climatic landscape, buildings and facilities, sports tourism commodities and cultural activities.

Based on this, for the feature extraction of sports tourism routes, this paper selects sports tourism route data as the basis, combined with the principles of sports tourism route design: sports tourism purpose, physical conditions of the sports tourism subject, tourists' economic conditions and travel time, tourists In terms of hobbies, special sports tourism subjects, etc., with the introduction of sports tourism classification as a reference, the characteristics of sports tourism routes are summarized into six categories: travel days, reference costs, route types, departure cities, and destination cities.

The interest characteristics of sports tourism routes are mainly summarized for tourists, because for the characteristics of sports tourism routes, this article only cares about the correlation between sports tourism routes and tourist interest characteristics. The types of routes are divided into: self-view.

Light sports tourism, humanities sports tourism, leisure and vacation sports tourism, shopping sports tourism, and experience sports tourism.

3.2 Extraction of Tourist Interest Features

In this paper, the understanding of tourists' interest characteristics has two aspects: on the one hand, based on the integration of sports tourist attractions and route characteristics, tourists extract the types of interest points of recommended content; on the other hand, it analyzes the historical records of tourists who have selected sports tourist routes as an input factor of recommendation methods. On the other hand, it is a static and attribute standard, The latter aspect is dynamic and attribute content. According to the previous analysis of the attribute characteristics of sports tourist attractions and sports tourism routes, in the field of sports tourism, the interest characteristics of tourists include the above two characteristics. Because the sports tourism industry is originally oriented to tourists, the feature extraction of the first two is based on tourists' hobbies.

Therefore, this paper concludes that the common interest characteristic attributes of tourists and scenic spots are: Province, district and county, category, star, opening time and ticket price. Thus, the measurement of tourist attraction correlation can be quantified as the feature vector of tourist attraction common interest: (province, district and county, category, star level, opening time and ticket price). Similarly, the common interest feature vector of tourist sports tourism route: (travel days, reference cost, route category, departure time, departure city, destination city).

Based on this, the feature vector is quantized. The feature vector quantization types designed in this paper are divided into three categories: 0–1 relationship, single threshold range relationship, and double threshold range relationship. The specific quantification method is as follows:

(1) 0–1 relationship

Some characteristic relationships between tourists and sports tourism routes are 0–1 relationships, such as provinces and departure cities. The interest characteristics of tourists are either completely consistent with the characteristic attributes of the

sports tourism route, and the corresponding characteristic relationship value in the tourist-tourism route characteristic vector is recorded as 1; or completely different, the corresponding characteristic relationship value is recorded as 0 at this time.

(2) Single threshold range relationship

There is a single threshold range relationship between tourists and some characteristic relationships of sports tourism routes, such as the number of days spent playing, star ratings, and so on. Suppose the tourist wants to play m day, then the tourist's play days feature value is m, and the sport tourism route's play days is n, then the sport tourism route's play days value n. At this time, the relationship between the number of play days in the feature vector of the tourist and the sports tourism route The value λ is written as:

$$\lambda = \frac{\min(m, n)}{\max(m, n)} \tag{1}$$

(3) The relationship between the dual threshold range

There is a dual-threshold range relationship between tourists and some characteristic relationships of sports tourism routes, such as ticket prices and opening hours. Assuming that the fare range that the tourist wants to pay is $m_i - m_a$, and the fare of the sports tourism route is n, then the fare relationship value δ in the feature vector of the tourist and the sports tourism route is recorded as:

$$\delta = \begin{cases} 0, n > m_a \| n < m_i \\ \frac{|avg(m_i, m_a) - n|}{avg(m_i, m_a)}, m_i < n < m_a \end{cases} \tag{2}$$

In this way, visitors' interest characteristics are obtained. The specific tourist interest feature extraction process is shown in Fig. 1.
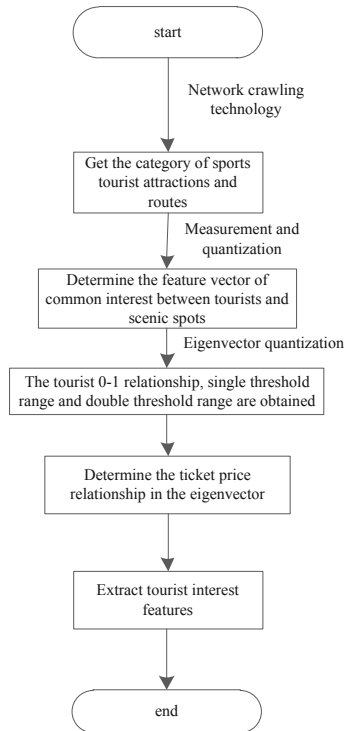
Fig. 1. Tourist interest feature extraction process

According to the analysis of Fig. 1, this paper obtains the categories of sports tourist attractions and routes through network crawling technology. Through measurement and quantification, according to the common interest feature vector of tourists and scenic spots, the interest feature types of tourists are divided into three categories through feature vector quantization, 0–1 relationship, single threshold range and double threshold range, and the relationship between double threshold range determines the ticket price relationship value in the feature vector of sports tourist routes, Finally, tourists' interest feature extraction is realized.

4 Route Recommendation Method Design Based on Recurrent Neural Network

4.1 Sports Tourist Attractions and Routes and Feature Clustering of Tourists

Based on the features extracted from sports tourist attractions, routes and tourists' interests, the above section clusters the features of tourist attractions, routes and tourists respectively. The clustered sports tourist attractions, routes and tourists are quickly indexed to speed up the query based on the similar set of interest features, so as to improve the recommendation efficiency.

The above extracted features cluster sports tourist attractions, sports tourism routes and tourists respectively, which greatly improves the data reading speed. The feature clustering of sports tourist attractions is divided into two parts. One part is to cluster the existing scenic spot database, and the other part is to analyze the features of new sports tourist attractions and add them to the corresponding feature cluster set. Here, a scenic spot may have multiple features, so a scenic spot may be clustered into multiple feature sets. The similarity of features is defined as:

$$p = \sum_i D_i Y_i \tag{3}$$

Among them, i respectively represent the existence strength of different features, D_i represents the fixed factor of the feature, and Y_i represents the non-fixed factor of the feature. This method is used to analyze the extracted data results.

And through formula (3), the clustering criterion between the features of scenic spots is judged, which can be expressed as:

$$C = \frac{p}{E} \tag{4}$$

where, E represents the sum of the intensity of the factors that occur in the scenic spot. When the similarity of features determined by fixed and non fixed features is greater, the degree of clustering is higher.

The clustering of sports tourism routes and tourist information is similar to that of scenic spots.

4.2 Build a Recommendation List

After clustering the features of scenic spots, routes and tourists, the corresponding feature clustering table is established in the database to reduce the time occupied by the recommendation method in data search and improve the recommendation efficiency. It can be seen from the discussion in the previous section that in real life, tourists think more about what is the most suitable next scenic spot at the current location, and don't care too much about others. Although the neural network method can also be used to solve the problem of route selection, the neural network method can not be directly used in the problem solving of this paper, because our scenic spot scoring is a dynamic process. Every time a tourist arrives at a scenic spot, he will re score the other scenic spots based on the scenic spot, so as to make a choice. In addition, the most important thing is that the direct use of greedy method can only get one path, which can not meet the requirements of multi-objective, that is, it can not provide tourists with multiple paths for tourists to choose. However, in our problem, if the order of sports tourist attractions is different, the scores of sports tourism paths are also different. At the same time, we hope to provide tourists with multiple alternative paths based on the use of cyclic neural network. Therefore, we propose a cyclic neural network to solve the problem of generating multiple paths. The method is described as follows:

Each time a tourist arrives at a scenic spot, based on their location, they select the next top N scenic spot with the highest score to form N paths. When $N > K$, by calculating

the TPS of each path, keep the first K TPS values larger Paths, after cutting out $N - K$ paths, and following this, keep the first K paths with the largest TPS each time. The final method obtains K paths, and then through the diversity evaluation of the K paths, the best path is obtained. Youlu recommended to tourists. The specific execution process is as follows:

   Input: the current latitude and longitude position of the tourist $u_L(l_{at}, l_{on})$, departure time st, end time et, budget b. The number of scenic spots expected to be visited M.

   Output: k paths with higher scores and each path score $TPS[]$.

1) Initialization: $VT[] = \emptyset$, mark the scenic spots $SS[] = 0$ visited by each path.
2) Calculate SS of each scenic spot at time t according to tourist $u_L(l_{at}, l_{on})$.
3) For the current scenic spot, select the first l points with higher scores as the scenic spots to be expanded in the next step, that is, extend l paths.
4) Choose the first N path with the highest score and add them to $VT[0]$ to $VT[k]$ respectively. If $N > k$, then cut the $N - k$ paths with lower TPS. Reserve the first k paths with higher TPS. l represents the length of the path, and N represents the number of scenic spots that can be selected in the next step each time the method is executed based on the current point.
5) Determine whether each path meets the time constraint $et - st$ and budget constraint b. If not, delete the path.
6) For the reserved path, repeat steps 2 ~ 4 based on each current point.
7) When $M = 1$, or $t > st$ or TPC $> b$. End the method, and finally get k paths and TPS scores for each path.

   According to actual needs, the multiple routes before the score are taken as the basic recommendation list.

4.3   Collaborative Filtering Based on the Recommendation List of Tourists

Collaborative filtering recommendations based on the recommendation list of tourists. First, use the correlation between tourists to obtain a group of "neighbors" similar to the target tourists, and then calculate the target tourists' predicted scores for unrated items based on the historical preferences of this group of tourists. And recommend to the target tourists based on the predicted score.

   The basic principle of collaborative filtering based on tourists: Assume that tourist A likes scenic spots 1 and 3, tourist B likes scenic spots 2, and tourist C likes scenic spots 1, scenic spots 3 and 4. From the historical information, we can see that tourist A and tourist B do not Commonly liked scenic spots, but both tourist A and tourist C like scenic spot 1 and scenic spot 3. That is to say, tourist A and tourist C have greater similarity, and tourist C likes scenic spot 4 in addition to scenic spot 1 and scenic spot 3, so it is speculated Tourist A may also like scenic spot 4, so we recommend scenic spot 4 to tourist A. User CF, a collaborative filtering recommendation method based on tourists. Based on this, the main steps involved in the process of using this method to achieve collaborative filtering of recommendation lists are as follows:

(1)   Tourist data preprocessing

Using the methods of data mining and preprocessing, the original data of tourists, including tourist attribute data, behavior data and scoring data, are preprocessed and modeled to form a tourist item scoring matrix, so that it can be processed and calculated by using the recommendation method, and the recommendation results can be quickly obtained in combination with the characteristics of sports tourist attractions and routes obtained above.

(2) Calculate the nearest neighbor set

The calculation of the nearest neighbor set is a particularly important step in the recommended method, and it is also a key point that affects the performance of the method. It is generally believed that similar tourists are also more similar in preference. First, calculate the tourist set $R = (r_1, r_2, ...r_i)$ that is most similar to the target tourist A, then calculate the score of the target tourist A for the unrated items through the scores of the items in the similar tourist set R, and finally recommend the target tourist from high to low according to the predicted score.

There are three existing methods for calculating similarity: cosine similarity, modified cosine similarity and Pearson correlation coefficient. This article adopts the method of cosine similarity to realize this process. The tourist item rating matrix is regarded as a vector in an n-dimensional space, where the value of the unrated item is initialized to 0, and the similarity between tourists is calculated by calculating the cosine of the angle between the vectors. Let vector $\vec{u}$ represent the score of tourist u in the space, and vector $\vec{v}$ represent the score of tourist v in the space, then the similarity formula between tourist u and tourist v is shown in formula (5):

$$sim(u, v) = \frac{\vec{u} \bullet \vec{v}}{\|\vec{u}\| * \|\vec{v}\|} \tag{5}$$

(3) Calculate forecast score

On the basis of the nearest neighbor set, the non scored items of the target tourists are calculated according to the similarity between the nearest neighbor set and the target tourists. User CF selects formula (2.46) to predict the score of non scored items:

$$p(u, i) = \sum_{v \in M(u,k)} S_{uv} R_{vi} \tag{6}$$

Among them, $p(u, i)$ represents the predicted score of tourist u on item i, S_{uv} represents the tourist similarity between tourist u and tourist v, R_{vi} represents the real score of tourist v on item i, and $M(u, k)$ represents the k nearest neighbors of tourist u.

Thus, the collaborative filtering of recommendation list based on tourists is realized.

4.4   Intelligent Selection of Recommended Routes

When using the recurrent neural network to search the initial population, this paper adopts the coding method of real number matrix, and carries out a separate real number coding for each filtered route. A single route map is regarded as a matrix chromosome,

route map is used as a vector, route type, route location, route opening time, route number attribute map is used as a gene location, a represents the number of recommended routes, b represents a route attribute, and the chromosome matrix is expressed as $(a + 1)(b + 1)$. Then the route coding matrix model is

$$
\begin{pmatrix}
X_{11} & \dots & X_{1b} \\
\vdots & \ddots & \vdots \\
X_{a1} & \cdots & X_{ab}
\end{pmatrix}
\tag{7}
$$

In the above route coding matrix model, the fitness function is set to screen the recommendation results and obtain the optimization of the optimal solution. The larger the fitness function value, the stronger the individual's ability to adapt to the environment and the greater the opportunity to reproduce. On the contrary, the smaller the individual function value, the smaller the individual's ability to adapt to the environment and is likely to be eliminated.

This article uses a piecewise function to design the fitness objective function, B represents the attribute of the target, and ε represents the maximum allowable error. Then the calculation method of the objective function value is:

$$
f_a = \begin{cases}
\left| \frac{b-B}{B} \right|, & \frac{b-B}{B} \leq \varepsilon \\
1, & \frac{b-B}{B} > \varepsilon
\end{cases}
\tag{8}
$$

Suppose the corresponding value of the route attribute is λ_a, and the number of routes included in the initial population is A, then the fitness objective function of the entire route can be expressed as:

$$
f = f_{\max} = \sum_{a=1}^{A} \lambda_a f_a
\tag{9}
$$

In this way, the continuous optimization of the recommendation results is realized, and the degree of fit between the route and the recommendation requirements is improved. The recommendation process of specific tourist routes is shown in Fig. 2:

According to the analysis of Fig. 2, the characteristics of sports tourist attractions, routes and tourists are obtained through data clustering, and the tourist data is preprocessed; Calculate the nearest neighbor set and calculate the route recommendation prediction score; The recommendation list of tourists is determined based on collaborative filtering, and the circular neural network model is constructed to realize the intelligent screening of recommended routes; The fitting degree of fitness objective function is calculated by piecewise function, and the recommendation effect of tourism route is obtained.
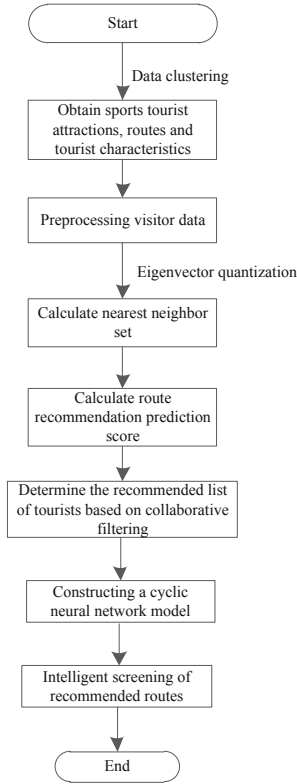
Fig. 2. Recommendation process of tourist routes

5 Experimental Results and Analysis

5.1 Experimental Design

The paper chooses to test the recommendation effect of the scenic spot recommendation method from the aspect of accuracy. In order to ensure that the experiment is true and effective, the design crawler grabbed the real score data and basic information of 2389 scenic spots through the network as experimental data. Since the paper sets a loop

Fig. 3. Experimental scene

parameter when building the tourist preference model, the loop parameter value 0 is analyzed experimentally.The specific experiment is shown in Fig. 3.

5.2 Experimental Index

Assuming that the true value of a group of time series is $y = \{y_1, y_2, ..., y_n\}$ and the predicted value is $y' = \{y'_1, y'_2, ..., y'_n\}$, the mean absolute value deviation (MAE) is taken as the experimental index,

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y'_i - y_i| \tag{10}$$

The value of *MAE* is between $[0, +\infty]$, and the greater the error, the greater the value of *MAE*.

5.3 Result Analysis

(1) When conducting experiments, under different values of N in the recommended list Top N, set the cycle coefficient 8 to 0, 0.3, 0.5, 0.7, and judge the influence of different cycle coefficients on the accuracy of the method, as shown in Fig. 4.
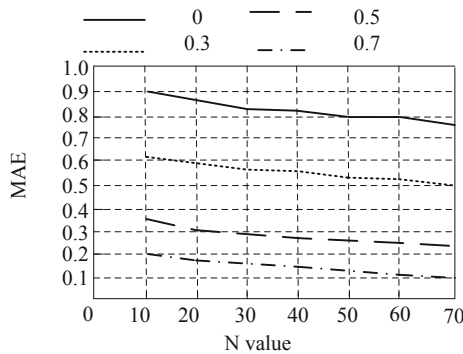

Fig. 4. Different cycle coefficients for the accuracy of the method

It can be seen from Fig. 4 that the accuracy of the method changes greatly under the influence of different cycle coefficients. When the circulation coefficient is, that is, tourists' preference is not affected by time, the high Mae value means that the accuracy of the method is not high. The greater the circulation coefficient, the greater the influence of time on tourists' preference. It can be seen from the experimental results that no matter what value n is, the greater the cyclic coefficient, the higher the accuracy of the method. When the cycle parameters are 0, 0.3 and 0.5, the MAE value decreases greatly. When the cycle parameter changes from 0.5 to 0.7, the MAE value decreases, but the range

is small. That is, when the cycle parameter is 0.7, the recommendation accuracy of the recommended method is high.

When the loop parameter is constant, and the value of N is 10, the MAE value is too high, that is, the accuracy of the method is not high. As the value of N increases, the value of MAE gradually decreases, and the accuracy of the method gradually increases. When the value of N changes from 60 to 70, the change of the MAE value tends to be stable, that is, the accuracy of the method does not change much.

(2) In order to verify the performance of the improved method, when the cycle parameter is 0.7, the paper method is compared with the method proposed in literature [8] and the method proposed in literature [9], and the performance of the paper method is verified from the aspect of MAE value. For the three methods, the MAE values are compared when the values of N in the top n of the recommendation list are 10, 20, 30, 40, 50, 60 and 70 respectively. The experimental results are shown in Table 1.

Table 1. Mae comparison of recommended results of different methods

N value	Reference [8] method	Reference [9] method	Paper method
10	0.75	0.66	0.12
20	0.69	0.58	0.08
30	0.62	0.56	0.07
40	0.60	0.52	0.05
50	0.58	0.49	0.05
60	0.76	0.63	0.05
70	0.69	0.70	0.05

It can be seen from Table 1 that the MAE value of the general collaborative cycle method is significantly higher than that of the other three methods, and the MAE value decreases with the increase of N, that is, the accuracy of the method continues to improve. There is little difference in Mae value between the method in literature [8] and the method in literature [9].The MAE value of the method in document [8] is not affected by N, that is, the accuracy of the method does not change significantly when n is taken, while the MAE value of the method in document [9] decreases with the increase of n. The MAE value of the scenic spot recommendation method in this paper is much smaller than that of the other three methods, and the MAE value of this method decreases with the increase of n value. When the n value changes from 40 to 50, the MAE value tends to be stable.That is, the accuracy of the scenic spot recommendation method in the paper is higher than the other three comparison methods, and with the increase of N, the accuracy of the scenic spot recommendation method will gradually stabilize.

6   Conclusion

This paper designs an intelligent recommendation method of sports tourism route based on cyclic neural network. Determine the recommendation target of sports tourism route, extract the characteristics of sports tourism attractions and routes and tourists' interest, collaborative filter the tourism recommendation list through data clustering, and finally optimize the recommended route by using cyclic neural network. The experimental results show that the MAE value of the recommended route of this method decreases continuously, and the MAE value tends to be stable when the n value changes from 40 to 50. It shows that the route recommendation effect of this method is good.

However, there are still some deficiencies in this method. The unified time calculation method between scenic spots is adopted in this paper, but in practice, there may be traffic jams and other problems between scenic spots, which will lead to the scenic spots of one day can not be visited in the original time period. Therefore, according to the preferences of tourists when playing in different types of scenic spots, the fluctuation of playing time and the consideration of traffic conditions between scenic spots when designing sports tourism routes will be the next research content of the paper.

References

1. Lizana, M., Carrasco, J.A., Tudela, A.: Studying the relationship between activity participation, social networks, expenditures and travel behavior on leisure activities. Transportation **47**(03), 1765–1786 (2020)
2. Chen, J., Qi, K., Zhu, S.: Traffic travel pattern recognition based on sparse global positioning system trajectory data. Int. J. Distrib. Sens. Netw. **16**(10), 15501477209 (2020)
3. Arif, A., Du, J.T.: Understanding collaborative tourism information searching to support online travel planning. Online Inf. Rev. **43**(3), 369–386 (2019)
4. Petersen, N.C., Rodrigues, F., Pereira, F.C.: Multi-output bus travel time prediction with convolutional LSTM neural network. Expert Syst. Appl. **120**(15), 426–435 (2019)
5. Malik, S., Kim, D.H.: Optimal travel route recommendation mechanism based on neural networks and particle swarm optimization for efficient tourism using tourist vehicular data. Sustainability **11**(12), 1–26 (2019)
6. Ma, L., Li, X., Bo, J., et al.: From subjective and objective perspective to reconstruct the high-quality tourism spatial structure-taking gannan prefecture in China as an example. Sustainability **12**(3), 1–17 (2020)
7. Chi, Y., Li, R., Zhao, S., et al.: Measuring multi-spatiotemporal scale tourist destination popularity based on text granular computing. PLoS ONE **15**(4), 1–33 (2020)
8. Lee, G.H., Han, H.S.: Clustering of tourist routes for individual tourists using sequential pattern mining. J. Supercomput. **76**(8), 5364–5381 (2020)
9. 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
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

Plant Landscape Configuration Method of Regional Characteristic Rainwater Garden Based on Deep Learning

Qian He[1,2(✉)], Jing Lin Ng[1], and Nur Ilya Farhana Md Noh[3]

[1] Department of Civil Engineering, Faculty of Engineering, Technology and Built Environment, UCSI University, 56000 Kuala Lumpur, Malaysia
`lingxi1109@163.com`
[2] Department of Civil and Hydraulic Engineering Institute, Xichang University, Xichang 615013, China
[3] School of Civil Engineering, College of Engineering, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia

Abstract. With the acceleration of urbanization and the continuous expansion and construction of cities, more and more cities in China are facing increasingly serious urban rainwater problems. As a rainwater management measure under the low impact development system, rainwater garden can manage and utilize rainwater resources. Therefore, a plant landscape configuration method of regional characteristic rainwater garden based on in-depth learning is proposed. By analyzing the types and characteristics of rainwater landscape facilities; Constructed rainwater garden runoff management system; Design the plant configuration scheme of wet area, semi-humid area, arid area and semi-arid area of rainwater garden; The plant landscape characteristic parameters are calculated based on the deep learning algorithm to complete the plant landscape configuration. Experiments show that the waterlogging tolerance index D of this rainwater garden landscape configuration method for different kinds of plant landscape is higher than that of the traditional configuration method, which can manage rainwater and improve the ecological environment at the same time.

Keywords: Deep learning · Area · Rain garden · Plant landscape · Configuration method

1 Introduction

As a traditional best management measure, the rain garden is a rainwater management facility under a low-impact development system. The common form is a shallow recessed green space with landscape effects for planting shrubs, flowers, grass, and trees. It absorbs from urban roofs. Rainwater in impervious areas such as roads, sidewalks, parking lots, and impervious lawns, and through the purification, filtration and infiltration of soil and plants to manage rainwater runoff [1]. After more than two decades of development,

related research and applications of rainwater gardens have become more abundant. The form and structure of rainwater gardens have also evolved and expanded with the continuous enrichment of application ranges and scales. The original more primitive rain garden was a kind of planting and land injection around the building. Its effect was mainly about the retention of rainwater, the purification of rainwater by plants and soil, and the infiltration of groundwater. The form and effect were relatively simple and single [2, 3]. With more and more abundant practice and exploration, modern rainwater gardens are becoming more and more efficient, scientific and reasonable in the function of rainwater management, and more flexible and rich in application scales, which in turn brings more changes in the form of landscape expression, landscape art The design space has also been greatly improved. Today's rainwater gardens are more scientific in the control of rainwater volume and water quality, and the rainwater management system surrounding rainwater gardens is also more systematic [4].

The rich application scale and changeable expression forms also expand the landscape expression space of today's rainwater garden. Therefore, the rainwater garden has developed from an engineering measure for site development of simple rainwater management to a landscape infrastructure, which has been greatly improved in both functionality and artistry. It can be said that the original rainwater garden is a site rainwater management measure with certain landscape value, while today's rainwater garden is more landscape design works integrated with the concept of rainwater management [5]. During the operation of rainwater garden, rainwater is collected from the site and treated in combination with its own soil, vegetation and related facilities. Deep learning (DL) is a subset of machine learning (ML). Inspired by human brain, DL adopts multi-layer interconnected artificial neural network algorithm. Modern deep learning usually adopts tens or even hundreds of layers of neural network structure. Each layer is gradually abstracted on the basis of the previous layer, and finally features are extracted from the training data. The deep learning algorithm is used to realize the automatic design of plant images, and the algorithm is embedded in mobile app and applied to the configuration of garden plant landscape [6–8].

In summary, this article proposes a deep learning-based method for configuring the landscape of regional rainwater garden plants. By using the multi-media effects of vegetation, microorganisms and soil, it provides storage space for rainwater runoff in and around the site to achieve collection and purification. The purpose of rain. In terms of ornamental value, the landscape form of the rain garden presents the beauty of nature. The appearance of the rain garden has the beauty of the traditional garden. It provides citizens with a comfortable place to relax, relax, and relieve fatigue. It can also use the power of nature to change the previous rainwater gathering. The phenomenon of water deterioration and mosquitoes in the land has promoted the closeness of the citizens to the natural environment.

2   Regional Featured Rainwater Garden Plant Landscape Configuration Method Based on Deep Learning

2.1   Analysis of the Types and Characteristics of Rainwater Garden Landscape Facilities

From the perspective of the functional objectives of rainwater management, rainwater garden is firstly an engineering measure. Technically, the construction of rainwater garden is a very specific and detailed work. Excellent and outstanding rainwater garden design requires the close cooperation between landscape designers and environmental engineers, and the engineering and landscape should be combined with each other [9, 10]. Reasonable engineering design and construction is the basis of exerting the efficiency of rainwater management and landscape optimization design. Firstly, taking the ground part of the main structure of the rainwater garden as the reference, this paper summarizes and divides the elements that have an important impact on the design of the rainwater garden.

Topography

The topography is manifested as the undulating changes of the topography on the surface. The topography is closely related to the formation of surface runoff. The topography and landform determine the natural drainage mode of the site. Rainwater that falls to the ground will form surface runoff when it has not evaporated and penetrated into the soil. Rainwater runoff, runoff direction and runoff speed are closely related to topography [11, 12]. Regulating surface drainage and guiding the direction of water flow are an important and inseparable part of the garden landscape design. Therefore, for the construction of rainwater gardens, the topographical planning and design is the key to the effective collection and management of rainwater. A comprehensive analysis and research of the site topography is an important prerequisite for deciding what rainwater management measures to take and the planning and layout of corresponding facilities in the site. From the perspective of runoff collection and management, the vertical design of the site and the planning and layout of rainwater management related facilities are also carried out based on the original topography of the site.

Rainwater Garden Facilities

Facilities are the main structural support of rainwater garden and the main carrier of rainwater runoff guidance and collection. The facilities in the rainwater garden are mainly divided into structural facilities and rainwater management functional facilities. The structural elements are mainly the main body of the form of the rainwater garden and the landscape functional facilities, such as the form boundary of the rainwater garden and the facilities for visitors and maintenance personnel to enter the garden. Stormwater management functional facilities include relevant facilities that can collect, guide, retain and overflow runoff, such as transmission facilities, overflow facilities, etc. In the construction of rainwater garden, the construction of facilities is first based on meeting

the functional requirements, and then combined with the landscape design of materials and forms to achieve the effect of optimizing the landscape quality.

Rainwater Garden Vegetation

The growth forms of plants that cover most of the ground are diverse. Plants are the most important landscaping element in garden landscape design. Different regions have different plant varieties and plant communities. The shape, color, fragrance, habits and habits of plants are Seasonal landscape performance is the main ornamental characteristic of plants. Plant design and configuration are also an important part of site landscape construction. In addition to the value of landscape, the selection and configuration of plants in rainwater gardens also have important rainwater management functions.

Plants in rain gardens can alleviate many environmental problems, such as air purification, soil and water conservation, water conservation, temperature adjustment, and so on. The surface vegetation has the function of intercepting and conserving precipitation. The leaves and roots absorb part of the moisture from rainfall, dew and fog, and the remaining part of the precipitation penetrates through the soil to maintain the groundwater level or supplement the mouth of the underground aquifer. Plants have a variety of effects on the ecological environment of the site. The respiration of plants can absorb moisture in the soil and then dissipate it into the air in the form of water vapor through transpiration, replenishing air moisture; plants can also improve the microclimate in summer. Can resist the drying effect of wind and sun to maintain a cool environment.

Rainwater Management Function Layer

The rainwater infiltration and purification function layer refers to the structural layer with purification effect experienced by the rainwater in the rainwater garden in the process of infiltration before groundwater supplement or collection and utilization. The common corresponding structural functional layers include planting soil layer, filling layer and rolling stone layer.

Due to different regional characteristics and rainwater management objectives and systems, there are certain differences in the composition of specific functional layers and the main structural components of each layer. In the area with good soil permeability, the filler layer does not need to add artificial materials to improve the permeability, and its main components are generally consistent with the planting stop layer. For sites with poor soil permeability, the surface planting soil should be mixed and improved to ensure that it is suitable for plant growth and has certain permeability. The filler layer needs to be filled with natural or artificial materials to ensure the timely infiltration of rainwater.

The characteristics of rainwater runoff also affect the structural composition of the functional layer of rainwater garden. The rainwater garden with large amount of rainwater resources to be collected and managed or heavy rainwater pollution needs a larger scale and deeper functional layer to prolong the infiltration time of rainwater and fully filter and purify the rainwater. At this time, the composition and thickness of the filler layer are very important design elements. The growth characteristics of plants directly affect the thickness of planting soil layer. In rainwater gardens, vegetation with relatively developed roots is often selected to improve the purification effect of plants on rainwater. Therefore, the corresponding thickness of planting stop layer should ensure the normal growth of

plants. Generally, the minimum thickness of planting King layer in rainwater gardens with trees is 120 cm, The minimum thickness of planting stop without trees is 60 cm.

2.2   Build a Runoff Management System for Rainwater Gardens

In the early stage of the design, the rainwater runoff management system of the rainwater garden should be constructed according to the site characteristics. The runoff management system of the rainwater garden constructed in this paper is shown in Fig. 1.
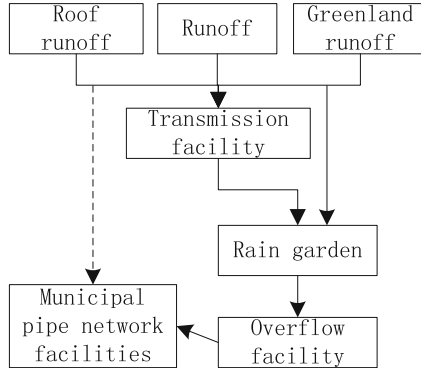

Fig. 1. Runoff management system structure of rainwater garden

As shown in Fig. 1, the rainwater garden mainly collects runoff from building roofs, ground paving, and surrounding green spaces. Part of the rainwater runoff from these areas flows directly into the rainwater garden, some directly into the municipal pipe network facilities, and most of the runoff directly or Indirectly combined with the guidance of the corresponding rainwater transmission facilities to flow to the rainwater garden, and the rainwater exceeding the treatment limit of the rainwater garden will also be discharged into the municipal pipe network facilities by the overflow facilities.

Therefore, the first step in the design of rainwater gardens needs to be combined with the runoff management system to determine the corresponding facilities and the horizontal position distribution of the rainwater garden in the site, so as to build a complete stormwater runoff control system in the site. The layout of facilities should follow the process of surface runoff conforming to the current terrain flow, and arrange transmission facilities, rainwater gardens and overflow facilities in sequence according to the runoff movement process, so as to gradually realize the orderly control of rainwater runoff in stages. Determine the location and distribution of the facilities and gardens. Further work is to clarify the design goals of the rainwater garden and the scale and scale of the garden based on the related characteristics of the site's rainwater runoff. Finally, before construction, the soil characteristics of the site should be carefully inspected, and the soil-related characteristics should be tested to determine the soil improvement plan.

The site selection of rainwater garden design should first consider the area that can conveniently and fully collect and utilize the rainwater runoff around the site. The location of rainwater garden should be located between the source of rainwater runoff

and the area where it naturally ends in the landscape, rather than the farthest point that rainwater runoff can reach, because the main purpose of rainwater garden is to collect rainwater as much as possible before rainwater runoff reaches the farthest point, and the farthest point of runoff often means that the regional catchment energy is strong but the drainage performance is poor. In general, the site selection of rainwater garden should pay attention to the following aspects: make sufficient research on the underground pipe network facilities in the site to avoid the harm of rainwater infiltration to the pipe network facilities. The rainwater garden around the building shall be at least 2.5 m away from the building foundation to prevent infiltration runoff from affecting the building foundation. Rainwater from the roof of the building can be guided into the rainwater garden by downpipes; The rainwater garden should not be built in the area close to the water supply system or around the well, so as not to pollute the water source by the rainwater that is not purified in time; The rainwater garden is not a wetland and should not be located in the low injection land with frequent ponding. If the rainwater garden is selected on the site with poor soil drainage, a large amount of rainwater will gather in the rainwater garden, and the rainwater cannot infiltrate in time, which is not only unfavorable to plant growth, but also easy to breed mosquitoes; Choosing a flat site to build a rainwater garden will be easier and easy to maintain; Try to keep the rain garden on the sunny side, which is conducive to the growth of plants and the evaporation of rain; Coordinate the relationship between the location and landscape performance of the rainwater garden and the surrounding landscape environment.

The area and scale of the rainwater garden is determined by many factors, and the rainwater runoff that the site needs to collect and manage is the most critical factor that affects the scale of the rainwater garden. Generally speaking, the designed rainwater garden needs to absorb all the rainwater runoff that may reach it, so the drainage area of the entire drainage runoff into the rainwater garden can be estimated, and then the approximate area of the rainwater garden to be designed can be estimated.

The area of the drainage basin that collects rainwater runoff from the surface of the building should be determined by the distance between the rainwater garden and the building and the amount of water discharged from the downpipe into the rainwater garden. If the rainwater garden is less than 10 m away from the building, when all downpipe water is discharged into the rainwater garden, the approximate area of the building itself should be regarded as the basin area. If only part of the downpipe water is discharged into the rainwater garden, only the amount of runoff will be estimated Part of the roof area. If the distance between the rainwater garden and the building is greater than 10m, the drainage surface is not only the roof area, but also the approximate drainage area of the runoff formed. The area value can be estimated by measuring its length and width. There are many calculation methods for more accurate size calculation at home and abroad. The commonly used methods for calculating the area of rainwater gardens abroad are: infiltration method based on Darcy's law, effective aquifer volume method and proportional estimation method based on catchment area.

2.3   Landscape Configuration Design of Rainwater Garden Plants

Configuration of Runoff Control Plants in Humid Areas
The pollutants contained in rainwater are as follows: suspended solids (SS), organic pollutants (COD), chlorine, total phosphorus (TP), dissolved phosphorus, total nitrogen (TN), ammonium nitrogen, total iron, lead, zinc, etc., and are mainly suspended solids (SS) and organic pollutants (COD). It has the following characteristics: the variation range of pollutants is large and the randomness is strong; The pollutant concentration shows a downward trend with the duration of rainfall, and the initial rainwater quality is poor, especially the SS, COD and other indicators exceed the standard seriously; D. There is a good linear correlation between Pb, Zn and SS. SS is not only a pollutant itself, but also the particle surface that makes up it provides attachment conditions for other pollutants.

Due to the large rainfall in the humid area, a certain catchment area will be formed. Therefore, special consideration should be given to the decontamination capacity of wetland plants and aquatic plants. Some studies have shown that the purification ability of wetland plants to water is generally submerged plants > floating and floating leaf plants > emergent plants. Wetland plants with nitrogen and phosphorus removal function in the wet area of rainwater garden are shown in Table 1.

Table 1. Wetland plants with nitrogen and phosphorus removal function in humid areas

Plant type	Plant name
Emergent plants	Cress, Rush, Calamus, Canna, Water onion, Calamus, Yellow calamus Cattail, Di, Lythrum chinensis, Flat stalk recommended grass, Zaili flower, Papyrus, Carex, Luzhu, Barracuda, Cigu, Shilongwei, Alisma, Rumex sorrel, Feng Car grass, Vetiver, Saccharum, Broadleaf cattail, Reed, Wattle, Duck Commelina, Rice, Cattail, Wild rice
Floating plants	Water lily, Yellow water dragon, Gold anemone, Dadang, amphibians, Sophora japonica, Water Turtle, Nymphaea, Manjianghong, Ziping, Eichhornia crassipes, Hydrilla verticillata, Potentilla vulgaris, Potentilla glabra
Submerged plants	Ceratophyllum, Vallisneria, Potamogeton crispus, Hydrilla verticillata, Myriophyllum sp Pickled vegetables

As shown in Table 1, calamus and elephant grass have the best nitrogen removal effects, and windmill grass and calamus have the best phosphorus removal effects. Considering the decontamination ability, stress resistance and ornamental properties of wetland plants, calamus, elephant grass, windmill grass, cattails, jasmine rice, yam, spring yam, pike grass and red eggs are suitable for planting in humid areas. Terrestrial plants should also be matched with plants with strong decontamination ability as much as possible. Such as weeping willow, hibiscus, triangular maple, etc., have a certain

adsorption effect on sulfur dioxide. Ficus microcarpa, Cinnamomum camphora, Luan tree, etc. which are resistant to oxygen fluoride, Ailanthus altissima, ash tree and Populus chinensis, etc., which are resistant to chlorine gas.

Reasonable matching design according to the purification effect of different plants can enhance the purification effect of plants. For example, plants with vigorous roots and good oxygen transport capacity are the best plant choices for removing BOD and N from water; for the purification of pollutants such as N, P, heavy metal ions and other organic matter, it is necessary to select plants with rapid growth ability and Plants with strong accumulation ability; the composition of urban precipitation becomes very complicated due to the different pollutants in the air. Therefore, when selecting plants, it is necessary to match plants with different root growth depths, so as to perform more comprehensive rainwater. Treatment: Plant density is directly proportional to the infiltration time of rainwater runoff. High-density planting will slow down the runoff speed and also help improve the purification effect. This kind of rain garden is suitable for selecting plant species that can grow quickly, have a well-developed root system, luxuriant branches and leaves, and have more sprouting peaks a year.

The rainfall in the humid area is large, so it is necessary to effectively control the rainwater runoff through the configuration of rainwater garden plants, increase the rainwater infiltration time and avoid siltation. Therefore, trees and shrubs with developed roots and flood resistance should be selected as the main plants. Such as Metasequoia, weeping willow, duying, mulberry, triangular maple, etc. Tall fescue, Pennisetum, iris, horseshoe gold and other waterlogging resistant plants with developed roots can be selected as the ground cover, which can not only slow down the runoff speed, but also achieve the natural connection with the surrounding environment and roads.

The rainwater collection tank can be in the form of a reservoir, i.e. a precast concrete tank. The reservoir collects rainwater in the rainy season, and the plants planted in it can be used as an ornamental planting tank outside the rainy season. The two plants planted in the house, flat rushes and multi flower blue fruit grass, have the habit of moisture and drought tolerance. The root system of Juncus latifolia is developed, which can slow down its flow rate in the process of rainwater flow and facilitate the infiltration of rainwater in the rainwater garden area.

Plant Landscape Configuration in Semi-Humid Area
Judging from the existing rain garden engineering practice, in winter, except for plants in warm areas that did not appear to die due to low temperature, wet plants in most areas of my country will suffer from the problem of above-ground die. The basic reason for the problem is that the currently used rain garden plants are mainly single herbaceous aquatic plant types such as reed, canna, windmill grass, cattail, water onion, water axe, ginger flower, calamus, and barracuda. In the cold and low temperature season in winter, most herbaceous plants will inevitably die in winter. Due to the loss of winter plant effect, it will affect the water purification and landscape effect of the winter rain garden. Therefore, the selection of cold-tolerant wetland plants is very important for this area. According to research, the purification capacity of wetland plants is mainly divided into the following three categories: The first category has strong purification potential, including canna, reed, windmill grass, water onion, Zai Lihua, Lythrum chinensis, Canna Canna, etc.; The purification potential of the major categories is medium, including Pueraria lobata,

Arundina striata, Cattails, Barracuda, etc.; the third category has a weak purification potential, including wild taro, warbler tail, rush, green onion, Alisma sinensis, and flower head According to their attributes, these wetland plants can be used in humid and semi-humid areas.

The selection of runoff control plants in semi humid area mainly considers their cold resistance, drought resistance, water humidity resistance and deep root, such as weeping willow, metasequoia, Populus microphylla, maple poplar, juniper, Ligustrum lucidum, etc. The ground cover can be tall fescue, ryegrass, iris, etc.

Plant Landscape Configuration in Arid and Semi-Arid Areas
Since the rainfall in semi-arid areas is not suitable for ponding, the purification of water quality is mainly through the purification of rainwater by trees, shrubs and terrestrial plants in the process of runoff and infiltration. The purification effect of xerophytes needs to be considered. At the same time, most plants are not suitable for survival in winter. In order to ensure the purification effect in winter, evergreen plants must be properly matched to ensure the purification function and landscape effect in winter. The optional plants in arid areas include weeping willow, metasequoia, triangular maple, Bauhinia, small leaf boxwood and other plants, which have a certain adsorption effect on heavy metal elements in rainwater.

In the semi-arid area, plants with drought tolerance and developed roots are mainly selected, trees can be strange willow, Platycladus orientalis and juniper, shrubs and ground covers can be Lespedeza, iris, Amorpha fruticosa and Sabina mongolica, which have developed roots, have good water conservation effect and delay the flow rate of rainwater. At the same time, it can be properly matched with Begonia and Tianmu Qionghua to increase the landscape color.

As one of the most important elements in the design of sponge city, rainwater garden should firstly focus on its functional benefits. After the water purification and runoff control are optimized, its aesthetic benefits should be considered on this basis, because rainwater gardens are used as urban green spaces., Need to bring people entertainment and sensory feelings. In short, the rain garden is an urban green space design that takes functional benefits as the main and aesthetic benefits as a supplement, and combines the two perfectly.

2.4   Calculating Plant Landscape Feature Parameters Based on Deep Learning Algorithms

Plant Landscape Data Collection and Preprocessing
Firstly, 1000 common garden plants were photographed to collect data. Because the shape, color and texture of plants change in different periods, it is necessary to select cloudy and sunny days in spring, summer, autumn and winter, and take 50 photos of each plant every day, a total of $1000 \times four \times two \times 50 = 400k$ original photos. Then ask landscape plant experts to screen the original photos, eliminate the photos that do not meet the requirements, mark each photo, and finally form about 300K photos of 1000 kinds of plants. These photos are stored on the hard disk in JPEG format. Each plant has a directory, and the directory name is the name of the plant.

The labeled data is divided into three sets: training set, verification set and test set. The training set accounts for 60% of the total data and is used to train the deep learning model; The validation set accounts for 20% of the total data and is used to adjust the super parameters. The adjustment process needs to use the performance of the model on the validation set as feedback; The test set accounts for 20% of the total data and is used to evaluate the training results.

Since the data set is limited, over-fitting may occur during training, resulting in poor generalization performance. In order to reduce the over-fitting phenomenon in the neural network training process, the data is preprocessed. Preprocessing will randomly generate more data from the training set data. There are 5 steps in preprocessing, and the parameters of each step are randomly generated. Among them, the first step of image rotation, the parameters include rotation center and rotation angle; the second step of image cropping, the location of the cropped area is randomly selected, and the size after cropping is 80% of the original image; the third step is mirroring, including X Axis mirroring and Y Axis mirroring; Step 4 γ correction, γ correction is performed on each of the three color channels of RGB; Step 5 Gaussian white noise, add Gaussian white noise of $N(\mu, \alpha^2)$, where $(\mu = 0, \alpha = 10)$; Finally, format the data as a tensor.

Calculate the Characteristic Parameters of Plant Landscape

The separable convolution of deep learning decomposes the standard convolution into deep convolution and a 1×1 point-by-point convolution. The deep convolution uses 1 filter for each input channel to filter, and the point-wise convolution uses a 1×1 convolution operation to combine the outputs obtained by all the deep convolutions. The depth separable convolution divides the convolution into 2 layers. This decomposition can effectively reduce the amount of calculation and the size of the model. Set the convolution kernel to K, and the calculation formula for the size of the convolution kernel is:

$$K = D_K \times D_K \times M \times N \tag{1}$$

Among them, D_K is the spatial dimension of the convolution kernel; M is the number of input channels; N is the number of output channels, and the plant landscape output characteristic parameter G is:

$$G = \sum Kij \times F_{k+i-1, l+j-1} \tag{2}$$

Among them, i and j respectively represent the input space dimension; k represents the space dimension constant. The calculation formula of the deep learning convolution feature parameters of the plant landscape is:

$$G_T = \sum K_{ij} \times i \frac{F_{k+i-1, l+j-1}}{j} \tag{3}$$

$$i = \frac{K_{i+1}}{D_K} \cdot F_{k+i-1} \tag{4}$$

$$j = \frac{K_{j+1}}{NG} \cdot F_{k+j-1} \tag{5}$$

The output of depth convolution is linearly combined by convolution. In deep learning, most problems are difficult to calculate the global optimal solution, so the iterative optimization method is usually used to calculate the local optimal solution. Applying deep learning technology to the subject of plant landscape configuration and training by using a large number of labeled data can not only effectively solve the number of species identified in plant landscape configuration, but also greatly improve the efficiency of plant landscape configuration in rainwater garden.

3 Experiment and Analysis

3.1 Experiment Preparation

In order to verify the effectiveness of the regional feature rain garden plant landscape configuration method based on deep learning proposed in this paper, the following experiments were carried out. The plants selected in the experiment belong to 8 common garden ground cover plants in 8 families and 8 genera. Among them, 4 kinds of herbaceous plants are water ghost banana, fragrant color finch, four season begonia, and purslane; 4 shrubs, namely Brazilian wild male, red banana, dragon boat flower, hibiscus. The native grass is 0.5-year-old seedlings, and the shrubs are 1–2 year-old seedlings. The test materials were purchased from the market. Plants with healthy growth, no pests and diseases, and uniformity were selected. They were planted in plastic flower pots with a height of 20 cm, a pot diameter of 25 cm, a bottom diameter of 15 cm, and a hole at the bottom. The planting soil was loess and peat. Soil 7:3 is mixed with a layer of ceramsite at the bottom; the seedlings of the test materials are slowed for 3 months after changing pots. The waterlogging tolerance test was carried out in a rain garden test base in Jiangsu Province using the potted water control method. For each plant, 4 pots of plants with the same morphology and growth are selected. Starting from June 10, 2019, the plants and flower pots are placed in large plastic pots, with the water surface 2cm above the upper edge of the plant pot as the benchmark. During the treatment, the water was changed every six days. At 0, 7, 14, 21, and 28 days of flooding stress, the morphological changes of plants were observed and physiological indexes were measured, and each index was repeated 3 times. The growth performance of 8 plants during the stress period, such as flowering, leaf yellowing and new leaf sprouting, were recorded every 7 days. At the same time, the leaves with normal growth and the same size were randomly selected for sampling. The OPTI-SCIENCESOS1p fluorometer was used to determine the PSII original light energy conversion efficiency of the plant leaves (Fv/Fm) value; the relative water content of the leaves is measured according to the saturated water content method; the relative conductivity is measured by a conductivity meter; the content of malondialdehyde (MDA) is measured by the thiobarbituric acid method; the content of proline (Pro) Measured with sulfosalicylic acid method. The obtained data was analyzed by SPSS one-way analysis of variance, and the average value, standard error and difference significance of each physiological index data were obtained, and the Excel software was used for drawing. Principal component analysis is used to obtain the contribution rate of the comprehensive index. At the same time, the membership function method is used to calculate the membership function value of the comprehensive index. Finally, the

comprehensive evaluation index D value is calculated to comprehensively evaluate the waterlogging tolerance of 8 plants.

3.2   Result Analysis

Set the plant landscape configuration method of regional characteristic rainwater garden based on in-depth learning proposed in this paper as the experimental group and the traditional landscape configuration method as the control group. Compare the D value of plant waterlogging tolerance index of the two methods. The comparison results are shown in Table 2.

Table 2.   Plant waterlogging tolerance index D values of the two methods

Plant species	Test group	Control group
Hibiscus	5.034	2.058
Water ghost banana	8.549	3.649
Fragrant finch	8.319	6.157
Four seasons begonia	10.057	6.199
Dragon boat flower	12.308	9.628
Brazilian wild peony	6.628	3.336
Purslane grandiflora	7.023	4.152
Zhu Jiao	4.168	2.058

As shown in Table 2, the plant landscape configuration method of regional characteristic rainwater garden based on in-depth learning proposed in this paper has higher waterlogging tolerance index D values for different types of plant landscape than the traditional configuration method, indicating that the plant landscape has less litter, the plant landscape grows well in the rainwater garden, and beautifies the environment while managing rainwater, Promote balanced ecological development.

In the experiment, the management efficiency of the runoff management system in this paper is analyzed. By comparing the management time cost of this method, traditional method 1 and traditional method 2, the results are shown in Table 3:

By analyzing the experimental data in Table 3, it can be seen that there are some differences in the time cost of management using this method, traditional method 1 and traditional method 2. Among them, the management efficiency of this method is the shortest, and the shortest time cost is 1.2 s. The time cost of traditional method 1 and traditional method 2 is longer, and is always greater than that of this method. This is because this paper systematically analyzes the types and characteristics of rainwater landscape facilities in the design; Construct rainwater garden runoff management system; Design the plant configuration scheme of wet area, semi-humid area, arid area and semi-arid area of rainwater garden; The plant landscape characteristic parameters are calculated based on the deep learning algorithm, the plant landscape configuration is completed, and the efficiency of system management is improved.

Table 3. Management time overhead of different management systems (s)

Iterations/time	The method of this paper	Traditional method 1	Traditional method 2
20	1.2	1.3	1.5
40	1.3	1.5	1.8
60	1.2	1.9	2.0
80	1.3	2.1	2.3
100	1.3	2.3	2.5

4 Conclusion

This paper studies the design of rain garden with regional characteristics from the perspective of deep learning, and focuses on the methods and approaches of landscape environment construction in rain garden design and application. Analyzed the influential elements of rainwater garden design, and detailed analysis of the influence relationship between each element and rainwater garden design, which provides a solid foundation for the rational design of rainwater garden. Building a runoff management system and a landscape design system, and focus on the landscape design system. From the aspects of site planning and topography design, plant landscape design, construction and construction of facilities, and the use of rainwater elements, the design of the rainwater garden guided by the landscape is studied. The method and approach are of great significance to the landscape configuration of the regional rainwater garden.

References

1. Cadavid-Florez, L., Laborde, J., Zahawi, R.A.: Using landscape composition and configuration metrics as indicators of woody vegetation attributes in tropical pastures. Ecol. Ind. **101**, 679–691 (2019)
2. Song, X., Gao, X., Weckler, P.R., et al.: An in-situ rainwater collection and infiltration system to improve plant-available water and fine root growth for drought resistance. Appl. Eng. Agric. **36**(5), 807–814 (2020)
3. LI, H., Guo, et al.: Landscapes in the building entrance space: a case study of Huacheng square in Guangzhou. J. Landsc. Res. **11**(05), 8–11+15 (2019)
4. An, Q., Ding, J., Tu, J.: Simulation of plant landscape image layout feature point optimization extraction method. Comput. Simul. **37**(11), 207–210 (2020)
5. Aslan, M., Akan, H.: A study of natural woody plants of forest in anlurfa–determination of detection and landscape values of parks and garden plants. Biol. Divers. Conserv. **12**(1), 50–65 (2019)
6. 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
7. Guneroglu, N., Bekar, M., Sahin, E.K.: Plant selection for roadside design: "the view of landscape architects." Environ. Sci. Pollut. Res. **26**(33), 34430–34439 (2019)
8. 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

9. Odusanya, T.I., Owolabi, C.O., Olosunde, O.M., et al.: Propagation and seedling growth of some species used as ornamental hedges in landscape design. Ornam. Horticult. **25**(4), 383–389 (2019)
10. Wu, Y., Wu, J.: Analysis of key frames of square greening landscape pattern data under VR technology. Comput. Simul. **38**(3), 336–340 (2021)
11. Al-Hayanni, M., Rafiev, A., Xia, F., et al.: PARMA: parallelization-aware run-time management forenergy-efficient many-core systems. IEEE Trans. Comput. **15**(21), 15–20 (2020)
12. Adil, O., Russell, J.L., Khan, W.U., et al.: Image-guided chest tube drainage in the management of chylothorax post cardiac surgery in children: a single-center case series. Pediatr. Radiol. **25**(02), 163–169 (2021)

Human Resource Scheduling Control Method Based on Deep Reinforcement Learning

Shaoping Zhang[(✉)] and Bo Sun

School of Labor Relations and Human Resource, China University of Labor Relations,
Beijing 100048, China
kaka5411000@126.com

Abstract. With the increasing importance of human resource management in project process management, human resource scheduling and control methods also need to keep pace with the times. Continuing to use the traditional human resource scheduling control method will waste a lot of potential value of human resources. Therefore, a human resource scheduling control method based on deep reinforcement learning is proposed. By constructing the human resource scheduling control model, the minimum human cost expenditure under various constraints is obtained; Design the deep reinforcement learning algorithm, and design the scheduling algorithm for specific scheduling objectives for the human resources scheduling control center; Create a model-based human resource scheduling management and control evaluation system, and improve the relationship between the comprehensive evaluation value and the advantages and disadvantages of multi project human resource scheduling. Experiments show that this human resource scheduling control method can guide different problems to allocate goals. Considering the time factor, the optimal solution of human resource scheduling and control can be obtained.

Keywords: Deep reinforcement learning · Human resources · Dispatching control

1 Introduction

Resource management generates a resource list for a single resource or a group of resources. Resource allocation strategies for various tasks can be configured, but the complexity of human resources is not considered. At present, scholars at home and abroad have many research schemes about human resource scheduling. These studies regard the human resource scheduling problem as a combinatorial optimization problem, and can roughly divide the methods to solve the human resource scheduling into three categories: heuristic algorithm, local search and optimization algorithm [1]. In recent years, scholars at home and abroad have expanded based on the above three solutions and developed more theoretical and practical results. In China, there are solutions to solve the human resource model according to the genetic algorithm to obtain the optimal solution, and there are also solutions to using the input scheduling strategy. The relevant

human resource models can obtain a probability distribution of the project cost, and the corresponding resource adjustment solution can be obtained by using the stochastic optimization. Or the human resource scheduling method based on process agent is used to model human resources, and the task and consortium bidding model are described and defined [2, 3].

The scheduling plan is generated according to the user's preference for the target. In foreign countries, one kind of modeling is based on dividing the software process into four stages: coding, rework, testing and project closure to construct the resource model and design the algorithm to solve the scheduling [4]. Another model scheduling strategy is to set personnel as fuzzy variables based on their professional skill level and uncertain activity cycle, and solve the scheduling for such a resource model. The time constraint grid model of task relationship is established, the model is modeled with mathematical formula, the relationship between tasks is judged according to the time constraint conditions, and finally the conclusion of whether resources conflict is obtained [5–7]. The resource conflict is detected based on the rule conflict of bit vector intersection operation. The algorithm is based on asbv algorithm and uses the divide and conquer idea and bit vector technology. Only one bit vector intersection operation is required for one-dimensional rule component processing, which greatly improves the processing efficiency [8].

Reinforcement learning, also known as reinforcement learning, is one of the paradigms and methodologies of machine learning. It is used to describe and solve the problem that agents maximize rewards or achieve specific goals through learning strategies in the process of interaction with the environment. The common model of reinforcement learning is the standard Markov decision process [9]. According to the given conditions, reinforcement learning is divided into model-based reinforcement learning (model-based RL) and model-free reinforcement learning (model-free RL). Reinforcement learning also includes more complex research directions, such as reverse reinforcement learning, hierarchical reinforcement learning and Multi-Agent Reinforcement Learning. The basic principle of reinforcement learning is to let the agent interact with the environment continuously, update the strategy by using interactive samples and feedback information, and use the strategy to finally obtain the optimal strategy [10, 11].

To sum up, this paper proposes a human resource scheduling control method based on deep reinforcement learning, which combines human resource scheduling with deep reinforcement learning to improve the management efficiency of enterprises.

2 Human Resource Scheduling Control Method Based on Deep Reinforcement Learning

2.1 Construction of Human Resource Scheduling Control Model

The purpose of human resource management is to improve production efficiency, reduce enterprise costs and increase profits. In order to achieve the established production indicators, it is necessary to assign manpower to each stage of task implementation, and allocate manpower scientifically and reasonably. The optimal allocation and effective utilization of manpower need to carry out targeted training to improve the ability of employees

to deal with various professional management work and improve their comprehensive quality. In terms of quantitative overall arrangement and application analysis of various resources, many industries have been using linear programming to realize the optimal allocation of limited resources such as human, financial and material resources in the economic management system. Among the research branches of operations research, linear programming has the characteristics of early exploration, rapid progress, easy understanding, universal application and mature algorithm research [12, 13].

As long as we engage in economic activities such as economic management, transportation, industrial and agricultural production, there will be a need to improve economic benefits. There are usually two ways to improve economic benefits: first, technological improvement. For example, select new materials, new equipment and improve the production process. Second, scientifically allocate human, financial and material resources and optimize production plans.

As a mathematical method to assist scientific management of practical problems, linear programming studies how to complete the optimal allocation of various resources and improve economic benefits according to the limitations of various conditions. Objective function, decision variables and constraints are the three elements of linear programming. Usually, the linear programming problem is to solve the maximum or minimum value of the objective function under linear constraints. The feasible solution is the solution that meets the linear constraints. A feasible region is a set containing all feasible solutions. Integer programming usually needs to round the set variables (part or all). If the variables in the linear model are limited to integers, the integer programming is called integer linear programming. Its general form is: list the objective function and constraints; Draw the feasible region (when there are many variables, it can be analyzed with the help of programming software). The model is established and solved within the feasible range to obtain the optimal value and optimal solution of the objective function. The calculation formula is as follows:

$$Maxz = \sum_{j=1}^{n} c_j x_j \tag{1}$$

where, z represents the optimal solution of the programming function; j represents linear variable of objective function; c represents integer linear programming vector; n represents the feasible region constant.

Establishing a mathematical model to solve practical problems usually has the following three steps: first, put forward assumptions, find decision variables, and determine them according to the factors affecting the purpose to be achieved; Determine the objective function, which is determined by the functional relationship between the decision variable and the goal; The constraints to be satisfied by setting the decision variables are determined by the constraints on the decision variables. Before the application and popularization of electronic computers, the solution and calculation of linear programming were quite complex. With the maturity of computer application technology, it is more and easier to use the powerful computer processing systems to solve linear problems. As a tool for solving linear and nonlinear optimization problems, LINGO software itself has a simple and easy to learn built-in modeling language. It can express complex logical relations with the help of various internal functions, making it possible to round decision

variables (i.e. integer programming), which is convenient to use and has high running speed. In terms of data interaction, the connection between the model and external data (such as Excel spreadsheet) can be realized by using input and output interface functions.

Enterprises adapt to the fierce competitive market, formulate real-time, accurate and reasonable personnel allocation schemes, and optimize the allocation of personnel, which all put forward practical requirements for the application of linear programming knowledge. In the past, the planning of enterprises was more complex. On the one hand, it was necessary to comprehensively consider the needs of customers of the construction party, on the other hand, it was also necessary to balance the profits of enterprises. Manual calculation was inaccurate and time-consuming. It was much easier to use linear programming and then use computers to complete the calculation and solution, which could improve the reliability and scientificity of enterprise decision-making. The decision-making theory is formed on the basis of strict theory, which can be obtained by analyzing and applying large-scale accurate basic data and strict mathematical solution. This model introduces the human capital theory, based on the concept of project management, divides the supervision projects of a certain scale, and realizes the multi project management under the restriction of human resources. Next, we use linear programming to obtain the optimal scheme for personnel scheduling from the aspects of the company's human resources arrangement in the cost standard of professional supervisors, project progress, project scale and customer needs, and the relationship between projects and supervisors, so as to achieve the goal of reducing the human cost of supervision projects.

In order to ensure that the model framework of this study is in line with practical application and scientific and reasonable, the following assumptions are put forward: due to the different supervision work contents and supervision contract requirements, different project supervision work needs to be carried out, and different supervision personnel need to be configured. The time is in months. The number of project personnel per month will change due to the progress plan of each project. It is necessary to use linear programming software to solve all kinds of supervisors required by the enterprise. The enterprise supervisors have the background of civil engineering, electromechanical or other housing and municipal public works, have learned relevant professional knowledge, and the work content is consistent with their major. For unrestricted work, all supervisors can exchange posts without obstacles, and there are only differences in work execution efficiency. It is necessary to assume that similar supervisors have the same monthly cost and work output value, calculate the monthly cost of various supervisors and the monthly output value of each type of supervisors respectively, and then summarize the total labor cost. The supervision work of this model does not consider the difference of project benefits and overtime factors. If it is necessary to refine this aspect, new constraints shall be added. If the parameters or constraints change due to irresistible factors, it is necessary to set new conditions to replace the human resource scheduling.The flow chart of human resource scheduling control model is shown in Fig. 1.
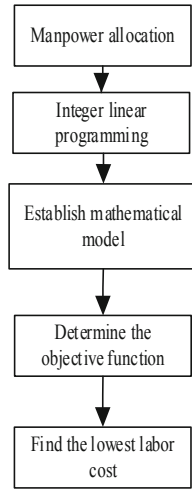
Fig. 1. Flow chart of human resource scheduling control model

The objective function expression of the human resource scheduling control model constructed in this paper is:

$$Minz = \sum_{j=1}^{n} c_j x_j z_j \tag{2}$$

$$WPij(i, j, k) = WPi(i, k) \times Ujk(j, k) \tag{3}$$

$$XPij(i, j, k) = WPij(i, j, k)/W_j(j) \tag{4}$$

$$Xij(i, j) = \sum_{k=1}^{e} XPij(i, j, k,) \tag{5}$$

where, i refers to the i month of planned work execution; j refers to the j supervisor; k refers to e projects in total in the k project, and e refers to the total number of supervision projects to be implemented within the planned time; Xij refers to the total number of supervisors of type j required for all projects in the i month; $XPij$ refers to the number of j supervisors required for the k project in the i month. The constraint expression for building the model is:

$$Xi(i) = \sum_{j=1}^{n} Xij(i, j) \tag{6}$$

$$Xj(j) = \sum_{i=1}^{n} Xij(i, j) \tag{7}$$

The purpose of this study is to find the minimum labor cost under various constraints. In order to carry out the supervision work of each project, different kinds of supervisors must be dispatched to carry out different supervision work under the condition of meeting various constraints.

2.2 Algorithm Design Based on Deep Reinforcement Learning

After the completion of the human resource scheduling control model constructed above, in terms of the agent, the required status information includes the agent type and whether the agent has selected the target. Here, considering the time factor, when the agent has selected the target, it is assumed that the movement of the agent has been completed. The agent's non motion attribute value is 1 and the already motion attribute value is 0. The status information required by the target includes the target type, target residual value and whether the target currently has obstacle areas. The target obstacle area attribute value is 1, and the non-existent attribute value is 0. In this way, there are 8 agents in total, each agent needs 2 values to represent the state, there are 9 targets in total, and each target needs 3 values to represent the state.

Considering the time factor, the definition of action selects the m-th target for the current n-th agent and moves, where $n \in [1, 8]$ and $m \in [1, 9]$, so the size of action space is 72.Because this action space is too large, the design here is divided into two actions. The space size of action 1 is 8, indicating which agent selects the target, and the space size of action 2 is 9, indicating which target the agent selects. In this way, the size of action space can be reduced from 72 dimensions to 17 dimensions, reducing the parameters of deep reinforcement learning, making the calculation of deep reinforcement learning network faster and making the design of action more meaningful [14]. After each action decision, the environment needs to use the path planning intelligent decision model for decision-making, calculate the arrival success rate according to the determined path, and then interact according to the environmental success rate. Therefore, the success rate and value coefficient cannot be directly used to calculate the score, and the real score under the current episode can only be calculated according to the real environment interaction. According to the success rate, when the agent can reach the specified goal, the reward is set to obtain a score multiplied by 100, and when it cannot reach the specified goal, the reward is set to - 10. Considering the time factor, there will be a special case, that is, the agent determined by the current action has selected the target and completed the motion, so the current action is meaningless. When the decision makes such an action, set the reward to - 10. In this way, actions are continuously selected according to the state, and each episode ends until all agents move.

Fitness function is the evaluation of individuals. An individual is a set of objective allocation solutions. Therefore, the final score obtained by the agent is directly used as the fitness function to evaluate the individual. The score is calculated by the target allocation scheme corresponding to the individual, the movement success rate of the agent and the value coefficient. Deep reinforcement learning combines the perception ability of deep learning with the decision-making ability of reinforcement learning, which can be controlled directly according to the input image. It is an artificial intelligence method closer to human thinking mode. It can solve more complex tasks that are closer to the current situation. It uses a depth network to represent the value function. According to

Q-learning in reinforcement learning, it provides the target value for the depth network and updates the network continuously until convergence. A fixed target network with better training stability and convergence.

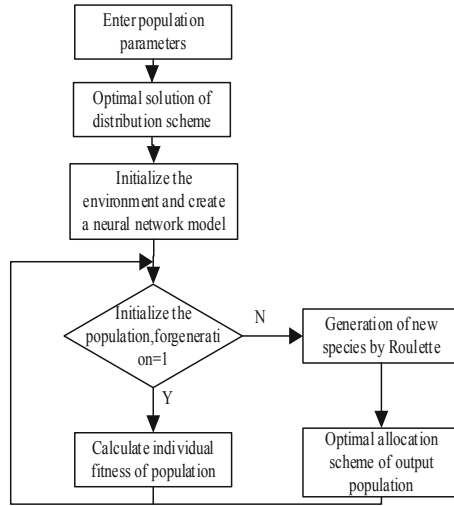The flow based on the deep reinforcement learning algorithm is shown in Fig. 2.

Fig. 2. Flow chart of deep reinforcement learning algorithm

As shown in Fig. 2, input population size, population size, crossover rate, mutation rate, genetic algebra, DNA size, disease and gene location boundary. Optimal solution of output allocation scheme; Initialize the environment, and create a neural network model according to the parameters of the environment initialization algorithm; Initialize the population pop, forGeneration = 1, translate the individuals in the population into allocation scheme, calculate the score in the environment according to the allocation scheme corresponding to the individuals as the fitness of the population individuals, and record the optimal individual fitness of the population; According to the fitness function, the roulette method is used to select and generate a new population pop, generate random numbers, and randomly select cross loci; Output the current population pop optimal individual corresponding allocation scheme. The policy trajectory of scheduling all job sets is obtained through the initialization model; Calculate the baseline at each time of all job sets; The difference between baseline and score function is used to update the parameter weight of the model. Next, the implementation details of imitation learning are described in combination with the code.

```
Time register level (void * ARG)
struct multiboot_ uinfo*mb=（struct multiboot_ Uinfo *) parameter;
EDF_ uregister_ ulevel（EDF_ Uenable all); / / level 0: EDF
CBS_ uregister_ ulevel（CBS_ Uenable all, 0); / / level 1: CBS
RR_ uegister_ ulevel（RRTICK，RR_ .MAIN_ Yes, MB); / / level 2:
loop
dummy_ uregister_ Ulevel(); / / Level 3: Virtual
Register module ();
1 / resource access protocol
CABS_ uregister_ umudule（）；
//Resource access protocol
Warning sound;
```

The multi project scheduling problem with limited human resources involves several parallel projects and a shared human resource database, which contains human resources with limited supply (a kind of renewable resources). Due to the complexity of several projects involved, it is assumed that the projects are independent of each other except for sharing the human resource database, there is no tight relationship between projects and between different project tasks, and there is no constraint relationship on other types of resources except human resources, that is, Competition for limited human resources is the only relationship between these parallel projects.

After the above implementation steps of in-depth reinforcement learning, the scheduling algorithm of specific scheduling objectives can be designed for the human resources scheduling control center.

2.3 Evaluation System of Human Resource Scheduling Management and Control Based on Model

The evaluation of multi project human resource scheduling based on the model refers to the evaluation of multi project human resource scheduling based on the model and suitable for the model. It is an evaluation matched with the model, which aims to find the situation of enterprise multi project human resource scheduling. It is not difficult to see that the evaluation is in the position of the supervisor of the implementation of the model.

The scientific principle of the evaluation index system means that the evaluation content should be scientific and standardized. The concept of each index should be scientific, accurate, with accurate connotation and extension, and the calculation scope should be clear without ambiguity; The index system should reflect the essential characteristics of the evaluation object as comprehensively and reasonably as possible. It is not a simple combination of indexes, but should reflect the overall effect and internal relationship of the project and the quantifiable degree of each index; The establishment of index system should minimize the subjectivity of evaluators and increase objectivity. Therefore, expert opinions should be widely solicited; The establishment of the index system must be guided by advanced scientific theory, which can reflect the objective and actual situation of the evaluation object. No matter what evaluation method is adopted

and what evaluation mathematical model is established, the index system must be an abstract description of objective reality.

The principle of system optimization requires that the number of indicators and the structural form of the index system should be based on the principle of comprehensively and systematically reflecting the evaluation objectives, and the evaluation index system should be established from the overall perspective. The indicator system shall be composed of several indicators. Indicators of different types in the system structure shall not be combined with each other, and the main indicators and accompanying indicators shall not be juxtaposed. Systematization first requires to avoid the index system being too complex to avoid the evaluation being difficult to implement, and to avoid too few indicators ignoring some important factors and difficult to reflect the internal essence of the project. Therefore, we should neither be all inclusive nor lose sight of one or the other. We should build a reasonable index system with fewer indicators as far as possible to achieve the purpose of optimizing the overall function of the index system. Secondly, it is required to take into account the relationship between current and long-term, overall and local, qualitative and quantitative.

The principle of comparability requires that the indicators must reflect the common attributes of the evaluated items, and the quantities of different indicators must be transformed into the same unit before they can be compared and calculated. The stronger the comparability, the more credible the evaluation results will be. In order to make the evaluation indicators universal, the designed evaluation indicators should adopt domestic and international standards or recognized concepts as far as possible. In the first mock exam, the factors that should be eliminated and the influence of environmental factors under specific conditions should be eliminated. The factors that can not be compared to the factors can be transformed into comparable factors, and the data of evaluation can be transformed into a unified equivalent value or dimensionless value, so that the evaluation indexes can be compatible with the same model and make them comparable.

The principle of practicability requires that the index system must have clear meaning, standardized data, moderate complexity and easy calculation. The requirements specified in the evaluation index shall conform to the actual situation of the evaluated object, that is, the specified requirements shall be appropriate, that is, they shall not be too high or too low. In order to facilitate practical use, a specific and measurable index system must be designed to characterize the main aspects of the goal. There should be enough information for the contents specified around the indicators, and there must be practical quantitative methods for use. The practicability also requires that the established indicators should have levels and key points, the qualitative indicators can be quantified, and the quantitative indicators can be measured directly, so as to make the evaluation work simple, save time and cost, and facilitate computer processing.

The independence of evaluation indexes means that the indexes in the evaluation index system should be independent of each other and can not be subordinate to and overlap with each other. This is because if the two indicators in the index system reflect the same attribute of the evaluated object, it will cause the repetition of the evaluation content, increase the workload of the evaluation, and even reduce the feasibility of the evaluation. Moreover, the repeated calculation of the content indicators twice is equal

to increasing the weight of the evaluation indicators of the project, which will naturally affect the scientificity of the whole evaluation.

The goal oriented principle means that the designed evaluation index system must fully reflect the evaluation objectives and fully reflect the basic principles based on the goal center, which requires that all indicators in the evaluation index system should be consistent with the objectives.

The core of multi project management is how to reasonably allocate various resources among various projects; The goal of multi project management is to solve the competition between multiple projects in terms of capital, time and resources by allocating enterprise resources and optimizing enterprise resource allocation, so as to reduce the project cost and improve the profit margin of the enterprise. The goal of multi project human resource scheduling in software development enterprises is to reasonably schedule all kinds of human resources among multiple projects to maximize the interests of the enterprise. In order to keep consistent with the model, the goal of maximizing enterprise interests is still adopted here.

From the four elements of modern project management, the goal of enterprise multi project human resources scheduling should cover three elements: quality, progress and cost, and satisfy all stakeholders. Therefore, the criterion layer of the evaluation index system is: quality, progress, cost and stakeholders, as shown in Fig. 3.
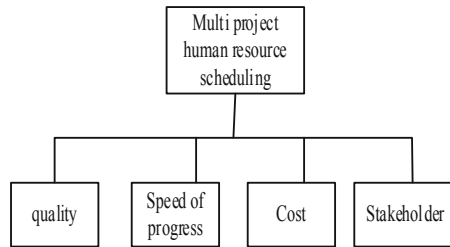

Fig. 3. Original index architecture

As shown in Fig. 3, it is the original index system structure established in this paper. The progress goal of the model established in this paper is the progress goal on the premise of ensuring the quality: the ranking of the relative importance of multiple projects makes the relationship between the main stakeholders (customers, teams, governments and enterprises) and some expenses (breach penalty) It is also included in the progress. Here, the indicators of multi project human resource scheduling evaluation of software development enterprises are simplified to the progress indicators that include the relevant interests of the enterprise (reflected in the different relative importance of multi projects). Comprehensively considering the relative importance of multiple projects and the model, the following evaluation scales are determined, as shown in Table 1.

Table 1. Evaluation items and corresponding symbols

Evaluation items	Corresponding symbol
Proportion of completed projects with weight coefficient of 9 to the total number of projects with weight coefficient of 9	N9
Proportion of completed projects with weight coefficient of 8 to the total number of projects with weight coefficient of 8	N8
Proportion of completed projects with weight coefficient of 7 to the total number of projects with weight coefficient of 7	N7
Proportion of completed projects with weight coefficient of 6 to the total number of projects with weight coefficient of 6	N6
Proportion of completed projects with weight coefficient of 5 to the total number of projects with weight coefficient of 5	N5
Proportion of completed projects with weight coefficient of 4 to the total number of projects with weight coefficient of 4	N4
Proportion of completed projects with weight coefficient of 3 to the total number of projects with weight coefficient of 3	N3
Proportion of completed projects with weight coefficient of 2 to the total number of projects with weight coefficient of 2	N2
Proportion of completed projects with weight coefficient of 1 to the total number of projects with weight coefficient of 1	N1

As shown in Table 1, each evaluation item is replaced by corresponding symbols, and each value corresponding to each evaluation item is the lower bound of the completion proportion range corresponding to the corresponding evaluation scale. After the evaluation scale of the evaluation project is determined, the weight of the evaluation project needs to be calculated. Here, the weight is calculated by the pair by pair comparison method. Like the determination of evaluation scale, the weight of evaluation items also has a certain subjectivity. The evaluation of comprehensive evaluation value is divided into two cases: one is the evaluation of single multi project comprehensive evaluation value, and the other is the evaluation of multiple multi project comprehensive evaluation values. The evaluation of single multi project comprehensive evaluation value refers to the evaluation of the scheduling results of a group of multi projects with human resource conflicts. The evaluation system is set as follows: if all projects are completed on schedule, the comprehensive evaluation value is 6.507. If 90% of all projects are completed on schedule, the comprehensive evaluation value is 5.508. If 80% of all projects are completed on schedule, the comprehensive evaluation value is 4.536. If 70% of all projects are completed on schedule, the comprehensive evaluation value is 3.591. If 60% of all projects are completed on schedule, the comprehensive evaluation value is 2.7.

The comprehensive evaluation value can be associated with the advantages and disadvantages of multi project scheduling. Due to the great difficulty of multi project practice

management, in order to encourage enterprises to adopt more suitable scheduling methods, 4 + 1 evaluation indexes are adopted here: excellent +, excellent, good, medium and poor, as shown in Table 2.

Table 2. Correlation between comprehensive evaluation value and multi project schedule

Serial number	Comprehensive evaluation range	Advantages and disadvantages of multi project team progress
1	5.508–6.507	Excellent +
2	4.536–5.508	Excellent
3	3.591–4.536	Good
4	2.7–3.591	Medium
5	0–2.7	Subalternation

As shown in Table 2, it is the correlation table between comprehensive evaluation value and multi project progress. The evaluation of multiple multi project comprehensive evaluation values can be completed by comparing their comprehensive evaluation values. The multi project human resource scheduling with large comprehensive evaluation value is better than that with small comprehensive evaluation value.

3   Experiment and Analysis

3.1   Experimental Preparation

In order to verify the effectiveness of the human resource scheduling and control method based on deep reinforcement learning, the following experimental operations are carried out. In this experiment, the depth reinforcement learning PPO algorithm is used, the attenuation rate is 0.99, the target KL divergence parameter is 0.01, game lambda is 0.97, and the shear objective function parameter is 0.2. The experimental algorithm model is divided into actor network and critical network. In this paper, a 4-layer fully connected neural network is used to build the actor network. The size of the state space of the input layer and the model is the same as 43 dimensions. The hidden layer has 2 layers and 64 neurons. The relu6 activation function is used. Because there are two actions in action modeling, the output layer of actor network has two outputs with dimensions of 8 and 9 respectively. Four layers of fully connected neural network are used to build a critical network. The input layer is the same as 43 dimensions, the hidden layer is the same as actor network, and the output layer is only 1 dimension. The output value estimates the value of the actions output by actor network. During network update, both actor network and critical network are optimized by Adam optimizer. The learning rate of actor network is 0.00002 and critical network is 0.00005. In this experiment, 10000 epochs are trained, and 1000 decisions are made in each epoch, so the batch size of the neural network is 1000. During the training, the reward sum of each episode and the loss function at each update are saved. After the trained intelligent decision model is

obtained, because the score expectation cannot be calculated directly, the environment is allowed to interact as much as possible and use the model for decision-making to calculate the score expectation of the decision model.

3.2   Result Analysis

The experiment in this paper adopts the deep reinforcement learning PPO algorithm, which does not plan according to the model, but trains the model through the samples generated by the environment. The score expectations of the deep reinforcement learning algorithm proposed in this paper are compared with those of ergodic method, genetic algorithm and microbial genetic algorithm. All the experimental results of target allocation are shown in Table 3.

Table 3. Comparison of experimental results of target allocation

Method	Is time taken into account	Score expectation
Ergodic method	No	5.3895411
Genetic algorithm	No	6.1582693
Microbial genetic algorithm	No	6.0214698
Deep reinforcement learning algorithm	Yes	4.1023458

As shown in Table 3, the following conclusions can be obtained by comparing and analyzing the experimental results of genetic algorithm and deep reinforcement learning. Genetic algorithm is very fast and can find the optimal solution of target allocation. However, the time factor is not considered, so there may be a better solution when the time factor is considered. Secondly, the genetic algorithm directly calculates the optimal solution according to the problem. When changing a problem, it needs to be recalculated, so it can not generate a general intelligent decision model. Deep reinforcement learning can save the decision model through neural network. When the state modeling includes location information, it can train a general intelligent decision model to guide different problems for target allocation. Considering the time factor, it can obtain a better solution, the number of samples required for training is less, the time is shorter, and the ideal effect is achieved.

In order to further verify the convergence of this method to human resource scheduling, test the performance of the algorithm, and the comparison results are shown in Table 4.

It can be seen from the above table that the number of iterations of this method is only 4 and 5, and the response time is 1.36 s and 2.45 s. Under different experimental conditions, this method can always respond in a short time. It shows that the optimization time of human resource scheduling in this method is short and the response ability is improved.

Table 4. Comparison of algorithm results

Method	Parameter performance	Experiment 1	Experiment 2
Ergodic method	Number of iterations	14	15
	Response time/s	3.56	12.45
Genetic algorithm	Number of iterations	11	14
	Response time/s	3.86	13.45
Microbial genetic algorithm	Number of iterations	17	18
	Response time/s	5.56	18.45
Deep reinforcement learning algorithm	Number of iterations	4	5
	Response time/s	1.36	2.45

4 Conclusion

The human resource scheduling and control method based on deep reinforcement learning proposed in this paper not only helps project managers identify the key human resources that determine the success or failure of the project, but also enables project managers to understand the overall ability level of existing human resources and formulate corresponding training and development plans. Build the human resource scheduling control model, set new conditions and replace the human resource scheduling. Considering the time factor and based on in-depth reinforcement learning, the scheduling algorithm with specific scheduling objectives can be designed for the human resource scheduling control center. When managing each link of optimal allocation of human resources in multi project environment, the more mature management model in each link is used for reference, which improves the efficiency and operability of management. In the evaluation of project priority, the index value established in this paper is limited, and there is no index of financial dimension. In the actual project management, capital is also an issue that must be considered. Therefore, in the future research, the index database will be continuously enriched to make the project priority management more perfect.

References

1. Otoo, F.: Human resource management (HRM) practices and organizational performance. Empl. Relat. **41**(5), 949–970 (2019)
2. Chen, J., Tong, S., Xie, H., et al.: Model and algorithm for human resource-constrained R&D program scheduling optimization. Discrete Dyn. Nat. Soc. **2019**, 1–13 (2019)
3. Hwang, K.J., Kang, P.K., Dong, J.J.: The Association between human resources investments in the internal accounting control system and unfaithful disclosure designations. Korean J. Manag. Account. Res. **20**(2), 55–78 (2020)
4. Imran, M.: Influence of human resources, information technology, internal control system and regional financial supervision on value of financial reporting information. Al-Kharaj J. Islamic Econ. Bus. **2**(1), 49–65 (2020)

5. Ali, M.C., Islam, K., Chung, S., et al.: A study of green human resources management (GHRM) and green creativity for human resources professionals. Int. J. Bus. Manag. Future **4**(2), 57–67 (2020)
6. Zaichenko, Y.O.: The current state of human resources management at enterprises: challenges and prospects. Bus. Inform **11**(514), 436–441 (2020)
7. Krpk, G.: The concept of employee's happiness in human resources management. Bus. Manag. Stud. Int. J. **8**(3), 2750–2775 (2020)
8. Paun, B., Raevi, I., Raketi, O.: IT solutions for human resources management. Serbian J. Eng. Manag. **4**(1), 15–21 (2019)
9. Parsehyan, B.G.: An overview of organizational behavior within the framework of human resources management in the art institutions. Turk. Online J. Des. Art Commun. **9**(2), 104–110 (2019)
10. Tahiri, A., Kovai, I., Krasniqi, A.: Human resource management, performance management and employee performance appraisal by SME managers in Kosovo. Int. J. Econ. Bus. Adm. **VIII**(Issue 4), pp. 288–298 (2020)
11. Huralska, V., Sharkova, H., Skrypnyk, N.: Adaptation and development of human resources as an instrument of increasing the competitiveness of the organization. Econ. Finances Law (6/2), 21–23 (2021)
12. 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
13. Liu, S., Bai, W., Liu, G., et al.: Parallel fractal compression method for big video data. Complexity **2018**, 2016976 (2018)
14. Zhao, N., Cheng, Y.Q., Liu, Z.H., et al.: Deep reinforcement learning-based channel intelligent access method and NS3 simulation. Comput. Simul. **385**, 5 (2021)

Research on Network Personal Data Vulnerability Detection Technology Based on Deep Learning

Lei Ma[1]([✉]) and Yunwei Li[2]

[1] Beijing Polytechnic, Beijing 100016, China
malei235@tom.com
[2] Beijing Youth Politics College, Beijing 100102, China

Abstract. Traditional data vulnerability detection technology has low detection accuracy in practical application. Therefore, this study designs a network personal data vulnerability detection technology based on deep learning. Firstly, the personal data set of network dynamic link library is established. In the data set, the vector characteristics of network personal data vulnerabilities are extracted by using deep learning network, and the vulnerability detection process is optimized. The test results show that this technology has high detection accuracy and low loss rate in the same data set. In the comparison of F value, this technology is also better than the traditional detection technology, which verifies that the network personal data vulnerability detection technology based on deep learning has high reliability in practical application.

Keywords: Deep learning · Network personal data · Data vulnerability detection · Vector features

1 Introduction

In the era of big data, the use of computers has greatly facilitated people's daily life and made people's communication more unimpeded. The utilization of online banking account, game account, credit card account and e-mail account has greatly improved people's quality of life, enabled people to enjoy the fun of life without leaving home, and the services of many network operators on the network have become richer [1]. But at the same time, it also brings opportunities for criminals.

As we all know, when registering an account, users need to fill in their own real information, which is often interrelated. It has great application value for both businesses and criminals, which enables some criminals to steal the user's account information through security holes, This will bring great security risks to the user's account security, resulting in the loss of user's account theft, credit card theft and so on.

The computer is the crystallization of the continuous development of human beings. The stability of the computer system is directly related to the user's experience of using the computer. However, it is precisely because the computer is invented by human beings

that the computer system will inevitably produce some security vulnerabilities, which will bring convenience to criminals, Many criminals will use the security vulnerabilities in the system to attack the user's computer, and then unknowingly obtain the user's personal information account accessed in the computer, which will undoubtedly pose a great threat to the user's personal information security [2]. At the same time, some criminals do not want to obtain illegal interests, and even attack the user's computer with the psychology of revenge, which will even implant viruses in the user's computer. These viruses will not only seriously affect the stability of the computer system, but also cause serious consequences such as computer system collapse and hardware burning.

Nowadays, in the era of big data, people no longer rely solely on computers to obtain data. The use of new intelligent terminal tools such as smart phones and industrial pads not only enriches people's daily life, but also lays a huge hidden danger for personal information security, Criminals or hackers often steal the user's personal information through security vulnerabilities in the intelligent terminal. For example, some software installation packages from unknown sources may contain various viruses. Once the user installs, these viruses will attack the intelligent terminal and steal the personal privacy stored in the intelligent terminal.

Against the above background, the relevant scholars have designed a web-based situational data mining leak detection technology, the technology by analyzing the data preparation, mining and evaluation shows three steps to do clustering processing data, provide help for vulnerability of data mining, the computation, and clustering results association mining method is applied to make tracking, clear scope of loopholes. Then design the overall structure framework of the technology, establish the detection database, and realize the design of the vulnerability detection technology from three aspects of static module, vulnerability scanning and page display module. Based on traditional research, this paper designs a network personal data vulnerability detection technology based on deep learning.

2 Detect Network Personal Data Vulnerabilities Based on Deep Learning

2.1 Establish Personal Data Set of Network Dynamic Link Library

Data set is a key and difficult point for personal data vulnerability mining based on deep learning. First of all, we often encounter the problem of unbalanced sample collection, because the number of vulnerable samples is much smaller than that of normal samples [3, 4]. Secondly, different types of data sets have different contributions to deep learning models. Data sets that perform well are often complex in the extraction process and expensive in time.

Aiming at some of the shortcomings of the current data set, this research proposes a data set with the granularity of assembly basic blocks and gives an idea of extraction. Aiming at the problem that there may be a large number of loops in the code block sequence, a descending loop method based on basic block clustering is proposed. Finally, a random oversampling method is proposed to alleviate the problem of unbalanced sample collection.

In the research of vulnerability mining based on machine learning for binary programs, the granularity of data sets used by predecessors includes program level, function call sequence level and function level. Different data sets will also have different effects on the performance of vulnerability mining models. The fine-grained data set is more conducive to the identification of the location of the vulnerability, and it also increases the difficulty of calculation and processing. In the process of establishing the dynamic data set of the personal link library, the format of each feature is set to "function name: parameter index = the parameter status at runtime" [5, 6]. First, store the declaration information of a large number of API functions locally, including information such as function names, parameter types, and function return value types. When a function breakpoint is captured in the process of tracing the program, the state type is subdivided according to the type of each parameter. If it is a pointer parameter, it is subdivided into a more precise pointer type through memory mapping information, such as Heap pointer, stack pointer, global pointer, etc.; if it is a data type parameter, the parameter x in the data set is represented by $Num32B_n$, and the value that meets the following conditions is required:

$$2^n \leq x < 2^{n+1} \tag{1}$$

After calculating the value of n in Eq. (1), it can be used as the basis and basis of vulnerability detection according to the actual situation.

2.2 Extract Vector Features of Network Personal Data Vulnerabilities

After feature vectorization, for a binary file, each word in the underlying language instruction becomes an N dimensional vector, and this binary file is composed of many N dimensional vectors [7–9]. However, the number of words contained in binary files of
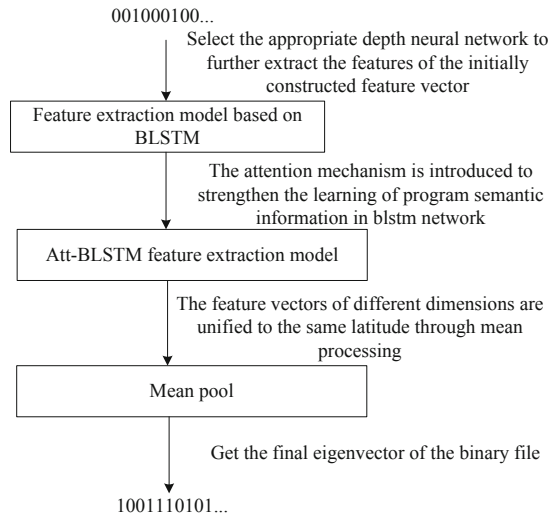

Fig. 1. Binary feature vector extraction process

different sizes is not the same, so the dimensionality of the feature vector representing each binary file is also different. Subsequent machine learning algorithms for vulnerability classification require that the feature vector of each sample has the same dimension, so this article divides the underlying language of the binary file according to the integer multiple of the sentence, and finally uses a mean pool to divide the characteristics of each binary file The vectors are unified into a fixed-size dimension.

The main solution to the problem is shown in Fig. 1.

Through the research on the construction of binary feature vector, this paper extracts the program information containing the context of words (in the underlying language instructions) from the binary file. Next, this paper selects the appropriate deep neural network to further extract the program information based on the context relationship between sentences (underlying language instructions).

With the segmentation of the underlying language, a binary file can be thought of as a text composed of serialized words. The process of extracting semantic information from such an article can be regarded as a natural language processing process. Therefore, the neural network suitable for natural language processing is also suitable for the vulnerability detection system in this paper. According to the characteristics of low-level language serialization, this paper first thinks of a cyclic neural network RNN that can learn sequence information [10]. However, RNN can not solve the problem of long-term dependence and can not remember the information of long-term sequence.

The propagation of vulnerability related data stream in the program may be a long process, especially the assembly code after binary disassembly is much larger than the source code, so a neural network that can remember long-term sequence information is needed. Therefore, this paper selects a variant of RNN that solves the long-term dependence problem, the long-term and short-term memory network LSTM, for feature extraction.

However, there is still a problem when using LSTM network. LSTM is a one-way network. It can well learn the front to back information in the sequence, but it can't learn

```
#include < stdio.h >
#include < ststring.h >

void    vulfunc(char*    str){
      char    src[10];
      strcpy(src, str);
}
int    main( ){
      char* str = "AAAAAAAAAAAAAAAAAAAAA"
      vulfunc(str);
      return;
}
```

Fig. 2. Stack overflow vulnerability

the back to front information. However, usually, the vulnerable part of a function may be affected by the previous program or the later program. Figure 2 shows a typical stack overflow vulnerability.

Figure 2 shows a typical UAF (Use After Free) vulnerability. The key to understanding this vulnerability is the system's memory management. After buf 1 is freed, the memory management will consider the address of buf 1 when it allocates buf 2 again. If the memory at the location is no longer useful, the memory will be allocated to buf 2, that is, the address of buf 2 points to the previous address of buf 1. The problem is that after the memory of buf 1 is released, no operation is performed to make the pointer of buf 1 NULL, which causes buf 1 to become a "wild pointer" and still point to the original memory address. In this way, if the pointer to buf 1 can be manipulated later, unexpected results can be produced.

The former is because there is no parameter check before calling the string copy function, and the latter is because there is no pointer nulling operation after calling the memory release function, which indicates that the forward and backward information of the program is related to the vulnerability, so It is not enough to use a one-way LSTM network. This article finally chooses to use a two-way LSTM network (BLSTM) for vulnerability feature extraction.

BLSTM is actually an extended structure of LSTM. Combining a forward LSTM and a backward LSTM network generates a BLSTM network. The BLSTM network can remember context information in both directions at the same time. Since the standard LSTM can only remember the above dependent information when processing the sequence, and the vulnerability may be related to the previous program or the subsequent program, it is necessary to add the following information of the program to the feature extraction training. BLSTM adds a reverse LSTM layer on the basis of the standard LSTM, and the information flow direction of the reverse LSTM layer is opposite to that of the standard LSTM layer. The schematic diagram of the BLSTM network is shown in Fig. 3.
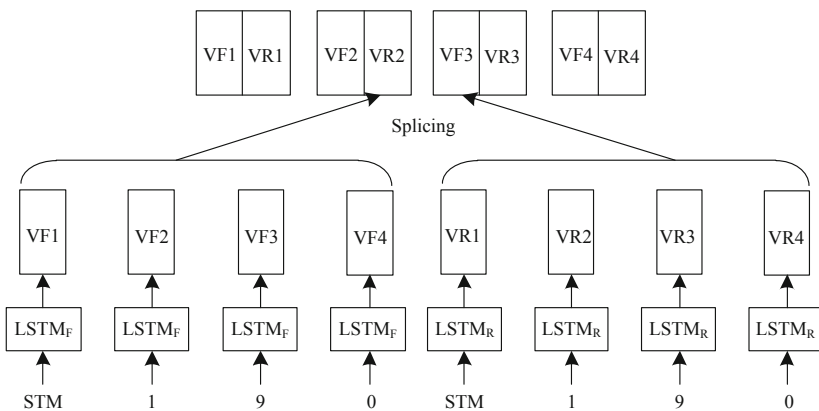

Fig. 3. BLSTM network principle diagram

Forward LSTM$_F$ input "STM", "1", "9", "0" in turn to get four vectors {$VF1$, $VF2$, $VF3$, $VF4$}. Input "0", "9", "1", and "STM" in the following LSTM$_R$ to get four vectors {$VR1$, $VR2$, $VR3$, $VR4$}. Finally, concatenate the forward and backward vectors to get:

$$\{[VF1, VR1], [VF2, VR2], [VF3, VR3], [VF4, VR4]\} \tag{2}$$

Which is:

$$\{v_0, v_1, v_2, v_3\} \tag{3}$$

In this paper, the training of the feature extraction model based on BLSTM is a supervised training. By training the BLSTM network to predict the next sequence of a long-term sequence in the underlying language to learn the context relationship between each assembly instruction of the underlying language, the purpose is to generate a representative The feature vector of each sample containing program semantic information. The feature extraction model diagram based on BLSTM is shown in Fig. 4.
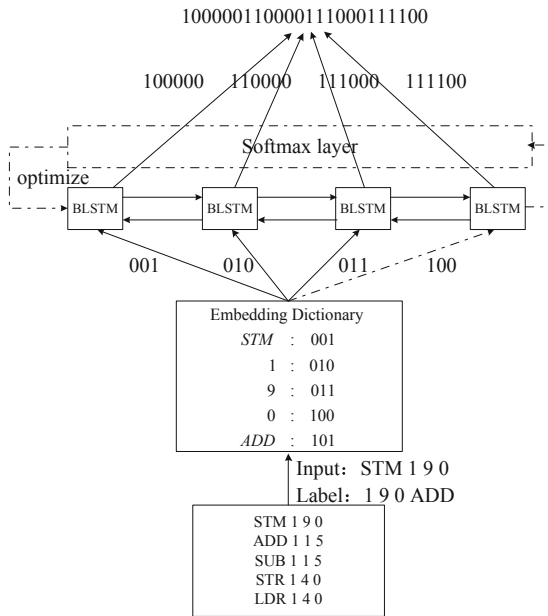

Fig. 4. The feature extraction model diagram based on BLSTM

Binary files of different sizes contain different numbers of words, so the dimensionality of the feature vector representing each binary file is also different. Some large binary files also contain a large number of words, which cannot be sent to the BLSTM network all at once. Learn. Therefore, this article first divides each binary sample. Each sentence of the underlying language contains 4 words, so this article divides each binary sample into pieces of data according to an integral multiple of 4, that is, in units of 4 × L words.

For each piece of data, $4 \times L$ is used as the window size of the sliding window, and the data obtained by sliding back one grid is used as the label of this piece of data. Then according to the word embedding dictionary trained based on the feature quantization model of the underlying language, the word vector of each word in each data is searched in sequence order, and then sent to the BLSTM network layer for learning. The word vector corresponding to the label is sent to the Softmax layer, and the Softmax layer compares the predicted sequence with the label, calculates the deviation and optimizes the BLSTM network. Finally, the BLSTM network will output the state vector of each word, and splice the state vectors of these words together, which is the feature vector extracted from this piece of data that contains the context relationship between the underlying language instructions.

2.3   Optimize the Vulnerability Detection Process

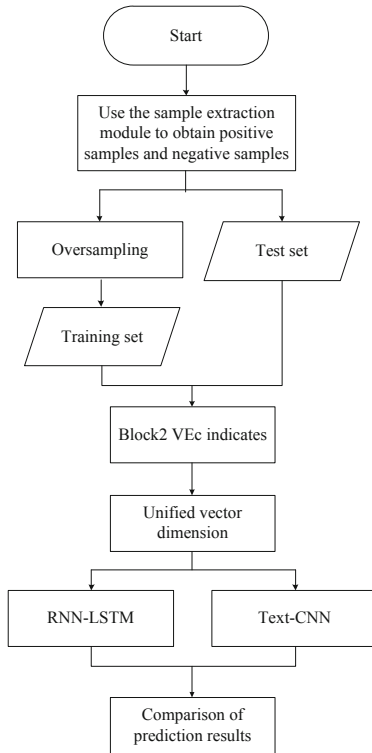The flow of the whole experiment is described from the perspective of the overall flow, as shown in Fig. 5.

Fig. 5. Flow chart of vulnerability mining based on code block

The whole process involves the acquisition, preprocessing and training of data sets. The core idea is to convert the program into code block sequence, and then into recognizable word vector. In this way, vulnerability mining can be carried out through machine learning. The specific scheme is as follows:

(1) Collect normal programs and vulnerability programs, obtain the program execution process through the tracking program, and obtain the assembly code block sequence of descending cycle through static analysis as the characteristics of corresponding samples;
(2) The training set and test set are divided, and the improved random oversampling method is used to expand the vulnerability samples in the training set;
(3) Use Doc2Vec to learn the word vectors in assembly language, to B1ock2Vec representation of the assembly code blocks in the sample, and to unify the vector dimensions of the sample;
(4) The RNN-LSTM model and the Text-CNN model are used for training respectively, and the two models are tested on the test set, and evaluation indicators are given.

3   Experimental Test

In order to verify the practical application performance of network personal data vulnerability detection technology based on deep learning, the following experiments are designed.

3.1   Deep Learning Model Training

In the training process, this article uses Google's deep learning library TensorFlow and deep learning top-level library Keras as tools to build deep learning vulnerability detection models. Keras is an independently developed deep learning library, which encapsulates commonly used neural network layers, simplifies the model building and training process, and lowers the threshold of deep learning. Keras is a high-level neural network API, written in pure Python and based on Tensorflow, Theano and CNTK backends. Keras was born to support speedy experiments, simple and fast prototyping. Keras has the characteristics of high modularity, minimalism and scalability, and supports CNN and RNN or a combination of the two. Due to the excellent performance of Keras, this study chooses Keras to construct and train the deep learning model.

Since vulnerability detection is a binary classification problem, this article uses binomial cross loss as the objective function:

$$L(Y, P(Y|X)) = -logP(Y|X) \tag{4}$$

In addition, this article uses an efficient Adam optimizer to optimize the model. In each training step, 32 small batches of sampled data are randomly selected from the data set to update the model, so that the training process can be stabilized. In addition, this article uses the glorot uniform function to initialize the weights of the convolutional layer and LSTM layer.

The experimental training process runs on a computer with 4G Hz Intel i5-8250U CPU and 8 GB memory. After each vulnerability detection model is constructed, in order to better analyze the performance of the model, this paper compares the detection performance of the traditional multi-layer perceptron (MLP) model. The multi-layer perceptron (MLP) model used in this study has three hidden layers, and the number of neurons in each hidden layer is 1024, 512, and 128, respectively. The number of specific parameters and output dimensions of each network layer of the multilayer perceptron model are shown in Table 1.

Table 1. The structure and parameters of the MLP model

Network layer	Output size	Number of parameters
Fully connected layer 1	(None, 1024)	208245
Discard layer 1	(None, 1024)	0
Fully connected layer 2	(None, 512)	548020
Discard layer 2	(None, 512)	0
Fully connected layer 3	(None, 128)	66645
Discard layer 3	(None, 128)	0
Fully connected layer 4	(None, 1)	192
Active layer	(None, 1)	0

The prediction accuracy curves in the training process of different models are shown in Fig. 6. It includes four deep learning models and a traditional multi-layer perceptron model.
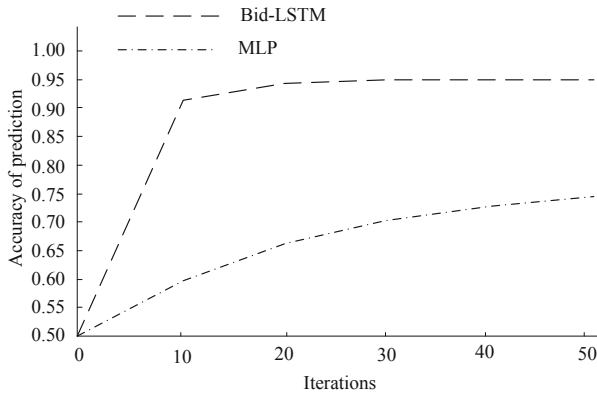

Fig. 6. The prediction accuracy curves of different models during the training process

In the first iteration, the accuracy curve of the deep learning model rises very fast. After training once on the training set, the prediction accuracy can reach more than 90%.

In the later iteration, the accuracy curve rises slowly and remains basically unchanged after the second iteration. Finally, the average prediction accuracy of the four deep learning models in this paper is about 90%.
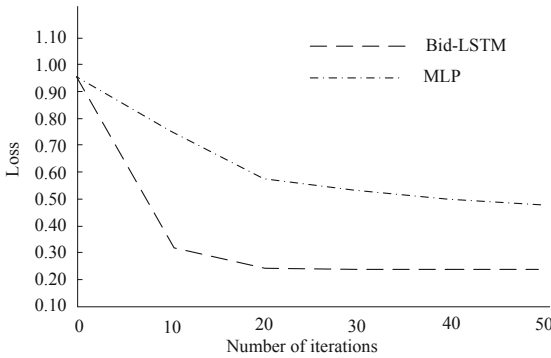

Fig. 7. Loss value curves of different models during training

Figure 7 shows the potential curves of loss function in the training process of different models. In this paper, the loss function bit of the deep learning model decreases significantly in the first iteration, reaching 0.2–0.3. In the following training process, the value curve of loss function decreases slowly and converges gradually, and finally stabilizes at about 0.18.

3.2   Test Results and Analysis

In order to avoid the uniformity of experimental results, the performance of the proposed method is compared with that of the traditional data vulnerability detection method based on network situation mining (Traditional method). The prediction accuracy and loss function values of the training set, verification set and test set in the whole training and testing stage are shown in Table 2.

Table 2. Experimental results of different detection technologies

Data set	Technology of this article	Traditional technology
Training set accuracy	1.0000	0.8261
Validation set accuracy	0.9865	0.746
Test set accuracy	0.9722	0.7366
Training set loss	0.2865	0.61
Validation set loss	0.4446	0.8349
Test set loss	0.5119	0.7068

According to the data in Table 2, after many iterations, the prediction accuracy of traditional detection technology finally reached 0.7366. Compared with the deep learning vulnerability detection technology proposed in this paper, the prediction accuracy is about 24% points lower, and the value of loss function is much higher than that of this method. According to the above experimental results, the detection accuracy of the traditional vulnerability detection technology in practical tasks is far lower than that of the technology in this paper. In conclusion, the application effect of the network personal data vulnerability detection technology based on deep learning designed in this paper is better.

In addition, the test data are divided into two categories: vulnerability (positive example) and no vulnerability (negative example), and the above vulnerability detection models are used to test the two types of test sets. The test results of the proposed technique and the traditional technique are shown in Table 3.

Table 3. Experimental results of different test data sets

Detection technology	There are loopholes (positive example)	No loopholes (counter-example)
Technology of this article	0.87	0.89
Traditional technology	0.72	0.59

According to Table 3, it can be seen that the detection results of the method in this paper are better than the traditional methods on the test sets with loopholes and without loopholes. In order to fully describe the performance of the model, it is not necessarily reliable to describe the detection accuracy and loss alone in the binary classification problem. Therefore, this paper calculates and compares the false positive rate (FPR, also known as false positive rate) of the model, in this paper Uniformly use "true and false, positive and negative" to express) and true rate (TPR, also called recall rate), and F-score (F-Score) as shown in Table 4, false positive rate (FPR) and true rate (TPR) It is a pair of elements that check and balance each other, and merge them into F-Score.

The specific calculation process of each result is as follows:

(1) True example (TP): The positive example predicted by the model to be positive, that is, the example whose label is positive and the predicted result is also positive.
(2) True Negative Example (TN): A negative example predicted by the model to be negative, that is, the label is negative and the predicted result is also negative.
(3) False positive example (FP): A negative example predicted by the model to be positive, that is, an example with a negative label and a positive prediction result.
(4) False Negative Example (FN): The positive example predicted to be negative by the model, that is, the example whose label is positive and the predicted result is negative.
(5) False positive rate (FPR): Solve by the above results, as follows:

$$FPR = FP/(TP + FP) \tag{5}$$

The number of negative sample results that are predicted to be positive accounts for the proportion of the number of results that are predicted to be positive. The lower the value, the better the performance of the model.

(6)   True rate (TPR): The calculation result is as follows:

$$TPR = TP/(TP + FN) \tag{6}$$

The number of positive sample results that are predicted to be positive/the actual number of positive samples. The higher the value, the better the performance of the model.

(7)   F-Score: As shown in formula 4.

$$F - Score = 2TP/(2TP + FP + FN) \tag{7}$$

Through the above calculation methods, the calculation results of different technologies are shown in Table 4.

Table 4. F-Score numerical calculation results

Detection method	False positive rate(FPR)	True rate(TPR)	F value(F-Score)
Technology of this article	0.1329	0.8956	0.8642
Traditional technology	0.4337	0.6551	0.635

As can be seen from Table 4, the false positive rate of the detection technology in this paper reaches 0.1329, and its true rate is also relatively high, reaching 0.8956. In the comparison of F values, the test results of the proposed technique are also better than the traditional technique, which highlights the application advantages of the proposed technique.

4   Conclusion

In this paper, a deep learning model is optimized for vulnerability detection tasks, and the network structure and hyperparameters of each layer of the model are designed. Through training on the collected 40000 data sets and finally testing on 5000 data sets, the deep learning vulnerability detection model can achieve high detection accuracy. Compared with the multi-layer perceptron model of traditional machine learning methods, the accuracy of this paper is greatly improved. By comparing the true rate, false positive rate and F value of different models on the test set, it can be seen that the F value of convolutional neural network model is the highest.

By analyzing the prediction accuracy curve and loss function value curve of the training process, it is found that the parameters of the deep learning model based on convolutional neural network are generally less, and the training speed is significantly faster than that based on long-term and short-term memory network. The comparison of the final performance of each model shows that the convolutional neural network model has the best comprehensive performance. To sum up, this paper has achieved effective results.

References

1. Shen, G.: Vulnerability of vulnerability defense control simulation of network system resource data. Comput. Simul. **37**(04), 308–311 (2020)
2. Wang, J., Liu, J., Ma, Y., et al.: An automated detection and verification method for webview component vulnerabilities. Trans. Beijing Inst. Technol. **40**(02), 169–174 (2020)
3. Gong, K., Zhou, Y., Ding, L., et al.: Vulnerability detection using bidirectional long short-term memory networks. Comput. Sci. **47**(05), 295–300 (2020)
4. Shao, S., Wang, M., Chen, D., et al.: Analysis of android application component exposure vulnerability based on machine learning. Trans. Beijing Inst. Technol. **39**(09), 974–977 (2019)
5. Li, Y., Feng, D.: Laser sensor network malicious code active detection system design. Laser J. **40**(06), 212–215 (2019)
6. Qi, L.: Mobile network hybrid security vulnerability detection simulation under static defense. Comput. Simul. **36**(04), 282–285 (2019)
7. Chen, B., Li, H., Li, B.: Application research on pseudo measurement modeling and AUKF in FDIAs identification of distribution network. Power Syst. Technol. **43**(09), 3226–3236 (2019)
8. Xia, Z., Yi, P., Yang, T.: Static vulnerability detection based on neural network and code similarity. Comput. Eng. **45**(12), 141–146 (2019)
9. Deng, Z., Lu, Y., Huang, Z, et al.: Network program vulnerability detection technology based on program modeling. J. Beijing Univ. Aeronaut. Astronautics Astronau **45**(04), 796–803
10. Li, Y., Cui, Y. ,Lv, J., et al:.Combined deep learning method for open source software vulnerability detection. Comput. Eng. Appli. (11), 52–59 (2019)

Competency Model of College Students' Innovation and Entrepreneurship from the Perspective of Deep Learning

Shuang Wang[1(✉)], Yunpeng Zheng[1], and Yang Li[2]

[1] College of Humanities and Information, Changchun University of Technology, Changchun 130122, China
`wangshuang012545@163.com`
[2] Guangzhou Huali College, Guangzhou 511325, China

Abstract. In order to avoid problems in the process of innovation and entrepreneurship, a competency model is created to help further development. Therefore, this paper constructs and designs the college students' innovation and entrepreneurship competency model from the perspective of deep learning. Set the goal of creating competency model under deep learning, build the competency model matrix under deep learning, and SDNN algorithm realizes the construction of competency model. The test results show that the normalized completion ratio of the designed deep learning competency application model test group is relatively high, indicating that the effect of this innovation and entrepreneurship competency model is relatively good, stable and comprehensive, and has practical application significance.

Keywords: Deep learning · College students' innovation and entrepreneurship · Competency model · Entrepreneurial competence · Model structure · Competency summary

1 Introduction

The rapid economic development urgently needs professionals with innovative thinking and entrepreneurial ability. Innovation and entrepreneurship has gradually become a new engine and new driving force for regional economic development. Colleges and universities undertake the important task of Cultivating College Students' innovation and entrepreneurship ability and talent incubation [1]. Cultivating innovative and entrepreneurial professionals in Colleges and universities is an important mission to actively support the adjustment of regional economic structure and complete the supply of regional talents [2]. For college students of professional courses in China, entrepreneurship is one of the possible choices of personal career [3]. By cultivating college students' entrepreneurial knowledge, entrepreneurial ability and entrepreneurial quality in professional courses, students can master the law of identifying, evaluating and utilizing entrepreneurial opportunities, which will not only help students broaden

their vertical career path, make rational choices between employment and entrepreneurship, and form a diversified and personalized career, It also helps students to plan the internal talent resources of the enterprise prospectively in the workplace, and stimulate the sustainable innovation vitality and entrepreneurial power of the enterprise. It is necessary to build the entrepreneurial competency model of college students in professional courses and improve the entrepreneurial competency of professional courses [4]. In terms of the current entrepreneurial practice of college students, with the support of national policies, college students' entrepreneurship has solved the employment problems of some college graduates and realized the combination of college education theory and practice. There are numerous successful cases of College Students' entrepreneurship, but compared with 20% of the average entrepreneurial success rate of international universities, The entrepreneurial success rate of college students in China is still at a very low level. The reasons are as follows: the imprisonment of traditional culture, the lack of support from parents, the backward employment concept of college students, the lack of entrepreneurial education and the low entrepreneurial competence of college students [5]. To some extent, the most fundamental reason affecting the success of entrepreneurial activities is the entrepreneurial competence of entrepreneurs, that is, the very low entrepreneurial success rate of college students in China lies in the low level of entrepreneurial competence of college students. Therefore, in the process of paying close attention to college students' entrepreneurship, we should focus on Improving College Students' entrepreneurial competence and strive to improve college students' entrepreneurial competence, Solve the problem of low entrepreneurial success rate of college students from the root [6].

For college students, innovation and entrepreneurship can help them know whether they have entrepreneurial competence and which entrepreneurial competence they have, so as to help them choose whether to start a business or to find employment. Relatively, the innovation and entrepreneurship model can, to a certain extent, help them rationally analyze their own competencies and enhance their comprehensive strength [7]. If you have entrepreneurial intention but lack entrepreneurial competence, which entrepreneurial competence should be improved? It can help students better plan their career and improve the success rate of entrepreneurship [8]. For the school, it can understand the competency characteristics required by college students in entrepreneurship, so as to determine what kind of entrepreneurship education plan needs to be made according to the specific situation [9], so as to help students improve their entrepreneurial competence and ultimately improve the success rate of entrepreneurship.

For the society, it is convenient to know what support and services are needed to improve the entrepreneurial competence of college students, so as to improve their entrepreneurial enthusiasm and reduce the obstacles in the entrepreneurial process of college students. However, in the process of practice, there are also some problems and obstacles, so it is necessary to add some information technology to assist. Deep learning is a widely used analysis and processing technology, which can quickly summarize and integrate huge data and complete target analysis at the same time. It is often applied in the construction of college students' innovation and entrepreneurship competency model [10]. In the model, entrepreneurial motivation, entrepreneurial characteristics and self-recognition are also included, which are the intrinsic and deep-seated

entrepreneurial potential of entrepreneurs, which cannot be easily changed by external influence and difficult to measure. Therefore, the innovation and entrepreneurship competency model of college students based on the perspective of deep learning is constructed. The competency model creation goal under deep learning was set up, the competency model matrix under deep learning was constructed, and the competency model was constructed by SDNN algorithm. In a more realistic environment, deep learning technology is associated with the competency model to improve the overall level of innovation and entrepreneurship.

2   Construction of Innovation and Entrepreneurship Competency Model from the Perspective of Deep Learning

2.1   Competency Model Creation and Goal Setting Under Deep Learning

The construction of college students' innovative entrepreneurial competency model, usually need to set the corresponding target, and combined with the corresponding requirements, combined with reaction time lenovo's psychological changes, can be smoothly to student's comprehensive ability test, aptitude test, personality test, etc., to a certain extent promote the development of individual psychology. The theory of individual difference mainly focuses on cognitive difference, personality difference and gender difference. Only with the corresponding knowledge, by differences in gender, age, reveal the difference of individual, can be gained by measuring the difference of the difference of individual intelligence, for example through a survey of personality can be concluded that the individual is more suitable jobs, can identify and evaluate leaders, managers, employees in the work of personal qualities, In this way, the initial goal of competence can be set.

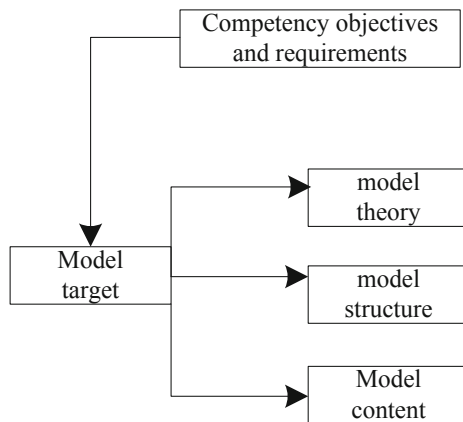First, you need to set the target hierarchy, as shown in Fig. 1:

Fig. 1. Hierarchical structure setting of objectives

According to the structural setting of the target in Fig. 1, the setting range of the target can be finally formed. On this basis, the leadership theory is added. Entrepreneurial competency is not only related to leadership, but also essential for leadership in enterprise management. The entrepreneurial process of entrepreneurs is also the process of leadership. The research on leadership and leadership theory in organizational behavior provides a theoretical reference for competency modeling. The theory of leadership characteristics, also known as the theory of leadership quality, focuses on the personal characteristics or personal quality of leaders in order to find, cultivate and use qualified leaders. The theory of leadership characteristics holds that the characteristics of leaders are not innate, but acquired. In the long-term study and hard practice, the characteristics of leaders gradually appear and finally become leaders. Foreign research on core competitiveness includes coordination and integration view, knowledge view, resource view, combination view and cultural view. Domestic research on core competitiveness includes resource view, ability view, asset and mechanism integration view, consumer surplus view, system view and culture and values. The research of foreign scholars focuses on the core competitiveness of enterprises, which is at the organizational level; Domestic researchers gradually focus on the individual level. The research on individual abilities and skills has aroused people's interest in competence. The core competitiveness plays an increasingly important role in the development of enterprises and individuals, and has attracted more and more attention.

When the research of competency develops to a certain stage, competency model comes into being. Competency model is composed of multiple projects, and each project has different dimensions. This paper holds that the competency model is the sum of competency characteristics, which are required for a specific post and role.

Behavioral event interview is the most effective method to build competency model. Its advantages include the following two aspects: first, it is an open questionnaire for the respondents, and requires the respondents to describe the causes and consequences of things in detail as much as possible, so the interviewers get more information and have higher authenticity; Secondly, the interviewees are required to describe key events during the interview, which are often the key factors affecting job performance. Behavioral event interview also has some limitations. First, it requires high professional knowledge and skills of model builders, and interviews need a large number of events and the cooperation of respondents; Secondly, it is a review of the respondents' past behavior, without considering that the enterprise is in a changing situation; Third, it requires respondents to describe key events, which may omit other relatively important responsibilities. Therefore, on this basis, the relevant data of College Students' innovation and entrepreneurship can be described and analyzed, and the actual comparison coefficient can be calculated, as shown in formula (1):

$$K = (2\alpha + 1) - m \tag{1}$$

In formula (1): K represents the actual contrast coefficient, α represents the comprehensive assignment, and m represents the actual action range of the model. Through the above calculation, the actual contrast coefficient can be obtained. Bring the innovation and entrepreneurship knowledge into the training goal of professional curriculum teaching systematically, and promote the reconstruction of professional curriculum system.

Organize the course content in the form of entrepreneurial projects, so that the design of entrepreneurial projects emphasizes practicality, is close to life and work, and reflects interest and operability. Taking tourism management as an example, the primary course focuses on the basic knowledge and skill training of College Students' Entrepreneurship and computer, aiming to popularize universal science and technology and cultivate students' thinking ability, practical ability and professional quality. Intermediate courses can introduce tourism marketing, tourism e-commerce and other projects with a certain professional foundation, in order to further improve students' knowledge level and practical ability. Advanced courses can be provided with comprehensive, innovative and practical work problem-solving courses such as tourism planning, tourism commodity development and design, tourism planning and development, so to improve students' ability of innovation and entrepreneurship and serve the industry. Based on the above, it is also necessary to set the change limit value of the target, as shown in formula (2):

$$M = \Im + \alpha \tag{2}$$

In formula (2): M represents the change limit value, $\Im$ represents the level change conversion value, and α represents the change error. Through the above calculation, the actual change limit value can be obtained. Take it as the change standard of the goal and set it in the initial model to complete the setting of the goal of creating the competency model.

2.2 Construction of Competency Model Matrix Under Deep Learning

For complex classification regression problems, the generalization ability is restricted. Deep learning attempts to break the constraints of shallow learning on levels. It can realize the approximation of complex functions by learning a deep-seated nonlinear network, showing a strong ability to learn the essential characteristics of data in a small number of samples [11]. Compared with the shallow learning structure, this kind of neural network has a deep structure, so it is called deep learning. It is pointed out that the deep learning structure can simply express the complex function mapping relationship. Through the review of deep learning, this paper analyzes the necessity of introducing deep learning, and points out that the shallow structure neural network has certain limitations in the ability of network to express complex functions, which can not represent high-dimensional complex functions, but can be expressed effectively by deep-seated neural network, At the same time, the multi-level features obtained by deep learning can be reused in similar or the same different classification predictions, which provides more useful information for the solution of the objective function of new tasks. It is pointed out that the high-level features learned in deep neural network do not change with the change of actual data, and have stronger robustness to practical problems. On this basis, the actual robustness can be calculated, as shown in formula (3):

$$P = \chi + s \tag{3}$$

In formula (3): P represents the actual robustness, χ represents the change value of comprehensive processing, and s represents the range of neural network. Deep learning still adopts the idea of neural network, but the middle layer contains multiple hidden

layers, which forms a multi-layer network. The deep learning model has more hidden layers, has stronger data fitting ability, and can mine the deep abstract meaning in the data. Deep neural technology, convolutional neural network, recursive neural network and other models are constantly proposed. Deep learning is one of the most cutting-edge fields of machine learning and statistical learning methods. Deep learning learns the features of each layer from the original data of College Students' innovation and entrepreneurship, and learns to express higher-level and more abstract features through layer by layer feature transformation, so as to learn the complex laws in high-dimensional complex data. Facing the challenge of today's big data, college students' innovation and entrepreneurship deep learning, as a cutting-edge method of machine learning, has been proved to significantly improve the effect compared with shallow learning.

In the above setting of hidden hierarchy, you also need to add DNN instruction. DNN is an advanced deep learning method. It is a nonlinear combination of multi-layer representation learning methods. Representation learning is a method to learn features from data and extract useful information from data during classification and prediction. Compared with shallow learning, DNN has better feature learning and prediction ability, especially in complex classification and regression problems, it can create a matrix of hierarchical model, as shown in the following formula:

$$P = \chi + \frac{\sqrt{R}}{3} - \Re \tag{4}$$

$$Y = 2\chi + \frac{\sqrt{R}}{3} - \Re \tag{5}$$

$$U = 3\chi + \frac{\sqrt{R}}{3} - \Re \tag{6}$$

In formulas (4), (5) and (6): P, Y and U represent the range of the normalized level, χ represents the unbalanced vector, R represents the characteristic coefficient, and $\Re$ represents the deep weight ratio. Through the above calculation, the actual normalization level range can be obtained. Combined with the DNN model, through multiple nonlinear fitting of the training data, it can mine the potential features in the data, and has a strong learning effect in the complex big data environment. The key to the calculation of DNN model is to minimize the loss function, and the weight of deep neural network can be trained by random gradient descent. SGD learns by updating the weight through error layer by layer back propagation, which can be divided into forward propagation stage from input layer to output layer and top-down error back conduction stage. Forward propagation calculates the activation function of each layer through the training data. During the error back propagation, the error between the actual output and the expected is calculated from the top down, and the weight of each layer is updated from the top down according to the gradient descent method. In the big data environment, deep learning can have strong learning ability for features, but there are still many problems in CTR prediction: there are many parameters, and parameter adjustment is a difficult problem. The number of layers of DNN and the adjustment of neuron nodes in each layer, that is, the setting of DNN model structure, is a difficult problem. To solve this problem,

the research idea of this paper is to determine the model structure first, then determine the key parameters, and carry out comparative experiments with various methods on the basis of optimal combination. Long training time. DNN is a powerful machine learning model, and the weight of the network can reach millions or more. It can learn the deep meaning of data features through the deep-seated deepening of the model. However, the deeper level means that the model needs to learn more weights, and the training time of DNN will be very long. Firstly, the resampling technology of unbalanced data is used to eliminate the impact of unbalanced characteristics on the prediction effect, and then the data features after resampling are deeply studied by DNN to calculate the feature depth neuron, as shown in formula (7):

$$q = \frac{\aleph + 1}{2} - \eta \tag{7}$$

In formula (7): q represents feature depth neurons, $\aleph$ represents the number of deep nodes, and η represents the number of mining systems. Through the above calculation, the actual characteristic depth neuron can be obtained. It is used as the implementation standard of the deep learning model. At the same time, DNN can mine the complex information that can not be simulated by shallow learning. Finally, two neural nodes are used as the prediction neurons of the overall model. The DNN model is constructed on the resampled data set that eliminates the influence, and the output of SDNN is obtained, that is, the prediction value probability and the corresponding prediction label. The core idea of SDNN is to construct a DNN model to eliminate the data imbalance, which can not only eliminate the impact of data imbalance on DNN, but also have the learning ability of high-level abstract features in DNN model, and finally complete the construction of competency model matrix under deep learning.

2.3 Implement the Competency Model Based on SDNN Algorithm

The construction experiment of deep classifier has been proved to be simple and easy to parallelize, but its disadvantages are also obvious. In the face of more and more complex massive data, the correlation between data features is also very complex. It is difficult for us to use manual feature selection and extraction to design an intelligent feature extraction method - deep learning, and study the construction of classification prediction model and the implementation of algorithm. Through the theoretical elaboration of SDNN model in the above, SDNN model will be constructed and SDNN algorithm will be discussed. SDNN model is different from DNN and can be considered as a lightweight balancing algorithm of DNN. SDNN will be more effective for data with unbalanced categories. The data predicted by CTR has been proved to have the imbalance of College Students' Entrepreneurship and innovation data. According to the practical problems of College Students' entrepreneurship, an improved algorithm of DNN - SDNN will be designed for the background of big data. Calculate the description range of the model and set the algorithm, as shown in the following formulas (8), (9) and (10):

$$E = \sqrt{2W + 1.25} - \kappa \tag{8}$$

$$G = \sqrt{4W + 1.25} - \kappa \tag{9}$$

$$D = \sqrt{6W + 1.25} - \kappa \tag{10}$$

In formulas (8), (9) and (10): E, G and D represent the description range of the model, W represents the execution instruction value of the model, and κ represents the deep learning protocol coefficient. The above calculation shows the flow of SDNN algorithm, and finally obtains the corresponding model processing results to complete the construction of the model.

3   Model Test

The test is mainly to verify and analyze the application effect of College Students' innovation and entrepreneurship competency model under deep learning. The test is divided into two models. One is the traditional regression competency application model, which is set as the test group of the traditional regression competency application model; The other group is the competency model designed in this paper, which is set as the deep learning competency application model test group. The two models are tested at the same time, and the results are compared and analyzed under the same test environment.

3.1   Test Preparation

Select school a as the main target of this model test, and summarize and integrate the innovation and entrepreneurship data information related to college students. Due to the imbalance of the manually summarized data, the SDNN model is constructed according to the redesigned experiment. After resampling, the label proportion of the training data is approximately equal to 1:1. The validation set is divided by 10% on the training set, and the parameters are optimized by cross validation. When the loss function value on the validation set does not change or the change range is very small, the iterative training process is terminated and the model training is completed. Since the data sampled each time are different, so the results of the prediction model are different, we calculate 20 times and take the average value for each method. And set corresponding DNN parameters, as shown in Table 1:

Table 1. DNN parameter setting

Parameter item	Parameter setting	Unit ratio
model structure	2022-1024-1024-800-2	1.25
objective function	Mean_squared_error	0.25
Maximum number of training iterations	200	2
Activation function	Relu	4.25

According to the data information in Table 1, the DNN parameters can be set finally. The depth neural network based on random undersampling of data O 25% of the samples were randomly sampled, and the sampled balanced data set resample_ The DNN model is established on data, and the parameter settings of DNN model are the same as those above.

Based on the training data, random undersampling is carried out, gbdt is used for feature selection, and then the DNN of m-500-500-2 structure is used to mainly represent the feature dimension after dimension reduction. The depth neural network classification prediction is carried out independently, and the parameter configuration of each stage is the same as the above setting in this paper. The parameter setting of gbdt has not changed. Table 1 shows the AUC of various DNNS and their improved methods. After random sampling, the AUC of DNN has been improved. It can be seen that repeated sampling of competency model is of great significance to eliminate the impact of data imbalance on prediction model, and category balanced data can improve prediction performance. Due to the large data scale of the innovation and entrepreneurship competency sample model, only random samples are compared and analyzed this time. In reality, the sampling method will consume computing memory and reduce the speed of the algorithm, so it is not practical. The sampling method based on the similarity between samples will produce a lot of computing work on the computer, which is time-consuming and inefficient. Therefore, the method considering random undersampling is more suitable for the real environment.

In addition, it is found that after the competency model in sampling is associated with the initial processing and analysis model of deep learning, DNN is better than gbdt's DNN, which proves that the deep learning model can mine high-level abstract features between features, and gbdt's feature selection is slightly underperformed because its shallow pattern can only mine shallow information. DNN has the performance of automatically mining feature meaning. Due to its hierarchical feature abstract expression ability, DNN shows strong feature mining ability, which is better than the prediction effect of shallow learning.

After completing the construction of the above model test environment, it is necessary to set the corresponding test equipment and corresponding index parameters. The experiment runs in a 64g memory, 2.25 GHz processor, the operating system is random, and the experimental programming tool is Python 2.6. In the construction experiment of CTR college students' innovation and entrepreneurship competency model with shallow learning, the experiments of logistic regression, decision tree and gbdt are completed based on Python's scikit learn machine learning library, and the experiments of neural network based on theano's keras library. The sampling algorithm in the CTR prediction experiment of deep learning is completed based on Python's balanced learning library, the design experiment of DNN is completed based on theano's keras deep learning framework, and the main experimental platform is GPU cluster. Complete the above test preparation, check whether the test equipment and environment are in stable operation state, and there are no external factors affecting the final test results. In order to ensure the fairness of the experiment, it is necessary to set the same experiment to swap in, and start the test after completion.

3.2 Test Process and Result Analysis

Through the above built model detection environment and test equipment, then start the test. Through exploratory factor analysis and confirmatory factor analysis, this paper determines the entrepreneurial competency model of college students. The survey objects of College Students' entrepreneurial competency are college students, but these college students also have more or less differences, such as gender, age, education and major. In order to further explore the impact of different groups of college students on the evaluation of College Students' entrepreneurial competency, this chapter tests the difference of the mean by taking the four factors of gender, age, education and major as independent variables and the five factors of College Students' entrepreneurial competency index as dependent variables.

In school a, 100 boys and 100 girls are selected as the test objects to test whether there are differences between men and women in the evaluation of College Students' entrepreneurial competence, and set the corresponding test standards, as shown in Table 2:

Table 2. Test standard setting table

Competency standard indicators	Mean difference	DF	Mean square deviation	Significant range	Change ratio
Conceptual power	0.25	4.2	5	1– 2.5	0.32
Opportunity power	0.34	4.16	4.5	1– 3.5	0.21
Organizational power	0.42	5.21	5.05	0.21– 4.25	2.1
Innovation	0.51	4.15	4.31	1– 0.68	0.11
Achievement ability	0.44	4.31	3	0	0.326
Guiding force	0.61	3.19	4.56	1.5–6.25	0.219
Comprehensive analytical power	0.65	4.2	5.65	2 –10.25	0.3

Complete the settings in Table 2 above, conduct actual tests, and compare and analyze the test results, as shown in Table 3:

Table 3. Comparison and analysis of test results

Test group	Normalized completion ratio of traditional regression competency application model test group	Normalized completion ratio of Deep Learning Competency application model test group
Test group 1	85.21	90.42
Test group 2	73.16	94.16
Test group 3	70.54	95.17
Test group 4	76.43	92.33
Test group 5	72.25	93.15

According to the data information in Table 3, the actual test results can be obtained, as follows: in the same test environment, compared with the traditional regression competency application model test group, the final normalized completion ratio of the deep learning competency application model test group designed in this paper is relatively high, indicating that the effect of this innovation and entrepreneurship competency model is relatively good, At the same time, it has stronger stability and comprehensiveness, and has practical application value.

4   Conclusion

To sum up, it is the design and analysis of College Students' innovation and entrepreneurship competency model from the perspective of deep learning. Through research, investigation and interview, this paper constructs the entrepreneurial competency model of college students, and the research hypothesis is verified. The competency model obtained in this study includes four competency characteristics: craftsman spirit, personality characteristics, market potential and entrepreneurial spirit. The weight of the four characteristics is determined. The results show that personality characteristics have the most far-reaching impact on College Students' entrepreneurial competency, followed by entrepreneurial spirit. Market potential and craftsman spirit need to be strengthened. This paper discusses the influencing factors of College Students' entrepreneurial competence. From the perspective of internal factors, gender, age and professional knowledge all have more or less influence on the exertion of College Students' entrepreneurial competence; From the perspective of external factors, region, family support, socio-economic status and school education have an impact on College Students' entrepreneurial competence, while family income has no impact on College Students' entrepreneurial competence. Improve the overall entrepreneurial quality level.

References

1. Xin, Y., Wang, C., Dong, Y., et al.: Management and entrepreneurship management mechanism of college students based on support vector machine algorithm. Comput. Intell. **25**(12), 1–13 (2020)

2. Hero, L.M., Lindfors, E.: Students' learning experience in a multidisciplinary innovation project. Educ. Train. **61**(4), 500–522 (2019)
3. Wu, Y., Wu, T., Li, Y.: Impact of using classroom response systems on students' entrepreneurship learning experience. Comput. Hum. Behav. **92**(5), 634–645 (2019)
4. Dong, D., Wang, X.: Human-computer system design of entrepreneurship education based on artificial intelligence and image feature retrieval. J. Intell. Fuzzy Syst. **39**(4), 5927–5939 (2020)
5. Zaheer, H., Breyer, Y., Dumay, J., et al.: Straight from the horse's mouth: founders' perspectives on achieving "traction" in digital start-ups. Comput. Hum. Behav. **95**(6), 262–274 (2019)
6. Jin, C., Luo, Y., Sun, H., et al.: Development of college students' innovation and entrepreneurship ability under the model of personalized education. Asian Agric. Res. **11**(3), 90–92 (2019)
7. Zhang, B., Yuan, H., Wang, Y.: Research on the training model of application-oriented undergraduates from the perspective of "innovation and entrepreneurship." Int. J. Inf. Educ. Technol. **9**(8), 584–588 (2019)
8. Zuo, L., Gong, M.: Research on the problems and countermeasures of the cultivation of college students' innovation and entrepreneurship quality. Open J. Soc. Sci. **8**(6), 261–266 (2020)
9. Wang, C., Zhang, Y., Moss, J.D., et al.: Multilevel factors affecting college students' perceived knowledge transferability: from the perspective of self-determination theory. Res. High. Educ. **61**(2), 1002–1026 (2020)
10. Yuan, Y., Liu, H., Yang, C., et al.: The influence of college students' character strengths on the recovery experience from the perspective of strengths use. Educ. Study **3**(1), 20–30 (2021)
11. Ge, M.Y., Yu, C.C., Zhou, L., et al.: Deep learning image classification algorithm based on semi-supervised collaboration training. Comput Simul. **36**(2), 206–210 (2019)

Research on Modeling of Adaptive Allocation of Labor Resources Based on Deep Reinforcement Learning

Bo Sun[✉] and Shaoping Zhang

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

Abstract. Based on the problem of the unreasonable allocation of labor resources in my country, a modeling method for adaptive allocation of labor resources based on deep reinforcement learning is proposed, combined with deep reinforcement learning algorithms to calculate the distortion of labor resource allocation in my country's primary, secondary, and tertiary industries degree. Analyzed the changing trend of labor resource allocation in urban and rural areas, and proposed an adaptive allocation plan of labor resources based on my country's industrial development structure in recent years to optimize the allocation structure of labor resources. Finally, it was confirmed by experiments that the adaptive allocation model of labor resources based on deep reinforcement learning It has high practicability and can better integrate the actual situation for effective allocation of labor resources.

Keywords: Deep reinforcement learning · Labor resources · Self-adaptation · Resource allocation

1 Introduction

The labor resource market in China is severely fragmented. Before the reform and opening up, it was mainly reflected in the segmentation of the urban and rural labor resource market [1]. After the reform and opening up, with the continuous development of the market economy and the flow of urban and rural labor resources, the segmentation of my country's urban and rural labor resource markets has weakened. Complex market segmentation. The market segmentation of labor resources directly hinders the effective flow of labor resources, causes distortions in the allocation of labor factors and resources, and hinders the healthy development of the economy. Therefore, paying attention to the segmentation of China's labor resource market, the distortion of labor factor allocation, and the impact on economic development is the meaning of the question [2]. Throughout the existing relevant literature, it has conducted a certain research on the state of the labor resource market segmentation in my country, the test of the existence of the labor resource market segmentation, and the impact of labor resource allocation on total factor

productivity, and the research on the degree of distortion of labor resource allocation in my country Relatively weak, it is generally only involved in the overall research on the distorted variables of labor resource allocation, and simple estimates are used to measure and replace them. There is a lack of systematic research. Therefore, it is necessary to explore the degree of distorted labor factor allocation in my country. Therefore, deep reinforcement learning is proposed. Modeling method for adaptive allocation of labor resources.

This paper firstly analyzes the labor resource allocation mechanism, adopts the labor resource allocation distortion algorithm based on deep reinforcement learning to allocate labor resources adaptively, and establishes a theoretical model of deep reinforcement learning algorithm with reference to the deep reinforcement learning algorithm method to allocate capital and labor. The static analysis of the situation is carried out, and the adaptive allocation model of labor resources is constructed.

2   Modeling of Adaptive Allocation of Labor Resources

2.1   Labor Resource Allocation Mechanism

The macro daily standard for the operation of the labor resource allocation mechanism is an indicator system, not a single indicator. The optimal allocation of human resources is to balance the total amount of human resources production and use, the production structure of human resources is consistent with the demand structure, and the collocation of human resources is reasonable. In order to better guarantee the rationality of labor resource allocation, first, a comprehensive analysis of population resources, human resources, labor resources, and the relationship between people and people is carried out, as shown in Fig. 1:
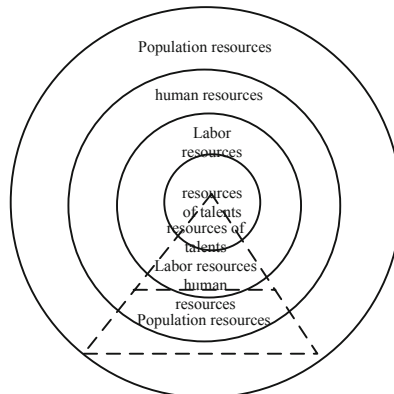

Fig. 1. Population resources, human resources, labor resources, and human relations

Based on the research of the labor resource allocation mechanism of deep reinforce-
ment learning based on the above figure, some people in the theoretical circle believe that
only economic benefits can be used as the operating goal of the labor allocation mech-
anism, ignoring the positive significance of full employment [3, 4]. Under the socialist
planned commodity economy, the operating goal of the labor resource mechanism is not
a single index, but an index system composed of economic benefits and full employ-
ment. Full employment does not necessarily exclude economic benefits, there are many
combinations between them. Figure 2 shows the L-S relationship between economic
benefits and full employment:
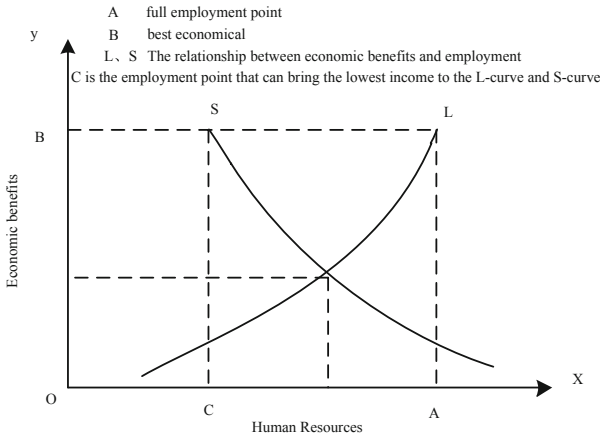
Fig. 2. The L-S relationship curve between economic benefits and human resources

In the figure, the x-axis represents human resources, the y-axis represents economic
benefits, point A is the full employment point, point B represents the best economic
benefit, and the L and S curves are the relationship between economic benefit and
employment under the conditions of two different labor resource allocation mechanisms.
For the combination curve, the point C is the employment point that can bring the lowest
benefit to the L curve, and the employment point that can bring the most benefit to the
S curve. The figure shows that there are countless combinations of employment and
economic benefits under each labor allocation mechanism. The best economic benefits
may be realized under the condition of less employment, and may also be realized under
the condition of full employment [5].

Society and enterprises also play an important role in the optimal allocation of labor
resources. The change of employment concepts, the emancipation of the mind, the cre-
ation of a cultural atmosphere, the emphasis on skills and knowledge by enterprises,
and the adoption of social organizations, non-profit organizations and even for-profit
organizations in society Provide employment services, training and recruitment con-
sulting activities, effectively increase employment services, and promote the optimal
allocation of labor resources during the transition period [6]. From the analysis below, it
can be seen that population management, employment services, training and education,
technological innovation, etc. also affect the allocation of labor resources. Government

departments are constantly exploring new management methods and macro policies in accordance with the needs of economic development and the trend of population development. Two-way interaction, making full use of the favorable factors provided by internal and external conditions to promote the optimal allocation of labor resources [7]. The specific influencing factors of the optimal allocation of labor resources are specifically discussed from several aspects in Fig. 3.
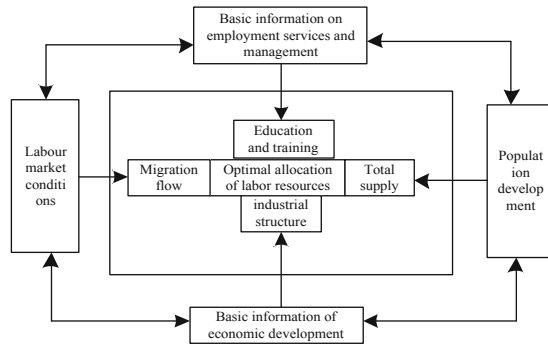

Fig. 3. Micro-influencing factors of optimal allocation of labor resources during the transition period

Economic benefits and employment may be positively correlated or negatively correlated. The key depends on the choice of labor resource allocation mechanism. The specific factors that affect this choice include: accumulation rate, current economic structure and target economic structure, total population and population composition, employment mechanism operation mode and adjustment system. If the employment operation mechanism is temporarily withdrawn, then to achieve higher economic benefits and full employment, it is necessary to choose the accumulation rate and target economic structure suitable for employment according to the specific situation [8].

2.2 The Distortion Degree Algorithm of Labor Resource Allocation Based on Deep Reinforcement Learning

The deep reinforcement learning algorithm is used to calculate the difference between the marginal output of labor and the marginal cost to measure the efficiency of labor resource allocation at the enterprise level. Use the ratio of the difference between the marginal output of labor and the marginal cost to the marginal cost, namely:

$$MISA_{it} = \frac{|MPL_{it} - B_{it}|}{B_{it}} \qquad (1)$$

Among them, MPL_{it} is the marginal output of labor employed by firm i in year t, and B_{it} is the marginal cost of labor hired by firm i in year t, where the sum of the average wages and welfare expenditures paid by the enterprise to individuals is substituted [9]. When the resource allocation factors are in the optimal state, the marginal output of labor

equals the marginal cost. When the relative difference between the marginal output of labor and the marginal labor cost of an enterprise is greater, that is, the greater the MSA, the higher the level of labor resource misallocation and the lower the efficiency of resource allocation. Therefore, in this article, we mainly focus on the deep reinforcement learning estimation model.

First, we must combine the deep reinforcement learning algorithm to calculate the marginal labor output m of enterprise i in the m-th year. Here, suppose that the production function of the company under the deep reinforcement learning algorithm is in the form of Cob-douglas, and the logarithm on both sides is taken to obtain methods such as deep reinforcement learning, and the theoretical model of deep reinforcement learning algorithm is established to statically analyze the allocation of capital and labor., That is, use the classic Solow model to measure China's TFP, and calculate γ as the balance from the known stock of labor resources and capital and the corresponding output level. Use i to represent a specific area, that is, there are m areas, $i = 1, 2, 3, \ldots, m$ respectively. To facilitate analysis, this article assumes that all different provinces have the same output elasticity. The specific output formula is:

$$Y_m = A_m K_m^{\alpha_n} L_m^{\beta_n}, \, Y_n = A_n K_n^{\alpha_n} L_n^{\beta_e} \tag{2}$$

Among them, the regression estimation coefficient of A_m is β_m. If β_m is positive, it means that the reduction of intermediate goods tariff L (increased trade liberalization of intermediate goods) reduces the mismatch level of labor resources β_n of the enterprise and improves the efficiency of labor allocation $K_n^{\alpha_n}$, otherwise it reduces the efficiency of labor resource allocation of the enterprise $A.K_n^{\alpha_n}$ In order to obtain the labor resource allocation efficiency of enterprise a in the nth year.

$$\beta = \sum_i \beta_i \frac{\gamma_i}{Y} \tag{3}$$

β_i represents the labor output elasticity of department i, and γ_i represents the output of department i. The degree of distortion of department m salary w_m with department n salary w_n as the reference frame [10]. Based on the above algorithm, the distorted degree of labor resource allocation in the entire society is decomposed into the distortion contribution of the internal labor resource market within each industry and the distorted contribution of inter-industry allocation. Similarly, the degree of distortion in the allocation of labor resources in urban and rural areas and cities can be decomposed into the distortion of the internal labor resource market within the industry and the distortion of the allocation between industries, which will not be repeated here.

2.3   Construction of an Adaptive Allocation Model of Labor Resources

Build an open regional labor resource optimal allocation model, analyze the changes in the labor market supply and demand, so as to realize the effective allocation of human resources.

With reference to the deep reinforcement learning algorithm method, the theoretical model of deep reinforcement learning algorithm is established, and the allocation of capital and labor is statically analyzed. That is, the classic Solow model is used to

measure China's TFP, which is based on the known labor resources and capital stock. And the corresponding output level is calculated as L_i^a as the margin. Use i to represent a specific province, that is, there are m provinces, i = 1, 2, 3, ..., m. For ease of analysis, this article assumes that all different provinces have the same output elasticity, and the specific output formula is

$$Y_i = A_i L_i^a K_i^{(1-a)}, 0 < a < 1 \tag{4}$$

Among them: A_i and $K_i^{(1-a)}$ respectively represent the output GDP, total labor resource status, capital stock, labor output share and TFP of province i, and the values of variables ω_i and $1 - \sigma$ are real values. Because it is necessary to measure the TFP of a specific province, for the convenience of analysis, it is assumed that the total output of each province satisfies the constant substitution elastic function between the total output and the total output, namely:

$$Y = \left(\sum_m^{i=1} \omega_i Y_i^{1-\sigma} \right)^{\frac{1}{1-\sigma}} \tag{5}$$

Among them: σ is the substitution elasticity of different provinces, and σ is greater than zero; ω_i is the ratio of the output of province i to the total output, and A_i^* and K_i represent the output level GDP of province i respectively. Capital and labor, which are factors of production, satisfy the condition that the sum of the production factors of each province is equal to the total production factors. The specific formula is:

$$\begin{cases} \Sigma_i L_i = L \\ \sum_i K_i = K \end{cases} \tag{6}$$

Among them, L_i, K_i, L and K represent the labor and capital stock levels of i province and respectively. The resource configuration that meets the following conditions is called effective resource configuration:

$$\begin{cases} \max_{L_i,K_i} Y \dfrac{L_i}{L} = \dfrac{K_i}{K} = \pi_i \\[2mm] \pi_i = \dfrac{\omega_i^{\frac{1}{\sigma}} \left(A_i^* \right)^{\frac{1-\sigma}{\sigma}}}{\sum_m^{i=1} \omega_i^{\frac{1}{\sigma}} \left(A_i^* \right)^{\frac{1-\sigma}{\sigma}}} \\[4mm] A^* = \left[\sum_m^{i=1} \omega_i^{\frac{1}{\sigma}} \left(A_i^* \right)^{\frac{1-\sigma}{\sigma}} \right]^{\frac{\sigma}{1-\sigma}} \end{cases} \tag{7}$$

The industrial structure is an important factor that affects the changes in the structure of labor resource allocation. When the level of economic development is at different stages, the industrial structure and employment structure will also be at the corresponding stage. Although the changes in the employment structure lag behind the changes in the industrial structure, the overall trend of the two shows a consistent trend [11–13].

Changes in the industrial structure are the precedent for changes in the employment structure, and adjustments in the industrial structure can lead to continuous optimization of the distribution structure of labor resources among industries. The high-end industrial structure can drive the transfer of workers from low-end industries to high-end industries, and can drive the improvement of the quality of labor resources. The optimal allocation model of labor resources in this paper is

$$S = f(R, E, C, J, A, T, Q, M, P, Z, L, D) \tag{8}$$

Among them: R is a comprehensive factor (factors not listed in the model), E is economic growth, C is urban planning, J is employment policy, A is investment growth, T is technological innovation, Q is reform and opening up, and M is market economy. P is population migration, Z is comprehensive evaluation, L is industrial structure, and D is talent policy. The research on the optimal allocation of labor resources needs to be carried out in the context of economic and social transformation, which is the key point to enhance the significance of the research. The labor resource market, the capital market, and the transformation of government functions are all formed and perfected during the economic and social transformation. Optimizing the allocation of labor resources has the responsibility of the government, and it is inseparable from the role of the market. If the optimal allocation of labor resources is regarded as a system, then the optimal allocation of labor resources is inseparable from the coordination of its own system, and it is also inseparable from the coordination of the external environment of the system and the communication between the internal and external environments of the system [14]. The external environment plays an important role in the optimal allocation of labor resources. Internal and external factors such as economic growth, opening to the outside world, talent policies, infrastructure, employment policies, population policies, technological innovation capabilities, government organization coordination capabilities, and industrial structure are all to a certain extent. The above restricts the advancement of the optimal allocation of labor resources. Based on this, the framework of the optimal allocation of labor resources in the open area is further analyzed as shown in Fig. 4:

In Fig. 4, the influence of population on the quantitative relationship of human resources is mainly manifested in its influence on ordinary labor resources—labor resources. Because ordinary labor resources are the main manifestation of the number of human resources and constitute the main part of human resources. Special labor resources, such as various talents, are a special part of human resources. Their knowledge and talents can only be obtained through long-term accumulation and specialized education and training, which is reflected in the further changes in labor resources. Therefore, changes in the supply of human resources are mainly reflected in changes in labor resources. The supply of labor resources is a major variable that affects the changes in the relationship between supply and demand in the labor market. The supply of labor resources includes actual labor resources and potential labor resources. In this way, the effective allocation of human resources can be realized.
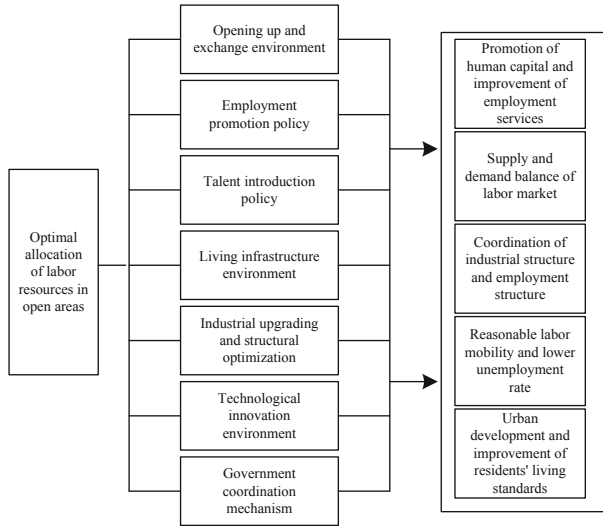
Fig. 4. Model framework for optimal allocation of labor resources in an open area

3   Analysis of Experimental Results

The human resource allocation in a province's statistical yearbook is used to affect the data of various variables, and after standardized collation, it is brought into the model for verification. Among them, human capital is expressed by the weighted product of the laborer's education level, average life expectancy, and skill level; the industrial structure is expressed by the proportion of the tertiary industry; the degree of marketization is expressed by the proportion of the number of employees in non-state-owned enterprises in the total employees. The specific test results are as follows (Table 1).

Table 1. Departure model test of human resource allocation in industrial structure

Classification	Coefficient	Std. error	T-Statistic	Prob.
Human capital	0.039	0.009	9.856	0.000
Industrial structure	0.089	0.277	-7.757	0.000
Degree of marketization	0.356	0.019	5.786	0.001
The degree of distortion in the allocation of rural labor resources	0.209	0.227	13.187	0.000

Table 2 reports the distortion degree of overall labor resource allocation, the degree of distortion of labor resource allocation in urban and rural areas, and the degree of distortion of labor resource allocation in cities in representative years. A further comparison of the overall distortion of the allocation of labor resources, the distortion of the allocation of labor resources in urban and rural areas, and the degree of distortion of the allocation

of labor resources in cities shows that the degree of distortion in the allocation of labor resources in my country's urban and rural areas is relatively serious, which is the main factor leading to the distortion of the overall allocation of labor resources; but this is not the case. At the same time, the distortion of the allocation of labor resources in our cities should also arouse great attention, and its influence on the degree of distortion of the allocation of labor resources will be greater and greater.

Table 2. The degree of distortion of overall labor resource allocation, the degree of distortion of labor resources in urban and rural areas, and the degree of distortion of labor resources in cities

Particular year	2017	2018	2019	2020
population	0.653	0.573	0.467	0.469
urban and rural	0.551	0.416	0.385	0.398
Proportion of urban and rural areas (%)	84.321	72.580	82.451	78.961
Within the city	0.103	0.158	0.083	0.100
Proportion in city (%)	15.680	27.420	17.550	21.310

Table 3 reports the decomposition factors of the distorted degree of overall labor resource allocation. The distorted decomposition of labor resource allocation in urban and rural areas and cities are similar and will not be reported. It can be seen from the table that the difference in the marginal productivity of labor between departments directly reflects the distortion of the allocation of labor resources, and has become the main factor of the distortion of the allocation. From an internal perspective of each industry, the wages of labor resources are not determined by the marginal productivity of labor. To varying degrees, it reflects the characteristics of the internal labor resource market in the industry. The reflected labor resource allocation is distorted. For example, in the representative years of the primary industry, the average wage is actually higher than the marginal productivity of labor, while in the secondary industry, the wage is actually lower than the marginal productivity of labor. To a certain extent, the amount of free flow of labor resources between different departments caused by the marginal productivity difference of labor is reduced, and the degree of distortion in the allocation of labor resources is also reduced.

Furthermore, various distortions of labor resource allocation considering the human capital situation are given. Compared with ignoring human capital, when considering human capital, the degree of distortion of overall labor resource allocation, urban and rural labor resource allocation, and labor resource allocation in cities has all decreased. The human capital level of the secondary industry is higher than that of the tertiary industry, and the human capital level of the secondary industry is significantly higher than that of the primary industry, which means that the secondary industry actually uses more labor resources. The primary industry actually uses more labor resources. Fewer labor resources, therefore, the degree of distortion in the allocation of labor resources decreases. Labor resources chase capital, and investment can effectively change the existing industrial structure. Generally speaking, the labor resource allocation structure

Table 3. Decomposition of contributing factors to the degree of distortion of overall labor resource allocation

Particular year	2017	2018	2019	2020
Distorted contribution in the primary industry	−2.085	−0.588	−1.049	−1.068
Proportion of primary industry (%)	−43.521	102.760	−225.071	−228.341
Distorted contribution in the secondary industry	−0.903	−0.777	−0.578	−0.555
Proportion of secondary industry (%)	−138.231	−165.800	−123.891	−118.211
Distorted contribution in tertiary industry	0.0002	−0.071	−0.449	−0.490
Proportion of tertiary industry (%)	0.030	−12.251	−96.411	−104.660
Intersectoral distorted contribution	1.839	2.009	2.539	2.582
Proportion of departments (%)	281.730	350.830	545.390	551.220

is subordinate to the industrial structure. From a regional perspective, due to different development stages, the efficiency of resource allocation in different urban markets in the province is also different, which can be clearly seen in the data in the table. In order to better understand the impact of the distortion of the allocation of production factors in the capital and labor market on TFP, the more efficient resource allocation is selected as a reference. Using the model in this paper and the data used by Garofalo, etc., the distortion of the allocation of resources in the capital and labor market was measured, and the results of the calculation are shown in Fig. 5.

(a) Distortion of China's inter provincial resource allocation

(b) Factors influencing the distortion of inter provincial resource allocation in China
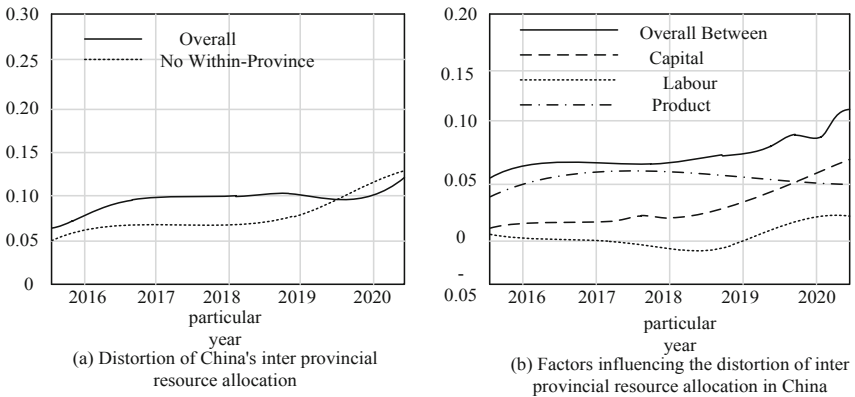
Fig. 5. Distortion of human resource allocation changes with market trends

The overall distortion of capital and labor allocation is relatively small. Based on the above content, the efficiency of resource allocation has experienced a process of initially deteriorating, slightly improving, and then deteriorating. Since the improvement in the middle 10 years is small, the overall allocation efficiency is deteriorating. of. Based on the above information, the data is summarized, and the method in this paper and the traditional method are used for experimental comparison, and the configuration effects of the two methods are recorded. The specific results are shown in Fig. 6.
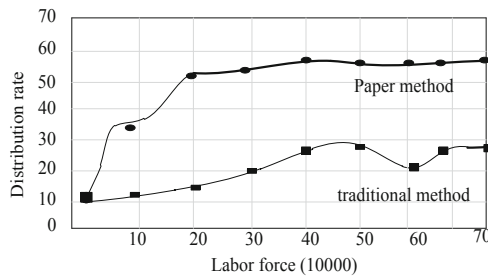
Fig. 6. Different methods of human resource allocation effect detection

Based on the analysis of the detection results in the above figure, it is not difficult to find that, compared with traditional methods, the adaptive configuration modeling of labor resources based on deep reinforcement learning proposed in this article can better realize the reasonable allocation of human resources in the actual application process. Fully meet the research requirements.

4   Concluding

This paper analyzes the impact of China's capital and labor allocation distortions on total factor productivity from 1993 to 2017 by building a model that includes different regions. The current household registration system. The current household registration system hinders the flow of labor resources and causes severe distortions in inter-provincial distribution. On the other hand, the allocation of capital resources among provinces has been deteriorating, mainly because my country's current financial market is not very developed. To solve the problem of distortions in the allocation of resources such as labor and capital between regions in my country, it is necessary to intensify the reform of the household registration system and create a favorable environment for the free flow of labor and the market. Other factors of production.

In future research, it is necessary to further deepen the financial supply. Improve the financing mechanism that meets the development needs of small and medium-sized enterprises to improve the relative shortage of funds in eastern my country.

References

1. Liu, N., Liu, Z., Cui, L.: Deep reinforcement learning based robotic assembly in simulation. Comput. Simul. **36**(12), 296–301 (2019)

2. Sun, H., Demanet, L.: Extrapolated full-waveform inversion with deep learning. Geophysics **85**(3), 1–71 (2020)
3. Nair, B.B., Krishnamoorthy, S., Geetha, M., et al.: Machine vision based flood monitoring system using deep learning techniques and fuzzy logic on crowdsourced image data. Intell. Decis. Technol. **9**, 1–14 (2021)
4. Prado, S.A., Rodriguez-Ruiz, B., García-Sampedro, M.: Working women and digital competence in the Spanish labor context. Revista Iberoamericana de Tecnologias del Aprendizaje **16**(1), 61–69 (2021)
5. Vedhathiri, T.: Collaborative dissertation based on the human resources needs of MSMES to improve their competitiveness and to overcome the disruption. Procedia Comput. Sci. **172**(4), 551–558 (2020)
6. Haghighi, N.F., Bijani, M.: A gap analysis between current and desired situation of economic factors affecting human resources development in Iran. GeoJournal **85**(4), 1175–1190 (2020)
7. Elsayed, M.S., Liu, G., Mostafa, A.M., et al.: Fault-recovery and robust deadlock control of reconfigurable multi-unit resource allocation systems using siphons. IEEE Access **PP**(99), 1 (2021)
8. Kaur, R., Gupta, A., Srivastava, A., et al.: Resource allocation and QoS guarantees for real world IP traffic in integrated XG-PON and IEEE802.11e EDCA networks. IEEE Access **PP**(99), 1 (2020)
9. Dai, Y., Zhang, K., Maharjan, S., et al.: Edge intelligence for energy-efficient computation offloading and resource allocation in 5G beyond. IEEE Trans. Veh. Technol. **PP**(99), 1 (2020)
10. Ahmad, K., Jasimuddin, S.M.: The linkage between communication satisfaction, human resources management practices, person-organization fit, and commitment: evidence from Malaysia. IEEE Trans. Prof. Commun. **PP**(99), 1–15 (2021)
11. 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
12. 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)
13. Liu, S., Li, Z., Zhang, Y., Cheng, X.: Introduction of key problems in long-distance learning and training. Mob. Net. Appl. **24**(1), 1–4 (2018). https://doi.org/10.1007/s11036-018-1136-6
14. 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

Real-Time Detection and Recognition of License Plates for Traffic Monitoring

Nam Van Nguyen[1,2]([✉]) and Quan Minh Vu[3]

[1] Data Governance Department, Viettel Group, Alley 7, TonThatThuyet Street, CauGiay District, Hanoi, Vietnam
`namnv78@viettel.com.vn, nvnam@tlu.edu.vn`
[2] Thuyloi University, 175 Tayson, Dongda, Hanoi, Vietnam
[3] Viettel CyberSpace Center, Viettel Group, 7 TonThatThuyet Street, CauGiay District, Hanoi, Vietnam
`quanvm4@viettel.com.vn`

Abstract. We address the task of real-time detection and recognition for heterogeneous license plate images of diverse vehicles with characters arranged in multiple lines and captured in all day and night conditions. This paper presents MixLPR (Mixed License Plate Recognition), a framework to develop a real-time system deployable in dense urban traffic to fill that gap. MixLPR consists of two components, a license plate detector, and an OCR. The plate detector includes new Mish-enhanced residual nets equipped with geometric transformations to deal with view-induced distortions. The OCR component is a new segmentation-free method based on the transformers, which work directly on the 2D character block. We trained and validated MixLPR on two large public and private datasets. Our results exhibit both improvements in accuracy and inference speed compared to state-of-the-art approaches.

Keywords: License Plate Detection and Recognition · Convolutional neural networks · Attention mechanism and transformer

1 Introduction

Real-time, scalable Automatic License Plate Recognition (ALPR) is essential for traffic monitoring in modern smart cities. A typical ALPR system first detects the license plates which occupy only a very small proportion of the wide camera view. Then each plates are run through an OCR subsystem which recognizes the characters. The problems are highly challenging in real traffic monitoring practice due to multiple vehicle types and variations in license plate formats of moving vehicles captured by fixed monitoring cameras. The challenges are also amplified by the low quality of images under different lighting conditions, blurring due to fast motion, distortion due to camera angle, and occlusion in typical dense urban traffic (Fig. 1).
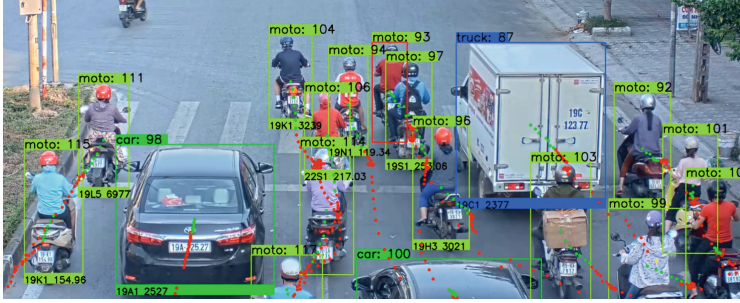
Fig. 1. Recognition of heterogeneous license plates in real-time under various lighting conditions. Vehicle bounding boxes are computed by a module not reported in this paper.

Many methods have proposed for recognising characters in single-line license plates, using numerous segmentation-based [2,3,6,7] and segmentation-free with CTC and attention mechanism [13,17,18,20]. However, there is no work on the recognition of license plate characters in multiple lines for small vehicles like motorbikes from live camera streams in crowded roads. Our main contributions are two-fold. First, we propose a robust neural network using our Mish-based residual blocks for accurate detection and rectification of license plates of cars, trucks, buses and bikes in both daytime and nighttime. Second, we propose an effective solution for the task of optical character recognition on multi-line license plates using Transformers.

We trained and validated MixLPR on two public datasets (CCPD-2020 and the Stanford-Cars) for cars only, and three large private datasets consisting of more than 200,000 images for a mix of bikes, cars, buses and trucks in total. Our results exhibit both accuracy and speed improvements over state-of-the-art approaches. MixLPR serves as the basis for a real-time traffic monitoring system for urban traffic in Vietnam.

The rest of paper is organised as follows. Section 2 reviews related works. Section 3 presents our main contributions – the MixLPR framework that can effectively work on mixed-type license plates. The effectiveness and efficiency are then evaluated in Sect. 4 on public and private datasets. Finally, Sect. 5 concludes the paper.

2   Related Work

Our work draws on recent works in object detection and optical character recognition (OCR) especially for license plate of vehicles.

2.1   License Plate Detection

Many ALPR are based on the state of the art object detection models [8,12, 15,21] that output rectangular bounding boxes for license plates. Our detection

approach, on the other hand, works on skewed non-rectangular shapes. We are inspired from recent works [5,13] which first detect four vertices of the quadrilateral bounding boxes of license plates and then rectify them to the original rectangular ones. These CNNs are also more light-weight than the state-of-the-art object detectors in license plate detection. However, these still produce high ratio of false positive predictions especially at night where car lights are detected as licences plates due to insufficient feature extraction of small objects. We improve from these baselines by introducing deep residual blocks [4] and Mish activation function [19].

2.2   Optical Character Recognition (OCR)

OCR has a long history. Early methods required character segmentation followed by single character classification [2,3,6,7,10]. The recent combination of CTC-based RNNs and CNNs have resulted in many efficient methods [21] for recognizing characters from the entire single-line license plate images. Our approach, on the other hand, tackles the problem of recognition of multi-line license plates. We combined a CNN [14] with a transformer as this model uses the multi-head attention mechanism over an unordered set, and thus in theory can handle arbitrary spatial character arrangement.

3   Methods
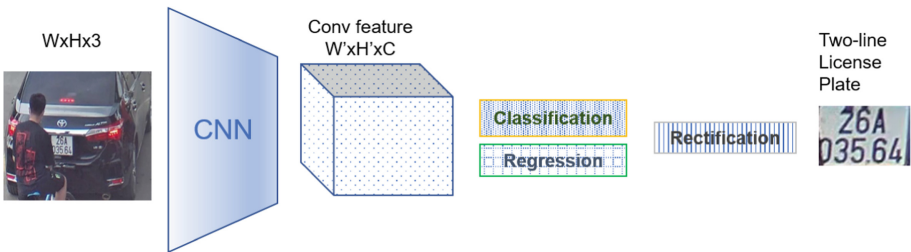
Fig. 2. Detection of license plate.

Our goal is to detect and rectify heterogeneous license plates of vehicles and then recognize their characters in urban traffic using fixed monitoring cameras. We assume that appropriate image pre-processing has been done to detect vehicles. Inspired by Vietnamese license plates, we develop generic methods for a mixture of single and multiple-line plate formats. Here the physical plates are rectangular metal panel containing a one-line or two-line character string. However, under the wide view angle from the cameras, plates appear non-rectangular. Those plates of moto-bikes are two-line attached at the back in Vietnam[1]. Cars,

---

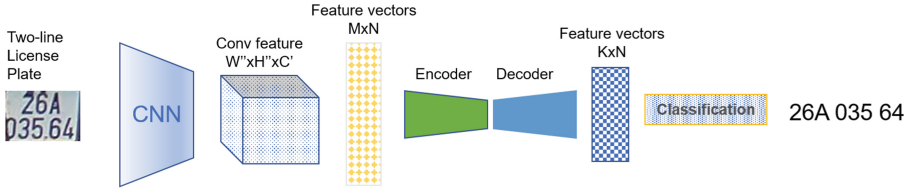[1] Plates can also be attached at the front of moto-bike in other countries.

Fig. 3. Recognition of multi-line license plates.

buses and trucks carry both kinds of plates at both the front and back of the vehicles. The plates have between 7 to 11 characters.

Our solution to the problem is a framework dubbed MixLPR (which stands for Mixed License Plate Recognition), which is graphically illustrated in Figs. 2 and 3. MixLPR consists of two connected modules: License plate detection, and multi-line character recognition.

The plate detection takes as input the vehicle bounding box and produces the rectified licence plate image as output. The modules processes the input image using a CNN to produce a feature map. The map then goes through a regression network and a classification network to predict the bounding box of the license plate and its geometrical transformation parameters. The parameters are then used to rectify the license plate.

This is then followed by the recognition module which generates the string of characters in the plate. A CNN is first applied to produce a feature map, which is sequenced into a set of position-encoded feature vectors. This serves as input for a transformer-based encoder-decoder architecture to generate the characters.

In what follow we present the two modules in more detail.

3.1   License Plate Detection

License plate detection can be considered as an object detection task but with unique characteristics that demand special treatments. This is because most of the license plate captured from camera at different angles are polygonal planar which can not be precisely bounded by rectangular boxes used in current object recognition models. Inspired from WPOD-NET [13], the quadrilateral shape of captured license plates can be seen as affine transformation of their original rectangular form. This is because the affine transformation can be built from the basic transformations such as translation, scaling, rotation and shearing.

Affine Transformation of License Plates and Its Inverse. We consider a point-to-point transformation resulting from a rotation and a translation. Given a two-dimension point $p(x_p, y_p)$, its affine transformed $q(x_q, y_q)$ is as follows:

$$q = Rp + T \tag{1}$$

where $T(x_T, y_T)$ is the translation vector and $R = \begin{pmatrix} r_1 & r_2 \\ r_3 & r_4 \end{pmatrix}$ is the rotation matrix. The reverse affine transformation from q to p is as follows:

$$p = R^{-1}q - (R^{-1}T) \tag{2}$$

where $R^{-1} = \frac{1}{r_1 r_4 - r_2 r_3} \begin{pmatrix} r_4 & -r_2 \\ -r_3 & r_1 \end{pmatrix}$ is the inverse of R.

In our models for license plate detection, the regression sub-network outputs six parameters $(r_1, r_2, r_3, r_4, x_T, y_T)$. The affine transformation is used for training and the inverse is deployed for inference. In the training phase, an affine transformation with the six parameters is applied on every rectangular cell of the input image to propose quadrilateral license plates. The six affine parameters are estimated by minimizing a square loss induced by the gap between the proposed license plates and the ground truths. By contrast, in the inference phase, an inverse affine transformation with the six output parameters is applied on the predicted quadrilateral license plate to produce the rectangular one. Thus, using this regression sub-network, we can both detect and align the distorted license plates.

Mish-Based Residual Blocks. The backbone neural network plays the critical role in license plate detection. A key parameter of the network is its depth – the number of feature transformation layers, usually the more layers the more powerful feature extraction. However, more layers can cause gradient vanishing, and thus lower layers are not updated during training, and thus reducing the performance of the networks. An effective solution is through skip-connections, which have been found to work well in recent residual neural networks (ResNets) [4]. The network is built from multiple residual blocks using a connection that skips several layers.

However, we found the rectified linear unit (ReLU) used in the original ResNets to be less effective in noisy images often seen in traffic cameras, causing overfitting. We propose to use a recent alternative activation function known as Mish [9] to combat the noise. Mish is a non-linear monotonic function which transforms the input as follows:

$$f(x) = x. \tanh\left(\ln\left(1 + e^x\right)\right) \tag{3}$$

We design our Mish-enabled residual backbone network as shown in Fig. 4. The first three initial, intermediate and alignment blocks include 2D convolutional filters, batch normalizations and Mish activation functions to extract the features of the input but do not reduce its size. The intermediate block is formed by adding a Mish function at the beginning. A convolutional filter and a batch normalization layer are also inserted to the intermediate block to create an alignment one. The name of the block indicates where it is integrated to the whole backbone network. Meanwhile, the feature of the input can be extracted and its size is also twice reduced by the down-sampling block thanks to its 2D convolutional filter with stride of 2. The last block also contains a maximum pooling at

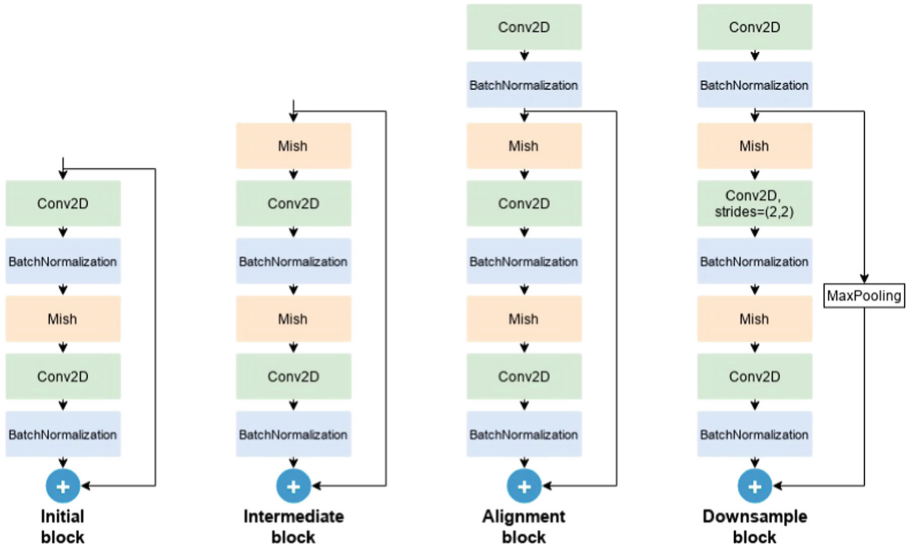the shortcut connection so that it can be aggregated to the main one with the same feature size.

Fig. 4. Mish-based residual blocks.

Simple Network for Heterogeneous License Plate Detection (SHELP). Based on the proposed Mish-based residual blocks, we designed our **S**imple network for **HE**terogeneous **L**icense **P**late detection (called SHELP) as shown in Fig. 5. Our SHELP takes as input an image containing a vehicle with a size of $208 \times 208 \times 3$ for training and of $416 \times 416 \times 3$ for inference. The backbone network contains two down-sampling residual blocks so that the size of input will be four times reduced. It also consists of one initial, one alignment and four intermediate blocks for efficient featuring. With these residual blocks, contextual information can be aggregated via shortcut connections with the local receptive field of license plate making the backbone network of SHELP more robust even with less number of layers. For detection of license plates, SHELP ends with regression and classification sub-networks and a concatenation layer with eight output parameters including six affine transformation parameters and two for {object, non-object} indicators.

Loss Function In the training phase, the input size is $208 \times 208 \times 3$ and the last output shape is $13 \times 13 \times 8$. This means that the input's size is 16 times reduced to form a grid of 13×13 cells. Given four corners $p_k(x_k, y_k), k = \overline{1,4}$ of a ground truth license plate, they then corresponds to $p'_k(\frac{x_k}{16}, \frac{y_k}{16}), k = \overline{1,4}$ in the grid. Suppose that the center of a cell $(i,j), i,j = \overline{1,13}$ can be seen as an origin
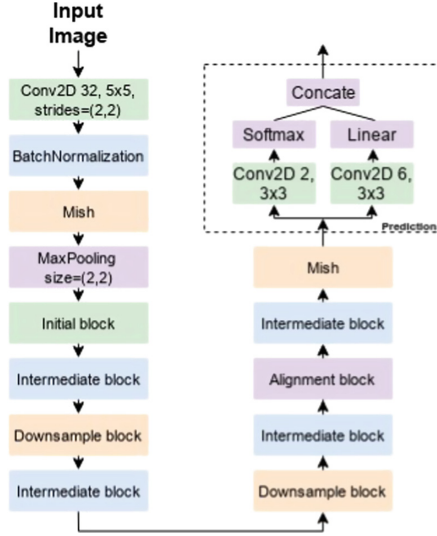
Fig. 5. The architecture of SHELP.

of the coordinate axes, the four corners of the ground truth license plate will be $q_k^{(i,j)}(\frac{x_k}{16} - i, \frac{y_k}{16} - j), k = \overline{1,4}$.

Given four corners of an unit box $u_k, k = \overline{1,4}$ originated at the center a cell $(i, j, i, j = \overline{1,13})$, at each epoch of training, the regression sub-network proposes six affine parameters which transform the unit box to a polygonal licence plate $v_k^{(i,j)} = Ru_k + T, k = \overline{1,4}$.

Inspired from SSD and WPOD-NET, the difference between the two boxes is

$l_{loc}^{(i,j)} = \sum_{k=1}^{4} smooth_{L_1}(v_k^{(i,j)} - q_k^{(i,j)})$, where $smooth_{L_1}(x) = \begin{cases} 0.5x^2, \text{if } |x| < 1 \\ |x| - 0.5, \text{otherwise} \end{cases}$.

We denote $x^{(i,j)} = 1$, if the IoU (Intersection over Union) between the cell (i, j) of size 16×16 in the input image and the ground truth license plate is greater than or equal to 0.3. Otherwise, $x^{(i,j)} = 0$. The classification sub-network outputs two parameters (c_1, c_2). Using softmax function, we have (o_1, o_2) where $o_k = \frac{e^{-c_k}}{e^{-c_1} + e^{-c_2}}, k = 1, 2$. In this case, $o_1 + o_2 = 1$. Then, the confidence loss for this sub-network is the binary cross-entropy function: $l_{conf}^{(i,j)} = -x^{(i,j)}.o_1 - (1 - x^{(i,j)}).o_2$

Finally, the loss function between the predicted and the ground truth license plate to be minimized is as follows:

$$L = \sum_{i=1}^{13} \sum_{j=1}^{13} \left(x^{(i,j)}.l_{loc}^{(i,j)} + l_{conf}^{(i,j)} \right) \tag{4}$$

This loss function is used for both SHELP and since they have the same end-block. Both networks output $13 \times 13 \times 8$ parameters – there are 169 cells, each of which has six affine parameters and two probabilities of confidence.

In the inference phase, the size of the input images is $416 \times 416 \times 3$ which is double of the training images. This is because the ground truth license plates in the training images have been zoomed, centered and resized. After predicting six affine parameters and two object/non-object probabilities for every cell, the non-maximal suppression algorithm will be used to select the most accurate predictions. Using the reverse transformation with the predicted affine parameters as in Eq. (2), the detected polygonal bounding boxes of license plate are re-transformed to the rectangular ones and their inside characters are re-aligned which are then fed to the optical character recognition model.

3.2   Recognition of Multi-line License Plates
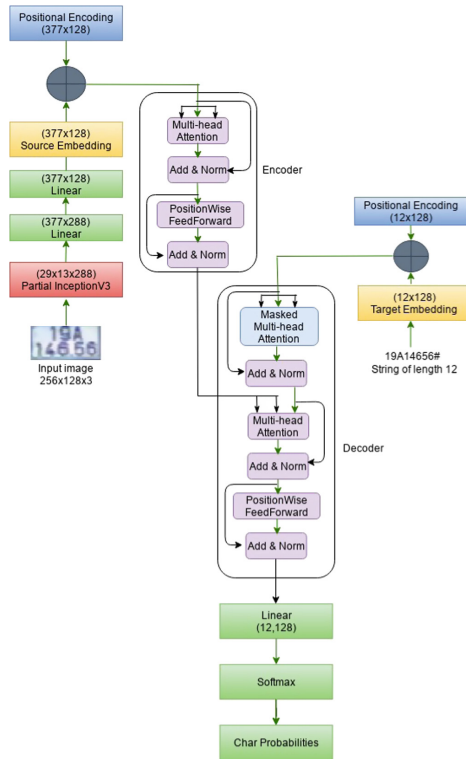
Fig. 6. MCR model architecture.

We present the second module of our MixLPR called MCR (Multi-line License Plate OCR), which translates an image of multi-line license plate into a string of characters. MCR employs a CNN for feature extraction and a transformer-based encoder-decoder for generation of characters.

In what follows we present the components in more detail: The CNN in Sect. 3.2, and the encoder-decoder and the loss function in Sect. 3.2.

CNN-Based Feature Extraction. As described in Fig. 6, all heterogeneous license plate images are resized to $256 \times 128 \times 3$. First, their heights are resized to 64 and 128 for one-line and two-line ones, respectively. Second, their left and right sides are all padded up to a width of 256. Third, the bottom of one-line images is filled up with a box of size $256 \times 64 \times 3$. The license plate images with longer width than 256 are only resized to 256 without padding. In other words, the CNN takes as input image of $256 \times 64 \times 3$ and outputs a multi-channel feature map of $29 \times 13 \times 288$.

For CNNs we use the Inception-v3 [14] thanks to its state-of-the-art performance on the ImageNet classification challenge [1,11]. The mixed-5d layer of the network is chosen for feature extraction due to its fast processing and high accuracy [16].

Encoder-Decoder and Loss Function. The core of our recognition network is the transformer-based encoder-decoder architecture to map the 377 location embedding of the source into characters in the target sequence of length 12. The encoder computes the similarity between source locations and refines the location embedding, making the features ready to be used in the decoder. This similarity is computed through a multi-head self-attention mechanism. The decoder produces one character at a time in the target sequence. At each decoding position $t = \{1, 2, ..., 12\}$, the decoder pools relevant source features and the previous decoded character embedding, also using the multi-head self-attention. The pooled information is then used to compute the probabilities of 37 possible characters (26 alphabet letters, 10 digits and a null):

$$y'_{t,i} = \frac{e^{o_{t,i}}}{\sum_{j=1}^{37} e^{o_{t,j}}} \tag{5}$$

where $o_{t,i}$ is the output of the linear layer corresponding to the ith character in the possible set at position t. This is depicted as the softmax layer in Fig. 6.

For training, the network uses the cross-entropy loss function for classification:

$$L_{OCR} = -\sum_{t=1}^{12} \sum_{i=1}^{37} y_{t,i} \log(y'_{t,i}) \tag{6}$$

where $y_{t,i}$ is the one-hot coding of the groundtruth character at position t.

4   Experiments

In this section we evaluate our proposed MixLPR on both public and private datasets, and compare its performance against strong rivals reported in the literature.

4.1   Experimental Settings

Datasets and Pre-processing. We use two public datasets for license plate detection (CCPD-2020 and Stanford-Cars) and three private datasets for both license plate detection and OCR. **CCPD-2020** (Chinese City Parking Dataset 2020) containing over 10,000 car images, with license plate location annotations. Similarly, **Stanford-Cars** is a dataset of cars with only 3,400 license plate annotations. Since these datasets contain only cars, they do not represent the real traffic in many Asian cities, which has many vehicle types, including bikes, cars, buses and trucks.

For that reason, we collected three private datasets from monitoring cameras in a densely populated Vietnamese city. The first dataset is for license plate detection (**uTVM-LP**), the second is for license plate OCR (**uTVM-OCR**), and the third is for end-to-end testing (**uTVM-N2N**). The **uTVM-LP** dataset has 100,000 images of vehicles (bikes, cars, buses and trucks) annotated with polygons of license plates. Similarly the **uTVM-OCR** dataset contains 100,000 images of license plates annotated with character strings. Lastly the **uTVM-N2N** dataset consists of 6,000 vehicles images which are annotated with their corresponding license plate characters, serving as a testbed for end-to-end system evaluation. Unless otherwise specified, datasets are divided into 70/30% for training and testing, respectively.

Evaluation Metrics. We measured the license plate detection and recognition models using the average precision (AP) and the accuracy of sequence (AOS).

Average Precision is a popular metric for evaluating the object detection models. This is the mean of precision values of the detector over multiple recall rates from 0 to 1. The precision and recall are formulated as follows:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

where TP, FP, FN is the true positive, false positive and false negative value, respectively. These values are determined according to an IoU threshold (intersection over union) between the detected polygonal bounding box and the ground truth license plate.

Accuracy of Sequence (AOS) for license plate recognition is defined as follows:

$$AOS = \frac{n - \#error}{n}$$

where $n, \#error$ are the total number of tested license plates and of unrecognized ones, respectively. If any of the characters in license plate string is not correctly recognized then the license plate recognition is erroneous.

4.2   Results on License Plate Detection

We trained, evaluated and compared SHELP and WPOD-NET on three datasets Stanford-Cars, CCPD and uTVM-LP.

Table 1. Average Precision of license plate detection models.

Dataset	IoU	WPOD-NET	SHELP
Stanford-Cars	0.5	81.2	82.3
Stanford-Cars	0.65	72.6	74.9
uTVM-LP	0.6	96.9	98.2
CCPD-2020	0.6	92.9	93.1

Accuracy of SHELP. The overall results on **all datasets** are presented in Table 1. On all datasets and IoU thresholds, our proposed method (SHELP) are better than WPOD-NET.

The results on **uTVM-LP** dataset partitioned into bikes and big vehicles in daytime and nighttime are shown in Table 2. With IoU of 0.5 all methods perform exceptionally well on bikes, with more than 99.2% AP, where our proposed method SHELP still show better accuracy. The high performance is because the license plates account for a bigger ratio of area in unique bike images than in unique four-wheel ones. Moreover, there are normally lots of noises in four-wheel images such as the poster of advertisements, the reflection in the car mirror or even the license plate images of the neighboring vehicles.

On the **uTVM-LP** dataset, the performance gap for big vehicles is wider among methods. Again, our method is better than WPOD-NET with a margin of 2.29 points in daytime and 6.74 points in nighttime. Interestingly, our methods work better in nighttime than daytime, when SHELP can reach 98.09%. This contrasts with the expected behaviour of WPOD-NET, which drops by 3 points when going from daytime to nighttime.

Table 2. Average Precision of license plate detection models on **uTVM-LP** dataset (IoU = 0.5).

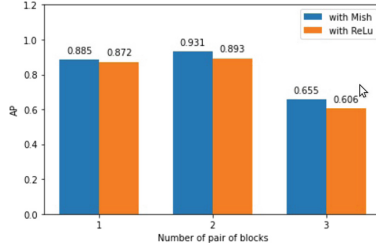| Vehicle	Time	WPOD-NET	SHELP
Big vehicles	Daytime	94.32	94.20
	Nighttime	91.35	**98.09**
Bikes	Daytime	99.24	99.48
	Nighttime	99.30	99.31

Fig. 7. Mish versus ReLU activation functions in SHELP license plate detection with various number of pair of blocks, evaluated on the **CCPD-2020** dataset.

Contribution of Mish Activation Functions. Figure 7 shows the comparison in Average Precision between the Mish and ReLU activation functions on the test set of CCPD-2020. The x-axis represents the number of pair of residual blocks (intermediate and down-sample) used in SHELP. It can be seen that Mish achieves a higher accuracy that the ReLU.

4.3   Results on License Plate OCR

The multi-line OCR results on the **uTVM-OCR** dataset are reported in Table 3. MCR achieved a very high AOS of 95.17%, 93.02%, 91.28% and 86.73% for recognition of license plate of four-wheel vehicle images and of bikes captured in day-time and in night-time, respectively. This accuracy is very significant given the fact that 83.7% of these license plates are two-lines.

Table 3. Accuracy of MCR license plate recognition on **uTVM-OCR** dataset.

Vehicle	Time	No. of images	AOS
Cars, trucks, buses	Daytime	8,419	95.17%
	Nighttime	1,579	93.92%
Bikes	Daytime	8,538	91.28%
	Nighttime	1,462	86.73%

4.4   End-to-End Results

So far we have evaluated MixLPR modules (license plate detection and OCR) separately. Since MixLPR is evaluated by the end user, it is important to know how license plate detection and recognition methods work in the joint system. We combine WPOD-NET and MCR as a baseline for comparison. Table 4 reports the results. As expected WPOD-NET + MCR is not as good as our MixLPR option (SHELP + MCR).

Table 4. Accuracy of sequence for end-to-end license plate detection and recognition methods on **uTVM-N2N** dataset.

Method	AOS
WPOD-NET + MCR	90.6
SHELP + MCR	**93.5**

4.5 Performance

For real-time systems, MixLPR uses SHELP (Sect. 3.1) for license plate detection due to the smaller model size. The combination of SHELP and MCR also exhibits higher accuracy for end-to-end recognition of license plate as shown in Table 4. These models were developed in Pytorch and Keras and were then converted into NVIDIA TensorRT for faster inference without loss of accuracy. In a workstation of 24-core CPU, 64G RAM, 2 RTX 2080Ti GPUs, the inference time of the models with different batch sizes is reported in Table 5. SHELP takes only 1.26, 15, 36 and 75 ms for inference with batch size of 1, 16, 32 and 64, respectively. On these batch sizes, the inference time of MCR is 6.2, 22.8, 64.8 and 86.8 ms, respectively. Together, these add up to the total running time of MixLPR of 7.46, 37.8, 100.8, and 161.8 ms, respectively. The speed of MixLPR is therefore faster than the end-to-end RPNet which is about 16.4 ms on Quadro P400 GPU.

Table 5. Inference time (in ms) of models with different batch sizes.

Model	Batch size			
	1	16	32	64
SHELP	1.26	15	36	75
MCR	6.2	22.8	64.8	86.8
SHELP+MCR	7.46	37.8	100.8	161.8

5 Conclusion

We have addressed the challenges of real-time detection and recognition of license plates in intelligent transportation systems. The system has to operate reliably in continuously changing contexts with high accuracy. Existing deep learning methods based on the convolutional and recurrent neural networks can only be efficient with grid and sequential data. Therefore, they are less effective or even inapplicable in heterogeneous environments such as those traffic in many of dense Asian cities, where license plates can be text block rather than a single line. We have presented a new framework dubbed MixLPR, which has been developed and validated in real scenarios. MixLPR is composed of a license plate

detector SHELP and a multi-line character recognizer (MCR). A comprehensive suite of experiments on two public and three private datasets demonstrated that MixLPR is fast and reliable, and it achieved competitive accuracy for both plate detection task and OCR task compared to state-of-the-art rivals. The end-to-end integration can also run fast at rate of 2.5 ms per image when run on a batch size of 64.

We plan several future works. One is to study the accuracy and scalability of MixLPR in the context of truly end-to-end traffic monitoring, starting from raw images from cameras to licence plate recognition of every vehicle in the visual field. It is also interesting to train the two modules of MixLPR in an end-to-end manner, rather than separately as in the current work.

References

1. Canziani, A., Paszke, A., Culurciello, E.: An analysis of deep neural network models for practical applications. CoRR arXiv:abs/1605.07678 (2016)
2. Goel, S., Dabas, S.: Vehicle registration plate recognition system using template matching. In: 2013 International Conference on Signal Processing and Communication (ICSC), pp. 315–318 (2013). https://doi.org/10.1109/ICSPCom.2013.6719804
3. Gou, C., Wang, K., Yao, Y., Li, Z.: Vehicle license plate recognition based on extremal regions and restricted Boltzmann machines. IEEE Trans. Intell. Transp. Syst. **17**(4), 1096–1107 (2016). https://doi.org/10.1109/TITS.2015.2496545
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
5. He, M.X., Hao, P.: Robust automatic recognition of Chinese license plates in natural scenes. IEEE Access **8**, 173804–173814 (2020). https://doi.org/10.1109/ACCESS.2020.3026181
6. Hendry, Chen, R.: Automatic license plate recognition via sliding-window darknet-YOLO deep learning. Image Vis. Comput. **87**, 47–56 (2019). https://doi.org/10.1016/j.imavis.2019.04.007
7. Wu, H.-H.P., Chen, H.-H., Wu, R.-J., Shen, D.-F.: License plate extraction in low resolution video. In: 18th International Conference on Pattern Recognition (ICPR 2006), vol. 1, pp. 824–827 (2006). https://doi.org/10.1109/ICPR.2006.761
8. Laroca, R., et al.: A robust real-time automatic license plate recognition based on the YOLO detector. In: International Joint Conference on Neural Networks (IJCNN), pp. 1–10, July 2018. https://doi.org/10.1109/IJCNN.2018.8489629
9. Misra, D.: Mish: a self regularized non-monotonic neural activation function. ArXiv arXiv:abs/1908.08681 (2019)
10. Montazzolli, S., Jung, C.: Real-time Brazilian license plate detection and recognition using deep convolutional neural networks. In: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 55–62 (2017). https://doi.org/10.1109/SIBGRAPI.2017.14
11. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
12. Selmi, Z., Ben Halima, M., Pal, U., Alimi, M.: DELP-DAR system for license plate detection and recognition. Pattern Recognit. Lett. **129**, 213–223 (2019). https://doi.org/10.1016/j.patrec.2019.11.007

13. Silva, S.M., Jung, C.R.: License plate detection and recognition in unconstrained scenarios. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11216, pp. 593–609. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01258-8_36

14. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016). https://doi.org/10.1109/CVPR.2016.308

15. Wang, Q.: License plate recognition via convolutional neural networks. In: 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 926–929 (2017). https://doi.org/10.1109/ICSESS.2017.8343061

16. Wojna, Z., et al.: Attention-based extraction of structured information from street view imagery. In: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, 9–15 November 2017, pp. 844–850. IEEE (2017). https://doi.org/10.1109/ICDAR.2017.143

17. Xu, Z., et al.: Towards end-to-end license plate detection and recognition: a large dataset and baseline. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 261–277. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_16

18. Yang, Y., Li, D., Duan, Z.: Chinese vehicle license plate recognition using kernel-based extreme learning machine with deep convolutional features. IET Intell. Transp. Syst. **12** (2017). https://doi.org/10.1049/iet-its.2017.0136

19. Zhang, Z., Yang, Z., Sun, Y., Wu, Y., Xing, Y.: Lenet-5 convolution neural network with mish activation function and fixed memory step gradient descent method. In: 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, pp. 196–199 (2019). https://doi.org/10.1109/ICCWAMTIP47768.2019.9067661

20. Zherzdev, S., Gruzdev, A.: LPRNet: license plate recognition via deep neural networks. CoRR arXiv:abs/1806.10447 (2018)

21. Zou, Y., et al.: A robust license plate recognition model based on Bi-LSTM. IEEE Access **8**, 211630–211641 (2020). https://doi.org/10.1109/ACCESS.2020.3040238

Pressure Compensation Method of Check Valve in Aircraft Hydraulic System Based on Artificial Intelligence

L. I. Yaping[(✉)] and Q. U. Mingfei

College of Aeronautical Engineering, Beijing Polytechnic, Beijing 100176, China
liyaping_12345@163.com

Abstract. In view of the poor application effect of the traditional valve pressure compensation method, this study designed an artificial intelligence-based pressure compensation method for one-way valve of aircraft hydraulic system. Based on the functional block diagram of aircraft hydraulic system, the pressure conflict of aircraft hydraulic system is judged and the feasibility of pressure compensation is analyzed. Based on this, in order to analyze the fault conditions, the unified modeling of multiple systems, including hydraulic energy system, integrated management control system and hydraulic user system, was completed based on Modelica. Then the supplementary design was completed by designing the pressure compensation model of aircraft one-way valve. The experimental results show that compared with the traditional method, the pressure compensation effect of the proposed method is better, and it has higher practical application value.

Keywords: Artificial intelligence technology · Aircraft hydraulic · Check valve · Pressure compensation

1 Introduction

With the continuous improvement of aviation design and manufacturing level, modern aircraft flight tasks and flight conditions are gradually diversified, which also puts forward higher requirements for aircraft power demand and safety. The power of aircraft hydraulic system also increases, and the structure of hydraulic system is becoming more and more complex [1].

The aircraft hydraulic system provides hydraulic energy for hydraulic users on the aircraft, such as main flight control, auxiliary flight control, landing gear retraction and retraction, wheel braking, front wheel turning and engine backstepping. In order to ensure the flight safety of aircraft, modern aircraft system usually adopts system redundancy design to increase the safety and reliability of aircraft operation.

The hydraulic system of civil airliner is composed of multiple sets of independent and backup hydraulic energy systems to provide hydraulic energy for hydraulic users. Each set of hydraulic system is composed of hydraulic energy system and its corresponding hydraulic users. At the same time, there is an integrated management controller to

monitor and control the working state of the hydraulic system [2]. There will be some differences in the design and layout of the hydraulic system of different types of aircraft according to the passenger capacity and purpose of the aircraft.

The aircraft hydraulic system is closely related to flight control, landing gear, avionics and other systems. The process from scheme design to integrated development is very complex. It is necessary to comprehensively consider the influence of various aspects such as mechanical electrical hydraulic control, as well as the interaction with various systems [3]. In the design process of the hydraulic system, the traditional design method mainly carries out physical experiments through the ground hydraulic test platform, uses the system platform to carry out a large number of hydraulic component tests and ground whole machine tests, and then modifies the design scheme through the test data, iterates continuously, and finally completes the finalization of the scheme [4]. However, it is difficult to adjust parameters through this design method, and each test requires a lot of time, capital and technical resources, which leads to the increase in the cost of aircraft hydraulic system design and development, and the development cycle becomes longer. At the same time, failure conditions such as single failure and double failure cannot be evaluated at the initial stage of scheme design. Therefore, the use of digital functional prototype technology and computer simulation technology in the design and development of aircraft hydraulic system has a great advantage in reducing test workload, reducing research and development costs, and improving the efficiency of system research and development [5].

In recent decades, with the continuous advancement of digital functional prototype technology and multi-domain unified modeling and simulation technology, the use of simulation models to simulate and analyze aircraft hydraulic systems has been continuously applied in the design process of aircraft hydraulic system schemes. The aircraft hydraulic system is a system involving different disciplines such as mechanics, hydraulics, electrical and control. The simulation analysis of a single field or a single hydraulic user cannot comprehensively consider the mutual coupling relationship between the systems, and cannot analyze the simultaneous action of multiple hydraulic users on the hydraulic system. The impact of the system, and it is difficult to verify the impact of hydraulic user load changes on the hydraulic control logic [6]. With the continuous development of the multi-domain unified modeling language, the digital functional prototype technology of multi-domain unified modeling has been widely used in hydraulic, mechanical, electrical, control and other fields, and the complex synthesis of multi-domain systems by simulation designers. The product can carry out overall system design and analysis. Therefore, in the process of designing the aircraft hydraulic system scheme, it is of great engineering significance to construct the functional prototype of the aircraft hydraulic system based on the multi-domain unified modeling technology [7].

With the rapid development of artificial intelligence, based on the idea of bionics, it attempts to create an intelligent machine that can replace human work. At present, artificial intelligence technology has gradually matured in many industries. Among them, the support vector machine, a classic method in machine learning, realizes the two classification of data in a multi-dimensional space by creating a hyperplane. It can reasonably

use the two classification capabilities of the support vector machine in machine learning, that is, over. The idea of categorizing the ADs-B data of the two aircraft realizes the detection of flight conflicts. This method is also a classic in the pattern recognition method. As the most widely used neural network in the field of artificial intelligence, it has a working principle that is more in line with human brain thinking, memory and other behaviors. Therefore, it has more powerful potential for decision-making, recognition and learning issues [8]. In the traditional flight conflict detection, the model pays more attention to whether the flight conflict occurs, but rarely pays attention to the severity of the flight conflict and the future occurrence time, and the classification ability of the neural network can be used to make short-term predictions of the time of the flight conflict. For decision-making problems such as giving flight conflicts to resolve solutions, the neural network has shown its ability to quickly determine, and a reasonable design classification model for the problem can provide a solution for the two aircraft at the same time as the flight conflict occurs.

2   Method Design

2.1   Detect the Pressure Conflict of the Aircraft Hydraulic System

Aircraft hydraulic system is generally composed of hydraulic energy system, integrated management control system and hydraulic user system. The hydraulic energy system provides hydraulic energy for hydraulic users according to the control signals of cockpit and integrated management control system, and the integrated management control system controls the main components of hydraulic energy system according to aircraft status signals and hydraulic status signals, The hydraulic user completes the corresponding flight function driven by the control command and hydraulic energy [9].

The aircraft hydraulic system is also equipped with a hydraulic integrated management controller that can automatically control the main hydraulic components. During the working process of the aircraft hydraulic system, HIMC will continuously collect the working status signal of the hydraulic system, and receive the aircraft flight status signal at the same time, make judgments through the hydraulic control logic in the HIMC, and analyze the current flow pressure demand and failure conditions of the hydraulic system, so as to follow The flow requirements of different flight profiles of hydraulic users automatically turn on or off hydraulic electric pumps or PTUs.

The hydraulic user system mainly includes: aileron, elevator, rudder and spoiler actuation system, landing gear retractable system, front wheel turning hydraulic system and wheel brake hydraulic system and other 7 sub-systems. The hydraulic user receives the maneuvering instructions, and is driven by the hydraulic to realize the control of the flight attitude and the related maneuvering of the landing gear. The aircraft hydraulic system involves multiple fields such as hydraulics, machinery, control, etc.

The mutual coupling between each system must be considered comprehensively when designing the system. Its functional block diagram is shown in Fig. 1.

As shown in Fig. 1, the hydraulic energy system is one of the important energy systems on the aircraft, which mainly provides corresponding hydraulic energy for the hydraulic users on the aircraft. The aircraft hydraulic energy system is composed of 3 sets of mutually independent hydraulic energy systems, which are respectively connected
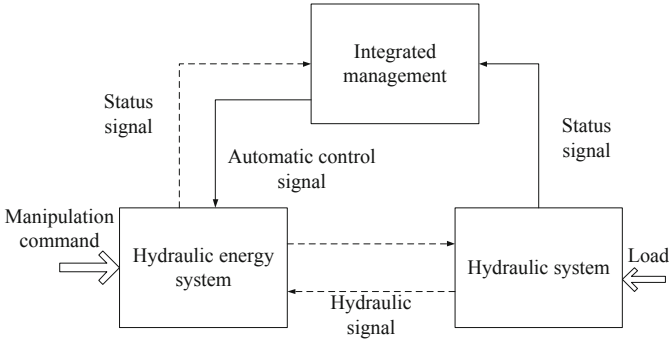
Fig. 1. Functional block diagram of aircraft hydraulic system

with the corresponding hydraulic users to realize the reasonable distribution of hydraulic load, and ensure that the pilot can still control the aircraft in the event of a single point of failure to achieve a safe landing of the aircraft.

2.2 Feasibility Analysis of Pressure Compensation

The system reliability needs to be analyzed before the system design, and the experimental data of the components that jointly constitute the aircraft system are calculated, which can respectively ensure the completeness of the aircraft hydraulic system design and serve as a reference for subsequent aircraft improvement. Functional reliability is the content that the aircraft hydraulic system needs to be analyzed in order to realize its function [10, 11].

This chapter analyzes the realization of pressure function, logic function and failure rate parameters of aircraft hydraulic system. The test platform is mainly used to simulate different pressure conditions of aircraft flight profile. Taking the aircraft flight process as the time axis, the aircraft from take-off to landing can be mainly divided into several stages: horizontal taxiing, take-off taxiing, rotary takeoff, take-off climbing, climbing, normal flight, aircraft descent, aircraft approach, landing flight, landing taxiing and aircraft taxiing. For different aircraft States, different pressure support is required under normal conditions, and the pressure requirements at each stage are shown in Table 1.

As shown in Table 1, the pressure required during takeoff is generally large, and the pressure is generally small during aircraft landing. In order to realize the above changes, necessary pressure simulation is needed to meet the signal requirements. The use of servo pressure control is not only a classic way of pressure control, but also a common method of aircraft hydraulic system. In order to realize this mode, the pressure and spool displacement are used as feedback in different states, and the pressure is controlled through PLC signal acquisition. Analog pressure load control is to realize the functional reliability of the aircraft. It is a semi physical simulation to simulate the pressure required by each hydraulic subsystem of the aircraft. It is an important premise to ensure the demand of the whole flight state of the aircraft. At the same time, it also meets the requirements of the aircraft hydraulic energy system for the pressure load simulation of each hydraulic user of the aircraft [12]. By analyzing different aircraft flight profile

Table 1. Definition of flight phase

Flight mission	Pressure demand (N/min)
Smooth line	7.84
Take off	81.36
Spin and lift off	82.68
Climb	82.78
Cruise	71.12
Decline	62.81
Go around	82.95

signals, the aircraft pressure demand can be divided into general step response and continuous change response. Different control parameters are used to simulate different pressure signals.

The load module is used to simulate the load pressure during aircraft flight and movement. Different pressure requirements need to be met for various stages and actions in the flight process of the aircraft. As the pressure simulation of actuator, on the one hand, load simulation module needs to realize the mutual backup of the system and can be used. On the other hand, it needs to meet the given system pressure to carry out the actual simulation of the system. Pressure simulation using servo valve.

Since the oil pressure of the aircraft hydraulic system rises very quickly, in order to reduce the high temperature of the hydraulic system caused by the excessive temperature rise and reduce the system reliability, the back pressure valve is returned to the rear of the fuel tank and the logic is also required. Controlled temperature cooling control switch.

2.3 Design of Aircraft Check Valve Pressure Compensation Model Based on Artificial Intelligence

At present, most of the handling methods for flight conflict resolution rely on the manual operation of controllers and less use computer-aided system, which puts forward very high requirements for the ability and quality of personnel [13]. In route flight, especially in the complex airspace of the terminal area, it is difficult to carry out complex guidance for the aircraft. Therefore, the controllers usually formulate some command standards. Although the guidance for the aircraft is often not the optimal path, different controllers ensure the safe operation of the aircraft to the greatest extent through unified and standard command methods.

In the process of controller's command, the command command of guiding the aircraft is more intuitive, and the methods of adjusting altitude, speed and heading are often used to solve the flight conflict. Adjusting the vertical distance between aircraft is an important method to solve flight conflicts. Through height adjustment, flight conflicts due to the vertical distance less than the safety interval can be adjusted, and flight conflicts on the horizontal plane can also be solved in three-dimensional space through space expansion. Moreover, in the process of approach control, due to the close horizontal

distance between aircraft and the large turning radius of aircraft, the method of adjusting course to avoid flight conflict is difficult to achieve, and it is relatively easy to change altitude. Therefore, the method of adjusting altitude is often used to solve the conflict.

In actual operation, it is important to pay attention to whether the adjacent altitude is occupied by other aircraft. If it is not occupied, the altitude adjustment method can be used. In addition, if there is a conflict between the ascending and descending aircraft altitude changes, the aircraft can first move to a conflict-free intermediate altitude, and then continue to ascend or descend after the conflict is resolved by related methods. When the cause of the conflict is that an aircraft is rising or falling and the distance between the aircraft is less than the safe interval, the conflict can be resolved by issuing an instruction to stop crossing. At the same time, for conflict resolution instructions, it is required to meet the one-step principle as much as possible to avoid the impact on other fluctuations caused by multiple changes in height. In order to reduce the pressure of aircraft hydraulic system, the pressure compensation model is constructed as follows:

$$B = HH^T \tag{1}$$

$$B' = D \wedge D^T \tag{2}$$

$$H_j^T = aD^T \tag{3}$$

$$a = H_j^T D \tag{4}$$

$$p = \sum_{i=1}^{k} \lambda_i \tag{5}$$

In formula (1–5), B is the symmetric positive definite matrix; B' is the decomposed symmetric positive definite matrix; H is the matrix of $m \times n$; T, a, and i are constants; D is the eigenvector; D^T is the T mutually orthogonal The characteristic vector of the point; H_j^T is the intelligent analysis vector of the j column of aircraft pressure; p is the contribution rate of aircraft pressure; k is the pressure characteristic value; λ_i is the characteristic principal component. This article standardizes the model data as follows:

$$X_i = \frac{X_i - X_{i\,\max}}{X_{i\,\max} - X_{i\,\min}} - 1 \tag{6}$$

In formula (6), X_i is the calibration value; $X_{i\,\max}$ and $X_{i\,\min}$ are the maximum and minimum values of calibration respectively. In this paper, let $p = 1$, then:

$$x_i = a_i f_1 + a_{i2} f_2 + ... + a_{iq} f_q \tag{7}$$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_i \end{bmatrix} \tag{8}$$

$$A = \begin{bmatrix} a_{11} \ a_{12} \ ... \ a_{1i} \\ a_{21} \ a_{22} \ ... \ a_{2i} \\ ... \ ... \ ... \ ... \\ a_{i1} \ a_{i2} \ ... \ a_{ip} \end{bmatrix} \tag{9}$$

$$X = AB + \varepsilon \tag{10}$$

In formula (7–10), x_i is the common factor; f_1, f_2, f_q, a_i, a_{i2}, a_{iq} are observable variables; X is the factor analysis formula; x_1, x_2, x_i are the pressure variables of the aircraft; A is the common factor pressure Load matrix; ε is the factor vector.

By constructing the model, the effect of pressure compensation needs to be further improved. The coupling relationship between various systems should be considered in the design process of aircraft hydraulic system. However, in the traditional modeling and simulation design, two methods are generally used. One is to establish a model of a single domain subsystem in a single software, fully analyze the subsystem, and then modify the system according to experience or test data to consider the impact of other systems. Its disadvantage is that the impact on other systems is not considered comprehensively and it is difficult to calculate accurately. The second is to model and simulate the aircraft hydraulic system through joint simulation [14]. This method first needs to artificially separate the coupling relationship between various systems, then establish simulation models in different software, integrate different models into one software through the interface between various simulation software, and build a joint simulation model to simulate the whole system. This method has high requirements for designers. They should master the operation of each software and be familiar with the interface between the software.

2.4 Realize the Pressure Compensation of the One-Way Valve of the Aircraft Hydraulic System

The design of the aircraft hydraulic system not only needs to consider the working characteristics of the hydraulic system under normal conditions, but also consider whether the hydraulic system can realize the function of energy system reconstruction under fault conditions. On the one hand, traditional design methods are difficult to analyze under fault conditions; on the other hand, the current modeling simulation does not model the system as a whole, and it is difficult to conduct a comprehensive analysis and evaluation of fault conditions.

A complete aircraft hydraulic system model, including hydraulic power system, integrated management control system and hydraulic user system, is required to analyze the failure conditions. Its structure is shown in Fig. 2.

In this study, the unified modeling of multiple systems is completed based on Modelica, which lays a foundation for the simulation analysis of fault conditions. Modelica is an open, object-oriented, equation-based computer language that can easily model complex physical systems across different domains, including: mechanical, electronic, power, hydraulic, thermal, control, and process-oriented subsystem models. The open Modelica standard library includes 920 component models from different physical domains with 620 functions.
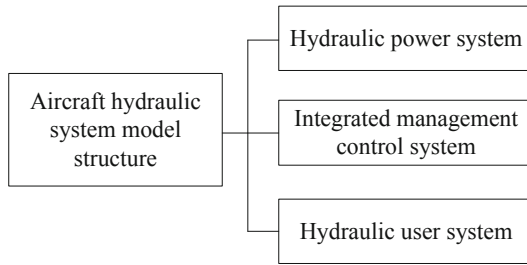
Fig. 2. Complete structure diagram of aircraft hydraulic system model

Due to aircraft hydraulic system of the power demand is bigger, strict in the volume and weight of the pump at the same time, as a result, the aircraft hydraulic system generally USES constant pressure variable piston pump as pump source, the constant pressure variable pump usually USES the axial plunger pump, axial type constant pressure variable piston pump with axial variable piston pump, the constant pressure valve and the variable cylinder.

Working principle of constant pressure variable pump: the setting pressure of constant pressure variable pump is determined by the preload spring of constant pressure valve. The setting pressure can be designed by changing the preload of spring. The pressure of the hydraulic system is determined by the load and regulated and limited by the constant pressure variable pump. When the system pressure is less than the set pressure, the constant pressure variable displacement pump outputs the maximum flow; When the system pressure reaches or exceeds the set pressure, the constant pressure variable pump enters the constant pressure model. In the constant pressure mode, the high-pressure hydraulic oil output by the pump controls the movement of the variable cylinder through the constant pressure valve. The change of piston displacement of the variable cylinder can change the inclination of the swashplate, so as to adjust the change of pump displacement. After the change of displacement, the output flow of the pump will also change, and finally achieve the adjustment of output pressure.

The control of the displacement of the constant pressure variable pump is the basis of constant pressure regulation. The displacement of the pump is controlled through the variable cylinder, so that the output flow of the pump is always consistent with the load flow demand, and the pressure of the regulating system is maintained near the working pressure.

The air-to-ground signal is used to determine whether the aircraft is in the air-flight state, and its high position indicates that the aircraft is in the air-state, including the aircraft's take-off, climb, cruise, approach, and preparation for landing stages. The signal is obtained by the pressure sensor installed on the landing gear struts of the aircraft. When the pressure of any sensor is not equal to zero, the wowC Weight On Wheels signal is true, and the aircraft is considered to be on the ground. The WOW signal is logically judged to produce the auxiliary signal of the aircraft hydraulic system, that is the air-to-ground signal. The air-to-ground signal is mainly judged by the control logic of the auxiliary hydraulic pump. Therefore, the actual working conditions of the aircraft hydraulic system must be considered comprehensively. Landing roll phase: When the

WOW signal changes from false to true, the aircraft hydraulic system is still in a high flow condition at this time, and the air-to-ground signal needs to be kept high for a period of time to ensure that the standby pump is turned on and the system has a large flow output; 2#hydraulic System low pressure condition: At this time, PTU needs to be turned on for pressure supply. If the aircraft is on the ground, PTU should not be turned on. When the above two conditions are met, the falling edge of the WOW signal is triggered, and the high bit remains for a period of time before it changes to the low bit.

The engine status signal is used to reflect the working status of the engine, including the aircraft's ground roll, take-off climb, cruise, approach and landing roll phases. The engine state signal can be generated by judging the positions of the left and right engine throttle levers. The engine signal is also the automatic control logic judgment of the user's standby pump. Therefore, the hydraulic system must be considered to meet the flow requirements of the hydraulic user in different working conditions.

In the ground idle phase: when the aircraft slides from landing to ground idle, the engine signal will change from true to false, but at this time, the hydraulic user is in the high flow condition, so it is necessary to keep the falling edge of the signal and delay the trigger for a period of time; Abort takeoff phase: when the aircraft needs to abort takeoff, the engine signal will change from true to false, but at this time, the hydraulic user is in high flow condition, so it is necessary to keep the falling edge of the signal and delay the trigger for a period of time. When the above two conditions are met, the falling edge of the engine status signal is triggered, and the high level remains for a period of time before it changes to the low level. The main function of the hydraulic pump control logic of the aircraft hydraulic system is to complete the energy reconstruction under the fault state of the redundant system. At the same time, it also has a certain function of energy optimal utilization, that is, it can automatically turn on or off the hydraulic electric pump according to the flow demand of the flight profile User system.

HIMC hydraulic pump control logic is composed of ACMP 1 B, ACMP2B, ACMP3A and ACMP3B control logic and PTU selection valve control logic_5. Due to the variety of control logics, the implementation and working principles of each pump control logic are the same. Therefore, the following mainly describes the automatic control logic modeling of the ACMPIB pump in detail, and the automatic control logic modeling process of other standby pumps is the same. The automatic control of the hydraulic pump is to automatically turn on the standby pump under high flow conditions and fault conditions, and automatically turn off the standby pump under low flow conditions and normal conditions. Large-flow operating conditions According to the principle of aircraft hydraulic system design, when the aircraft hydraulic system is in large-flow operating conditions and the No. 1 system is working normally, the electric pump ACMPIB acts as a backup pump to supplement the system's flow and stabilize the system pressure. When the aircraft is in the take-off or landing phase and the flap angle is not zero, the hydraulic system is in a high-flow condition. When the aircraft hydraulic system is in the normal flight profile, the system is under low pressure due to EDP 1. A failure, or the left engine is shut down. At this time, if the oil volume and oil temperature of the 1# fuel tank are normal, ACMPIB needs to be automatically turned on to supplement the pressure and flow of the system, To ensure the normal flight safety of the aircraft.

The electro-hydraulic servo valve is a key hydraulic component in the PCU. Its main function is to control the flow and flow rate of the hydraulic oil according to the electric control signal and the hydraulic servo signal, thereby controlling the extension or retraction of the actuator. The working principle is: when the control current is input, the control coil generates a deflection torque to make the jet tube deviate from the center position, the pressure in the two jet receiving tubes changes, the pressure on one side increases and the pressure on the other side decreases, so that the two ends of the main spool The pressure difference is generated, and the spool is driven to move relative to the valve body; meanwhile, the spool displacement generates a feedback torque through the feedback rod to make the jet tube return to the center position. When the torque generated by the control coil and the feedback torque are equal, the jet tube returns to the jet receiver. At this time, the main spool is in a stable position, and the corresponding flow and pressure control hydraulic oil are output.

3   Experiment and Analysis

In order to verify the practical effect of the one-way valve pressure compensation method of aircraft hydraulic system based on artificial intelligence, the following experimental process is designed.

3.1   Experimental Environment Parameter Design

In atmospheric air, even if the discharge voltage is up to 30 kV and the capacitor storage energy exceeds 4J, the test piece still cannot be ignited.

In the pure oxygen environment, as the pressure increases, the energy required to ignite the test piece decreases dramatically. Due to the limitation of test conditions, the pressure can only be raised to 3 Mpa in the experiment.

In the actual oxygenation process of the aircraft high-pressure oxygen check valve, the internal pressure of the valve can reach more than 10 Mpa, so it is estimated that the energy required for ignition is very small. In this environment, electrostatic discharge is easy to cause an explosion accident. In view of this situation, set the pressure parameters of the aircraft as shown in Table 2.

Table 2. Aircraft pressure parameters

Oxygen pressure/MPa	Falling pressure/MPa	Capacitors store energy/mJ
0.1	123.7	2808.45
0.5	113.4	897.80
1.0	105.0	12.00
2.0	92.1	48.05
3.0	81.2	7.20

In Table 2, the falling pressure of the aircraft will decrease with the increase of oxygen pressure. Therefore, the safety of the aircraft falling can be guaranteed to the maximum extent.

3.2 Experimental Results and Discussion

In the above experimental environment, the traditional method is compared with the method designed in this paper. The experimental results are as follows.

Table 3. Experimental results

Number of experiments/time	Compensation pressure by traditional method/MPa	Compensation pressure/MPa
1	121.2	251.3
2	134.5	278.2
3	142.3	302.4
4	156.8	327.9

According to the results shown in Table 3, the compensation pressure of the traditional method is small, which will lead to the problem of insufficient supply of aircraft hydraulic system, thus causing hidden trouble in flight. Compared with the traditional method, the pressure value compensated by the method in this paper is larger and always within the safety range, which proves that it can adapt to the supply effect of aircraft hydraulic system and meets the research purpose of this paper.

4   Conclusion

Although the probability of a large civil aircraft crash is low, it still exists. The analysis of the reliability of the aircraft hydraulic system can analyze the possible situations of the hydraulic system. Through prevention and key maintenance, the probability of aircraft crashes due to hydraulic system failures can be further reduced. The design of the hydraulic system test bench is an important process to ensure the successful development of the civil aircraft hydraulic system. It plays an important role in verifying whether the principle layout of the aircraft hydraulic system is reasonable and whether the various indicators meet the requirements. The necessary analysis can avoid unnecessary fatal errors in the aircraft development process.

In this paper, the necessary analysis is carried out in terms of reliability. Through the reliability analysis applicable to the general hydraulic system, the analysis is carried out with a focus on the failure rate parameters of the hydraulic system. The reliability of the aircraft hydraulic system should run through the entire aircraft design, not only the simulation settings of some reliability parameters need to be completed before the aircraft design, but also the reliability analysis of other parts during the aircraft installation process, until the aircraft completes the hydraulic system In the functional design of the

above, it is necessary to collect various parameters of the aircraft during the final flight test phase and the normal operation phase of the aircraft, and reserve them for each aircraft maintenance use. This analysis is a general analysis and use case. Generally, it takes a long time to test the parts. This part of the design shows the parameters that need to be paid attention to for reliability, provides comparison and certain standards for the design of future civil aircraft, and improves the safety and professionalism of our civil aircraft.

References

 1. Li, D., Dong, S., Wang, J., et al.: Thermal dynamics and thermal management strategy for a civil aircraft hydraulic system. Therm. Sci. **24**(4), 2311–2318 (2020)
 2. Jani, D.B., Ashish, S., Aditya, S., et al.: An overview on aircraft hydraulic system. Renew. Sustain. Energy Rev. **6**(5), 29–35 (2019)
 3. Shang, Y., Li, X., Qian, H., et al.: A novel electro hydrostatic actuator system with energy recovery module for more electric aircraft. IEEE Trans. Industr. Electron. **67**(4), 2991–2999 (2020)
 4. Aw, A., Rc, B., Nn, C., et al.: Adoption of a Bayesian belief network for the system safety assessment of remotely piloted aircraft systems. Saf. Sci. **118**, 654–673 (2019)
 5. Sheng, R., Guo, J.: Hydraulic piston pump in civil aircraft: Current status, future directions and critical technologies. Chin. J. Aeron. **33**(01), 20–34 (2020)
 6. Chu, Y., Yuan, Z., Chang, W.: Research on the dynamic erosion wear characteristics of a Nozzle flapper pressure servo valve used in aircraft brake system. Math. Probl. Eng. **20**(2), 1–13 (2020)
 7. Yang, J., Yan, H., Gu, C., et al.: Modeling and stability enhancement of a permanent magnet synchronous generator based DC system for more electric aircraft. IEEE Trans. Industr. Electron. **12**(09), 1 (2021)
 8. Shuai, L., Dongye, L., Khan, M., Ding, W.: Effective template update mechanism in visual tracking with background clutter. Neurocomputing **458**, 615–625 (2021)
 9. Zhan, C., Hu, H., Wang, Z., et al.: Unmanned aircraft system aided adaptive video streaming: a joint optimization approach. IEEE Trans. Multimedia **22**(3), 795–807 (2020)
10. Ghazi, G., Gerardin, B., Gelhaye, M., et al.: New adaptive algorithm development for monitoring aircraft performance and improving flight management system predictions. J. Aerospace Inf. Syst. **17**(10), 1–16 (2019)
11. Liu, S., et al., Human memory update strategy: a multi-layer template update mechanism for remote visual monitoring. IEEE Trans. Multimedia **23**, 2188–2198 (2021)
12. Stephenson, J.H., Weitsman, D., Zawodny, N.S.: Effects of flow recirculation on unmanned aircraft system (UAS) acoustic measurements in closed anechoic chambers. J. Acoust. Soc. Am. **145**(3), 1153–1155 (2019)
13. Gao, P., Li, J., Liu, S.: An introduction to key technology in artificial intelligence and big data driven e-learning and e-education. Mobile Networks Appl. **26**, 2123–2126 (2021)
14. Jiang, Z., Wang, Z., Liu, F.: Valve stiction detection and quantification based on hammerstein model. Comput. Simul. **38**(07), 391–396+432 (2021)

Multi-face Synchronous Recognition Method for Civil Aviation Security Inspection Based on Deep Migration Learning

Ning Zhang[1], Youcheng Liang[1], Tianyou Wu[1], and Jiangang Yin[2(✉)]

[1] Guangzhou Civil Aviation College, Guangzhou 510403, China
[2] State Grid Hubei Electric Power Co., Ltd, Wuhan 430000, China
jf125125@163.com

Abstract. There is a large flow of people in the civil aviation security inspection process, and the traditional face recognition method is inefficient in synchronously recognizing multiple faces. To this end, a synchronous recognition method of multiple faces in civil aviation security check based on deep transfer learning is designed. Collect multiple face image data, and perform face image feature processing. Using deep transfer learning to build a multi-face synchronous recognition model for civil aviation security inspection, and then realize the multi-face synchronous recognition of civil aviation security inspection. The comparison experiment is used to verify that the face synchronization recognition effect of the new method is better, which is of great promotion value.

Keywords: Deep migration learning · Civil aviation security check · Multiple faces · Synchronous recognition method

1 Introduction

Identity authentication technology is playing a very important role in modern society. Especially with the rapid development of the Internet, information security has become more and more important. Identity authentication is widely used in finance, security, justice, network transmission and other application fields [1]. Currently, the most widely used methods of identity authentication mainly include logos (represented by keys, chest cards, employee cards, identity cards, etc.), specific knowledge (such as passwords, passwords, passwords, etc.) and the combination of logos and specific knowledge (such as bank cards passwords, access cards + passwords, etc.) [2]. Identification numbers, magnetic cards, IC cards and other technologies have been widely used. Although these technologies are technically mature and can be protected by information encryption and other strategies, the essence of these technologies is to add additional distinguishing information to the individual, which is easy to be lost, forged or stolen, and it is difficult to distinguish who is the real user and who is the impostor of the system. Today, with the rapid development of the information superhighway and the expanding influence of cyberspace on human beings, the reliability and methods of traditional authentication

methods are challenged in the fields of electronic commerce, human-computer interaction, public security and network transmission, and the innovation of these technologies also put forward higher requirements for authentication methods and methods [3, 4].

Compared with fingerprint, iris and other biometrics, face recognition has many advantages, such as direct, friendly, convenient, stealthy, non-invasive, interactive and so on. Users have no any psychological barriers, and through the facial expression and posture analysis, they can get some information that other recognition systems can not get. As a result, academia and industry have focused on face recognition over the past 30 years [5]. Some scholars have used improved convolutional neural networks to recognize faces. On the basis of improving the CNN network, the ensemble learning strategy based on voting method is used to convexly combine the results of all individual learners into the final result to achieve more face recognition. However, the performance of this method in face recognition needs to be improved. There are also scholars who use the local ternary mode to extract low-resolution facial features, divide them into several blocks, and count the feature histograms of each sub-block. Face recognition is achieved by measuring the similarity of the histograms of the training set and the test set by the chi-square distance. However, the recognition of face details in this method needs to be optimized.

Face is a kind of highly non-rigid target, and there are a lot of details that reflect individual differences. First of all, the extremely complex face structure makes the face imaging itself a difficult problem, and the movement of facial muscles makes the face become a non-rigid object. Due to the influence of race, human face has different skin color models, contour models, but this difference in the same race is also different [6]. In a word, face recognition is a complex technology involving many subjects such as pattern recognition, computer vision, image processing, physiology and cognition. In order to improve the effectiveness of civil aviation security, this paper fuses the depth transfer learning with the multi-face synchronous recognition method.

2   Multi-face Image Preprocessing for Civil Aviation Security Inspection

2.1   Collect Multiple Face Image Data

The first step in face recognition is to collect face images. The quality of the collected face images directly affects the success or failure of subsequent face matching [7]. There are two types of image acquisition today: offline scanning and live scanning. In vivo scanning refers to the direct acquisition of human faces. In recent years, there have been many technological innovations in the sensors used in vivo scanning. The main types of sensors are optical sensors, silicon crystal sensors and ultrasonic sensors. Offline scanning, such as collecting the faces of criminals at the crime scene, collecting face and facial information, etc.

Optical Sensor: Optical Sensor is based on the principle of total reflection of light, light pressure on the face of the glass surface, CCD acquisition of reflected light to draw the face image, the amount of reflected light depends on the face pressed on the glass surface of the ridge and valley line depth and skin and grease between the

glass [8]. Silicon crystal sensors are small in size and low in power consumption, but are easily affected by static electricity, are easily damaged, and cost more than optical sensors. Ultrasonic sensor: Ultrasonic sensor is considered to be a very good way of face imaging technology, and its process is very similar to the laser in optical scanning. In an ultrasound scan, the dirt and oil accumulated on the skin have little effect on the obtained image, so such an image can be truly reflected. The face images collected by the three sensors are shown in Fig. 1.

| Optical sensor | Silicon crystal sensor | Ultrasound scan |
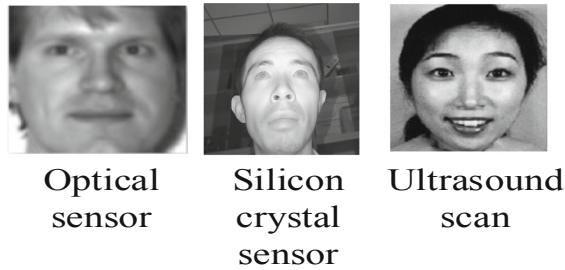
Fig. 1. Images captured by three sensors

2.2 Face Image Feature Processing

The process of face recognition is actually the process of face feature comparison, so the quality of face feature extraction determines the success or failure of the whole recognition system. Feature extraction includes three questions: what feature to extract, what method to extract feature, and whether the extracted feature can represent the true feature of the face [9]. Feature extraction is the core algorithm in the whole face recognition and identification process. The traditional feature extraction algorithm extracts and processes as many singular points and detail points as possible, because singular points are mainly used for rough classification in face database, and detail points are used for distinguishing different types of faces. In the face recognition process, we must first classify a face according to its macroscopic characteristics, and then match it with the corresponding face in the face database by using the detail feature parameters. Face feature extraction is an important premise to ensure the matching performance of automatic face recognition system. Face classification and matching are all based on face feature information. At present, most of the popular automatic face recognition systems are based on minutiae. This method simulates artificial face matching, determines the position of detail features in the face image, and then determines whether the face matches by comparing the correlations of features. Detail feature is the mutation of facial ridge, and its distribution and proportion are not the same in the face image. Endpoint and bifurcation are the most common detail features. Therefore, in most automatic face recognition systems, only endpoint and bifurcation points are used as detail features, and the pattern matching or point pattern matching is generally used.

Face image processing is a very important step in the whole automatic face recognition system. All kinds of noises usually accompany the face images acquired. Part of

it is caused by the picker, such as the stains on the picker, the parameter setting of the picker is not appropriate, and the other part is caused by facial features. The first kind of noise is relatively fixed error and is easy to recover. The second kind of noise is closely related to the individual face features, so it is difficult to restore them. Therefore, face image preprocessing is a key step in the face recognition process. It is unavoidable to be disturbed in the process of collecting face images. Under the present condition, it is impossible to ensure the quality of the captured face images. The aim of face image preprocessing is to remove all kinds of noises in the face image, eliminate the deformation of the face image and restore the clear and complete face feature structure. The processing flow is shown in Fig. 2.
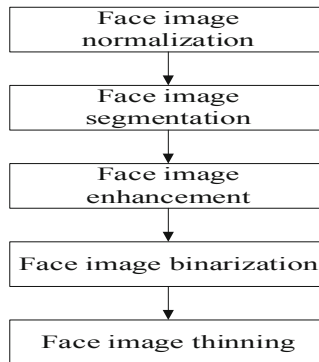
Fig. 2. Face image feature processing flow

As shown in Fig. 2, preprocessing is to process the low-quality face images with certain algorithms, so as to make the texture structure clearer and ensure the reliability of feature extraction. The accuracy and stability of feature extraction are directly affected by the processing results, which determines the performance of automatic face recognition. In general, face image recognition includes image normalization, segmentation, enhancement, binarization, thinning and so on.

Face image segmentation can not only reduce the processing time of redundant information, but also improve the accuracy of its subsequent processing algorithm. So far, many methods have been proposed for face image segmentation, which can be divided into two categories: gray-based method and direction-based method. The method based on gray level is to select an appropriate threshold to segment the foreground and background of human face on the statistical feature distribution of gray level. The direction based method mainly considers the consistency of the face texture direction. Firstly, the face region and background region are distinguished according to the direction of the texture on the face. In the face region with directional guidance, this method can identify the face region more accurately. But for the regions with discontinuous ripples and single gray level, and the regions around the center and triangle, the effect of this method is not ideal. Both methods have their own advantages and get good results. For the common face images with clear texture and uniform background, both methods can segment the face images accurately.

But the two methods also have some limitations: the gray-scale method is often too wet or too dry for the face image segmentation results are not accurate, and the direction based method for the face image of perspiration and other noise processing power is not strong. Using the local gray mean, local standard deviation and local consistency as features, the face image is segmented by line classification. Among them, local consistency characterizes the consistency of local image texture orientation, but these features are not enough to identify the fuzzy region. That is to say, the current method is not ideal for the poor quality image processing, so we need to propose a low-quality image segmentation method according to the actual needs.

At present, the face image enhancement methods are mainly based on the texture characteristics of the face lines, including direction continuity, distance approximation and so on. Firstly, a separation filter group is designed according to different grain directions and grain widths. Then different filters are used to filter the images and a group of filtered images are obtained. Finally, different filtered images are reconstructed using local texture orientation to reconstruct filtered images. When the number of filter banks is very large, the algorithm is very good, but the processing time is long, which will cause the performance of the algorithm to decline, reduce the number of filter banks and shorten the processing time, but the enhancement effect is not very ideal.

In order to enhance the face image, a filter is designed based on the pre-estimated local ridge direction and width, and then the filter and local face direction are coincided to achieve face image filtering. Therefore, if the actual grain width exceeds a predetermined range, this method will fail. Poor image quality can enhance the image quality, but it leads to a large number of false endpoints or bifurcation points. Feature extraction is to extract the required features from the input image. It is responsible for expressing in numeric form such features as the line orientation, line breakpoints, and intersections of a face image, which can fully represent the uniqueness of the face. In order to compare the accuracy, feature extraction algorithm is required to extract as many effective features as possible, while filtering out false features caused by various reasons.

2.3 Construction of Multiple Face Simultaneous Recognition Model for Civil Aviation Security Screening Based on Deep Transfer Learning

Global optimization Deep belief networks with multiple hidden layers are often difficult to deal with, and in order to achieve better optimization performance, greedy algorithms can be used here, that is, optimization layer by layer, learning only two adjacent RBM model parameters at a time, and by so greedy learning layer by layer to obtain the global DBNs. As for why learning greedily is effective for learning DBNs. The DRN obtained by DRN can be fine-tuned according to the final criteria of interest, that is, a DBNs network is stratified, unsupervised learning is performed on each layer, and the whole network is fine-tuned by supervised learning. Each module of a deep belief network consists of a number of nodes (often hundreds or thousands). The RBM model is shown in Fig. 3.

During the calculation, each layer learns the RBM parameters of this layer through the learning method of RBM, according to the input data (usually the output of the next layer), including the connection weights and the hidden layer node values. As the lowest input, it usually comes from observation variables, i. e. the original training data
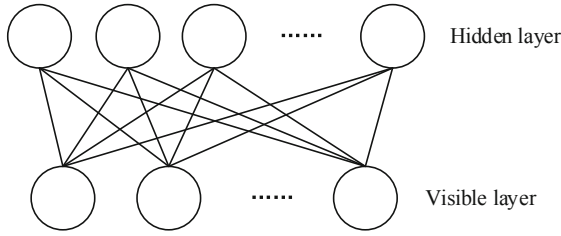
Fig. 3. Face image feature processing flow

of the object, such as the pixel gray of the image. That is, the visible layer (the training data) and the h form a typical RBM, making the RBM achieve energy balance. Then the output layer of the first RBM, the hidden layer, is used as the input of the second RBM, and the parameters are adjusted separately to make the current RBM structure energy tend to balance. Higher RBM's are trained in the same way until the top RBM is trained. This series of processes is called pre-training or pre-learning. After learning the RBM layer by layer, we adjust the whole network according to the maximum likelihood function using the original training data as the monitor data, which is called fine-tuning. That is the process of deep transfer learning [10]. In order to ensure the accuracy of face recognition and enhance the ability of multi-face recognition, this paper constructs a multi-face recognition model for civil aviation security screening based on deep transfer learning. The model is as follows:

$$P_i = \sum_{j \in C} P_{ij} \tag{1}$$

$$f(A) = \sum_{t=1}^{k} P_t \tag{2}$$

$$P_{ij} = \frac{\exp\left(-\|Ax_i - Ax_j\|^2\right)}{\sum_{t=1}^{k} \left(-\|Ax_i - Ax_t\|^2\right)} \tag{3}$$

$$P_{ij}' = \frac{\exp\left(-d_{ij}^2\right)}{\sum_{t=1}^{k} t\left(-d_{ij}^2\right)} \tag{4}$$

In formula (1–4), it is the construction rule of the model. Among them, P_i is the gradient of the random transformation of the face; P_{ij} is the learning efficiency between i and j; C is the identification label; $f(A)$ is the partial derivative function of the random data A; t is the learning time; x_i and x_j are training samples respectively; k is a constant; exp is the function definition formula; P_{ij}' is the learning efficiency between i and j in a specific area; d_{ij} is the Euclidean distance in the feature space; d_{ij} is defined, and its relevant rules are:

$$d_{ij} = \|F(x_i; W) - F(x_j; W)\| \tag{5}$$

In the formula (5), $F(x_i; W)$ is the multi-layer learning network of the weight vector W parameterized to x_i transition; $F(x_j; W)$ is the multi-layer learning network of the weight vector W parameterized to x_j transition. Let $j = k$, then:

$$P = c_i \sum_{k=1}^{t} kP_{ij} \qquad (6)$$

In formula (6–7), P is the identification data point; c_i is the label data. It is concluded that the constraints in the model designed in this paper are as follows.

$$O_{NCA} = \sum_{t=1}^{N} tP_{ij} \qquad (7)$$

$$\partial O_{NCA} = \partial F(x_j; W) \qquad (8)$$

$$\frac{\partial O_{NCA}}{\partial W} = \frac{\partial O_{NCA}}{\partial F(x_j; W)} \qquad (9)$$

In the Eq. (7–9), O_{NCA} is the objective function of NCA, NCA is the linear data point for deep learning, and ∂ is the partial derivative parameter. Under the model design, the simultaneous recognition of multiple faces can be performed.

3 Experiment and Analysis

In order to verify the effectiveness of the proposed method, the simulation platform is MATLAB 2019a, the traditional method is reference [5], and the face database is Yale B. The database, created by the University of Yale, contains more than 5,000 images of individuals photographed in a single light, from different poses and lighting conditions. The size of each picture is 640×480. Figure 4 shows some sample images of this database. The experimental process and results are shown below.

Fig. 4. Examples of some images of Yale B's face library

3.1   Experiment Procedure

First, analyze the passenger information table of civil aviation security check, as shown in Table 1 below.

As shown in Table 1, it is the passenger information table of civil aviation security check. According to the research on face recognition in the article, deep learning and face recognition are integrated, and the passenger information and face information are compared to ensure the safety of passengers and civil aviation.

Table 1. Passenger information form for civil aviation security check

Name	Type	Comment
Id	NUMBER	Serial number
FlightNum	VARCHAR	Flight number
Name	VARCHAR	Name
Age	VARCHAR	Age
Photo	Blob	Picture

3.2   Experimental Results and Discussion

In the above environment, the method designed in this paper is compared with the traditional method to verify the recognition effect of the two methods. The results of partial face recognition are shown in Table 2.

Table 2. Experimental results

Number of experiments	References [5] Simultaneous recognition of the number of faces/piece	The face recognition method designed in this paper recognizes the number of faces at the same time/piece
1	25	50
2	30	55
3	35	60
4	40	65

As shown in Table 2, the traditional method can recognize up to 40 faces at the same time. Under the same conditions, the method designed in this paper can recognize up to 65 faces at the same time, and the recognition effect is better, which is in line with the research purpose of this paper.

On the basis of the above experimental environment, we conducted a comparative test again. The experimental indicator of this test is the face information recognition

Table 3. Face information recognition time comparison

The number of experiments	Traditional method	The method of this paper
1	7.23	2.35
2	7.21	2.21
3	7.26	2.58
4	7.26	2.64
5	7.14	2.57
6	7.23	2.76
7	7.35	2.72
8	7.41	2.46
9	7.25	2.26
10	7.47	2.62

time. Randomly select 100 face images in the database, each group of 10, and conduct a total of 10 experiments. The experimental results are shown in Table 3.

It can be seen from Table 3 that the time for traditional methods to recognize face information is all higher than 7s, while the time for this method to recognize face information is all less than 3s. It can be seen that the method in this paper can realize the recognition of multiple face information in a relatively fast time, and improve the efficiency of civil aviation security inspection.

In order to further visually verify the effectiveness of the method in this paper, the effect of using the method in this paper to identify the face identity is shown in Fig. 5.

Multi-face recognition renderings
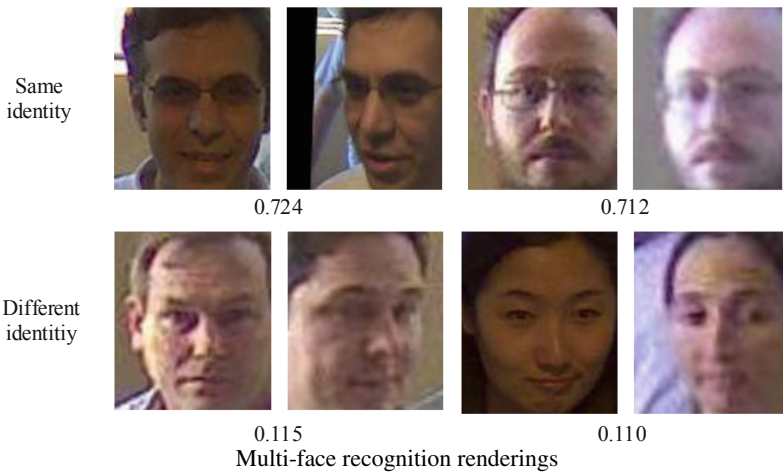
It can be seen from Fig. 5 that the similarity between face pairs with the same identity but different facial poses is high, while the similarity between face pairs with different

identities and different facial poses is very low. It shows that the method in this paper can extract highly recognizable face features from the multi-pose face images in the surveillance scene.

4   Conclusion

With the continuous development of society, ensuring information security through identification has become a hot topic today, and face recognition technology has gradually become a hot area of research. Based on the actual civil aviation security inspection system, this paper studies the preprocessing of face images, image feature extraction and image matching, and finally applies the research to actual projects.

Through the research of this article, we have a deep understanding of the field of face recognition, and face recognition technology needs to be completed by many aspects of cooperation. The algorithm also needs to be improved in practical applications to achieve the best results. The processing of face images involves several steps to work together to ensure accuracy. This paper completes the experiment of face recognition by extracting local feature information and matching faces in polar coordinates. Finally, it is applied through the inspection of the aviation security inspection system, and focuses on ensuring the effect of facial synchronization recognition, providing convenient conditions for civil aviation security inspection.

References

1. Liu, Z., Zhang, H., Wang, S., et al.: Reliability evaluation of public security face recognition system based on continuous Bayesian network. Math. Probl. Eng. **2020**(3), 1–9 (2020)
2. Wang, K., Hu, H., Li, L., et al.: Discriminative face recognition methods with structure and label information via $$l_2$$l2-Norm regularization. Neural Process. Lett. **51**(1), 639–655 (2020)
3. Zaman, F.: Locally lateral manifolds of normalised Gabor features for face recognition. IET Comput. Vision **14**(4), 122–130 (2020)
4. Wu, G.L.: Masked face recognition algorithm for a contactless distribution cabinet. Math. Probl. Eng. **2021**(2), 1–11 (2021)
5. Wu, Y., Hu, H., Li, H.: Age-invariant face recognition using coupled similarity reference coding. Neural Process. Lett. **50**(1), 397–411 (2019)
6. Chong, Y.L , Song, T., et al.: Feature fusions for 2.5D face recognition in Random Maxout Extreme Learning Machine. Appl. Soft Comput. **75**, 358–372 (2019)
7. Bala, A., Rani, A., Kumar, S.: An illumination insensitive normalization approach to face recognition using locality sensitive discriminant analysis. Traitement du Signal **37**(3), 451–460 (2020)
8. Liu, S., Bai, W., Srivastava, G., et al.: Property of self-similarity between baseband and modulated signals. Mobile Networks Appl. **25**(4), 1537–1547 (2020)
9. Liu, S., Bai, W., Liu, G., et al.: Parallel fractal compression method for big video data. Complexity **1**, 1–16 (2018)
10. Yi, S., Zhu, J.S., Jing, H.: Face recognition technology applies in railway scene based on MTCNN face occlusion technology research. Comput. Simul. **37**(05), 96–99 (2020)

Big Data and Signal Processing

Design of Human Resources Multi-dimensional Evaluation System Based on Big Data Mining

Haonan Chu[1] and Junlin Li[2(✉)]

[1] School of Labor Relations and Human Resource, China University of Labor Relations, Beijing 100048, China
[2] Beijing Union University, Beijing 100101, China
bfgb545@163.com

Abstract. In order to improve the load quota and throughput performance of human resources multi-dimensional evaluation system, this study proposes a human resources multi-dimensional evaluation system based on big data mining. The system includes four parts: data acquisition module, data preprocessing module, evaluation module and feedback incentive module; According to the demand of system mining human resource data, the decision tree algorithm is used to mine human resource data, and then according to the evaluation purpose, characteristics and principles, the evaluation index construction steps are designed to build a multi-dimensional evaluation index system. On the basis of determining the factor set of evaluation indicators, calculate the weight of evaluation indicators, so as to establish a multi-dimensional evaluation model of human resources and realize the multi-dimensional evaluation of human resources. The experimental results show that the load and throughput test indexes of the system are within the design standard, which proves that the system has achieved the expected goal.

Keywords: Big data mining · Human resources · Multi-dimensional evaluation · System design

1 Introduction

In the final analysis, the competition between enterprises is the competition between the pros and cons of corporate human resources. With the development of science and technology, the economic rate of return of human resources is significantly higher than that of other resources. Statistics in the United States in recent years show that the income gap caused by human resources and physical capital investment is still widening. The ratio of the two has reached 4:1, which fully confirms that the role of human resources in economic development is much higher than other resources, and Plays a vital role in the development of enterprises [1].

Organizational productivity depends on the use and control of three factors, these three factors are capital, methods and human resources. Effectively organizing the system requires information about what is happening, as well as a mechanism for correcting

or adjusting inputs [2]. The productivity gains obtained by capital can generally be measured through sophisticated accounting systems. The benefits obtained by the method can be evaluated by a control system similar to this one. However, the contribution of human resources to productivity is difficult to measure. This dynamic performance can only be evaluated by the work output or work behavior of employees in a period of time. Performance appraisal is to use productivity effects and efficiency criteria to assign values to employee behavior or work output [3].

At present, the performance evaluation methods commonly used by most enterprises in China are still based on the traditional performance evaluation methods based on empirical judgment. Therefore, there are some problems, such as large subjective randomness, poor scientificity, single evaluation means, lack of objective evaluation standards, closed, opaque and closed evaluation, resulting in large deviation in the evaluation results. It frustrates the work enthusiasm of managers and employees, resulting in low organizational efficiency and difficult to achieve enterprise goals [4]. Therefore, establishing and improving a scientific and reasonable performance evaluation system is an urgent task for enterprise human resources development and management [5].

Therefore, in order to improve the CPU and memory occupancy rate, system load capacity and throughput, this study designed a multidimensional evaluation system of human resources based on big data mining, in order to provide better technical support for human resources work.

2   System Design

2.1   System Overall Architecture Design

The design of the human resources multi-dimensional evaluation system is divided into four parts: data acquisition module, data preprocessing module, evaluation and assessment module and feedback incentive module, as shown in Fig. 1.
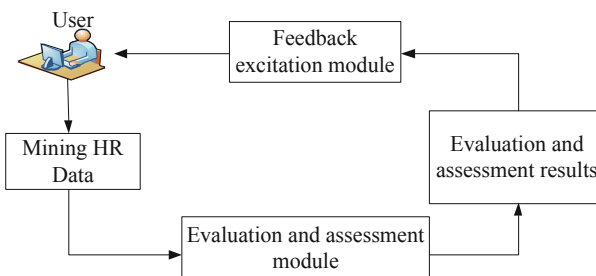
Fig. 1. Overall framework of the system

As can be seen from Fig. 1, the system in this paper takes employees as the main body of the system, and the final evaluation results will be fed back to employees, thus forming a closed loop for the whole system.

Firstly, data mining technology is used to process and analyze the relevant data of employees' work, and key characteristic indexes are extracted; Then, after preprocessing

the extracted key indicators, they are used as the input of artificial intelligence algorithm according to the weight, and analyzed by artificial intelligence algorithm after training; Finally, grade evaluation is given to employees' work; Due to the transparency of the algorithm, the evaluation process of the system can be output and used as a feedback incentive to the evaluated employees.

2.2   Big Data Mining Human Resource Data

In data mining classification technology, there are mainly decision tree method, Bayes method, neural network method and rough set method [6]. A decision tree is a tree structure similar to a flowchart, in which each internal node of the tree represents a test of an attribute (value), and its branches represent each result of the test; and each leaf node of the tree It represents a category. The highest node of the tree is the root node. Its core algorithm is ID3 algorithm. For this reason, the decision tree method of ID3 algorithm is adopted to mine human resource data.

Assuming that the data set S contains a collection of s data samples, the category attribute can take m different values, corresponding to m different categories C_i, where $i \in \{1, 2, \cdots, m\}$. Assuming that s_i is the number of samples in category C_i, then the amount of information I required to classify a given data object is:

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} p_i \log(p_i) \tag{1}$$

In formula (1), $p_i = \frac{s_i}{s}$ represents the probability of belonging to category C_i for any data object.

Assuming that an attribute A takes v different values $\{a_1, a_2, \cdots, a_v\}$, the attribute A can be used to divide the set S into v subsets $\{S_1, S_2, \cdots, S_v\}$, where S_j contains the data samples of the S sets of attributes A and a_j values. If the attribute A is selected as the test attribute (used to divide the current sample set), let s_{ij} be the number of samples of the attribute C_i category in the subset S_j. . Then the information (entropy) h required to divide the current sample set using the attribute A can be calculated as follows:

$$h(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + \cdots + s_{mj}}{s} I(s_1, s_{2j}, \cdots, s_m) \tag{2}$$

In formula (2), item $\frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s}$ is regarded as the weight of the j subset, which is the sum of the sample data of attribute A and a_j in all subsets divided by the total number of samples in the S set [7]. The smaller the $h(A)$ calculation result, the more "pure" (good) the result of the subset division. For a given sub-set S_j, its information is:

$$I(s_1, s_{2j}, \cdots, s_m) = -\sum_{i=1}^{m} p_{ij} \log(p_{ij}) \tag{3}$$

In formula (3), $p_{ij} = \frac{s_{ij}}{|S_j|}$ represents the probability that any data sample in subset S_j belongs to category C_i. The information gain obtained by using attribute A to divide

the corresponding sample set of the current branch node is:

$$Z(A) = I(s_1, s_{2j}, \cdots, s_m) - h(A) \tag{4}$$

In formula (4), $Z(A)$ represents the reduction of (information) entropy obtained by dividing the set according to the value of attribute A.

The decision tree induction algorithm calculates the information gain of each attribute, selects the attribute with the largest information gain as the test attribute of a given set, and generates the corresponding branch node. The generated nodes are marked as corresponding attributes, and corresponding (decision tree) branches are generated according to different values of this attribute, and each branch represents a divided sample subset [8].

The sample subset of the data set divided by the above process is the human resource data mined by the decision tree algorithm in the mass data. Next, based on these data, a multidimensional evaluation index system of human resources is established.

2.3 Construct a Multi-dimensional Evaluation Index System

The main purpose of this system is to achieve the following 8 purposes: to provide a basis for the promotion/demotion/transfer and resignation of employees, the organization's feedback on employee performance appraisal, to evaluate the contribution of employees and teams to the organization, and to provide a basis for employees' compensation decisions., Evaluate the decision-making of recruitment and selection and work distribution, understand the training and education needs of employees and teams, evaluate the effects of training and employee career planning, and provide information for work planning/budget evaluation/human resource planning.

Therefore, the system of this study needs to have five characteristics of conformity to reality, sensitivity, reliability, acceptability and practicality.

Based on this, the research index u is mainly composed of three parts: evaluation element a, evaluation mark b and status scale c, namely:

$$u = a + b + c \tag{5}$$

In formula (5), a represents the basic unit of the evaluation object; b is mainly used to reveal the key identifiable characteristics of the evaluation elements. It has a variety of forms. From the descriptive connotation, there are objective forms, subjective forms, semi-objective and semi-subjective forms. In terms of expression, there are short sentence, question, and direction indicators; c represents the degree of difference in evaluation elements or signs, and the state order and scale. It can be divided into quantifier, quantitative, hierarchical, symbolic, and graphical, Definition formula, comprehensive formula, etc. Therefore, multi-dimensional evaluation indicators have the functions of materialized connection, unified guidance, prevention of subjective one-sidedness, and deepening of understanding. To this end, according to the principle of homogeneity, reliability, universality, independence, completeness, and structure, a multi-dimensional evaluation index system is constructed according to the following steps:

1. Design of index content. It mainly includes the formulation of evaluation elements, the selection of evaluation signs and the division of evaluation scales. The design of each content has different ways and methods for candidates [9].
2. Classification, merger and design. Divide indicators by category.
3. Quantification of indicators. Quantification mainly includes three tasks: weighting, scoring and scoring. The specific contents are as follows:

    1) Weighting. Weighting refers to comparing the importance of different indicators in the "system". Weighting is actually a process in which the evaluation index system is vertically equalized and can be added. The weighting methods are: subjective experience method, A, B, C classification weighting method, expert survey method, comparative weighting method, Del non-weighting method, analytic hierarchy process weighting, multiple regression analysis weighting method, principal factor analysis weighting method, standard deviation Weighting method, etc.
    2) Assign points. That is, according to certain rules, a certain score is given to the "standard status" of each indicator and the degree of difference. There are many ways to assign points, including standard assignments, grade assignments, regular assignments, random assignments, precise assignments, fuzzy assignments, absolute assignments, relative assignments, secondary assignments, and statistical assignments., Decentralized scoring, these scoring methods have their own advantages and disadvantages, and the scoring method should be determined according to the specific evaluation situation.
    3) Scoring. Score cent is to point to when evaluating or after evaluating to evaluate result quantification and express, the form basically has statistical method, calculation method, judgment method, choice method.

4. Indicators are applicable.
5. Index test. Judge whether the index is qualified according to the inspection results. If not, return to step 1 to modify the index; If the indicators are qualified, the constructed indicators are output.

According to the above five steps, the indicators constructed in this study are shown in Fig. 2.

According to the multi-dimensional evaluation index system shown in Fig. 2, human resources are evaluated in six aspects: environment, personality, performance, ability, knowledge and morality.

2.4 Multi-dimensional Assessment of Human Resources

According to the index system shown in Fig. 2, the fuzzy comprehensive evaluation method is adopted to evaluate human resources in multiple dimensions. Based on this, the evaluation factor set is established as follows:

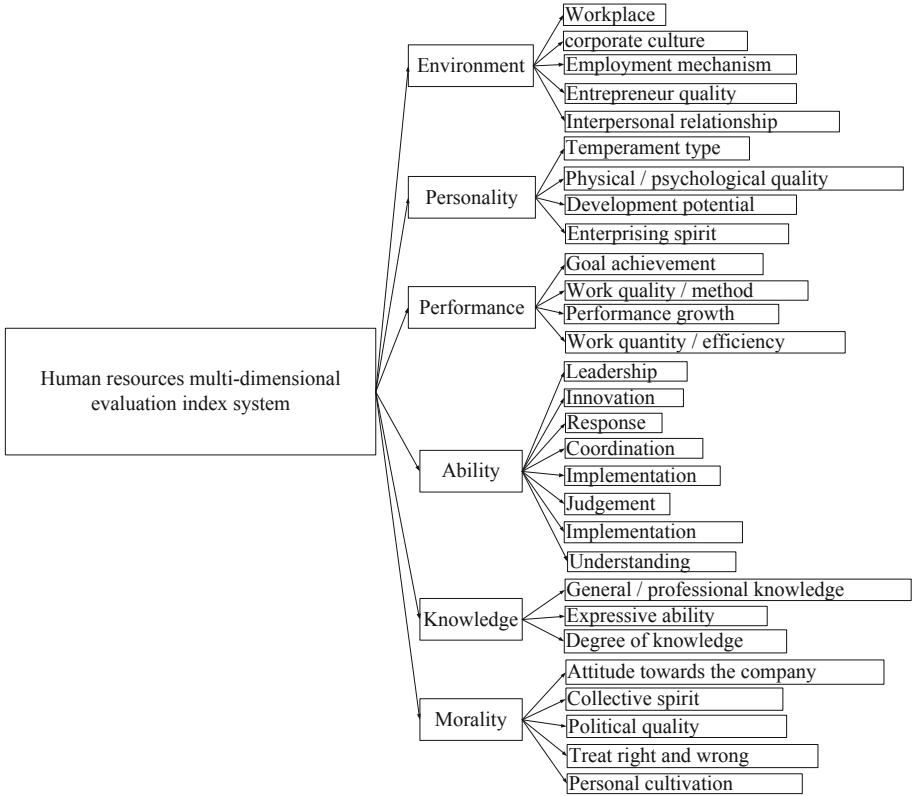$$U = \{u_1, u_2, \cdots, u_n\} \tag{6}$$

Fig. 2. Multi-dimensional evaluation index system

In formula (6), U represents the evaluation factor set; $u_i(i = 1, 2, \cdots, n)$ represents the i evaluation index, and n represents the number of indexes.

According to the evaluation factor set shown in formula (6), an initial matrix of initial data indicators of n evaluation indicators of m samples is established:

$$X = \left\{x_{ij}\right\}_{m*n} \tag{7}$$

In Eq. (7), X represents the initial matrix of dimension $m * n$; x_{ij} represents the element in row i and column j of the matrix. Since there are great differences in the dimension, order of magnitude and degree curve of each index, the initialization data shall be standardized:

$$x'_{ij} = \frac{x_{ij}}{\sum\limits_{i=1}^{m} x_{ij}} \tag{8}$$

In formula (8), x'_{ij} represents the data after x_{ij} normalization. According to formula (8), the standardized matrix X' of the data can be obtained:

$$X' = \left\{x'_{ij}\right\}_{m*n} \tag{9}$$

The j information entropy value H_j obtained from the unit entropy function H is:

$$H = -k \sum x_i' \lambda n x_j'$$
$$H_j = -k \sum x_i' \lambda n x_{ij}'$$

(10)

In formula (10), k represents Boltzmann's constant; λ represents the number of index states [10]. For a system with completely disordered information, the order is zero, and its entropy value is the largest, $H = 1$. When each sample of m is in a completely disordered distribution state, $x_{ij}' = \frac{1}{m}$, then:

$$H = -k \sum_{i=1}^{m} \frac{1}{m} \lambda n \frac{1}{m} = k \sum_{i=1}^{m} \frac{1}{m} \lambda n m = k \lambda n m = 1$$

(11)

So there is: $k = (\lambda n m)^{-1} \quad 0 \leq H \leq 1$.

Since the information entropy H_j can be used to measure the utility value of the information of the j index (the number of indicators), when it is completely disordered, $H_j = 1$, at this time, the information of H_j (that is, the number of the j index) is important for the comprehensive evaluation The utility value is zero. Therefore, the information utility value of a certain index depends on the information entropy value H_j of the index, and the difference $E_j = 1 - H_j$ of 1.

It can be seen that using the entropy method to estimate the weight of each indicator is essentially calculated by using the value coefficient of the indicator information. The higher the value coefficient, the greater the importance for evaluation. Therefore, the weight W_j of the j index is:

$$W_j = \frac{E_j}{\sum_{i=1}^{m} E_j}$$

(12)

Assuming that the comment set is V and there are n evaluation samples, then there are:

$$V = \{v_1, v_2, \cdots, v_n\}$$

(13)

According to the comment set shown in formula (13), the membership degree of the comment set is set to $\varsigma(u_n) = e^{-\frac{2n-1}{10}}$, where e represents the natural constant. If there are m evaluation factors, the fuzzy relationship between the evaluation factor set U and the comment set V can be expressed by the evaluation matrix R:

$$R_i = \{r_{ij}\}_{n*m}$$

(14)

In formula (14), $n * m$ represents the dimension of the matrix; r_{ij} represents the element in row i and column j of the matrix. There is the following relationship with the evaluation factor set U and the comment set V:

$$r_{ij} = ur(u_i, v_i)(0 \leq r_{ij} \leq 1)$$

(15)

When the weight term W and the fuzzy relationship matrix R are known, the compound operation of the fuzzy matrix can be used to establish the fuzzy comprehensive model P of the evaluation index system:

$$P = W_{il} * R_{il} \tag{16}$$

In formula (16), W_{il} represents the evaluation weight vector of the i index and the group l; R_{il} represents the evaluation matrix of the i index and the l group.

The calculation results of Eqs. (12) and (14) are brought into Eq. (16), and the multi-dimensional evaluation results of human resources can be obtained. Realize multi-dimensional evaluation of human resources.

3   System Test and Result Analysis

In order to verify the practical application performance of the above designed human resources multi-dimensional evaluation system based on big data mining, the following experiments are designed.

In order to verify the performance of the client, network communication and server at the same time, the LoadRunner stress test tool is used to generate a large number of users in the system, make the system in an overload state, and detect the usage of CPU and memory of the system server, as well as the maximum, minimum and average response time of the system page.

Considering the performance requirements of the design system in practical application, the following experimental environment is designed: Set the maximum number of resource access users to 2000, the maximum response time for resource requests to 1s, and the CPU and memory usage to less than 30%. If the response time exceeds 1s and the CPU and memory usage exceeds 30%, the system does not meet the design requirements and needs to be rectified.

The test steps are as follows:

1. Use the LoadRunner stress test tool to create a script on the computer and enter the website to generate the system stress test parametric setting script.
2. Connect the generated script with the system, create different resources in the system and record the script;
3. Set the number of virtual users loading resources in different resources to test the system pressure.

3.1   This Section Describes How to Test the System CPU and Memory Usage

First, test the load rate of the system, and the result is shown in Fig. 3.

It can be seen from Fig. 3 that when the virtual users in the system reach 10,000, the CPU and memory occupied by its running resources are 25% and 27%, respectively, which are 5% and 3% smaller than the maximum occupancy rate, which is already very high. Close to the limit of CPU and memory usage.
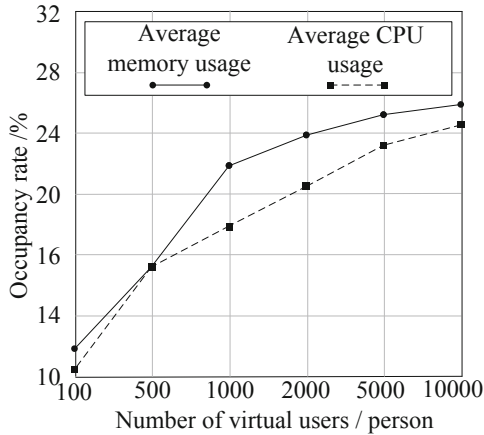
Fig. 3. CPU and memory usage

3.2 System Load Rate Test

The system runs the Vuser test script designed in this experiment 30 times to perform system pressure test. When the number of script runs reaches 30, after 5 min, the script runs are reduced by 5 times every 30s to perform system decompression test. The system load test result is shown as in Fig. 4.

It can be seen from Fig. 4 that under the load pressure shown in Figure (a), the system load test time is 11 min and 5 s, the average number of clicks is 160.54 s/time, and the average transaction response time is 4.467s, both in Within the scope of the system test standard designed in this experiment.

3.3 System Throughput Test

Based on the system load test results, test the system throughput, and the test results are shown in Fig. 5.

As can be seen from Fig. 5, the system throughput will change with the change of the number of concurrent users in the system, showing a direct proportional relationship with the system load. Thus, the server operation of this system tends to be stable gradually.

Based on the above three experimental results, it can be seen that the function and performance of the evaluation system designed in this paper have reached the expected goal, and it is suitable for practical application.

4   Conclusion

In this paper, combined with the theory of performance evaluation, according to human resources evaluation principles, objectives, characteristics, the establishment of performance evaluation index system. Then the fuzzy comprehensive evaluation model of performance evaluation is constructed by using fuzzy mathematics theory. From the

(a) System increase and decrease

(b) Hits per second

(c) Average transaction response time
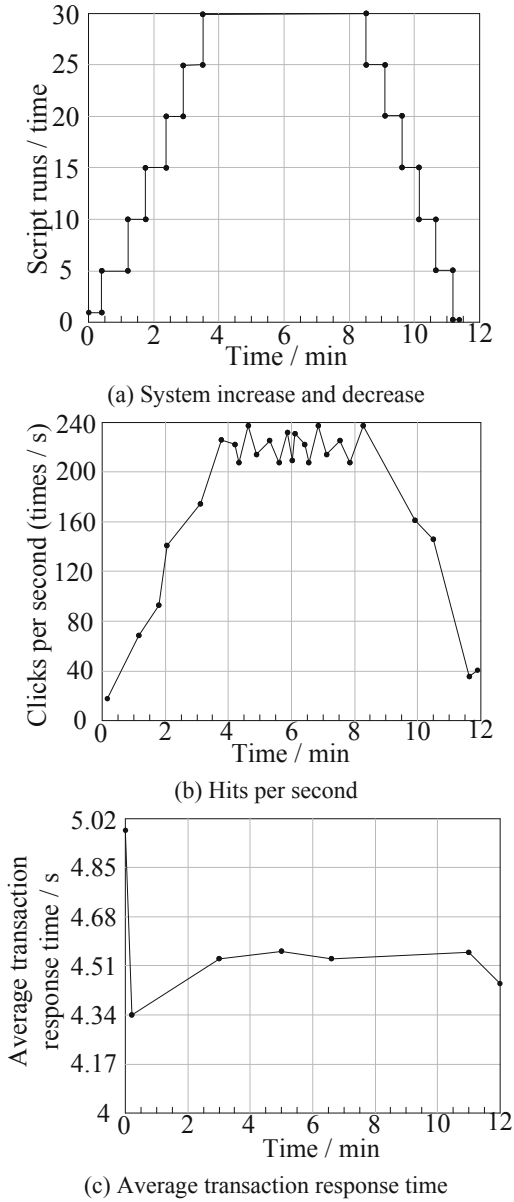
Fig. 4. System load test result graph

overall experimental results, the system in this paper has a good application effect in load and throughput.

Because the methods and theories based on in the research process have been quite mature and perfect, so the research on these theories and methods is grounded and
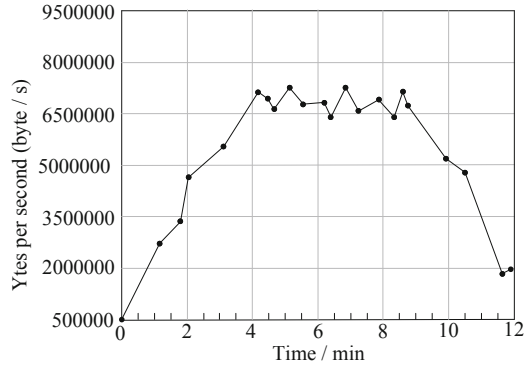
Fig. 5. System throughput test

scientifically feasible. At the same time, the method as an expert system can also provide a certain reference value for other enterprises system evaluation.

References

1. Xie, R.: Design of Java EE based human resource management system with friendly man? machine interaction. Mod. Electron. Techn. **44**(8), 114–118 (2021)
2. Su, B., Liu, S., Lu, Y., et al.: Evaluation of human resource allocation of primary healthcare in China: based on agglomera-tion degree. Chin. J. Health Policy **14**(4), 49–54 (2021)
3. Song, S., Wei, L., Liu, Y., et al.: Application of Data Envelopment Analysis in evaluation human resources efficiency of nursing in department of cardiology. Chin. Nurs. Manage. **21**(5), 745–749 (2021)
4. Feng, T., Zhao, M., Bu, Q.: Practice of the comprehensive evaluation system through combining the workload of RBRVS with specific performance. Chin. Hosp. Manage. **41**(8), 51–53 (2021)
5. Yi, W., Cheng, Z.: Research on the performance evaluation mechanism of new think tank: from the dual perspective of stimulation and constraint. J. Xiangtan Univ. (Phil. Soc. Sci.) **44**(3), 56–60 (2020)
6. Congman, W., Yunye, Z.: Research on the evaluation of human resource supply quality of domestic service in Beijing, Tianjin and Hebei Based on AHP-TOPSIS Model. Contemp. Econ. Manage. **42**(1), 78–83 (2020)
7. Zhu, Y., Yu, W.: Construction of evaluation index system of china's public sports resource allocation level in new era. J. Wuhan Inst. Phys. Educ. **54**(3), 5–12 (2020)
8. Weng, Y., Cheng, L., Bai, Y., et al.: Construction of evaluation index system of nursing human resources efficiency. Chin. Nurs. Res. **33**(11), 1855–1859 (2019)
9. Jia, H., Lijuan, X., Huan, J.: Development of indexes for evaluating nursing human resource efficiency in geriatric department based on data envelopment analysis. J. Nurs. Sci. **34**(14), 53–55 (2019)
10. Chen, L., Wang, C.: Research and simulation of failure-aware algorithms for trusted cloud resource scheduling. Comput. Simul. **37**(10), 334–337+356 (2020)
11. Zhu, B., Chen, G., Li, P., et al.: Design of human resource management system based on B/S mode and MySQL. Mod. Electron. Tech. **44**(14), 65–69 (2021)
12. Xu, J., Liu, C.: Design of evaluation system based on artificial intelligence and deep data analysis technology. Electron. Des. Eng. **29**(6), 179–183 (2021)

Railway Traffic Volume Prediction Method Based on Hadoop Big Data Platform

Pei Su$^{(\boxtimes)}$

Wuhan Railway Vocational College of Technology, Wuhan 430000, China
`udsfg68@163.com`

Abstract. In order to improve the accuracy and efficiency of railway traffic volume prediction, a railway traffic volume prediction method based on Hadoop big data platform is proposed. Firstly, the traffic big data preprocessing mainly includes three parts: redundant data processing, numerical abnormal data processing and missing data processing. Then the spatial cross-correlation characteristics of traffic flow are calculated. Finally, a combined prediction model based on multi features and multifractals is established to realize the railway traffic volume prediction based on Hadoop big data platform. The experimental results show that the prediction method in this study has high prediction accuracy, reduces the prediction time, and meets the needs of method design.

Keywords: Hadoop big data platform · Railway transportation · Transportation volume forecast · Redundant data · Threshold method · Relevance

1 Introduction

Railway transportation is a mode of transportation that uses railway trains to transport passengers and goods. It plays an important role in the process of social material production. It is characterized by large transportation volume, high speed, low cost and generally not limited by climate conditions. It plays an important role in China's economic development and residents' lives. Railway transportation often has problems such as congestion and equipment overload operation. Accurate prediction of railway passenger volume can not only quickly arrange train dispatching and prevent congestion and equipment overload operation, it can also reduce transportation in idle time and achieve the goal of energy conservation and emission reduction.

In order to solve the problems of congestion and equipment overload operation, many scholars have carried out research on the prediction method of railway traffic volume. Among them, reference [1] proposed a medium and long-term high-speed railway network passenger flow OD and channel traffic volume prediction method, which is based on Logit the model builds a passenger flow distribution model, and uses an iterative weighting method to solve the problem to realize the forecast of transportation volume. With the increasing amount of data involved in intelligent transportation, traditional methods cannot achieve better application effects.

The emergence of Hadoop can well analyze and process these data. Hadoop is a distributed architecture, which is studied and developed by the Apache foundation. Users do not need to thoroughly understand the implementation process at the very bottom of the system, so they can write corresponding applications in common programming languages. Use clusters for fast computing and storage. An important part of Hadoop is the system file distributed Hadoop (HDFS). One of the advantages of HDFS is its high fault tolerance and very low hardware requirements. It provides high data rate for application data and is suitable for applications with large data sets. HDFS has wide requirements for POSIX. In the file system, the data reading operation is carried out through streaming. In urban traffic, a large number of traffic information data is generated every day. The emergence of Hadoop HDFS can make good use of these traffic information to reasonably induce urban traffic. Therefore, this paper proposes a railway traffic volume prediction method based on the Hadoop big data platform to achieve high prediction accuracy of railway traffic volume, so as to alleviate the pressure of urban traffic congestion, provide convenience for people's daily life and work and travel.

2   Traffic Big Data Preprocessing

The traditional traffic data preprocessing method is to check whether there is redundancy and missing in the data by manual comparison or using conventional data processing tools, and then find out whether there is numerical abnormal data according to certain criteria. If problems are found, delete, modify and fill them manually according to relevant standards, so as to complete the preprocessing of traffic data. This method is completely feasible when dealing with conventional data, but it may not be fully applicable when dealing with traffic big data.

According to the definition of traffic big data, its scale is large and the data storage files may be scattered. Using traditional data restoration methods to detect and repair quality problems of traffic big data has a large workload and low efficiency, which can not meet the timeliness requirements of traffic management. At the same time, human operation errors may occur, resulting in problem data or threatening the security of data. In addition, due to the data protection policy of HDFS distributed file system, the data cannot be modified directly. Therefore, when processing traffic big data, we should rely on big data technology and traditional data preprocessing methods to detect and repair data.

According to the actual situation of the traffic big data file used in the research, this paper uses Hadoop big data technology to preprocess the traffic big data on the principle of maintaining the original characteristics of the data as much as possible. The processing sequence is carried out according to three steps: data redundancy processing, numerical abnormal data processing and missing data processing.

2.1   Redundant Data Processing

In traffic data, the form of redundant data is that each data item in one data element is exactly the same as the data items in other data elements. The redundant data detection

method for conventional traffic data with a small amount of data is to compare each row of data elements with the data elements of all other rows, and the time complexity is O(n2). When the amount of data is small, the processing method can work normally. However, when the amount of data is very large, this processing method will have serious problems in terms of memory overhead and time overhead. Therefore, the redundant processing of big data still needs to rely on big data technology [2]. Hadoop has its own character statistics function, which can be used to detect and repair redundant data.

1) Use data collection equipment number, collection date and time serial number as keywords for character statistics.
2) The data element with the count of 1 time is normally output once, and the redundant data element with the count of more than 1 time is only output for the first time, and a new data file is generated [3].
3) Replace the old data file with the newly generated data file to achieve the purpose of eliminating redundant data [4].

The specific flow of redundant data processing of traffic big data used in the experiment is shown in Fig. 1.
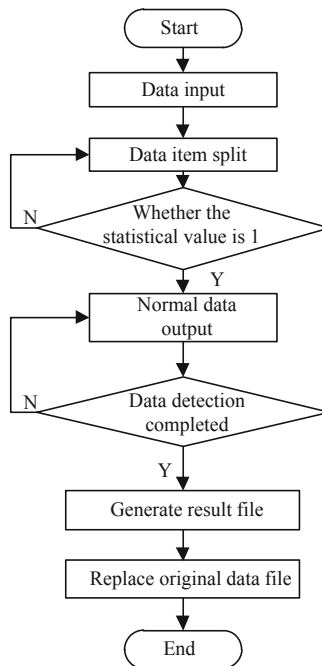
Fig. 1. Preprocessing method for redundant data of traffic big data

This method reduces the comparison times and improves the detection efficiency of redundant traffic data by classifying data items. At the same time, by generating

new data files to replace the original files, it not only conforms to the data protection strategy of HDFS system, but also avoids the problems of low efficiency and human operation errors that may occur in traditional redundant data processing methods. The corresponding pseudocode for redundant data processing is expressed as:

```
df.duplicated() #
df.drop_duplicates(inplace=True) #
```

2.2   Processing of Numerical Abnormal Data

Traffic data is an objective reflection of road traffic operation. In reality, the traffic flow of a specific section cannot exceed the saturation flow rate of its section. Similarly, the average travel speed of the traffic flow will not be much greater than the road speed limit. The data processing method of threshold method is relatively simple, that is, set a reasonable value interval [5] for the corresponding data items, and determine the data items exceeding the reasonable value interval as numerical abnormal data items. However, the threshold method can only detect the data items with significant abnormal values, and the recognition degree of abnormal data contained in the normal value range is not high. However, when the amount of data is large enough, using the threshold method to process data has three advantages. Firstly, it can eliminate the numerical abnormal data items with relatively large interference to the follow-up research, and meet the requirements of data reliability as a whole; Second, the original data characteristics can be retained as much as possible [6] to reduce the generation of artificial noise data; The third is that the algorithm has low complexity, improves the efficiency of data processing, and meets the requirements of practical application timeliness. Therefore, this paper uses the threshold method to determine the numerical abnormal traffic data.

The threshold method discriminates the abnormal vehicle speed data and defines the value range of the average vehicle speed v:

$$0 \leq v \leq c \times v_l \tag{1}$$

In formula (1), v_l represents the legal speed limit of the road and c is the average speed correction factor.

The threshold method discriminates the abnormal traffic flow data and defines the value range of traffic flow q:

$$0 \leq q \leq \varepsilon \times C \times t/60_v \tag{2}$$

Because the HDFS file system has a data protection strategy and cannot directly modify the value of abnormal data, a new method must be adopted for the processing of traffic big data. Firstly, threshold method and traffic flow mechanism method are used to test whether the value of traffic flow parameters is within a reasonable range. For the data elements determined as abnormal values, record the detector number, acquisition date and time serial number as keywords, generate abnormal data identification files, locate the abnormal traffic data, and deal with them together with the missing data in the

474   P. Su

next work. This method not only conforms to the data protection strategy of the system, but also improves the efficiency of traffic big data preprocessing. The corresponding pseudocode for numerical abnormal data processing is expressed as:

```
import pandas as pd
import numpy as np
test_dict = {'id':[1,2,3,4,5,6],'name':['Alice','Bob','Cindy','Eric','Helen','Grace'],'math':[90,'\\N',99,78,97,93],'english':[89,94,80,94,94,90]}

df = pd.DataFrame(test_dict)
df
df.loc[df['math']=='\\N']
df.loc[df['math']=='\\N','math'] = df.drop(1).math.mean()
df
```

2.3 Missing Data Processing

Judging whether the data is missing is the premise of missing data processing. Since the data used in the experiment is the data collected by the microwave detector at a fixed time interval of 2 min, it can be seen that under normal working conditions, the standard data collection volume of each detector per day should be 720 pieces. Therefore, the missing data detection method for traffic big data is to detect whether there is a null value in each row of data elements with the equipment number of microwave detector and data acquisition date as keywords after data redundancy processing. If none is empty, the statistical value is added by 1. The statistical value is based on 720. If it is lower than this value, it indicates that the node is missing data. There have been many studies on how to fix the problem of missing traffic information data. The common processing methods are historical average method, moving average method and exponential smoothing method. The moving average method uses adjacent data items to fill in the missing data, and the historical average method uses the periodic law of traffic flow parameters to repair the missing data with the data of the previous day or week. The exponential smoothing method uses the trend of time series to repair missing data. In the relevant research on missing traffic data repair methods, experiments have proved that in the case of a small amount of missing data, the exponential smoothing method is not much different from the moving average method and the historical average method [7]. Considering comprehensively, this paper adopts the moving average method and historical average method to deal with the missing data of experimental traffic big data.

The formula of moving average method is as follows:

$$y(t) = \frac{[y(t+n) + y(t+n-1) + \ldots + y(t-n)]}{n} \tag{3}$$

The missing data processing method in Hadoop is also a way to generate new files after data repair to replace old files, meet the system data protection strategy, reduce workload and improve efficiency. The corresponding pseudocode for missing data processing is expressed as:

```
df.isnull().sum()
#axis=0 means delete this line, =1 means to delete this column
df.dropna(axis=0,inplace=True)
df.fillna(0, inplace=True) #
df.fillna(df.mean(),inplace=True) #
df.fillna(value={'edu_deg_cd': train_tag['edu_deg_cd'].mode()[0], #
                          'deg_cd':train_tag['deg_cd'].mode()[0],
                               'atdd_type':
train_tag['atdd_type'].mode()[0]},inplace = True)
```

3 Calculation of Spatial Cross-Correlation Characteristics of Traffic Flow

Traffic flow data is a typical time series and spatial geographic data, which has strong correlation in time and spatial dimensions. Therefore, urban road traffic flow has strong temporal and spatial distribution characteristics, showing not only temporal variation characteristics, but also spatial variation characteristics. The spatial variation characteristics of traffic flow mainly include the transverse variation characteristics of the same section and the longitudinal variation characteristics of different sections. Lateral variation characteristics are also simply called spatial correlation, which refers to the variation characteristics of traffic flow in the same section and different lanes; Longitudinal variation characteristics, also known as spatial time lag characteristics, refer to the variation characteristics of traffic flow between upstream and downstream detection sections of the same lane or the same section. In this paper, spatial statistical analysis and correlation analysis will be used to study the spatial correlation and spatial time delay of traffic flow.

Spatial correlation [8] is an important property of spatial geographic data. Its concept is similar to the autocorrelation of time series. It describes the correlation characteristics of a spatial location and its adjacent spatial location in the value of research attributes. Spatial correlation analysis of traffic flow data refers to the spatial correlation characteristics of traffic flow attribute values (such as flow, density or speed). This paper also uses Pearson correlation coefficient to analyze the spatial correlation of traffic flow. If the adjacent spatial location is close in the value of the research attribute, the value of Pearson coefficient is large, indicating that the correlation between the two spatial locations is strong; Otherwise, the spatial location has weak or no relevance. Suppose $X\{x_i|i = 1, 2, \cdots n\}$ and $Y\{y_i|i = 1, 2, \cdots n\}$ represent two traffic flow sequences at different spatial locations respectively, $\bar{x}$ and $\bar{y}$ represent x and y mean values respectively, in which the spatial correlation number of d order delay is an extension of Pearson correlation coefficient [9], which is expressed as:

$$\rho_{xy} = \frac{\sum_{n}^{i=1}(x_i - \bar{x})(y_{i-d} - \bar{y})}{\sqrt{\sum_{n}^{i=1}(x_i - \bar{x})^2}\sqrt{\sum_{n}^{i=1}(y_{i-d} - \bar{y})^2}} \tag{4}$$

Based on the above process analysis, the spatial correlation analysis of traffic volume provides a basic basis for traffic flow prediction.

4 Realization of Railway Traffic Volume Prediction
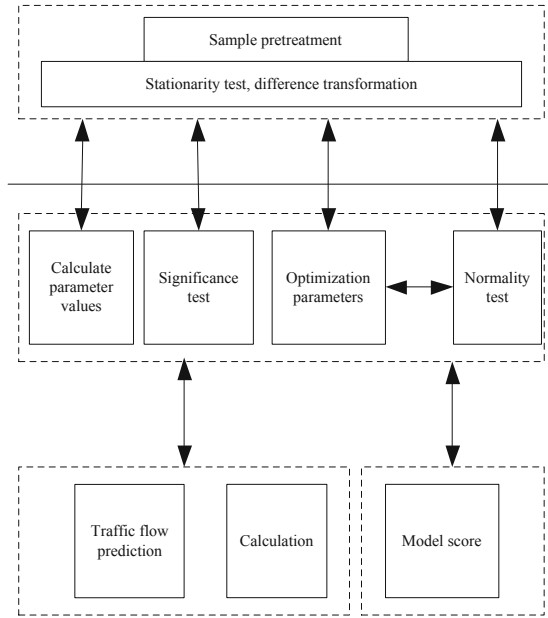
The modeling process is shown in Fig. 2:

Fig. 2. Modeling process

The conditional variance of traffic flow time series exists, that is, the possibility of Heteroscedasticity in the actual traffic flow series. The GARCH model conducts modeling research on residual variance, solves the modeling problem of residual variance, and can simulate and predict the nonlinear volatility of traffic flow sequence. The GARCH model was proposed by Bolseslev in 1986. It is an extension of the autoregressive conditional heteroscedasticity model ARCH. It solves the problem of the high order of the ARCH model and the difficulty of estimating the parameters due to the long-term autocorrelation of the residual sequence conditional heteroscedasticity in practical applications. The problem is a very important model for processing time series data [10]. Express the calculation formula as:

$$\begin{cases} \sigma_t^2 = \alpha_0 + \sum_{u_t=\sigma_t e_t}^{j=1} \beta_j \sigma_{t-j}^2 + \sum_{v}^{i=1} \alpha_i \varepsilon_{t-i}^2 \end{cases} \tag{5}$$

In formula (5), α represents the autoregressive order, β_j represents the residual sequence, σ is the lagging sample variance coefficient, and e is the lagging conditional variance coefficient.

The periodicity, trend, linear correlation and random volatility of traffic flow and the complexity of traffic system jointly determine that accurate prediction can not be solved by a single model or method. A single prediction model has its advantages and disadvantages, scope of application and application conditions under specific circumstances, it is impossible to deeply mine the characteristics of traffic flow and make more accurate prediction. So far, no prediction method can show better performance than all other methods. Combining each advantageous single model according to a certain method, comprehensively maximizing the use of the useful information of a single model and comprehensively understanding the predicted object will help to improve the prediction effect and improve the prediction accuracy. Moreover, to predict traffic flow more accurately, it not only needs methodological innovation, such as cloud computing, big data, deep learning, the internet of things and combined forecasting [11]; It also needs to have a deeper understanding of the internal characteristics of traffic flow, such as spatio-temporal analysis [12], multifractal analysis, statistical analysis and so on. Therefore, this paper proposes a combined traffic flow forecasting method that integrates the time series forecasting models ARIMA and SARIMA and the time series fluctuation forecasting model GARCH, using spectrum analysis, time series and statistical fluctuation analysis methods to fully explore the temporal and spatial characteristics of traffic flow. According to the characteristics of self-similarity, long-term memory, and self-similarity of the traffic flow itself, it is proposed to decompose the traffic flow time series into periodic items, trend items and random fluctuation items, and combine and forecast the characteristics of different items thought of. This method is different from the traditional method, which simply and subjectively assumes that the time series of traffic flow meets a certain mathematical model or regular distribution, but uses the periodic and random fluctuation characteristics of traffic flow itself, uses different models to predict respectively, and finally recombines. Because after the non-stationary time series are transformed into stationary time series by difference, the model is constructed by returning the dependent variable only to the present value and lag value of its lag term and random error term, which is very suitable for the prediction of non-stationary single variable time series. Therefore, ARIMA and SARIMA are combined with GARCH.

A combined forecasting model based on multi feature and multifractal is established. The model considers the periodicity, trend, linear and nonlinear characteristics of traffic flow caused by many factors, and analyzes the combined forecasting method of modeling for each sub item. The schematic diagram of the combined model is shown in Fig. 3.

Firstly, the method decomposes the traffic flow sequence into four sub items: the periodic sub item is represented by $P(t)$, the trend sub item is represented by $T(t)$, the linear sub item is represented by $L(t)$ and the nonlinear sub item is represented by $N(t)$. Then a traffic flow sequence can be represented by the cumulative sum of $X(t)$ sub items, as shown in the following formula:

$$X(t) = P(t) + T(t) + L(t) + N(t) + \varepsilon_t \tag{6}$$

In formula (6), $P(t)$ and $T(t)$ are the determined components of traffic flow, ε_t represents the error term, and $L(t) + N(t) + \varepsilon_t$ is the uncertain component of traffic flow [13], representing the random fluctuation of traffic flow.
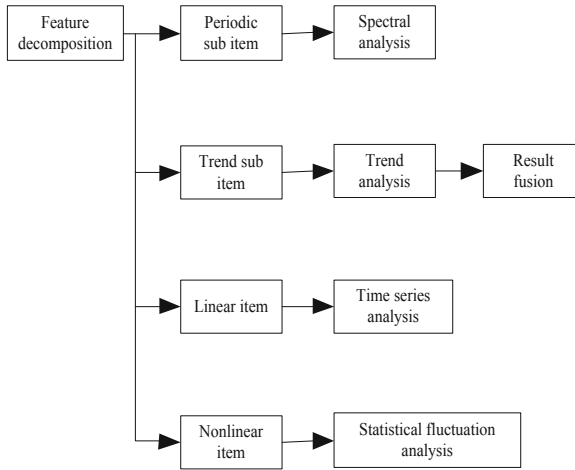
Fig. 3. Schematic diagram of combined model

Due to many random factors, a large amount of information with many influencing factors and complex periodic rules is divided into several modules, and a multi module weighted neural network prediction model based on the characteristics of time series is established. Different from the traditional RBF neural network, RBF neural network mixes the data information with different time series characteristics as an independent processing unit, which is input respectively for multi module comprehensive analysis and training learning. The feature layer output of each module is weighted to obtain the final prediction result. It avoids the nonlinear optimization problems such as blindness and local optimization in the design of single kernel function, improves the prediction accuracy and solution performance.

The principle of model improvement is as follows:

(1) Given historical data and preprocessed according to the predicted demand;
(2) Make statistical analysis on the change trend of historical data and specific cycle law, and consider the impact of various influencing factors on passenger flow;
(3) The input layer receives data samples with certain time series characteristics and establishes the corresponding prediction model single kernel function;
(4) Nonlinear analysis is carried out on the relevant feature data, and the weights between the input layer and the feature layer are partially connected, rather than all connected, that is, each module is trained and learned separately;
(5) The integration layer integrates the data and information with certain temporal characteristics. The number of integration layers is the arithmetic average of the number of neurons in the feature layer and the output layer, which is completely connected to the feature layer and the output layer to realize the exchange of information;
(6) The neurons in the integration layer are completely connected with the output layer to produce the final output of the network. The RBF network takes the predicted value of the modular model as the input value, and the actual output value is the training value of the RBF network until the predetermined error is reached. Finally, the multi modular network structure is determined.

The detailed prediction process of the model is as follows:

Step 1: Define error function:

$$E = \frac{1}{2} \sum_k \left(y_k - y_k' \right)^2 \tag{7}$$

In formula (7), y_k represents the expected output of the k sample point; y_k' represents the actual output of the k sample point.

Step 2: Given preset error.

Step 3: Select the radiation basis kernel function. On this basis, Gaussian function is selected as the kernel function:

$$\phi_i(x) = \exp\left[-(x - c_i)^2 / 2\sigma_i^2 \right], i = 1, 2, \ldots, h \tag{8}$$

In formula (8), x represents the input sample, c_i represents the center of the i unit in the feature layer, and σ_i^2 represents the width of the i unit in the feature layer.

Step 4: Model training. The improved model is trained to obtain the weight from the integration layer to the output layer. Specific algorithms:

$$y = \sum_m^{l=1} w_l \phi_1(x), y / w^T \varphi \tag{9}$$

In formula (9), y represents the expected output of sampling, φ represents the number of neurons in the composite layer, and w^T represents the weight of the T neuron output layer in the composite layer.

Through the training of the model, the trained model is used for prediction, and the prediction results are obtained. The model comprehensively considers the long-term invariance and short-term time-varying characteristics of traffic flow, and can show and mine the evolution law and fluctuation characteristics of traffic flow to a great extent.

5   Experimental Comparison

In order to verify the effectiveness of the proposed railway traffic volume prediction method, experiments are carried out. The data used in this experiment is China Statistical Yearbook network, which mainly includes railway passenger volume, total population, number of domestic tourists and other data. The experiment is carried out based on the support of these data. Due to the large amount of data, in order to improve the experimental speed and accuracy, the experimental environment is established, as shown in Table 1.

The experimental process is shown in Fig. 4.

Table. 1 Experimental hardware configuration

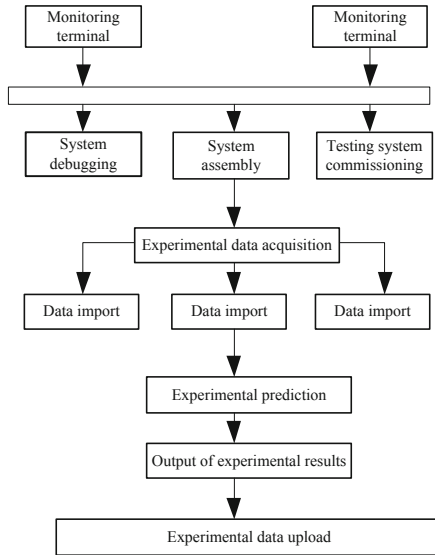| Database server	Client computer	Network environment	Isolation device
CPU: Pentium core4 I7	CPU: Pentium core2 I7	100M/10M network card	Reverse physical isolation device
Memory: DDR3–1667	Memory: DDR3–1667		
Hard disk: 2TG	Hard disk: 1TG		

Fig. 4. Experimental process

Based on the above unified experimental environment, the railway traffic volume prediction method based on Hadoop big data platform is taken as the research method, and the medium and long-term high-speed railway network passenger OD and channel traffic volume prediction method proposed in reference [1] is taken as the traditional method. According to the above process experiment, the research method is compared with the traditional method. The specific experimental contents are as follows.

5.1 Comparison of Vehicle Speed Prediction Results

Draw a broken line diagram between the predicted speed and the actual data, and the comparison of speed prediction results is shown in Fig. 5.

Based on Fig. 5, it can be seen that the road speed at the location of the detector shows an upward, downward and upward trend during the morning peak hours, indicating that there is traffic congestion, and then it gradually returns to smooth. The predicted speed is generally consistent with the actual situation, and the changes of traffic operation during morning peak hours can be predicted.
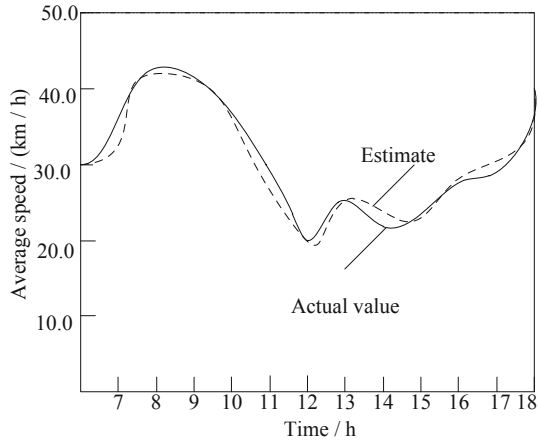
Fig. 5. Comparison of vehicle speed prediction results

5.2 Comparison of Prediction Deviation of Maximum Continuous Traffic Operation

The maximum continuous prediction deviation between the prediction results and the actual traffic operation conditions is shown in Table 2.

Table 2. Comparison of prediction deviation of maximum continuous traffic operation

Time	Predicted speed (km/h)	Actual speed (km/h)
9:04	47.70	57.4
9:06	60.55	53.1
9:08	49.68	66.2
9:10	54.69	60
9:12	49.51	64.5
…	…	…
9:30	52.91	63.1
9:32	49.75	63.9
9:34	55.91	64.3
9:36	55.57	57.5
9:38	53.04	64.4
9:40	50.93	59.1
9:42	45.30	59.6

Based on Table 2, it can be seen that there are 12 continuous time points of prediction deviation between the prediction results of traffic operation conditions and the actual situation, and the fluctuation trend of actual speed and predicted speed shows a sawtooth deviation.

The comparison results of the maximum continuous traffic operation prediction deviation between the studied method and the traditional method are shown in Fig. 6
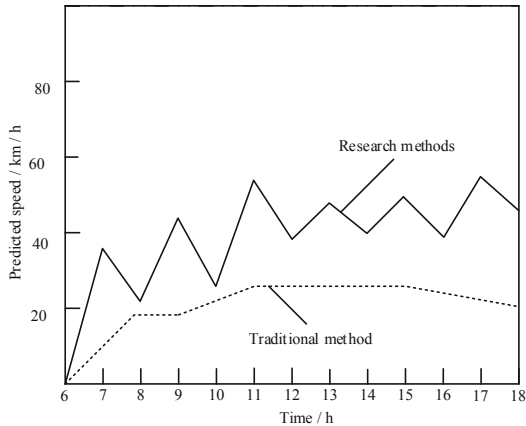
Fig. 6. Comparison results of prediction deviation of maximum continuous traffic operation

According to Fig. 6, the prediction speed of the research method is consistent with the actual speed, while the prediction speed of the traditional method is quite different from the actual speed. Therefore, the prediction deviation of the maximum continuous traffic operation condition of the research method is small.

5.3   Comparison of Prediction Errors of Traffic Volume

The comparison results of the prediction errors between the studied method and the traditional method are shown in Fig. 7.

Based on Fig. 7, it can be seen that the prediction error of the research method is relatively small, which can be less than 5, and has high prediction accuracy, while the prediction error of the traditional method is high, which has no good application effect than the proposed prediction method. Because the research method processes the data in advance and calculates the spatial cross-correlation characteristics of traffic flow, the prediction error is reduced.
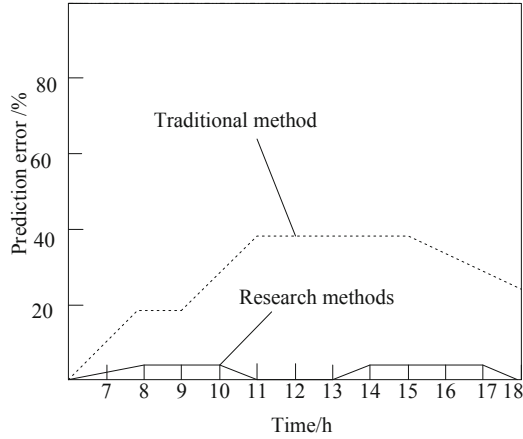
Fig. 7. Comparison of prediction errors of traffic volume

5.4 Comparison of Prediction Time

The prediction time of the research method and the traditional method is compared, and the comparison results are shown in Fig. 8.
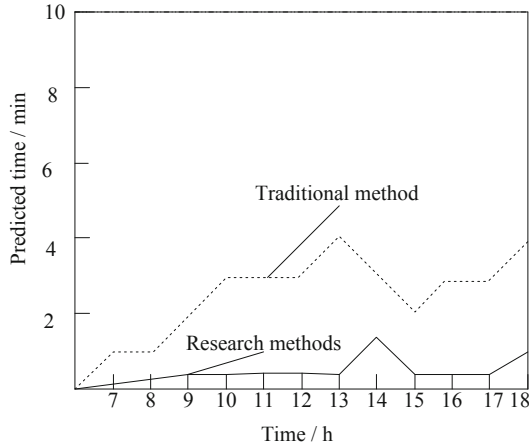

Fig. 8. Comparison of prediction time

Based on Fig. 8, the prediction time of the research method is less, and the prediction can ensure a shorter prediction time in each time period. The prediction stability of the traditional method is low, the time spent in some time is less, and the time spent in some time prediction is more, which is worse than the prediction effect of the proposed traffic volume prediction method. The reason for the poor effect of traditional methods may be that there is no preprocessing of redundant data, which is greatly disturbed by the data, thus reducing the application effect.

5.5   Comparison of Prediction Accuracy

The prediction accuracy of the research method and the traditional method are compared, and the comparison results are shown in Table 3.

Table 3. Comparison of prediction accuracy between the research method and the traditional method

Time	Predicted speed (%)	Actual speed (%)
9: 04	98.6	82.6
9: 06	97.8	81.9
9: 08	96.5	83.5
9: 10	96.8	82.3
9: 12	97.1	80.9
…	…	…
9: 30	98.2	81.4
9: 32	98.5	82.6
9: 34	97.6	82.4
9: 36	96.8	80.8
9: 38	96.4	81.6
9: 40	97.6	83.6
9: 42	98.5	84.7

According to Table 3, the prediction accuracy of the research method is higher, and the prediction can guarantee a higher prediction accuracy in each time period, while the prediction accuracy of the traditional method is lower. It can be seen that the research method can effectively improve the prediction accuracy.

6   Conclusion

This paper proposes a prediction method for railway traffic volume based on Hadoop big data platform. By preprocessing traffic big data and calculating the spatial cross-correlation characteristics of traffic flow, a combined prediction model based on multi-feature and multi-fractal is established to realize railway traffic volume prediction. It is verified by experiments that the research method not only improves the accuracy of forecasting, but also improves the efficiency of traffic volume forecasting, which has certain practical application significance.

References

1. Long, W.: Prediction method of high-speed rail passenger OD flow and traffic volume in medium and long-term high-speed railway network plan. J. Beijing Jiaotong Univ. **44**(4), 76–85 (2020)

2. Jiang, Y.: Simulation of multi-dimensional discrete data efficient clustering method under big data analysis. Comput. Simul. **36**(02), 205–208 (2019)
3. Iqbal, B., Iqbal, W., Khan, N., Mahmood, A., Erradi, A.: Canny edge detection and hough transform for high resolution video streams using hadoop and spark. Cluster Comput. **23**(1), 397–408 (2019). https://doi.org/10.1007/s10586-019-02929-x
4. Teng, L., Li, H., Yin, S., Sun, Y.: A modified advanced encryption standard for data security. Int. J. Network Secur. **22**(1), 112–117 (2020)
5. Chawla, S., Shahu, J.T., Gupta, R.K.: Design methodology for reinforced railway tracks based on threshold stress approach. Geosynth. Int. **26**(2), 111–120 (2019)
6. Sun, G., He, S., Fu, H., Xie, J., Zheng, L.: Study on shaking table test method for seismic responses of bridge-tunnel lapped structure in weak surrounding rocks. Tiedao Xuebao/J. China Railway Soc. **41**(1), 117–125 (2019)
7. Zhang, J.: Research on adaptive recommendation algorithm for big data mining based on hadoop platform. Int. J. Internet Protoc. Technol. **12**(4), 213–220 (2019)
8. Li, R., Huang, Y., Wang, J.: Long-term traffic volume prediction based on type-2 fuzzy sets with confidence interval method. Int. J. Fuzzy Syst. **21**(7), 2120–2131 (2019)
9. Gao, K., Han, F.R., Wen, M.F., Du, R.H., Li, S., Zhou, F.: Coordinated control method of intersection traffic light in one-way road based on v2x. J. Central South Univ. **26**(9), 2516–2527 (2019)
10. Yuan, W., Wang, J.: High mobility sparse channel estimation method based-on DCS-KF. Tiedao Xuebao/J. China Railway Soc. **41**(1), 74–79 (2019)
11. Liu, S.B., W, & Liu, G.: Parallel fractal compression method for big video data. Complexity **2018**, 2016976 (2018)
12. 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
13. 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

Intelligent Forecasting Method for Substation Operating Cost of Power Network with Nonstationary Characteristics

Shaohong Lin[1(✉)], Ying Wang[2], Xuemei Zhu[2], Ye Ke[2], and Meihua Zou[2]

[1] State Grid Fujian Electric Power Co, Ltd., Fuzhou 350007, China
oijsrg54@163.com

[2] State Grid Fujian Power Economic Research Institute, Fuzhou 350012, China

Abstract. With the rapid development of the national economy, the power system has been greatly expanded. The operation cost of power grid substation also presents a trend of sharp increase, which brings great challenges to power system. In order to manage and control the operation cost of substation effectively, the intelligent prediction method of power system operation cost with non-stationary characteristics is proposed. Based on the MCSSD method, this paper extracts the non-stationary features of substation operation information, analyzes the economic load rate of substation, determines the substation capacity, constructs the intelligent forecast model of substation operation cost according to the operation law of substation equipment life cycle, and obtains the forecast result of operation cost, and based on this, formulates the intelligent forecast and control measures of substation operation cost. The experimental data shows that after the method should be proposed, the average time for forecasting the operation cost of substation is 11.6, which is lower than the maximum limit, and the highest prediction accuracy is 83%. The results of the two indicators are in line with the standard, which fully confirms that the proposed method has a better forecasting effect of substation operation cost.

Keywords: Non-stationary characteristics · Power grid · Substation · Operating cost · Intelligence · Prediction

1 Introduction

In recent years, with the rapid development of national economy, the state has accelerated the construction of power facilities. By 2020, the State Power Grid will promote the optimal allocation of power resources in a wide range, complete the cross-regional and cross-provincial power transmission capacity of 605.489 billion kilowatt hours, complete the electricity-sharing tasks for 15,300 households and 4937,000 people without electricity, comprehensively narrow the power supply gap between urban and rural areas, and invest 80.74 billion yuan in completing rural network projects. Power enterprises not only actively fulfill their social responsibility, but also pursue the maximization of economic, social environment and comprehensive value under the requirement of promoting

sustainable development. Therefore, on the basis of ensuring the safety and reliability of power supply, substation projects should consider not only the initial construction cost of the project, but also the operation and maintenance costs, so as to minimize the investment cost [1].

For a long time, the fierce competition situation has been existing in the power grid construction industry, and the high quality and fast completion of tasks has become an important means for all transformer construction enterprises to participate in market competition. Due to the increase of labor costs year by year, it is difficult to improve the construction operation level to a large extent in a short term, so the profit margin of the whole industry is low [2]. Under the current market situation, the substation construction enterprise can only change the management idea and mode, on the one hand, it can adapt to the demand and change of the power grid project construction to enlarge the market share, on the other hand, it must choose the scientific and effective tools and methods of cost forecast and control, and combine with the actual management to apply, so as to enhance the cost leading competitiveness of the enterprise.

Substation is the process of power transmission through transformers. A transformer is a device that uses the principle of electromagnetic induction to change the alternating voltage. The main components are the primary coil, the secondary coil and the iron core (magnetic core). Therefore, the intelligent prediction of substation operation cost mainly aims at transformer. The collected substation information is non-stationary, which brings great difficulty to the forecast of substation operation cost. Therefore, the research on intelligent forecast method of substation operation cost is proposed. Based on the extraction of non-stationary features of substation information, the substation operating cost can be accurately predicted to provide accurate data support for the control of substation operating cost.

2   Research on Intelligent Forecasting Method of Power Grid Substation Operation Cost

2.1   Nonstationary Feature Extraction of Power Grid Substation Information

In order to accurately predict the operation cost of power grid substation, the first task is to extract the non-stationary characteristics of power grid substation information. According to the above requirements, this research based on Multi-scale Chirplet Sparse Signal Decomposition (MCSSD) method extracts the non-stationary features of power grid substation information as follows:

Based on the theory of MCSSD, it can be found that the multi-scale characteristic of LFM base function makes it match the analysis signal dynamically, and solves the problem of single base function in whole analysis time. The frequency slope information contained in the LFM basis function makes it suitable for analyzing the non-stationary signal whose frequency is curvilinear, and preserves the multi-scale property of the wavelet transform. The multi-scale LFM sparse signal decomposition method can decompose a multi-component signal into a number of single-component signals with physical significance. Multi-modal non-stationary signals in substation can be decomposed into single-modal signals according to frequency, so that non-stationary signals

can be decoupled in time domain [3]. The signal results decomposed by the MCSSD method are shown in Fig. 1.
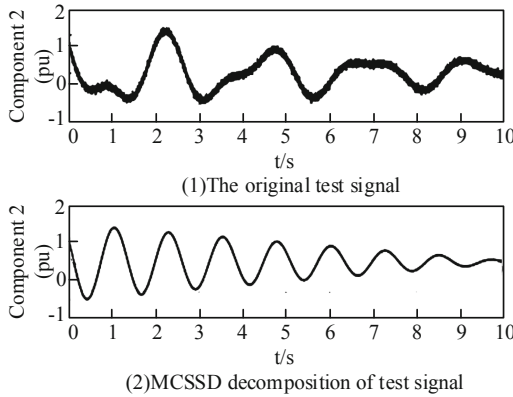

Fig. 1. MCSSD decomposed signal results

In summary, MCSSD is particularly suitable for the decomposition of multimodal coupled nonstationary signals. The power nonstationary signal of power system is essentially a response signal of multimodal time-varying vibration system. Therefore, MCSSD can be used to select single modal vibration component from multimodal nonstationary signals. The single-mode oscillation response can be approximately expressed as a formula (1) by the product of a pure FM signal and an envelope signal.

$$
\begin{aligned}
x_i(t) &= A_i(t)\cos(\omega_{di}(t)t + \theta_{i0}) \\
&= A_{i0}e^{-\xi_i\omega_i t}\cos\left(2\pi f_i\sqrt{1 - \xi_i^2}\,t + \theta_{i0}\right)
\end{aligned}
\tag{1}
$$

In Formula (1), $A_i(t)$ represents the amplitude envelope of the i substation signal, $A_i(0) = A_{i0}$. $\omega_{di}(t)$ represents the damped modal frequency of the i substation signal. t stands for time. θ_{i0} represents the initial phase of the i substation signal. ξ_i represents the damping ratio of the i substation signal. f_i represents the frequency of the substation signal of the i-th grid.

Take the logarithm of the amplitude envelope of the i substation signal, and get the formula (2).

$$
\ln A_i(t) = -\xi_i\omega_i t + \ln A_{i0}
\tag{2}
$$

The mathematical meaning of the expression (2) is: $\ln A_i(t)$ is a function of time t of the first degree, with a slope of $\xi_i\omega_i$. The $A_i(t)$ and $\omega_{di}(t)$ of the i-decomposed signal can be obtained by MCSSD decomposition. Subsequently, the frequency $f_i(t)$ and the modal damping ratio $\xi_i(t)$ of the i-th decomposed signal are obtained by combining formula (1) and formula (2), respectively, as shown in formula (3).

$$\begin{cases} f_i(t) = \frac{1}{2\pi}\sqrt{\omega_{di}^2(t) + \left(\frac{d\ln A_i(t)}{dt}\right)^2} \\ \xi_i(t) = \frac{\frac{d\ln A_i(t)}{dt}}{\sqrt{\omega_{di}^2(t)+\left(\frac{d\ln A_i(t)}{dt}\right)^2}} \end{cases} \tag{3}$$

It should be pointed out that because of the influence of noise, $\ln A_i(t)$ and time t are not strictly linear relations, $\ln A_i(t)$ approximation is a straight line, can be fitted by the least squares method. Using the relations mentioned above, the oscillation characteristic parameters $f_i(t)$ and $\xi_i(t)$ can be identified and extracted from the component signals derived from MCSSD, which is the non-stationary characteristic of substation signals.

Through the above, the non-stationary characteristics of power system substation signals are extracted, which provides support for the subsequent economic load analysis of substation [4].

2.2   Economic Load Analysis of Substation

The economical load rate of substations is always changing during the whole process from being put into production until it is finally withdrawn from operation, generally showing a law from small to large. The trend of running costs shows a change from large to small and then from small to large. In order to optimize the operation cost of transformer substation, the transformer substation shall operate within a certain load rate range [5]. The study of operation cost is based on reliability, and the relationship between economic load rate and operation cost is shown in Fig. 2.
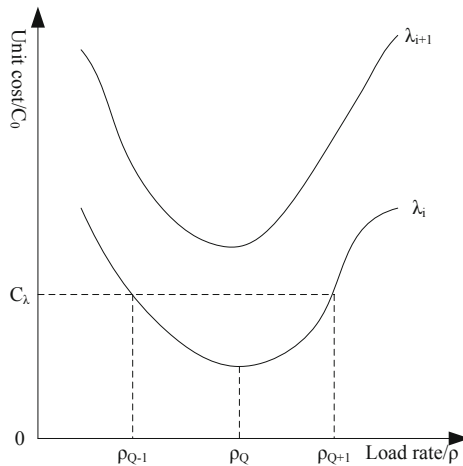
Fig. 2. Relationship curve between economic load rate and operating cost

As shown in Fig. 2, unit cost C_0 increases with the increase of reliability; under certain conditions of reliability λ_i, there exists load rate ρ_Q, making unit cost C_0 lowest and substation operation cost optimal; when load rate from 0 to ρ_Q gradually increases, unit cost C_0 gradually decreases; when load rate exceeds ρ_Q, unit cost C_0 gradually increases with the increase of load rate; within acceptable unit cost C_λ, economic load rate has a certain optimization range, the lower limit of economic load rate can prevent excessive waste of resources, and the upper limit can guarantee the safe and reliable operation of equipment [6].

The formula for calculating the economic load rate of power grid substation shall be:

$$\rho_Q = \sqrt{\frac{P_0 T + k_Q Q_0 T}{P_k \tau_{\max} + k_Q Q_k \tau_{\max}}} \tag{4}$$

In Formula (4), ρ_Q refers to the economical load rate of the power grid substation; P_0 refers to the no-load loss of the main transformer; T refers to the operation cycle; k_Q refers to the economic equivalent of reactive power, wherevers to the coefficient of converting the reactive power loss into active power loss according to the economic cost. In practical calculation, reactive power economic equivalent k_Q can be simply calculated as the ratio of reactive power price to active power price; Q_0 means reactive power no-load loss; P_k means load loss of main transformer; $\tau_{\max}$ means annual loss hours; Q_k means reactive power load loss.

According to the design requirements of the main transformer capacity in the Electrical Design Manual of Power Engineering, the capacity of other transformers should be 70%–80% of the total capacity when one main transformer is out of service. The load rate of the main transformer determined according to the design requirements is called safe operation load rate [7]. In general, for a substation with several transformers of the same capacity, if the load rate of the remaining transformers does not exceed 130% after a main transformer is switched off due to failure, the load rate of each transformer shall be:

$$\rho_a = \rho_k \times N/N - 1 \tag{5}$$

In Formula (5), ρ_a represents the safe operating load rate; ρ_k represents the short-time allowable overload rate; and N represents the number of transformers.

Through the above process, the calculation of the economic load rate of the power grid transformation is completed, and sufficient data support is provided for the subsequent determination of the power transformation capacity.

2.3 Determining Process of Substation Capacity

For the substation operation cost forecast, it mainly aims at the component main transformer. The capacity and the number of main transformers have a significant impact on the grid structure, power supply security and reliability and economy. The selection of the capacity and number of main transformers generally depends on the nature of the load in the planning area, the level of demand for reliability of power supply, the rate of

growth, the capacity of the next higher grid or power plant to provide the load and the technical and performance indicators of the distribution devices connected thereto, the cost per unit capacity and the system short-circuit capacity and transport and installation conditions [8].

At present, there is no specific provision on the selection of the number and capacity of transformers in the power grid system in China, which is usually carried out in accordance with the relevant rules and regulations and in combination with the experience of designers. Based on the application of the prediction theory of operating costs, the following circumstances shall be considered:

(1) The capacity of the main transformer is determined according to the nature of the planned load and the grid structure. The planned capacity of newly-built substation transformer should meet the demand of 5–10 years planning load development of power supply area, and avoid unnecessary expansion and capacity increase. In order to avoid the waste of resources when the initial load of a substation is relatively small, the time to put several main transformers into operation may be arranged according to the demand for load growth [9].

(2) Where there are important users in the planned power supply area of a substation, the power supply load of the first and second grade users shall be guaranteed within the allowable time for overload of other transformers under the condition that one main transformer is overhauled or fails. For a substation with normal load, any outage of a transformer shall ensure that 70%–80% of the total load of the power supply is not affected.

(3) The capacity of a single transformer should not be too large or too small, and the space for expansion due to load increase should be reserved. And in order to ensure the power supply mode of operation convenient, should consider the use of multiple transformers.

(4) Although the loss of large capacity main transformer is lower than the cost of unit capacity, but the requirement of matching equipment is also high, it may be difficult to compensate for the increased investment. Therefore, the economic operation mode of the main transformer should be calculated when the transformer capacity is selected, and the operation cost of the main transformer and its supporting devices should be considered. The reasonable design scheme of the main transformer capacity and the number of units is chosen to control the load rate at the economical load rate.

(5) The reasonable design scheme of the main transformer capacity and the number of units is chosen to control the load rate within a certain range of the economic load rate. A lower limit value of economic load rate can prevent excessive waste of resources caused by low load rate of main transformer, and determine the upper limit value to ensure the safe and stable operation of the system.

In this paper, a more economical selection method is proposed based on the reliability of operation and the theory of life cycle cost. The flow chart is shown in Fig. 3.

Use the process shown in Fig. 3 to determine the substation capacity and make intelligent predictions for the subsequent substation operating costs.
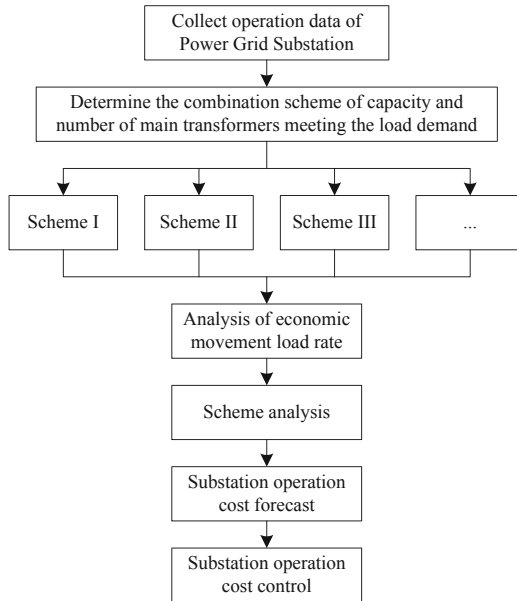
Fig. 3. Flow chart for determination of substation capacity

2.4 Intelligent Prediction of Substation Operating Cost

According to the operation law of the equipment life cycle, the standard operating status and key control points are the focus of the whole process management, according to the life cycle cost theory and the relevant expenses of the substation equipment, an intelligent prediction model of the substation operation cost can be constructed. Which is:

$$LCC = IC + OC + MC + FC + DC \qquad (6)$$

In formula (6), LCC represents the total operating cost of the substation equipment during the entire life cycle; IC represents the initial investment cost, which mainly includes equipment purchase costs, installation engineering costs, land acquisition costs, etc.; OC represents Operating costs, mainly including equipment energy costs, status inspection costs and labor costs, etc.; MC represents maintenance and overhaul costs, including daily maintenance costs, planned overhaul costs and personnel training costs, etc.; FC represents failure costs, including power outages Loss costs, troubleshooting costs and penalty costs, etc.; DC shows equipment scrap processing costs, including equipment decommissioning processing costs and equipment residual value.

Among them, operating costs are mainly variable costs of substation engineering projects, which mainly include three aspects, operating costs, overhaul and maintenance costs, and failure costs.

n-year operating costs $=$ annual daily inspection costs $+$ n-year equipment energy costs $=$ annual inspection labor costs $+$ n-year equipment energy costs;

Overhaul and maintenance costs $=$ periodic disassembly and overhaul costs $+$ various types of periodic overhaul and maintenance costs;

Failure cost = Troubleshooting costs + Failure cost.

The above process completes the intelligent prediction of substation operation cost, and provides accurate data support for substation operation cost control.

2.5 Cost Control of Substation Operation

Based on the forecast data of substation operation cost, the paper controls the substation operation cost by earned value method, so as to ensure the normal operation of substation project, minimize the operation cost and increase the competitiveness of substation enterprises.

The steps of substation operation cost control are as follows:

Step 1: Run work structure decomposition

Work structure decomposition is of great importance in project earned value management. Scientific and reasonable work structure decomposition is the basic criterion of operation plan, and also optimizes the allocation of resources to improve the work efficiency. The work breakdown structure decomposes the project level by level according to the scope of the construction project and optimizes it layer by layer [10].

The first floor of the substation operation project consists of six parts: earthwork, foundation works, tower works, wiring works, accessory installation and acceptance. The earth and stone works mainly aim at some basic civil engineering works; the basic works mainly include some basic works, such as digging, digging, pouring, maintenance and backfilling; the tower pole works mainly focus on the tower pole assembly, mainly aim at the tower pole selection, tower pole transportation, tower pole component, lightning protection, grounding device and the selection of relevant equipment; the wire erection works mainly include the selection of conductors and ground wires, the wear of conductors, the construction and placement of hydraulic tubes and the selection and configuration of metal tools; the acceptance link is relatively simple, mainly to check whether the whole project conforms to relevant standards.

Step 2: Operation schedule

In order to better achieve the specific requirements of operating cost control, the operation management department has made the overall planning of the operation schedule according to the requirements of construction period and production efficiency. Considering the natural conditions, social conditions, weather conditions, the working efficiency of the builders and the unpredictable emergencies, it is necessary to refine and decompose the target of the operation in order to ensure the smooth development and implementation of the operation.

Step 3: Analysis of operating cost control process

The cost control of substation operation needs to consider all kinds of factors comprehensively and meticulously. The operating cost and schedule deviation of earned value method are explained in detail and systematically, which has overall control over the operation and implementation situation. Through collecting and sorting out the deviation data in the course of actual operation, we can supervise and manage the construction period, the cost, the quality of the operation and so on according to the specific situation of the operation deviation, and finally find out the basic reasons of the deviation of all parties in the operation, and optimize the scheme of the operation on this basis, so as to achieve the expected final effect of cost control.

In order to achieve the ultimate goal of cost control, the implementation parties of substation operation have made a concrete analysis according to the actual operation situation before the project construction, and made a detailed cost control plan, as well as the corresponding emergency measures and solutions. In the operation cost control plan, the investment arrangement should not only accord with the actual situation of the operation, but also consider the unexpected situation. In the development of operating cost control plan should be refined to the month, or even in some operational links need to be refined to the day, such a cost control plan is more reasonable.

The main methods for monitoring operating cost expenditure and progress include routine monitoring and regular monitoring. Day-To-Day monitoring is the observation and recording of aspects of the day-to-day state of operations. It relates not only to the commencement time, completion time, duration and actual cost expenditure of the operations, but also to the specific allocation of human, material and financial resources. Regular monitoring is mainly in a fixed date for a certain period of time to monitor the state of operation. Regular monitoring is based on routine monitoring but is more systematic and standardized. In the combination of daily monitoring and regular monitoring of statistical data formed on the basis of a greater authority and scientific nature, more in line with the actual situation.

Through the above process, the intelligent prediction and control of the operation cost of the power grid substation is realized, which provides effective support for the stable operation of the substation. It also reduces costs for the overall operation of the power grid, boosts the overall development of the power system, and provides residents with better power supply.

3   Experiment and Result Analysis

3.1   Experiment Preparation Stage

Table 1. Power demand forecast table for planning area

Year	Average growth rate	Maximum load
2015	5%	265 MV
2016	5%	278 MV
2017	6%	290 MV
2018	8%	301 MV
2019	9%	310 MV
2020	8%	338 MV

In order to verify the application performance of the proposed method, the experimental objects are selected and their basic conditions are introduced, as follows:

A 220 kV substation is planned to be constructed, with a safe operation life of 30 years, a base load of 35MW and an annual load growth rate of $\gamma(t)$. The electricity demand forecast for the planning area is shown in Table 1.

The main equipment for power grid transformation is a transformer, as shown in Fig. 4.
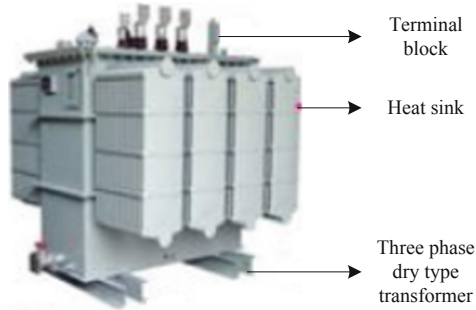

Fig. 4. Transformer schematic

Based on the above-mentioned prepared data and objects, the intelligent prediction experiment of power grid substation operation cost is carried out, and the application effect of the proposed method is shown through the prediction time and accuracy.

3.2 Forecast Time Analysis

Table 2 shows the estimated time of power grid substation operation cost obtained through experiments.

Table 2. Time for forecasting the operation cost of power grid substation

Number of transformers	Forecast time	Maximum limit
10	6.56 s	8.56 s
20	7.04 s	9.12 s
30	9.50 s	10.45 s
40	10.01 s	12.40 s
50	11.45 s	13.52 s
60	12.08 s	14.58 s
70	13.44 s	15.00 s
80	14.03 s	15.95 s
90	15.89 s	16.47 s
100	16.00 s	17.80 s

As shown by the data in Table 2, compared with the maximum prediction time, the prediction time of running cost obtained by the proposed method is lower. The average prediction time of the method in this paper is 11.6 s, which indicates that the proposed method has better real-time prediction.

3.3 Forecast Accuracy Analysis

Based on the experimental environment set up in Sect. 3.1, the prediction accuracy of the method in this paper, the substation cost prediction method based on the improved BP neural network (method 1) and the substation cost prediction method based on the least squares support vector machine (method 2) is compared.. The experimental results are shown in Fig. 5.
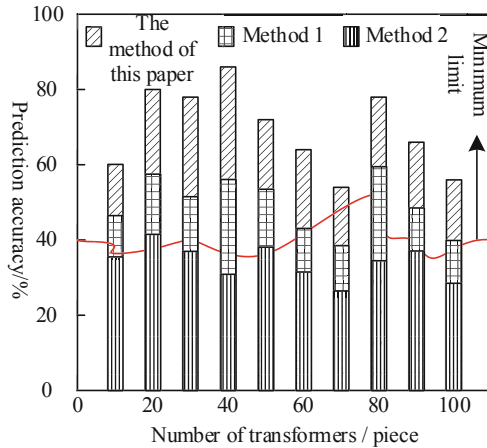
Fig. 5. Prediction accuracy of power grid substation operation cost

As shown in the data in Fig. 5, compared with the minimum limit of prediction accuracy, the prediction accuracy of operating cost obtained by the proposed method is higher, and the highest accuracy is 83%. The prediction accuracy of the method in this paper is higher than the two compared methods, indicating that the proposed method has higher prediction accuracy.

The above-mentioned experimental data shows that after the method should be proposed, the prediction time and prediction accuracy of the power grid substation operation cost are in line with the standard, which fully proves that the proposed method has a better substation operation cost prediction effect.

4   Conclusion

In order to effectively manage and control the cost of substation operation, a research on the intelligent prediction method of substation operation cost of power grid with non-stationary characteristics is proposed. The MCSSD method is used to extract the non-stationary characteristics of the power grid substation information, analyze the substation

economic load rate, and determine the substation capacity. According to the relevant expenses of substation equipment, an intelligent prediction model of substation operation cost is constructed, and the prediction result of operation cost is obtained. Based on this, the control measures of substation operation cost are formulated. It is verified by experiments that the proposed method greatly shortens the running cost prediction time, and the average prediction time is 11.6s. At the same time, the method improves the prediction accuracy of operating cost, and the highest prediction accuracy is 82%. It provides more effective support means for substation operation, and also provides some reference and reference for cost forecasting research.

References

1. Ahmad, W., Hasan, O., Tahar, S.: Formal reliability and failure analysis of ethernet based communication networks in a smart grid substation. Formal Aspects Comput. **32**(2), 71–111 (2020)
2. Heidari, M., Niknam, T., Zare, M., Niknam, S.: Integrated battery model in cost-effective operation and load management of grid-connected smart nano-grid. IET Renew. Power Gener. **13**(7), 1123–1131 (2019). https://doi.org/10.1049/iet-rpg.2018.5842
3. An, S., Lu, Q.: Simulation of LFM interference suppression in spread spectrum communication based on FRFT and blind separation. Comput. Simul. **38**(9), 150–154 (20210
4. Hu, H., Fang, M., Hu, F., et al.: A new design of substation grounding based on electrolytic cathodic protection and on transfer corrosion current. Electric Power Syst. Res. **195**(2), 107174 (2021)
5. Wang, X., Fu, Z., Wang, Y., et al.: A non-destructive testing method for fault detection of substation grounding grids. Sensors **19**(9), 2046 (2019)
6. Liu, S., Bai, W., Srivastava, G., et al.: Property of self-similarity between baseband and modulated signals. Mobile Networks Appl. **25**(4), 1537–1547 (2020)
7. Yang, N., Liu, Z., Yan, J., et al.: A planning method of substation considering main transformer adjustment utilization and safety efficiency cost. Zhongguo Dianji Gongcheng Xuebao/Proc. Chin. Soc. Electric. Eng. **40**(13), 4187–4200 (2020)
8. Liu, S., Chen, X., Li, Y., et al.: Micro-distortion detection of lidar scanning signals based on geometric analysis. Symmetry **11**(12), 1471 (2019)
9. Nassif, A.B., Torquato, R.: Field verification of autonomous anti-islanding schemes and grid support functions of an inverter-based microturbine distributed generator. IEEE Trans. Ind. Appl. **55**(6), 5652–5658 (2019)
10. Zhang, Y., Cai, H., Jia, L., et al.: Influence of integrated grounding system for high-speed railway on the grounding impedance measurement of traction substation. Gaodianya Jishu/High Voltage Eng. **45**(3), 723–729 (2019)

Research on Standard Cost Prediction of Intelligent Overhaul Based on Multiparticle Optimization

Li Huang[1]([✉]), Ye Ke[2], Fenghui Huang[2], Ying Wang[2], and Cong Zeng[2]

[1] State Grid Fujian Electric Power Co, Ltd., Fuzhou 350007, China
zjtjlq@163.com
[2] State Grid Fujian Power Economic Research Institute, Fuzhou 350012, China

Abstract. In order to effectively control the cost consumption in the process of intelligent overhaul of power grids, so as to maximize the saving of power supply cost, a standard cost forecast model based on multi-particle optimization algorithm is proposed. Starting from the global mode and local mode, the concrete calculation results of optimization operator are determined, and the cost statistics of power network based on multi-particle optimization algorithm is realized. On this basis, the cloud application concept of maintenance cost is defined, and the actual value of standard gray number is determined according to the numerical calculation law of cost characteristics. Experimental results show that MPSO can save the consumption of overhaul cost and meet the practical need of effectively controlling the supply cost of electricity under the same power supply.

Keywords: Multi-particle optimization algorithm · Power grid maintenance · Cost forecast · Global model · Local model · Predicted grey number

1 Introduction

Each algorithm is to solve some practical problems or better to solve these problems, particle swarm optimization algorithm is no more than the production of such. Multi-particle optimization algorithm is a stochastic optimization algorithm based on iteration. Many scholars have studied a lot of classical optimization algorithms before they are produced. For example, the classical simplex method, steepest descent method, Newton method, quasi-Newton method, conjugate gradient method and trust region method are suitable for solving unconstrained optimization problems, but less effective for more complex problems. In order to solve more and more complex practical problems, with the continuous development and growth of applications based on bionics, a series of optimization algorithms inspired by biological or biological group behavior characteristics and some natural phenomena have emerged through continuous research on artificial life, such as genetic algorithms, quasi-annealing algorithms, tabu search and artificial neural networks, which were proposed successively after the 1940s according to the actual

needs. These algorithms are derived from the actual laws of nature and the experience of practical problems [1].

Power grid plays a fundamental role in China's economic development, mainly in the transmission and distribution of electric power and electric power trading functions of these two aspects. The construction of power grids is facing new challenges and opportunities, and the construction of a well-off society in an all-round way puts forward new requirements for the breadth of the coverage of power grid construction; the promotion of the development of energy bases in the northwest puts forward new requirements for the regional density of power grid construction; and the demand for network access of multiple renewable energies puts forward new requirements for the intelligent regulation of power grids. In order to meet the needs of national economic development and strategic adjustment, power grids need to be constructed vigorously in the future. Prior to the upgrading of the power grid, it is faced with the problems of inadequate capacity for the allocation of power grid resources and inadequate security, reliability and economy [2]. In order to solve the above problems, the State Grid carried out the backbone grid construction and grid upgrading projects during the 11th Five-year Plan period, with a cumulative investment of 1.2 trillion yuan to increase transmission lines and transformer capacity, with growth rates of 8.0% and 14.5% respectively. Compared with the "Tenth Five-year Plan", the length of new transmission lines increased by 1.20 times and the capacity increased by 1.93 times. After ten years of transformation and upgrading of power lines and substations, the above-mentioned problems have been greatly improved, and the economic stability of the power grid continues to improve.

2 Power Grid Cost Statistics Based on Multi Particle Optimization Algorithm

2.1 Global Mode

The original particle swarm optimization algorithm is a global optimization algorithm, which searches for the best position in all the allowed search space until finding the global best position. In the optimization theory, we define the particle as the possible solution of the optimization problem. The optimization process can be seen as follows: first, a group of particles is determined randomly in the solution space, then each particle in the swarm is evaluated by the objective function, and then the optimal solution is found through continuous evolution by following the optimal position of its own understanding. The global mode flow chart of multi particle optimization algorithm is shown in Fig. 1:
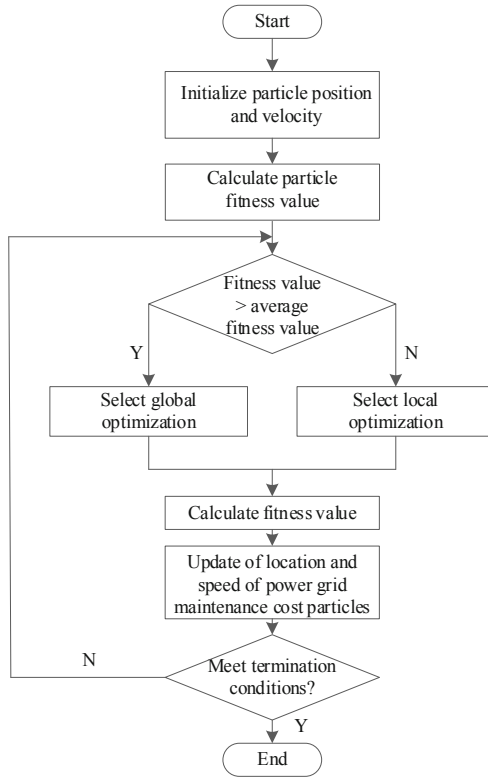
Fig. 1. Global mode flow chart of multi Particle Optimization Algorithm

N problem of a is set, and a power grid maintenance cost particle swarm with size M is set. The position of the i particle is $X_i = (x_{i1}, x_{i2}, \cdots x_{iN})$ and the speed is $V_i = (v_{i1}, v_{i2}, \cdots, v_{iN})$. The speed determines the distance and direction of each movement of the power grid maintenance cost particle. The current fitness value of the particle is calculated according to the objective function describing the problem to measure the quality of the particle. Let $pb_i = (p_{i1}, p_{i2}, \cdots p_{iD})$ be the best position experienced by the particle of power grid maintenance cost moving in the search space so far, and $pb_i = (p_{i1}, p_{i2}, \cdots p_{iD})$ be the best position found by all particles so far.

The location and speed of power grid maintenance cost particles are updated generation by generation according to the following formula:

$$\begin{cases} V_{in}(k+1) = V_{in}(k) + c_1 r_1 (pb_{in} - x_{in}(k)) \\ x_{in}(k+1) = x_{in}(k) + v_{in}(k+1) \end{cases} \tag{1}$$

In the formula, k represents the algebra of the current power grid maintenance cost particles, $i = 1, 2, \cdots, M$ represents the dimension of the problem, $n = 1, 2, \cdots, N$ represents the number of each particle in the group, N is the set power grid maintenance cost particle swarm size, and in the update formula of c_1 and c_2 speed, it affects the power of particles moving towards their own experience and group experience, so it

is often vividly called learning factor or acceleration factor. If these two factors are properly adjusted, we will find that they do not have a great impact on the convergence, but they can make the particles not close to the local minimum as far as possible, so as to improve the convergence speed. The parameters r_1 and r_2 are two random numbers between (0, 1). Their function is to maintain the diversity of particles, so as to reduce the algorithm falling into local optimization.

The second item of the first formula, which is related to all the experience of the particles in the process of optimization, is usually called the "cognition" part of the particles, which represents the self-learning ability of the particles. The third item of the formula is called the part of the particles' recognition in the "society", and these two parts also reflect the direct experience and indirect experience of the particles, which reflects the obvious advantage of the cooperation between the particles based on the multi-particle optimization algorithm.

2.2 Local Mode

In the global version of standard multi-particle optimization algorithm, particle swarm optimization algorithm, the cost of power network maintenance is to find the global optimal solution by tracking the optimal value and population. This method can quickly converge to find the optimal solution, but it is easy to fall into the local optimum (or local extremum) because the search is not careful enough, and the optimal solution in the specified solution space cannot be found accurately [3]. To solve this problem, in order to find the global optimal solution accurately, a local version of the optimization algorithm is produced. It stipulates that the position and velocity of particles are updated by referring to the optimal value in the "memory" of particles and the optimal value in the nearest particles. By comparing these two values, the position and velocity of particles are constantly adjusted to achieve full search and finally find the global optimal. Compared with the global version, this model can find the global optimum more effectively, but it also has the disadvantage of slow convergence speed.

The basic steps for the global and local versions are the same, except that the reference points for evaluating the particles are different, in short, the ways in which the particles move.

Set χ to represent the local optimization coefficient of the power grid maintenance cost particle swarm, c and a to represent the execution step value of two different optimization instructions, f to represent the local mode eigenvalue of the power grid maintenance cost particle, ΔD to represent the local mode change of the power grid maintenance cost particle per unit time, and the simultaneous formula (1) to represent the local mode of the power grid maintenance cost particle swarm based on the multi-particle optimization algorithm as follows:

$$S = \sqrt{\frac{1}{\chi} \sum_{c=1}^{a} \left(\frac{V_{in}(k+1)^2 - x_{in}(k+1)^2}{f \cdot \Delta D} \right)} \tag{2}$$

Since the local version was proposed, multi-particle optimization algorithms based on different topological structures have been proposed successively, including random

ring, wheel and random wheel. These topological structures put forward a new direction for the improvement of traditional particle swarm optimization algorithm.

Based on the trend of tax law and market price, the paper analyzes the relationship between project cost and income, that is, the payback period and the ability to repay the loan, and gives the benefit conclusion. In financial evaluation, the selection of price is the key of evaluation and analysis. Whether the price of each product can be selected accurately or not directly affects the accuracy of the final conclusion. But the product price is a dynamic attribute, which is changing with the market supply and demand [4]. Only by managing the price of key materials in the west market and fully analyzing the price sensitivity, can the conclusion be used to provide the basis for the decision-maker to make an accurate decision. But the project cost in view of this aspect correlation research and the application is quite lag.

2.3 Operator Design

Different from function optimization, the object of MPO is discrete, each discrete state corresponds to a set of discrete values, and each discrete state corresponds to a value of objective function.

In the global mode and local mode formulae given above, the maximum value of particle distribution determines the ability of particle to find the global optimal solution. If the particle movement step is too large, it is easy to skip the most position and can not find the global optimal solution. That is, the larger particle distribution condition can make the particle search faster and more sufficient, and improve the particle excavation ability, while the smaller particle distribution condition can make the particle search more careful, and improve the particle excavation ability.

The key of a successful optimization algorithm lies in a good balance between global search and local search, but the whole search process of particle swarm is a nonlinear search process from global to local, and the search from global to local can not be accurately grasped by linear decreasing inertia weight [5]. For different optimization problems, this improved particle swarm optimization algorithm may not be able to achieve good results. Especially for the dynamic optimization problem, the simple linear decreasing inertia weight strategy can not meet the requirements of problem solving.

For the simple optimization problem, the particle swarm can get the optimal solution quickly, but for the complex optimization problem, the particle swarm is easier to fall into the local optimum. Therefore, it is not enough to rely on group experience, but also rely on their own experience.

Let $\overline{L}$ represent the optimization mean value of power grid maintenance cost particles, f represent the particle swarm integration coefficient, ϕ represent the established optimization index parameters, and β represent the definition conditions of power grid maintenance cost. The operator design result based on multi particle optimization algorithm can be expressed as:

$$\mu = \frac{f \cdot \overline{L}}{\left| 2 - \phi - \sqrt{S^2 - 4\beta} \right|} \tag{3}$$

Through the simulation of the multi-particle optimization algorithm, we can be familiar with the basic process of the particles of the maintenance cost of the power grid [6].

It can be seen from the above that compared with the standard multi particle optimization algorithm, the application effect of local optimization algorithm is better, which can effectively solve the problem of falling into local optimization because the search of traditional algorithm is not careful enough. Through the optimization mean value of power grid maintenance cost particles, particle swarm integration coefficient, established optimization index parameters and definition conditions of power grid maintenance cost, the operator in particle algorithm is designed to solve the problems such as insufficient particle convergence behavior.

3 Standard Cost Prediction Method of Power Grid Intelligent Maintenance

3.1 Maintenance Cost Cloud Definition

When making investment decisions for power grid infrastructure projects, the analysis of the various stages of the future life cycle is obtained through the available historical data or the forecast of development prospects. This forecast often uses a fixed forecast value to indicate an uncertain factor in the future. The accuracy of this expression is not high. Often, the cost of the various stages of the life cycle of the power grid infrastructure projects is an interval number. The accuracy of this expression is far greater than the accuracy of a forecast value [7]. Therefore, we can reduce the loss of information to the greatest extent by describing the uncertainties in each phase of the life cycle of the power grid infrastructure project as a cloud model. The cost cloud is defined as follows.

Cost cloud is the stage cost domain that can be expressed by accurate values within the research scope of life-cycle cost of power grid infrastructure projects, and T is the qualitative estimate associated with l. The elements in l are random numbers with stable membership, and the distribution of membership ξ in universe l is cost cloud. The specific definition conditions are as follows:

$$\varpi = \frac{\sum_{c=1}^{b} l \times T}{\mu \xi^2} \tag{4}$$

Among them, c and b represent two different definition coefficients of intelligent overhaul cost c.

The basic construction projects of power grids have the characteristics of long life cycle, complicated factors to be considered, numerous indicators and so on. By analyzing the closeness between the to-be-estimated project and the historical project, the historical project most similar to the to-be-estimated project is selected by using the principle of maximum closeness, and the final full-life cycle cost of the to-be-estimated project is estimated by using the data of the historical project, thus ensuring the accuracy and rationality of the estimation [8]. But it also has some disadvantages, such as the calculation is too detailed and comprehensive, the inevitable calculation of complex, the

cost of human and material huge, which itself involves a cost issue. Moreover, there is a great uncertainty in the whole life cycle of the power grid construction project. Using historical data analysis can not fully reflect the uncertainty of the cost in the whole life cycle of the project to be estimated, and can not provide comprehensive cost information for the final decision.

3.2   Cost Characteristic Value

There are many uncertainties in the life cycle cost of power grid construction projects. How to express and transform these uncertainties into scientific theories is the key to solve these problems. The cloud model can well describe the uncertainty of the whole life cycle of the power grid construction project. Using the multi-particle optimization algorithm, the experts can estimate the cost of each phase of the construction project. Then the cost cloud can be transformed into the gray number of three parameter interval. Therefore, the cloud model can retain the uncertain information to the maximum extent and is more advantageous to the final cost accounting.

The concept of cost management in the management of infrastructure projects in our power grid is more lacking, mainly in three areas. Firstly, because of the limitation of cost management scope, the scope of cost management of power grid capital construction projects is only limited to the construction period, while the management of operation expenses, maintenance expenses and overhaul expenses after construction is largely ignored. The cost of power grid capital construction projects is characterized by one-time investment (construction cost) accounting for only about 40% of the life cycle cost of power grid capital construction projects, so the scope of cost management at present is limited [9]. Secondly, because of the ambiguity of cost management, the cost management of power grid enterprises is generally limited to how to reduce the cost, but not how much benefit it can bring to the enterprise. Third, the method of cost management is backward, power grid enterprises use traditional means to manage costs, and advanced theories such as lifecycle theory are not timely applied in power grid enterprises.

Set j_1 and j_2 to represent two different coefficients of power grid construction cost, $\dot{e}$ to the intelligent definition index of power grid maintenance cost, θ to the predicted scalar value of power grid maintenance cost, and the simultaneous formula (4). The standard cost eigenvalue of power grid intelligent maintenance based on multi-particle optimization algorithm may be represented as follows:

$$M = \frac{\sum\limits_{j_1}^{j_2} \varpi \, (1 + \dot{e})^{j_2 - j_1}}{\theta^2} \tag{5}$$

In the implementation of power grid maintenance projects, due to the time and region differences, will directly affect the cost of each link, and ultimately affect the entire construction phase of capital investment. Then the time and region adjustment coefficients can be introduced to correct the change of input caused by time and region by comparing the price of the key process of the estimated project with that of the key process of the historical similar project.

3.3 Standard Prediction Grey Number

The characteristics of power grid infrastructure projects determine that the life-cycle cost theory is of great value in cost management. However, the application of multi-particle optimization theory in power grid infrastructure projects is superficial and has great theoretical value. From the perspective of power grid enterprises, the cost of the whole life cycle of power grid infrastructure projects is analyzed and studied in detail, and the cost structure and estimation model of each link in the life cycle are determined by consulting historical data and using modern statistical methods. Through the analysis of the internal relations between the costs of each link, the cost improvement approach is found, and the cost optimization is finally realized. Through the application of life-cycle theory, the cost control of power grid construction project can achieve the harmony of the cost of each stage and the life-cycle cost [10, 11]. When the life-cycle cost is used to control the cost of power grid construction project, the cost of operation, maintenance, overhaul and scrapping should be minimized besides the reduction of one-time investment.

The development of network intelligent overhaul cost forecast can be regarded as the development of cost management research of construction enterprises. Through carding the development context of engineering cost, we can understand the exploration of cost control of construction enterprises. First of all, the definition of project cost, the main body of project cost is different, its meaning is not the same, the main body has a complete project, a single project, water and heating projects or installation projects. Its main body is different its project cost content also along with it changes.

Set up $g_{\min}$ to indicate the minimum overhaul cost forecast condition, $g_{\max}$ to indicate the maximum overhaul cost forecast condition, ψ to indicate the processing authority characteristic value of the grey number forecast, $\hat{h}$ to indicate the characteristic index of the power network overhaul cost optimization under the specific situation. With the support of the above physical quantities [12], the standard predictive grey value can be represented by the simultaneous formula (5):

$$ I = \frac{\psi \cdot M}{\sum\limits_{g_{\min}}^{g_{\max}} \hat{h}^{g_{\max} - g_{\min}} \Big/ \Delta U^2} \tag{6} $$

In the above formula, ΔU represents the power grid maintenance cost consumption per unit time.

The grey number of standard forecast is an index that can determine the trend of maintenance cost of power network. The cost accounting ability directly determines its bidding, construction and acceptance. Therefore, the study of the work cost has great practical significance, and the research results must have the characteristics of universality, convenience and accuracy, otherwise the project cost is of no practical significance.

Bidding for power grid infrastructure projects refers to the process whereby the bid inviter, after releasing the bidding information of power grid infrastructure projects, drafts bidding documents and puts forward specific construction measures for the relevant projects based on the bidding information, and selects the entities with the highest

bidding price within a reasonable scope and with the relevant qualifications required by the State and the highest personnel level according to the bidding documents of all bidding entities. In the evaluation process of a bidding document, it is not only necessary to analyze the bidding documents of the current project, but also to make a comprehensive analysis in combination with the bidding documents of the same kind or of the same bidder so as to comprehensively analyze the internal economy, management, qualification and other aspects of the bidding enterprise and provide the basis for decision-making. In order to select the enterprises with high capacity and high level, it is necessary to compare the quoted price of each stage with the internal quota, analyze the reasons, reflect the real project cost level on the premise of guaranteeing the construction quality, and finally get a project cost quota with basis, competitiveness and management level.

4   Case Analysis

In order to verify the practical value of the standard cost prediction model of power grid Intelligent Maintenance Based on multi particle optimization algorithm, the following comparative experiments are designed. The periodic estimation model and multi particle optimization algorithm are used to control the cost of power grid intelligent maintenance, in which the former is used as the control group and the latter as the experimental group.

Table 1 records the specific numerical changes of power grid work during the given experimental time.

Table 1. Work done by power grid

Project	Experimental time/(h)	Power grid work/(KWh)
1	2	104
2	4	205
3	6	311
4	8	420
5	10	532
6	12	608
7	14	703
8	16	826
9	18	904
10	20	997

Analysis of Table 1 shows that the experiment takes 2h as a unit duration, and the power value keeps increasing during the whole experiment process, and the increase rate is higher than the experiment.

The cost consumption of power grid overhaul can describe the implementation of power grid project. The smaller the cost consumption value is, the stronger the application

ability of power grid project in saving supply cost is, otherwise the weaker it is. In this experiment, the cost prediction method of substation maintenance and operation based on Improved BP neural network proposed in reference [1] and the cost prediction method of power grid project based on whole life cycle proposed in reference [2] are selected as the experimental control group, and compared with the experimental test results of the proposed method, The specific numerical changes of power grid maintenance cost consumption of different methods are shown in Fig. 2.
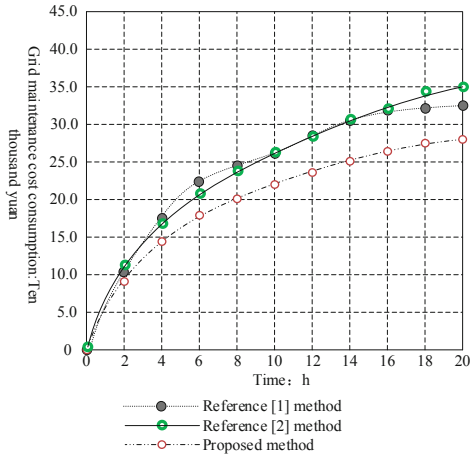
Fig. 2. Power grid maintenance cost consumption

Analysis of Fig. 2, with the extension of the experimental time, the experimental group, the control group of power grid maintenance cost consumption has shown a rising trend of numerical change. For the experimental group, the global maximum value is only 278,000 yuan, while for the control group, the global maximum value is 350,000 yuan.

Table 2 records the increment amplitude of the power grid maintenance cost consumption of the experimental group and the control group in unit time, and its numerical comparison with the actual power grid work.

Table 2 shows that the unit increment of power grid maintenance cost consumption in control group is always larger than that in test group in 0–8 h and 12–16 h, and is always smaller than that in test group in 10th and 18–20 h. During the whole experiment, the maximum value of the rising amplitude of the experimental group and the control group was 92,000 yuan and 102,000 yuan, respectively. The time of the rising amplitude was 2h.

Compare the cost prediction accuracy of the proposed method, the substation maintenance and operation cost prediction method based on Improved BP neural network proposed in reference [1] and the power grid project cost prediction method based on life cycle proposed in reference [2]. The experimental results are shown in Fig. 3.

Table 2. Unit rising amplitude of power grid maintenance cost consumption

Experimental time/(h)	Power grid work/(KWh)	Rising amplitude of experimental group/(Ten thousand yuan)	Rising amplitude of reference [1] method/(Ten thousand yuan)	Rising amplitude of reference [2] method/(Ten thousand yuan)
2	104	9.2	10.2	10.1
4	205	5.4	7.3	7.4
6	311	3.3	4.8	4.7
8	420	2.4	2.5	2.3
10	532	1.9	1.2	1.8
12	608	1.7	1.9	1.9
14	703	1.2	2.5	2.5
16	826	1.6	1.9	1.9
18	904	0.9	0.3	0.3
20	997	0.4	0.2	0.6
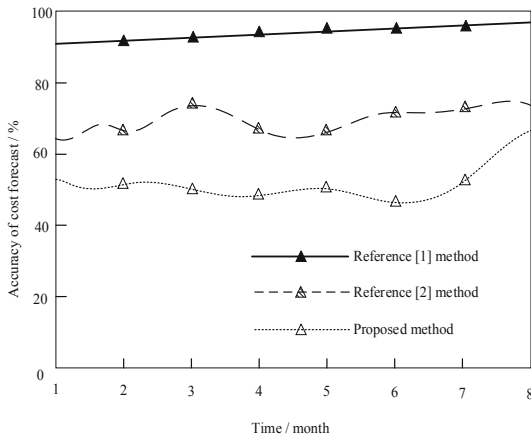
Fig. 3. Comparison of cost prediction accuracy

According to the comparative experimental results of cost prediction performance of various methods in Fig. 2, the cost budget accuracy of the proposed method is higher than that of reference [1] method and reference [2]. Therefore, with the application of multi particle optimization algorithm, the consumption value of power grid maintenance cost has been effectively controlled, and the cost prediction has higher accuracy, This has indeed played a promoting role in saving the cost of power supply.

5   Conclusion

Compared with the periodic estimation model, the standard cost prediction model based on multi-particle optimization algorithm can accurately calculate the parameters of the optimization operator, and the cost eigenvalue can directly affect the actual calculation result of the grey value of the standard prediction. The experimental results show that the increase amplitude of power grid maintenance cost under the proposed method is always relatively small, and the global maximum can only reach 278000 yuan; In contrast, the global maximum of power grid maintenance cost of the other two reference methods has reached 350000 yuan. The proposed method also has lower maintenance cost and consumption. The cost budget accuracy of the proposed method is always higher than 90%, which is always higher than that of reference [1] and reference [2]. From the point of view of practicality, the cost of maintenance and repair of power network has been effectively controlled, and the increase in the cost of consumption per unit time has also shown a trend of reduction, which can achieve the maximum savings in the cost of electricity supply.

References

1. Xiong, Y., Zhan, Z.H., Ke, F.C., et al.: Overhaul operation and maintenance cost prediction of substation based on improved BP neural network. J. Electric Power Sci. Technol. **36**(4), 44–52 (2021)
2. Chen, Q., Zhang, J.H., Li, X.L., et al.: Research on investment of power grid project based on life cycle. Mod. Electric Power **28**(5), 80–84 (2011)
3. Soltan, S., Mittal, P., Poor, H.V.: Line failure detection after a cyber-physical attack on the grid using bayesian regression. IEEE Trans. Power Syst. **34**(5), 3758–3768 (2019)
4. Tariq, M., Adnan, M., Srivastava, G., et al.: Instability detection and prevention in smart grids under asymmetric faults. IEEE Trans. Ind. Appl. **56**(4), 4510–4520 (2020)
5. Houssein, E.H., Saad, M.R., Hussain, K., et al.: Optimal sink node placement in large scale wireless sensor networks based on Harris' hawk optimization algorithm. IEEE Access **8**(99), 19381–19397 (2020)
6. Gomathy, M.: Optimal feature selection for speech emotion recognition using enhanced cat swarm optimization algorithm. Int. J. Speech Technol. **24**(1), 155–163 (2021)
7. Sekaran, K., Dinesh, K., Rajkumar, Y., et al.: An energy-efficient cluster head selection in wireless sensor network using grey wolf optimization algorithm[J]. TELKOMNIKA (Telecommun. Comput. Electron. Control) **18**(6), 2822–2833 (2020)
8. Liu, S., Bai, W., Srivastava, G., Machado, J.A.T.: Property of self-similarity between baseband and modulated signals. Mobile Networks Appl. **25**(4), 1537–1547 (2019). https://doi.org/10.1007/s11036-019-01358-9
9. Gupta, R., Alam, M.A., Agarwal, P.: Whale optimization algorithm fused with SVM to detect stress in EEG signals. Intell. Dec. Technol. **15**(1), 87–97 (2021)
10. Chaudhary, P., Gupta, R., Singh, A., et al.: Joint image compression and encryption using a novel column-wise scanning and optimization algorithm. Procedia Comput. Sci. **167**(1), 244–253 (2020)
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. Gao, C., Huang, Y.S., Ma, H.J.: Simulation of spatial load density distribution in medium voltage distribution network. Comput. Simul. **37**(03), 61–65 (2020)

Research on Medical Information Processing Based on Data Mining Technology

Zhiying Cao[(✉)]

The Affiliated Changshu Hospital of Soochow University (Changshu No.1 People's Hospital),
Suzhou 215500, Jiangsu, China
502921758@qq.com

Abstract. Big data construction has become a national strategic policy, and medical big data related to human health is an important part of it. Data mining uses computers to extract useful information from massive, incomplete, noisy, fuzzy and random data. This paper introduces that data mining technology is divided into association classification rule technology, cluster analysis and rough set theory. As well as the feature pattern polymorphism, data fuzziness, data timing and data redundancy of medical data mining. Data mining architecture is composed of preprocessing module, mining process module, result evaluation module, knowledge guidance module and mining object management module. It is a system integrating information management, retrieval analysis and evaluation, and data warehouse. The application prospect of data mining in medical field is very broad. With the deepening of research, the value of data mining is constantly reflected, which can better serve the public health.

Keywords: Data mining · Data warehouse · Medical information processing

1 Introduction

Big data construction has become a national strategic policy, and medical big data related to human health is an important part of it [1]. In June 2016, the Guiding Opinions of the General Office of the State Council on Promoting and Regulating the Application and Development of Big Data in Health Care clearly stated that "Big Data in Health Care is an important basic strategic resource of the country" [2, 3]. The application of medical big data is of great significance to clinical medical research, scientific management and the transformation and development of medical service mode, and its research is bound to become an important development direction in the next two decades. The value of big data center can not be realized without data mining technology [4–6].

Data Mining (DM) is an information processing technology developed in 1990s. It is a process of extracting potentially useful information and knowledge hidden in a large number of incomplete, noisy, fuzzy and random data by computer, involving knowledge in many fields such as database, artificial intelligence and statistics [7]. By applying data mining technology to medicine, we can find the rules and patterns of medical

diagnosis, thus assisting doctors in disease diagnosis and providing reliable basis for scientific management and medical research in hospitals [8–11]. The data mining process can be divided into nine stages: data preparation, data selection, data preprocessing, data reduction, data mining target determination, mining algorithm determination, DM, pattern interpretation and knowledge evaluation [12].

2 Research Status

(1) Association classification rules in 1998, association classification method cBA31 was put forward. CBA integrated the process of classification rule mining and association rule mining, and achieved better classification effect than decision tree classification algorithm C4.5 based on rule mining. Since then, some people have put forward various improvement methods aiming at the shortcomings of CBA, typically CARL2 and ARC. The basic idea of these methods is to use the existing association rule mining algorithm 3 to generate feature words or feature word itemsets that frequently appear in various categories, and use these frequent feature word itemsets to construct classification rules to classify test samples. The more frequent feature words the test sample contains and the higher the confidence level, the more likely it is that the test sample belongs to this category; otherwise, the less likely it is to belong to this category [13, 14]. Compared with other classification methods, association classification generates classification patterns in the training stage, and only needs to compare and match the documents to be classified with the classification patterns in the classification stage, so it has the advantages of short training time and classification time.

(2) Chaster analysis is also a commonly used technology in data mining, which refers to the process of dividing a data into several categories according to the principle that the distance between data objects in the same group is small, and the distance between data objects in different groups is large, among which there are many methods to define the distance. A group of abstract or physical objects are divided into several groups according to their similarity: the more similar objects are divided into one group, and this process is the clustering process. A collection of similar objects is called a cluster: objects in different clusters are not similar. Clustering is to search for valuable associations between data items from a given data set [15, 16].

Clustering is also called unsupervised induction in machine learning. The biggest difference between clustering and classification is that the classification problem is to classify data objects into different known classes when the classification attributes of training samples are known. In clustering problem, for unknown data objects, the final classification results need to be found in training samples. Cluster analysis has a wide range of applications, including business, insurance, biology, geography, medicine and so on. In business, cluster analysis can help market workers find different groups of customers and describe their different characteristics by purchasing patterns. In the field of biological research, cluster analysis can be used to obtain the hierarchical structure of animals and plants, and classify them according to their gene functions. Clustering analysis can also be used alone

as a tool to understand the characteristics of data and analyze the distribution of data, or as a preprocessing step of other algorithms, such as qualitative induction algorithm. In the medical field, cluster analysis has been widely used in DNA analysis, automatic analysis of medical image data, analysis of disease risk factors and so on [17]. Many researchers in China have been doing research in related fields, such as classification of Neisseria gonorrhoeae drug-resistant epidemic strains and clustering analysis of coronary atherosclerotic heart disease.

(3) In the actual system, there are uncertain factors of varying degrees in many cases. The collected data often contain noise, inaccuracy or even incompleteness, and roughness and theory are the mathematical tools to deal with such uncertain factors. In 1970s, scientists of Polish Academy of Sciences first put forward rough set theory. Rough set theory defines the concepts of fuzziness and uncertainty in the sense of classification, which can effectively analyze various uncertain information such as inconsistency, inaccuracy and incompleteness, and can also analyze and reason data to find hidden patterns and knowledge. The main goal of rough set theory is to get decisions and rules on the basis of keeping the classification ability unchanged by reducing knowledge. To understand the basic concepts of rough set theory, we must first understand a series of concepts such as knowledge attributes closely related to it. In rough set theory, "knowledge" can be regarded as an ability to classify realistic and abstract objects according to the attributes of features. People's behavior is based on the ability to distinguish real or abstract objects. For example, in ancient times, in order to survive, human beings had the ability to distinguish whether food was edible or not. Doctors must be able to distinguish which disease the patient was suffering from when diagnosing the patient. This ability to classify things according to their characteristics can be regarded as some kind of knowledge.

3 Characteristics of Medical Data Mining

Particularity of medical data Clinical medical data is the main object of medical data mining research, which contains all data resources of patients and diseases in the whole medical process. Compared with ordinary data, these data have the following particularity:

(1) pattern polymorphism: medical information includes pure document data manually entered by doctors, image data such as MRI and CT obtained by medical imaging equipment, signal data and voice data of otolaryngology, etc., which are polymorphic.
(2) Data fuzziness: the objective incompleteness of disease and case information and subjective inaccuracy in describing disease make it impossible for medical data to fully reflect the situation of patients or diseases, and at the same time, data design, data collection, data entry and other links may all lead to the missing of the final medical database, which leads to greater fuzziness of medical data.
(3) Timing of data: the patient's visit to a doctor or the onset of a disease has a progress in time, and the images obtained by medical equipment such as electrocardiograph are also a function of time. These medical data have higher timing than ordinary data.

(4) Data redundancy: Data redundancy means that the same information is stored in multiple places. There is a great deal of identical information in medical database, which may lead to wrong or meaningless patterns in medical data mining activities.

It is precisely because of these characteristics of medical data, and because it involves many ethical and legal issues, that medical data mining has its particularity, which requires some unique technologies in data mining.

4   Architecture of Data Mining

Data mining is based on artificial intelligence, machine learning and statistics. However, the data mining system does not simply combine these technologies, but adds a lot of auxiliary technical support to form a complete system. A typical data mining system is shown in Fig. 1, which has the following main components: It can be seen from Fig. 1 that the data mining system is composed of preprocessing module, mining process module, result evaluation module, knowledge guidance module and mining object management module, and is an application software system integrating information management, retrieval analysis and evaluation, data warehouse and so on. It can complete a series of tasks such as data collection, preprocessing, data analysis and result expression, and finally output the results to users. See Fig. 1.
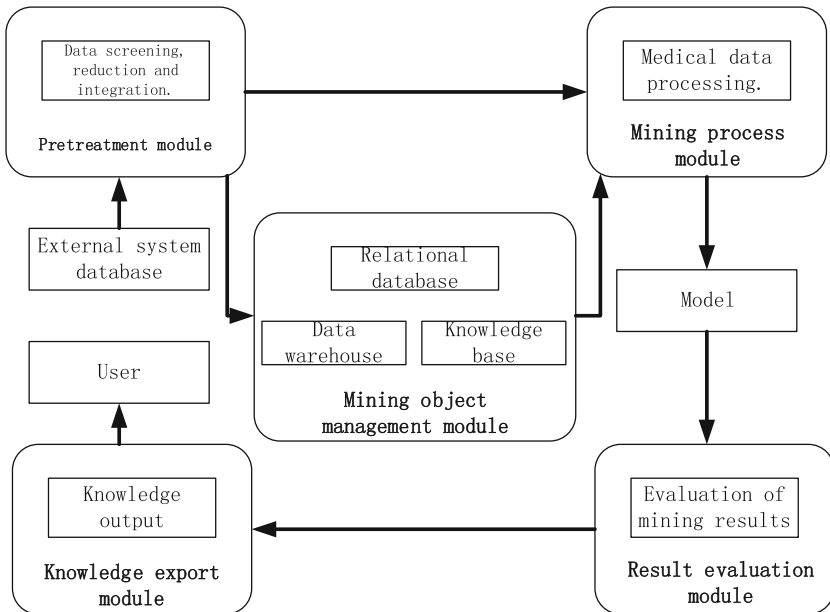
Fig. 1. Architecture of data mining

(1) Pre-processing module: The pre-mining processing module performs various processing on the collected data, including removing noise, integrating various data sources, selecting data related to the problem, changing the selected data into a mineable form, and then generating a data warehouse or data mining library. In the process of mining, if the pattern evaluation finds that the mining pattern is affected by data problems, it will return to the module for data processing before data mining.

(2) Mining process module: Mining operation module is the core part of the whole data mining system. It uses various data mining algorithms and technologies, such as decision tree induction, regression analysis, Bayesian classification, association analysis, online analytical processing, text and multimedia data mining technology, to mine and discover knowledge for databases and data warehouses, and by means of mining rules, experiences, methods and factual data in knowledge base.

(3) Result evaluation module: The main purpose of the mode evaluation module is to evaluate the knowledge and results obtained from data mining. Analyze and compare the user's interest degree with many patterns excavated, evaluate the value of the patterns, and analyze the defects. The difference between the user's interest degree and the mined module is too large, which needs to be returned to the corresponding module for re-execution. And some patterns that meet the user's interest will be directly transmitted to the knowledge output module.

(4) Knowledge export module: Knowledge export module is the interface and bridge between users and data mining system, which translates and explains the patterns obtained from data mining and provides them to decision makers in a way that users can easily understand. Users can directly interact with the system to provide information, formulate mining tasks, and carry out progressive and exploratory data mining according to the results of each step of data mining.

(5) Mining object management module: The database management module is responsible for maintaining and managing all kinds of databases in the system, including databases, data warehouses and mining knowledge bases. These internal databases are obtained by exchanging, cleaning and purifying with external databases, which is the basis of data mining. Mining knowledge base, also known as domain knowledge base, contains experience, methods, techniques, theories, rules, facts and knowledge used or obtained in the mining process. The main purpose of this module is to guide the mining process and evaluate the candidate patterns obtained from mining. From the above introduction, it can be seen that it is difficult to fully realize all functions of a complete data mining system. At present, many data mining systems are incomplete in a strict sense. If it can't handle a large amount of data, it should be called a machine learning system, or a statistical analysis tool, or an experimental system prototype. Similarly, if a system can only perform some data and information retrieval tasks, including summation operation or deductive query and answer, it can only be called an information retrieval system.

5   Summary

With the maturity of data mining technology, its application in medical field will be more and more extensive. Further research work has the following prospects.

(1)   In the horizontal aspect, other technologies such as decision tree classification and artificial neural network in data mining also have their unique advantages. It is an important research topic to give full play to these advantages and apply them to other medical fields including disease diagnosis and analysis, and prediction after medication guidance.

(2)   In the vertical aspect, the association rule mining and rough set mining in this paper are both single-layer mining, which has certain limitations and requires Boolean processing of features in advance, which to a certain extent affects the richness of extracted rules and the accuracy of classification by rules. If we can conduct in-depth research and multi-dimensional and multi-level image mining, the effect will be better.

In a word, the application prospect of data mining in the medical field is very broad. With the deepening of research, the value of data mining is constantly reflected, which can better serve the public health.

References

1.   Sirichanya, C., Kraisak, K.: Semantic data mining in the information age: a systematic review. Int. J. Intell. Syst. **36**(8), 3880–3916 (2021)
2.   Li, Z.: Research on the new path of internet of things data mining under the background of cloud computing. J. Phys. Conf. Ser. **1915**, 042089 (6 pp) (2021)
3.   Istratova, E., Sin, D., Strokin, K.: A comparative analysis of data mining analysis tools. In: Pattnaik, P.K., Sain, M., Al-Absi, A.A., Kumar, P. (eds.) Proceedings of International Conference on Smart Computing and Cyber Security. Strategic Foresight, Security Challenges and Innovation (SMARTCYBER 2020). LNNS, vol. 149, pp. 165–72. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-7990-5_16
4.   Chen, X., Zhao, D., Zhong, W., Ye, J.: Research on brain image segmentation based on FCM algorithm optimization. In: Fu, W., Xu, Y., Wang, SH., Zhang, Y. (eds.) Multimedia Technology and Enhanced Learning. ICMTEL 2021, LNICST, vol. 388, pp. 278–289. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-82565-2_23
5.   Gupta, P., Hoi, C.S.H., Leung, C.K., Ye, Y., Xiaoke, Z., Zhida Z.: Vertical data mining from relational data and its application to COVID-19 data. Big Data Analyses, Services, and Smart Data. Advances in Intelligent Systems and Computing (AISC 899), pp. 106–116 (2021)
6.   Rao, A.S., D'Mello, D.A., Anand, R., Nayak, S.: Clinical significance of measles and its prediction using data mining techniques: a systematic review. In: Chiplunkar, N., Fukao, T. (eds.) Advances in Artificial Intelligence and Data Engineering. Select Proceedings of AIDE 2019. AISC, vol. 1133, pp 737–59, Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-3514-7_56
7.   Marimuthu, V.K., Lakshmi, C.: Performance analysis of privacy preserving distributed data mining based on cryptographic techniques. In: Proceedings of the 7th International Conference on Electrical Energy Systems (ICEES 2021), pp. 635–40 (2021)

8. Mandan, N., Agrawal, K., Kumar, S.: Analyzing different domains using data mining techniques. In: 2020 International Conference on Computer Communication and Informatics (ICCCI), p. 6 (2020)

9. Chen, X., Zhao, D., Zhong, W.: Auxiliary recognition of alzheimer's disease based on Gaussian probability brain image segmentation model. In: Ning, H. (eds.) Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health. CyberDI CyberLife 2019. CCIS, vol. 1138, pp. 513–520. Springer, Singapore (2019). https://doi.org/10.1007/978-981-15-1925-3_37

10. Mahmud, H., et al.: Technologies in medical information processing. Advances in Telemedicine for Health Monitoring: Technologies, design and applications, pp. 31–54 (2020)

11. Kishor, A., Chakraborty, C., Jeberson, W.: Reinforcement learning for medical information processing over heterogeneous networks. Multimedia Tools Appl. **80**(16), 23983–24004 (2021). https://doi.org/10.1007/s11042-021-10840-0

12. Melnykova, N., Mukalov, P., Koziy, D.: The special ways of application of neural networks for medical information processing. In: 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). Proceedings, pp. 428–431 (2018)

13. Karali, E.: Novel approaches to medical information processing and analysis. In: Lambropoulou, S., Theodorou, D., Stefaneas, P., Kauffman, L. (eds.) Algebraic Modeling of Topological and Computational Structures and Applications. Springer Proceedings in Mathematics and Statistics. PROMS, vol. 219, pp. 453–482. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68103-0_23

14. Andrikov, D.A., Kuchin, A.S.: Development of a prototype of a medical information system for a clinical diagnostic center. Procedia Comput. Sci. **186**, 287–292 (2021). (14th International Symposium "Intelligent Systems", INTELS 2020)

15. Chen, X., Zhao, D., Zhong, W., Ye, J., Gao, F.: Research on early warning monitoring model of serious mental disorder based on multi-source heterogeneous data sources. In: Zhang, YD., Wang, SH., Liu, S. (eds.) Multimedia Technology and Enhanced Learning. ICMTEL 2020. LNICST, vol. 327, pp. 403–410. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51103-6_36

16. Xinlei, C., Xiaogang, R., Yue, W., Jiufeng, Y.: Design and realization of a comprehensive management system for severe mental disorders based on FLUX mode. J. Med. Imaging Health Inf. ASP **10**(2), 522–527(6) (2020)

17. Zhang, Y., Wu, L., Wang, S.: Magnetic resonance brain image classification by an improved artificial bee colony algorithm. Progress Electromagn. Res. **116**(2011), 65–79 (2011)

Modeling and Printing Technology Based on 3D Registration Algorithm of MIMICS Software Applied to Hip Fracture

Jinshun Ding[1], Kefeng Xu[1(✉)], Yu Ren[1], and Zhiying Cao[2]

[1] Changshu Meili Hospital, Changshu 215500, Jiangsu, China
dingjin5@163.com

[2] The Affiliated Changshu Hospital of Soochow University (Changshu No.1 People's Hospital), Changshu 215500, Jiangsu, China

Abstract. Hip fracture is the most common and serious fracture in the elderly. Because of the complicated types of acetabular fractures and the curved surface of acetabulum, the surgical treatment of two rows of acetabular fractures has proved to be challenging. Operation is the best treatment for hip fracture in the elderly. However, the elderly patients complicated with a variety of complex medical diseases and the decline of various organs increase the risk of operation, and the incidence of postoperative complications and mortality are also extremely high. The 3d reconstruction adopts MC algorithm. The basic idea is to treat the 2d slice sequence data as a 3d data field, process the voxels in the data field one by one, compare the value of each vertex of the voxel with a given threshold value to determine the construction form of the internal isosurface of the voxel, and connect it into a triangular patch to fit the surface in a certain topological form. Finally, connect the isosurface of each voxel to form the whole isosurface, which is used to represent the surface of the object. The hip joint model is established by imaging and 3D printing. The 3D solid model can truly express individual cases and the actual state and actual stress environment of individual cases, which makes the formulation of hip joint disease treatment plan faster, improves the prediction and prevention of long-term curative effect, and provides valuable clinical research.

Keywords: Hip fracture · Image segmentation · 3D modeling · 3D printing

1 Introduction

Hip fracture is the most common and serious fracture in the elderly. Because of the complicated types of acetabular fractures and the curved surface of acetabulum, the surgical treatment of two rows of acetabular fractures has proved to be challenging [1]. In America, the probability of hip fracture in women is equal to the sum of breast cancer, ovarian cancer and uterine cancer, while the probability of hip fracture in men is higher than that of prostate cancer [2, 3]. With the aggravation of population aging, the incidence of hip fracture in the elderly in the world is increasing at a rate of 1%–3%

every year. It is estimated that the number of hip fracture patients in the world will rise to 6.26 million in 2050, of which more than half will occur in Asia [4]. The incidence of hip fracture in China is also increasing year by year: during 1990–1992, the incidence of hip fracture was 83/100,000 for men and 80/100,000 for women over 50 years old, and it has increased to 129/100,000 for men and 229/100,000 for women during 2002–2006, with a sharp increase in medical expenses [5]. Therefore, hip fracture in the elderly has become one of the most important public health problems in the world [6].

With the aging of China's population, the proportion of osteoporosis in the elderly is increasing, which leads to an increase in the number of patients with hip fractures year by year. At present, there are non-surgical treatment and surgical treatment for elderly patients with hip fracture. However, due to poor resistance and many basic diseases, non-surgical treatment needs to stay in bed for a long time, which is prone to complications such as pressure sores, lung infection and deep vein thrombosis, which seriously affects the quality of life of patients and even endangers their life safety [7–9]. Therefore, surgery is the best treatment for hip fracture in the elderly. However, the elderly patients complicated with a variety of complex medical diseases and the decline of organs in various systems increase the risk of surgery, and the incidence of postoperative complications and mortality are also extremely high [10–12].

2   Research Status

The principle of imaging and the optimized modeling of 3D algorithm of imaging images: The development of digital imaging and the digital construction of hospitals are the products of the combination of electronic industry, computer technology and medicine, which is the inevitable development of imaging and the whole scientific development [13]. With the development of science today, electronic information and computer technology have been fully developed, and the product of their combination is the origin and foundation of the development of digital images [14]. The main advantages of digital imaging are as follows: it can turn simulated dead images into reusable or data, and further change two-dimensional planar images into multidimensional stereo images; It can make image quantitative diagnosis possible. Completely changed the traditional medical image viewing, use, storage and management methods [15, 16].

Digital image is to turn past analog images into reusable data. In the past, hospitals gave patients an X-ray film, which could only record the images of patients under current conditions, but could not see new things through it [17]. Digitalization turns the image into a kind of living data, which can change the past two-dimensional plane image into a multi-dimensional three-dimensional image, from only one plane and length and width in the past to a long, wide, high or front and back, left and right, up and down three-dimensional image [18].

Because of the different functions, medical imaging itself not only reflects the three-dimensional structure, but also includes elements such as time and resolution. In the functional change, we call it a four-dimensional image. In the past, we could only make qualitative judgments, and there was no exact data to make quantitative judgments on patients' films. With the help of digital images, we can make accurate measurements. For example, by measuring the CT value of the patient's image, the tissue type of the lesion can be clearly obtained, so as to make a diagnosis [19].

By introducing highly integrated and easy-to-use 3D image generation and editing software, it can input various scanned data (CT, MRI), build 3D models for editing, and then output general CAD (Computer Aided Design), FEA (Finite Element Analysis) and RP (Rapid Prototyping) formats, which can be used for large-scale data conversion processing on PC. Scan input data can be processed quickly, and the corresponding file format can be output for FEA (Finite Element Analysis) and CFD (Computer Simulation Fluid Dynamics). Users can build 3D model with scan data, and then mesh the surface for FEA analysis. The mesh re-division function in FEA module optimizes the input data of FEA to the maximum extent. Based on Heinz unit of scanned data, materials can be assigned to volume meshes [20].

Development of 3D printers and printing materials: In 1990, Sachs of Massachusetts Institute of Technology first proposed the concept of 3D printing. Since 1990s, 3D printing technology has made vigorous development. In 2003, Mironov of the University of Southern Carona put forward the concept of 3D printing of biological tissues and organs. Compared with the traditional plane printing, 3D printing makes the printed object three-dimensional. In the aspect of three-dimensional forming, 3D printing is a typical additive manufacturing, that is, stacking layer by layer to form a three-dimensional structure [21]. As one of the disruptive technologies that will determine the future economy and human life, 3D printing effectively integrates materials, machinery manufacturing, information processing, electronic equipment and engineering design, and breaks through the dilemma that traditional manufacturing processes are limited by structural complexity and difficult to process and manufacture. Customizing personalized products through 3D printing will reduce production costs and lead a new industrial revolution. At present, 3D printing is widely used in biomedicine, aerospace, architectural design, cultural industry, industrial manufacturing and military equipment, etc. Especially in biomedicine, 3D printing plays an increasingly important role.

The predecessor of 3D printing technology is rapid prototyping technology. The basic idea is based on the digital 3D model, the objects are digitized and layered, and the information such as the 2D processing path of each layer is obtained. By using appropriate materials and processes, the objects are printed layer by layer along the set path through automatic control technology, and finally accumulated into 3D objects. Application of 3D printing in biomedical field Because it is suitable for small batch and highly customized occasions, 3D printing is quickly combined with biomedicine to form biological 3D printing. Biological printing can be divided into broad sense and narrow sense. Generalized biological 3D printing refers to 3D printing that directly serves the biomedical field. In the narrow sense, biological 3D printing refers to the operation of biological ink containing cells to construct living tissue structure, which is the advanced stage and ultimate goal of biological 3D printing. From the development process, the generalized biological 3D printing can be roughly divided into four levels, namely, medical AIDS, non-degradable implants, degradable implants and cell-carrying printing.

3   Research Contents

Mimics is the abbreviation of Materialise's Interactive Medical Image Control System, which can provide user-defined input modules and support various file input formats such

as BMP, JPG, TIF and RAW. It has irreplaceable advantages in basic image processing, segmentation, extraction and visualization.

3.1   MIMICS Software Image Import and Preprocessing

The reconstruction of bone model has always been an important task in biomechanical engineering analysis, and it is also the prerequisite for the accuracy of model reconstruction. It is not only a requirement for the model, but also a special need for the subsequent finite element analysis and the following work.

Medical image three-dimensional reconstruction technology refers to the use of visualization technology to convert two-dimensional image data obtained from medical imaging equipment into three-dimensional data, so as to display the three-dimensional shape of human tissues and organs and make qualitative and quantitative analysis. In this study, based on DICOM format file of hip joint medical image of clinical patients, interactive medical image control system software (MIMICS) was used to reconstruct hip bone.

(1)   Raw data collection: collect DICOM format image data of a hip fracture patient, and apply MMCS new project guide command (import the image data into MIMICS software, which automatically virtualizes the cross-section, coronal plane and sagittal plane according to the fault sequence, and obtain the perfect angle of image processing by adjusting the visual angle and fault sequence.

(2)   After importing hip joint data into image processing, the software displays the image sequence of axial view, and at the same time, automatically virtualizes coronal view, sagittal view and three-dimensional display viewport. In different viewports, the indication information of images is expressed by tick marks, intesection lines, slice position, etc.

Image preprocessing: Image preprocessing is a process before sorting out CT images and handing them over to the segmentation model. This process is called image preprocessing. In image analysis, the quality of the image directly affects the design of the recognition algorithm and the precision of the effect, so preprocessing is needed before image analysis (feature extraction, segmentation, matching and recognition, etc.). The main purpose of image preprocessing is to eliminate irrelevant information in the image, recover useful real information, enhance the detectability of relevant information, and simplify data to the maximum extent, thus improving the reliability of feature extraction, image segmentation, matching and recognition. See Fig. 1.
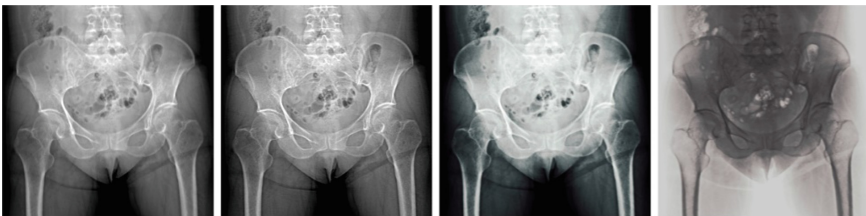
Fig. 1. Image preprocessing schematic diagram

3.2   Algorithms for 3D Modeling of Mimics Software

The 3d reconstruction adopts MC algorithm. The basic idea is to treat the 2d slice sequence data as a 3d data field, process the voxels in the data field one by one, compare the value of each vertex of the voxel with a given threshold value to determine the construction form of the internal isosurface of the voxel, and connect it into a triangular patch to fit the surface in a certain topological form. Finally, connect the isosurface of each voxel to form the whole isosurface, which is used to represent the surface of the object. The essence of 3D reconstruction is to fit the boundary of mask voxel, and the distance between contour lines is the distance between fault slices. Three-dimensional reconstruction is a process of calculating masks based on three-dimensional models, that is, voxelization. Three-dimensional reconstructed hip joint model. See Fig. 2.
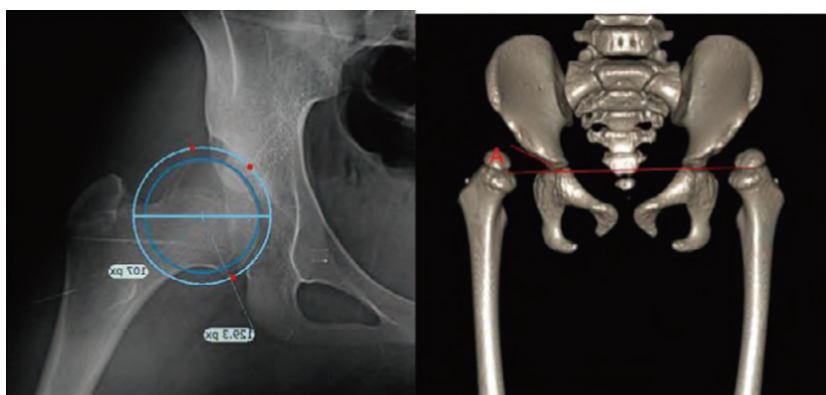
Fig. 2.   Schematic diagram of 3D model generation

In order to give doctors sufficient information for diagnosis of orthopedic diseases, this paper studies two ways of visualization technology and quantitative index calculation, and proposes an evaluation method for auxiliary diagnosis of orthopedic diseases based on visualization technology and quantitative index calculation, which displays useful information through quantitative index values, two-dimensional images and three-dimensional models to assist doctors in diagnosis. In order to realize 3D visual evaluation of multi-target areas, this study proposes an improved MC 3D reconstruction algorithm based on multi-region labels and a 3D model rapid prototyping file export algorithm. The former can provide 3D model images, while the latter can be used to print 3D solid models.

3.3   3D Printing of 3D Registration Model with Mimics Software

Through the export algorithm of 3D model rapid prototyping file of MIMICS software, the STL of hip bone model file is generated. The 3D printer can print 3D through 3D model CAD data. The 3D reconstructed image and 3D physical model before operation can help to achieve fracture reduction, plate bending and position determination

more accurately in clinic. Therefore, preoperative planning is conducive to the smooth operation, and can obtain satisfactory surgical results. See Fig. 3.
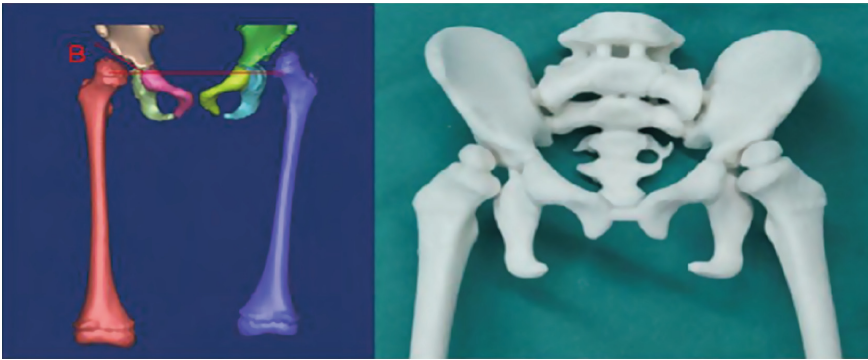
Fig. 3. 3D printing skeleton model schematic diagram

4   Key Technologies

(1) Through image preprocessing algorithm, the image is optimized to solve the problems of CT image quality, boundary registration, coordinate error and so on.
(2) 3-D registration of preprocessed images by software, so as to achieve accurate and rapid reconstruction of CT image skeleton solid model.
(3) The hip joint model is established by image and 3D printing, and the 3D solid model can truly express individual cases and the actual state and actual stress environment of individual cases, which makes the formulation of hip joint disease treatment plan faster, improves the prediction and prevention of long-term curative effect, and provides valuable clinical research.
(4) Three-dimensional solid model can guide patients to avoid secondary fractures caused by new imbalance of strength during rehabilitation training, as well as postoperative nursing care.
(5) Provide teaching guidance for clinical teaching of hip fracture research. The experience of hip fracture research can be extended to other orthopedic diseases.

5   Discussion and Summary

In this paper, an image preprocessing model is proposed, which can preprocess CT images to overcome the gray-scale inhomogeneity that often exists in images, and is also robust to noise and contrast, thus improving the accuracy of 3D modeling. Furthermore, the CT-based bone threshold segmentation is proposed, and the advantages and disadvantages of the Segment- Threshold algorithm and the Advance segment- CT Bone algorithm are compared. Aiming at the segmentation accuracy of the target region, the Advance segment- CT Bone algorithm is more suitable for bone image segmentation. The 3D modeling adopts MC algorithm, which regards the 2D slice sequence data as a

3D data field, processes the voxels in the data field one by one, compares the value of each vertex of the voxel with a given threshold to determine the construction form of the internal isosurface of the voxel, and connects them into triangular patches to fit the surface in a certain topological form. Finally, the isosurface of each voxel is connected to form the whole isosurface, which is used to represent the surface of an object. The 3D reconstruction is mainly carried out by surface rendering, which has the advantages of high 3D reduction and small error. Finally, the 3D model is printed by the generated STL file, and the 3D hip joint model is applied to clinical application.

References

1. Fatemeh, J.: Enhancing hip fracture risk prediction by statistical modeling and texture analysis on DXA images. Quenneville, Cheryl E. Med. Eng. Phys. **78**, 14–20 (2020)
2. Kleiven, S.: Hip fracture risk functions for elderly men and women in sideways falls. J. Biomech. **105**, 109771 (6 pp.) (2020)
3. Cordeiro, M., Caskey, S., Frank, C., Martin, S., Srivastava, A.: Atkinson, Hybrid triad provides fracture plane stability in a computational model of a Pauwels Type III hip fracture. Comput. Methods Biomech. Biomed. Eng. **23**, 476–483 (2020)
4. T. Computer Methods in Biomechanics and Biomedical Engineering, v 23, n 9, p 476–83, 2020
5. Carballido-Gamio, J., et al.: Hip fracture discrimination based on statistical multi-parametric modeling (SMPM). Ann. Biomed. Eng. **47**(11), 2199–2212 (2019). https://doi.org/10.1007/s10439-019-02298-x
6. Gang, L., Jian, J.: Convolutional neural network to explore the effect of the drug on postoperative POCD in elderly patients with hip fracture. J. Intell. Fuzzy Syst. Appl. Eng. Technol. **39**(4), 4989–4997 (2020)
7. Zuki, A.A.M., Mat, F., Daud, R., Kamaruddin, N.S., Ibrahim, I.: A review of hip fracture analysis subjected to impact loading. IOP Conf. Ser. Mater. Sci. Eng. **670**, 012026 (5 pp.) (2019)
8. Awal, R., Faisal, T.R.: Multiple regression analysis of hip fracture risk assessment via finite element analysis. J. Eng. Sci. Med. Diagn. Therapy **4**(1) (2021)
9. Alessandra, A., Mara, T., Alberto, L.A., Bignardi, C., Morbiducci, U.: Combining shape and intensity DXA-based statistical approaches for osteoporotic HIP fracture risk assessment. Comput. Biol. Med. **127** (2020)
10. Fatemeh, J., Cheryl, E.Q.: 3D analysis of the proximal femur compared to 2D analysis for hip fracture risk prediction in a clinical population. Ann. Biomed. Eng. **49**(4), 1222–1232 (2021)
11. Mishra, A., Srivastava, V.: Biomaterials and 3D printing techniques used in the medical field. J. Med. Eng. Technol. **45**(4), 290–302 (2021)
12. Xiao, J., et al. Large-scale 3D printing concrete technology: current status and future opportunities. Cement Concrete Composit. **122** (2021)
13. Schouten, M., Wolterink, G., Dijkshoorn, A., Kosmas, D., Stramigioli, S., Krijnen, G.: A review of extrusion-based 3D printing for the fabrication of electro- and biomechanical sensors. IEEE Sens. J. **21**(11), 12900–12912 (2021)
14. Blyweert, P., Nicolas, V., Fierro, V., Celzard, A.: 3D printing of carbon-based materials: a review. Carbon **183**, 449–485 (2021)

15. Mahdiyar, S., Henry, J.: Current status in the utilization of biobased polymers for 3D printing process: a systematic review of the materials, processes, and challenges. ACS Appl. Bio Mater. **4**(1), 325–369 (2021)
16. Kamiya, T.Y., Corrêa Maceno, M.M., Kleina, M.: Case study applying the methodology in a 3D printing process. In: Environmental and Financial Performance Evaluation in 3D Printing using MFCA and LCA. SAST, pp. 31–68. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69695-5_3
17. Alwazzan, M.J., Alkhfagi, A.O., Alattar, A.M.: Image segmentation algorithm based on statistical properties. In: Kumar, R., Quang, N.H., Kumar Solanki, V., Cardona, M., Pattnaik, P.K. (eds.) Research in Intelligent and Computing in Engineering AISC, vol. 1254, pp. 333–40. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-7527-3_32
18. Prasath, V.B.S., Thanh, D.N.H., Hai, N.H., Dvoenko, S.: Multiregion multiscale image segmentation with anisotropic diffusion. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12665, pp. 129–140. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68821-9_13
19. Jun, M.: Cutting-edge 3D Medical Image Segmentation Methods in 2020: Are Happy Families All Alike?. arXiv, p 13 (2021)
20. Chen, X., Zhao, D., Zhong, W.: Auxiliary recognition of alzheimer's disease based on gaussian probability brain image segmentation model. In: Ning, H. (eds.) Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health. CyberDI CyberLife 2019. CCIS, vol. 1138, pp. 513–520. Springer, Singapore (2019). https://doi.org/10.1007/978-981-15-1925-3_37
21. Wang, Y., Ding, J., Fang, W., Cao, J.: Segmentation-assisted diagnosis of pulmonary nodule recognition based on adaptive particle swarm image algorithm. In: Ning, H. (ed.) CyberDI/CyberLife -2019. CCIS, vol. 1138, pp. 504–512. Springer, Singapore (2019). https://doi.org/10.1007/978-981-15-1925-3_36

Fall Detection Based on Action Structured Method and Cascaded Dilated Graph Convolution Network

Xin Xiong[1,2], Lei Cao[1], Qiang Liu[1], Zhiwei Tu[1], and Huixia Li[1(✉)]

[1] Information Department, The First Affiliated Hospital of Nanchang University, Nanchang 330000, China
lihuixia0601@163.com
[2] Jiangxi Key Laboratory of Smart City, Nanchang 330000, China

Abstract. The research of fall detection is a hot topic in computer vision. Most existing methods only detect the fall in simple scenes of a single person. Moreover, these methods only extract fall action features from RGB images, and neglect to extract features from human joint coordinates, resulting in a decrease in recognition accuracy. In order to extract discriminative action features, a fall detection method based on action structured method and cascade dilated graph convolution neural network is proposed. The action structured method (ASM) is proposed to model the skeleton of human action through the pose estimation algorithm, which removes the interference of complex background. Besides, the object detection algorithm is utilized to locate multiple people to transfers the fall detection issue of multi-person to single person fall detection. The proposed cascaded dilated graph convolution network (CD-GCN) enlarges the receptive field by the dilated operation, effectively extracts action features from joint node coordinates, and fuses multichannel features with different dilation rates, then finally obtains the classification results. The proposed method achieves the best accuracy on three public datasets and one self-collected dataset, which is out-performing other state-of-art fall detection methods.

Keywords: Fall detection · Action structured method · Pose estimation · Multichannel · Cascaded dilated graph convolution network

1 Introduction

Fall detection has a wide demand of application and research significance in the field of safety monitoring in smart pension, smart city [1] and smart factory [2]. In the field of smart pension, real-time monitoring of the fall action of the elderly can effectively reduce the casualty rate, and enable the elderly to receive treatment in the first time. In the field of public safety monitoring [3] in smart city and smart factory, fall action is also an important detection action for safe production. An effective and rapid detection method for fall can improve people's quality of life and production level. Most existing visual-based fall detection methods can only recognize fall action in simple environments.

When there are moving pedestrians or objects in a complex background, the image-based method mixes the features of multiple people to detect falls. These extracted features are inaccurate to classify each person, which leads to failed detection. Some methods use the human skeleton to reduce the interference of background, such as scene or light changes. These methods rely on the pose estimation algorithms [4] to obtain the human skeleton representation. The scene in the background can be eliminated and reserve the skeleton information, but the pose estimation algorithm may incorrectly process some of the back-ground pixels as human skeleton, resulting in inaccurate extraction of action features by deep learning network. As shown in Fig. 1, cartons are mistaken for human skeleton, and the error processing representation is shown on the right part of the image, which makes the feature redundancy unable to discriminate the fall action.
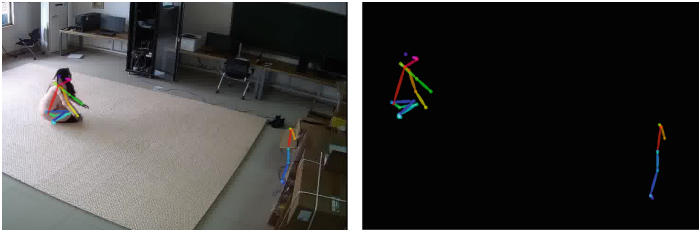
Fig. 1. The pose estimation algorithm error processes the skeleton representation.

In addition, the existing methods only extract action features from RGB images, but neglect to extract features from human joint coordinates, resulting in poor recognition accuracy. These methods use pixel-level features to detect falls, which leads to lack of feature scale invariance and loss of structured information. Human skeleton contains structured information which is hard to extract from image.

In this paper, a fall detection method based on action structured method and cascaded dilated graph convolution neural network is proposed. The contributions of the proposed method are summarized as follows:

(1) An action structured method (ASM) is proposed to model the skeleton of human actions by using pose estimation algorithms, which removes the interference of complex background. Also, the object detection algorithm is utilized to locate multiple people to transfers the fall detection issue of multi-person to single person fall detection.

(2) The cascaded dilated graph convolution network (CD-GCN) is proposed to enlarge the receptive field by the dilated operation, effectively extracts action features from the coordinates of joint nodes, and fuses the multichannel features with different dilation rates. Then finally obtains the classification results.

2   Related Works

Automatic fall detection technology has become a hot research topic in recent years [5–7]. With the aging of Chinese population and the development of public security

technology in smart city, the automatic detection of human falls is of great importance for protecting vulnerable groups such as the elderly and children and ensuring public safety. Falling is the second cause of accidental injury for people of all ages, and it is the primary cause of accidental injury for people over 79 years old [8]. Fall detection methods are mainly based on three categories of wearable sensors [9], environmental sensors [10] and machine vision [11]. In addition, many published methods focus on general motion recognition and video understanding, rather than specific fall detection. There is a problem of inaccurate extraction due to the small gap between feature classes. The wearable sensor method is to detect the acceleration and position of the human movement by wearing one or more sensors on the pedestrian to analyze whether it conforms to the motion characteristics of human falling [12–14]. These methods are uneconomical and can only detect the wearer of the sensor, which cannot cover all people. Environmental sensor method uses multiple sensors arranged in the environment to detect floor vibration, the current generated near the fall, falling sound and other information to judge human fall [15]. This method has poor anti-interference and high misjudgement rate, and cannot be widely used. The machine vision method judges human fall action through video images. Literature [16] designed local spatio-temporal points of interest were designed to represent features, and then the support vector machine (SVM) was used to classify and identify fall. In literature [17], human joint information is obtained by object detection algorithm, and feature vectors are formed, and then the integrated classifier is used to detect fall. Literature [18] proposed a multi-feature fusion detection method. In literature [19], a deep neural network-based method for identifying fall behaviour is proposed. Abobakr et.al [20] proposed a method based on skeleton information and random decision forest to extract fall features, and then a method using human skeleton information is proposed to remove background information through support vector machine classification. Xin et al [21] proposed a method of removing background information by using human skeleton information, and then extracting temporal and spatial features of actions through three-dimensional convolution neural network. Mastorakis et al. [22] proposed a human modeling method with other occlusion and used Hausdorff distance measure for fall detection. Panahi et al. [23] used support vector machine to classify the action features of depth images. When there are moving pedestrians in complex backgrounds, the image-based methods mixed the feature of multiple people to detect the fall. These extracted features are inaccurate, leading to detection failure. Moreover, the existing methods only extract action features from RGB images, neglecting to extract features from human joint coordinates, resulting in poor recognition accuracy. These methods use pixel-level features to detect falls, resulting in the invariance missing of feature scale and loss of structured information.

In order to solve the abovementioned problems, a fall detection method based on ASM and CD-GCN is proposed. The proposed ASM models the skeleton of human action through the pose estimation algorithm, which removes the interference of complex background. In addition, the object detection algorithm is utilized to locate multiple people to transfers the fall detection issue of multi-person to single person fall detection. The proposed CD-GCN enlarges the receptive field by the dilated operation, effectively extracts the action features from the coordinates of joint nodes, and fuses the multichannel feature with different dilation rate, then finally obtains the classification results.

3   Overview of Proposed Method

In this paper, a fall detection method based on ASM and CD-GCN is proposed. As shown in Fig. 2. The ASM is proposed to conduct skeleton model of human action through pose estimation algorithm [24] to remove the interference of background, and classify pedestrians by using the object detection algorithm [25] to transfer the fall detection issue of multi-person to single person fall detection. The proposed deep learning network based on the CD-GCN enlarges the receptive field and effectively extracts the fall features from joints coordinates. Finally, the classification results are obtained.
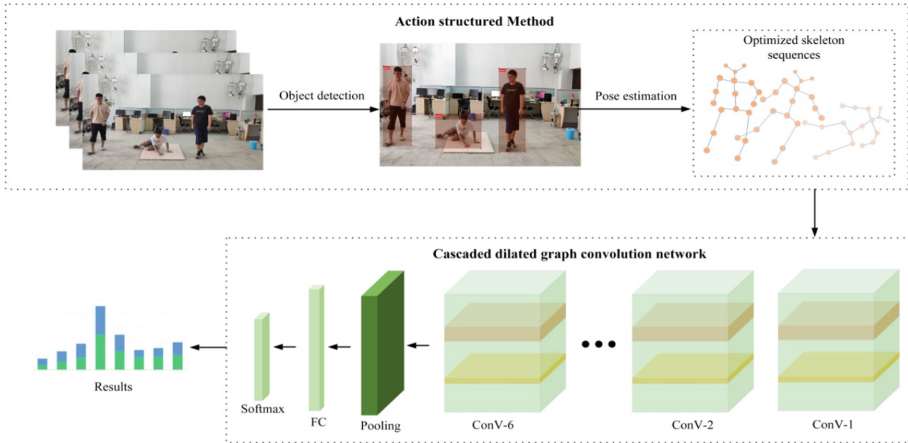
Fig. 2. The structure of the proposed method.

4   Action Structured Method

In order to remove the background interference in the image, as shown in Fig. 3, an action structured optimization method (ASM) is proposed in this paper. Firstly, the proposed method utilizes the YOLOv5 [25] object detection algorithm to classify the people in image, and uses the detection coordinates to reduce the image size. The ASM reduces the computational complexity of subsequent processing and improves the processing efficiency. Secondly, the coordinate information of each human skeleton in the image is obtained through pose estimation algorithm, OpenPose [24], and finally the skeleton sequence with spatio-temporal action features is obtained. The pose estimation algorithm has the problem of misrecognizing the background object as human body, which has been illustrated in Sect. 1, which results in redundant and erroneous human action information in the data. The object detection algorithm recognizes each person's region and then performs the pose estimation process, so that the skeleton information of each person can be obtained without the interference of background misjudgment, which transfers the fall detection issue of multi-person to single person fall detection and improves the feature extraction. Finally, the optimized skeleton sequences of each person with spatio-temporal features are obtained.
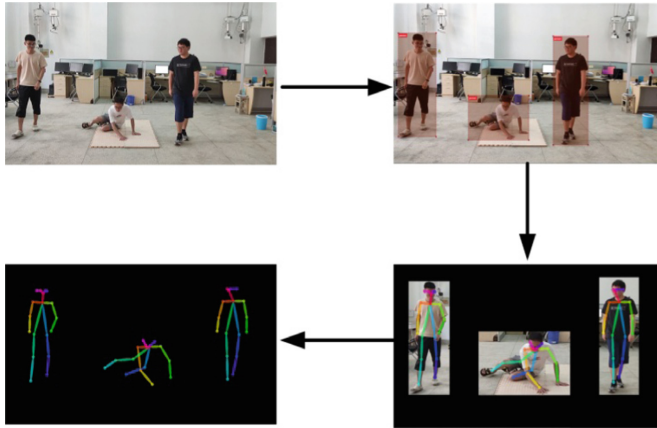
Fig. 3. The process of ASM.

5   Cascaded Dilated Graph Convolution Neural Network

In recent years, graph convolution network (GCN) theory has attracted more and more attention from researchers [26–29]. The advantage of graph convolution is that it can process data with non-Euclidean structure. The skeleton structure of human body is a natural non-Euclidean data, and the coordinates of each joint are native and original feature vectors. However, the existing methods extract action features from RGB images, which are not only susceptible to background interference, but also have a large gap in the extracted fall features, resulting in poor accuracy of detection methods. A cascaded dilated graph convolution neural network (CD-GCN) for feature extraction and fall detection is proposed. The proposed dilated graph convolution is shown in Fig. 4. The dilated theory is transformed from CNN to GCN.
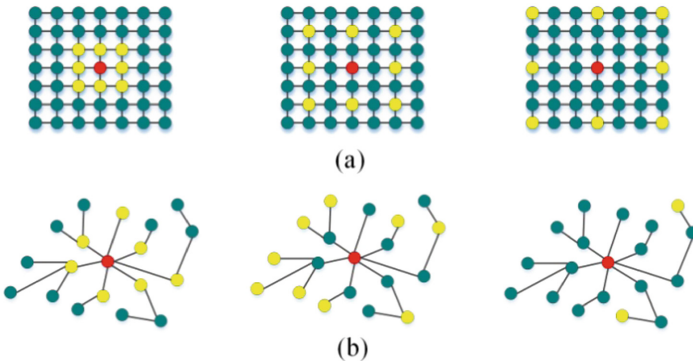
Fig. 4. Figure (a) is the traditional CNN with regular structure; (b) is the non-Euclidean structure. The red node is the central node, the yellow node is the aggregated node with different dilation rate. (Color figure online)

The idea of dilated graph convolution is enlightened from traditional convolution neural network. As shown in Fig. 4, (a) shows the dilated convolution under the grid structure data based on RGB pixel rules, (b) shows the dilated graph convolution of non-European structured data. It can be seen that the aggregated pixels or nodes (yellow) of the central node or pixel point (red) are gradually diffused around. The degree of diffusion is defined as dilation rate, and the distance from the central node to aggregated node is taken as the quantitative index. Dilation rate $dr = 1$ means that the distance from the central node to the aggregated node is 1. In this paper, the feature of a cascaded dilated graph convolution is an aggregated eigenmatrix of different dilation rate. As shown in Fig. 5, the structure of the cascaded graph convolution operation is proposed. Features of different dilated rate are cascaded and fused.
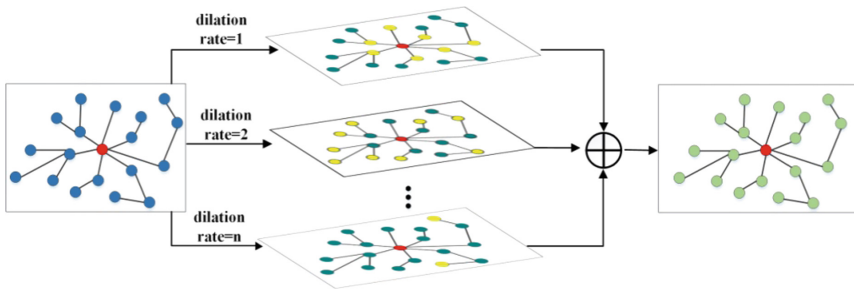
Fig. 5. The structure of cascaded dilated graph convolution operation.

As shown in Fig. 6, the proposed cascaded dilated graph convolution neural network in this paper consists of six convolution layers, one pooling layer, one full connection layer and one Softmax layer. The joint coordinates of the optimized skeleton sequence into the network as the input, and a $C \times T \times V$ tensor is designed, where T means the number of video frames, V means the number of human joints, and C means the joint coordinate data dimension. In this paper, $C = 3$, $T = 40$, $V = 18$ are set. The mathematical expression of graph convolution operation is shown in formula (1):

$$f = \Lambda^{-\frac{1}{2}}(A_m + 2I)\Lambda^{-\frac{1}{2}}f_{in}W \tag{1}$$

where f means the feature matrix, A_m means the adjacency matrix of human body structure, here is the cascaded graph data, and I means the identity matrix with the same dimension as A_m. In this paper, the mathematical expression of A_m is shown in formula (2):

$$A_m = A_1 + A_2 + \cdots + A_n \tag{2}$$

where A_1 means the adjacency matrix with dilation rate $dr = 1$, A_2 means the adjacency matrix with the $dr = 2$, and A_n means the adjacency matrix with $dr = n$. Then the formula (1) transfers to formula (3):

$$f = \sum_{m=1}^{n} \Lambda^{-\frac{1}{2}}(A_m + 2I)\Lambda^{-\frac{1}{2}}f_{in}W \tag{3}$$

In this paper, the accuracy performance with different dilation rate is studied. When the convolution strategy of $dr = 1$ and $dr = 2$ are fused to extract features, the optimal recognition accuracy is obtained.
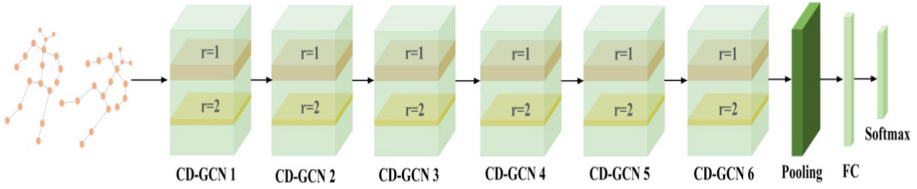

Fig. 6. The structure of proposed cascaded dilated graph convolution network.

6 Experiments

6.1 Datasets and Implementation Setup

Experiments and analyses are implemented on three public data sets and one self-collected dataset. These three public datasets are single-person fall datasets in a simple environment. UR dataset [30] contains 30 falling videos and 40 daily action videos. FDD dataset [31] contains 191 videos. MCFD dataset [32] contains 192 videos, including 7 types of actions such as fall, sit, walk and run. In this paper, the UR, MCFD and FDD datasets are divided into two categories: fall and daily behaviour. The self-collected dataset, for 5 classes, uses HIKVISION DS camera to collect video data. As shown in Fig. 7, we collected more than 150 videos with complex scenarios, with each video length of 150–250 frames and each video lasts 5–9 s. We also collected 300 daily non-fall videos with 4 classes, including normal walking, squatting, lying down and sitting.

Fig. 7. Example frames of self-collected dataset.

The experiment is implemented on Ubuntu18.04 system, Intel Xeon E2 CPU and NVIDIA 16 GB RTX5000 GPU. The program is implemented in Pytorch 0.4.1 and Pycharm. The initial learning rate of the cascaded dilated graph convolution network is 0.1, the weight decay is 0.0001, and the batch size is 64. A total 30 epochs are trained, and the learning rate decayed every 10 epochs.

6.2   Comparison with State-of-Art Methods

A fall detection method based on ASM and CD-GCN is proposed. The ASM models the skeleton of human action through the pose estimation algorithm, which removes the interference of complex background. Also, the object detection algorithm is utilized to locate multiple people to transfers the fall detection issue of multi-person to single person fall detection. The proposed CD-GCN enlarges the receptive field by the dilated operation, effectively extracts the action features from the coordinates of joint nodes, and fuses the multichannel feature with different dilation rates, then finally obtains the classification results.

Evaluation of ASM. The ASM is proposed to conduct skeleton modeling of human action through pose estimation algorithm, and to remove the interference of background. The pose estimation algorithm has the problem of misrecognizing background objects as human body, which leads to redundant and erroneous human action information in the data. The object detection algorithm recognizes each person's region and then performs the pose estimation processing, so that the skeleton information of each person can be obtained without the interference of background misjudgment, which transfers the fall detection issue of multi-person to single person fall detection and improves the efficiency of feature extraction. Experiments results on four datasets are shown in Table 1. As shown in the table, experiments were performed with and without ASM, respectively, and performance evaluations were performed. The experiments show that the proposed method achieves the highest accuracy on four datasets. When only the human skeleton is input and ASM is not used, there is the possibility of multi-person skeleton interference and pose estimation misrecognizing the limb skeleton. The accuracy of the proposed method is slightly improved on the single-person fall dataset in a simple environment, while it is greatly improved on the self-collected dataset in a complex multi-person scenario.

Table 1.   The accuracy evaluation of ASM (%).

Method	UR	FDD	MCFD	Self-collected
Without ASM	98.9	98.0	99.3	88.8
With ASM	**99.3**	**98.9**	**99.5**	**95.6**

Evaluation of CD-GCN. The proposed cascaded dilated graph convolution network (CD-GCN) is proposed to enlarge the receptive field by the dilated operation, effectively extracts the action features from the coordinates of joint nodes, and fuses the multichannel feature with different dilation rate. Then finally obtains the classification results. The evaluation of CD-GCN is shown in Table 2, the dilation combination of the cascade dilated graph convolution is evaluated. Dilation rate $dr = 1$ means the convolution without any dilation process. The experimental results show that when the features of the cascade dilation rate are 1 and 2, the proposed method achieves the highest accuracy.

Table 2. The accuracy evaluation of CD-GCN with different dilation rate. (%)

r	UR	FDD	MCFD	Self-collected
1	99.0	97.3	97.9	94.9
2	90.9	85.0	90.9	86.3
3	70.1	67.6	73.7	62.5
1+2	**99.3**	**98.9**	**99.5**	**95.6**
1+3	98.2	96.5	94.4	91.6
1+2+3	96.9	96.7	90.3	90.4

6.3 Comparison with State-Of-Art Methods

In this paper, the proposed method is compared with the state-of-the-art methods in recent years, as shown in Table 3. Literature [7–14] uses a method based on machine vision. [16–23]. These methods ignore the structured optimization of human body information, resulting in inaccurate detection, especially in a multi-person scene with a complex background, which is susceptible to interference from other action features. In addition, these methods only extract action features from RGB pixels, and neglect to extract features from human joint coordinates. The proposed method achieves accuracy improvement about 0.7%, 0.6% and 0.5% on UR, FDD and MCFD datasets, and 1.9% on self-collected datasets. The experiment proves the superiority of the proposed method.

Table 3. The accuracy comparison of proposed method with state-of-art methods. (%)

Method	UR	FDD	MCFD	Self-collected
Su [16]	96.3	97.5	98.1	93.7
Zhao [17]	97.2	98.3	99.0	89.9
Li [18]	83.5	87.9	90.0	80.9
Fan [19]	94.3	92.9	97.5	91.1
Abobakr [20]	96.1	95.2	96.1	92.3
Xin [21]	83.5	87.9	90.0	89.4
Mastorakis [22]	98.6	94.3	97.3	93.2
Panahi [23]	97.1	90.9	93.6	90.5
The proposed method	**99.3**	**98.9**	**99.5**	**95.6**

7 Conclusions

In this paper, a cascade dilated graph convolution neural network for fall detection in complex scenes is proposed. In order solve the problem that existing methods are suscep- tible to the interference of complex backgrounds, the proposed ASM uses pose estimation

algorithms to model human action skeleton, and combines object detection algorithm to remove the interference of objects and pedestrians in the background. Aiming at the problem of inaccurate extraction of action structure features in existing methods, a cascaded dilated graph convolution network is proposed to expand the receptive field and effectively extract the action features from joint point coordinates. In the future, human body occlusion and multi-limb overlap action recognition methods will be studied.

References

1. Buzachis, A., Celesti, A., Galletta, A., Fazio, M., Fortino, G., Villari, M.: A multi-agent autonomous intersection management (MA-AIM) system for smart cities leveraging edge-of-things and blockchain. Inf. Sci. **522**, 148–163 (2020)
2. Zhang, R.: Improved control for industrial systems over model uncertainty: a receding horizon expanded state space control approach. IEEE Trans. Syst. Man Cybern. Syst. **50**(4), 1343–1349 (2020)
3. Ahire, S.K., Wankhade, N.R.: Context-aware local binary feature learning for face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(5), 1139–1153 (2019)
4. Fang. H.S., Xie. S., Tai. Y.W., Lu. C.: RMPE: regional multi-person pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2353–2362 (2017)
5. Sadreazami, H., Bolic, M., Rajan, S.: Contactless fall detection using time-frequency analysis and convolutional neural networks. IEEE Trans. Industr. Inf. **17**(10), 6842–6851 (2021)
6. Tahir, A., Morison, G., Skelton, D.A., Gibson, R.M.: A novel functional link network stacking ensemble with fractal features for multichannel fall detection. Cogn. Comput. **12**(5), 1024–1042 (2020). https://doi.org/10.1007/s12559-020-09749-x
7. Mrozek, D., Koczur, A., Maysiak-Mrozek, B.: Fall detection in older adults with mobile IoT devices and machine learning in the cloud and on the edge. Inf. Sci. **537**(5), 132–147 (2020)
8. Mubashir, M., Shao, L., Seed, L.: A survey on fall detection: principles and approaches. Neurocomputing **100**, 144–152 (2013)
9. Qian, X., Chen, H., Jiang, H., Green, J., Cheng, H., Huang, M.: Wearable computing with distributed deep learning hierarchy: a study of fall detection. IEEE Sens. J. **20**(16), 9408–9416 (2020)
10. Liu, J., Tan, R., Han, G., Sun, N., Kwong, S.: Privacy-preserving in-home fall detection using visual shielding sensing and private information-embedding. IEEE Trans. Multimedia **23**, 3684–3699 (2020)
11. Khan, S.S., Hoey, J.: Review of fall detection techniques: a data availability perspective. Med. Eng. Phys. **39**, 12–22 (2017)
12. Medrano, C., Plaza, I., Igual, R., Sanchez, A., Castro, M.: The effect of personalization on smartphone-based fall detectors. Sensors **16**(1), 117 (2016)
13. Cola, G., Avvenuti, M., Vecchio, A., Yang, G.Z., Lo, B.: An on-node processing approach for anomaly detection in gait. IEEE Sens. J. **15**(11), 6640–6649 (2015)
14. Wei, W., Song, H., Li, W., Shen, P., Vasilakos, A.: Gradient-driven parking navigation using a continuous information potential field based on wireless sensor network. Inf. Sci. **408**, 100–114 (2017)

15. Rimminen, H., Lindstrom, J., Linnavuo, M., Sepponen, R.: Detection of falls among the elderly by a floor sensor using the electric near field. IEEE Trans. Inf Technol. Biomed. **14**(6), 1475–1476 (2010)
16. Su, S., Wu, S.-S., Chen, S.-Y., Duh, D.-J., Li, S.: Multi-view fall detection based on spatio-temporal interest points. Multimedia Tools Appl. **75**(14), 8469–8492 (2015). https://doi.org/10.1007/s11042-015-2766-3
17. Zhao, X., Hu, A., He, W.: Fall detection based on convolutional neural network and XGBoost. Laser Optoelectron. Progress **57**(16), 248–256 (2020)
18. Li, Y., Yang, B.: Fall detection method based on ViBe algorithm and multi-feature fusion. Chin. J. Electron Devices **42**(6), 1583–1589 (2019)
19. Fan, Y., Levine, M.D., Wen, G., Qiu, S.: A deep neural network for real-time detection of falling humans in naturally occurring scenes. Neurocomputing **260**, 43–58 (2017)
20. Abobakr, A., Hossny, M., Nahavandi, S.: A skeleton-free fall detection system from depth images using random decision forest. IEEE Syst. J. **12**(3), 2994–3005 (2018)
21. Xiong, X., Min, W., Zheng, W.-S., Liao, P., Yang, H., Wang, S.: S3D-CNN: skeleton-based 3D consecutive-low-pooling neural network for fall detection. Appl. Intell. **50**(10), 3521–3534 (2020). https://doi.org/10.1007/s10489-020-01751-y
22. Mastorakis, G., Ellis, T., Makris, D.: Fall detection without people: a simulation approach tackling video data scarcity. Expert Syst. Appl. **112**, 125–137 (2018)
23. Panahi, L., Ghods, V.: Human fall detection using machine vision techniques on RGB-D images. Biomed. Signal Process **44**, 146–153 (2018)
24. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D Pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2019)
25. Tan. S., Lu. G., Jiang. Z., Huang. L.: Improved YOLOv5 network model and application in safety helmet detection. In: Proceedings of the IEEE International Conference on Intelligence and Safety for Robotics, pp. 330–333 (2021)
26. Kipf. T.N., Welling M.: Semi-supervised classification with graph convolutional networks. In: Proceedings of the International Conference on Learning Representations (2016)
27. Albert, M.M., Javier, R.H.: 2D–3D geometric fusion network using multi-neighbourhood graph convolution for RGB-D indoor scene classification. Inf. Fus. **76**, 46–54 (2021)
28. Qin. L., Che, W., Ni. M., Li. Y., Liu. T.: Knowing where to leverage: context-aware graph convolution network with an adaptive fusion layer for contextual spoken language understanding. In: Proceedings of the IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 1280–1289 (2021)
29. Zi, W., Xiong, W., Chen, H., Chen, L.: TAGCN: station-level demand prediction for bike-sharing system via a temporal attention graph convolution network. Inf. Sci. **561**, 274–285 (2021)
30. Kwolek, B., Kepski, M.: Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput. Methods Programs Biomed. **117**(3), 489–501 (2014)
31. Charfi. I., Miteran. J., Dubois. J., Atri. M., Tourki. R.: Definition and performance evaluation of a robust SVM based fall detection solution. In: Proceedings of the IEEE Eighth International Conference on Signal Image Technology and Internet Based Systems, pp. 218–224 (2012)
32. Auvinet, E., Multon, F., Alain, S.A., Rousseau, J., Meunier, J.: Fall detection with multiple cameras: an occlusion-resistant method based on 3D silhouette vertical distribution. Proc. IEEE Trans. Inf. Technol. Biomed. **15**(2), 290–300 (2011)
33. Lu, N., Wu, Y.D., Feng, L., Song, J.B.: Deep learning for fall detection: three-dimensional CNN combined with LSTM on video kinematic data. IEEE J. Biomed. Health Inform. **23**(1), 314–323 (2019)

Surface Reconstruction Method of Color 3D Image Based on Independent Adjustable Sparse Coefficient

Jiangang Yin[1(✉)], Mingnian Zhang[2], and Ning Zhang[3]

[1] State Grid Hubei Electric Power Co., Ltd, Wuhan 430000, China
yjgg21@163.com
[2] State Grid Huanggang Electric Power Co., Ltd, Huanggang 430000, China
[3] Guangdong Polytechnic of Environmental Protection Engineering, Foshan 528000, China

Abstract. Color 3D images play an important role in many fields. The segmentation effect of traditional 3D image surface reconstruction methods is poor, resulting in unreliable reconstruction results. Therefore, a new surface reconstruction method of color 3D image is designed based on the independent adjustable sparse coefficient. Firstly, the color 3D image is enhanced, and then the point cloud registration algorithm is designed based on the independent adjustable sparse coefficient. The surface reconstruction of the color 3D image is realized by the point cloud registration algorithm. The experimental results show that the reconstruction effect of this method is good, which proves that it has high application value.

Keywords: Independently adjustable sparsity coefficient · Color three-dimensional image · Surface reconstruction · Image enhancement · Point cloud registration algorithm

1 Introduction

Vision is the main way for human beings to capture information from the real world. However, the information obtained from two-dimensional scene is limited. Therefore, three-dimensional technology has emerged [1]. Computer vision is to make the computer have the ability to perceive and analyze the information of the external environment, install "eyes" for the computer, and make it have the function similar to human eyes. This technology enables the machine to record the physical information of the object in the environment, and then analyze and calculate the object [2].

Relevant scholars summarized the research results of many disciplines in 1982 and put forward a series of computer vision theories. Its core idea is to reconstruct the three-dimensional structure through two-dimensional images. Computer vision has attracted more and more attention and has been applied to various fields, such as visual image processing, various inspection and monitoring, robot autonomous recognition and instrument navigation, commercial field, space technology and military simulation [3].

The research of 3D reconstruction technology involves many disciplines and is a hot research direction in the field of computer vision technology. It is to explore how to build the two-dimensional model of the object into a three-dimensional model in the computer, so that the computer can analyze and process it. In the process of reconstructing object model, depth image acquisition and model reconstruction are the most key parts. The depth image, also known as the distance image, can be obtained by a three-dimensional information sampling device. However, due to the limitations of the principle of the equipment itself, the obtained depth image will have problems such as black holes and noise, so it is necessary to enhance the depth image. The key step in model reconstruction is to convert the obtained 3D data point cloud data into the 3D surface model of the object, which is also called point cloud registration. The quality of reconstruction results will be directly affected by the accuracy of point cloud registration. Therefore, in order to obtain a 3D model with high reconstruction degree, it is necessary to realize high-precision point cloud registration [4].

With the development of computer hardware and software technology, the world presented by two-dimensional images can no longer satisfy people's visual experience, and thus the depth image is born, and then the three-dimensional real world is obtained. Nowadays, it has become easier to obtain better depth image information using depth sensors, making depth images the focus of researchers, and applications based on depth images are also developing rapidly [5].

In recent years, technological progress has promoted the development of reverse engineering research. In reverse engineering, you first need to use a laser scanner to obtain geometric data, and then obtain a three-dimensional model of the object through certain operations [6]. The update of laser measurement technology makes the laser scanner more efficient, lighter and more powerful, so that the acquired point cloud data is more accurate, providing a solid foundation for deeper application, and it has demonstrated powerful functions in all walks of life. The advancement of nature and technology.

Based on the above analysis, this paper designs a new surface reconstruction method of color 3D image based on independently adjustable sparse coefficients. In this study, the method firstly enhances the detail delicacy in color 3D image, and fundamentally improves the reconstruction effect by clarifying the detail information. Then, a point cloud registration algorithm was designed based on the independent adjustable sparse coefficients, and the surface reconstruction of color 3D images was realized through the point cloud registration algorithm.

2 Method Design

2.1 Color 3D Image Enhancement Processing

The pixel in the grayscale image represents the brightness value, and the pixel in the depth image represents the distance from the point to the camera, that is, the depth value or distance value, so the depth image is also called the distance image [7]. The relationship between depth image and grayscale image is shown in Fig. 1.

As can be seen from Fig. 1, the depth value represents the distance between the target point and the measuring instrument. Because the depth value is only related to
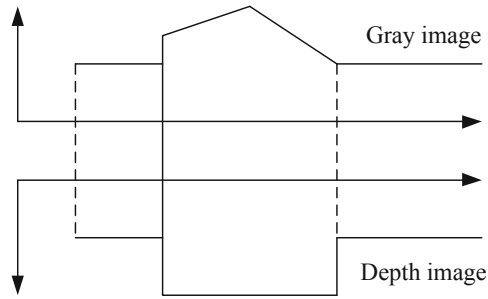
Fig. 1. The relationship between depth image and grayscale image

the distance and has nothing to do with external factors, the depth image can correctly reflect the physical depth information of the scene. By constructing the three-dimensional model of the object, it can provide a more solid foundation for higher-level computer vision applications.

The primary task in 3D reconstruction is to obtain the 3D information of the object, and the depth sensor has become an indispensable device to obtain the depth information of the object [8]. The depth sensor can be divided into active depth imaging sensor and passive depth imaging sensor according to the need for external emission light source. The active depth imaging sensor needs the help of external emission light source when collecting image information. The passive depth imaging sensor is just opposite to the active depth sensor. It only uses its own principle and the information existing in the environment when collecting image information, and does not need the help of external emission light source. The following is a brief overview of the technologies involved in the depth sensor.

Stereo vision technology is a passive depth imaging technology without external light source. Stereo vision technology reconstructs three-dimensional information of the scene by obtaining multiple images from different perspectives from multiple cameras at multiple locations. So far, stereo vision technology has a relatively perfect and mature system, which can be divided according to the number of images required in the process of three-dimensional reconstruction. If the number of images required is 2, two images taken from two different directions of the same scene are needed to obtain depth information. It is also called binocular stereo vision technology because of its similar structure to the information acquired by the human eye [9].

Lidar ranging technology requires the use of laser light sources to calculate depth information based on the time difference between laser emission and reception and the light propagation speed. Its core technology is light flight time ranging technology. When using laser ranging, the light source periodically emits a laser signal to the measured object, and then the signal hits the measured object and returns to the imaging device. Record the time required for this process, and then the distance between the camera and each surface of the object It can be calculated from this time. The advantages and disadvantages of these technologies are shown in Table 1.

Technology	Advantage	Shortcoming
Stereo Vision Technology	Suitable for short-distance high-precision measurement	Affected by camera performance, lighting and baseline distance
Lidar ranging technology	Simple and fast	Affected by background light and diffuse reflection
Structured light technology	Wide range and high precision	Susceptible to physical optics and occlusion

It can be seen from Table 1 that structured light technology belongs to active depth imaging technology, and its core idea is the idea of triangulation. The key to structured light imaging technology is to obtain fringes with recognizable codes. The formation of these fringes will be affected by the projection distance of the light source, and different fringes will be formed due to different surfaces of the object.

In June 2010, Microsoft designed a somatosensory device called Kinect for the Xbox360 game console. Kinect is mainly composed of three parts: color camera, infrared 3D depth sensor, microphone [10]. Based on the above-mentioned infrared camera and color camera, the device can acquire depth images and color images at the same time. Multi-array microphones can collect sound information, thus providing a more powerful way of human-computer interaction.

The core component of Kinect is Prime Sensor, and the core of this component is the PS1080 system-level chip. It is precisely because of this chip that Kinect has a powerful human-computer interaction method. And Prime Sensor equipment uses optical coding technology, the key is to use the camera to obtain structured light patterns. In the entire measurement process, a COMS sensor can be completed, reducing the cost of testing.

The traditional structured light ranging technology uses structured light patterns to estimate the distance. The specific process is: first use the active light source to project the distinguishable patterns onto the objects in the scene to form different structured light patterns, and then use the camera to receive these structured light patterns reflected by the object, and analyze the position and The structured light pattern of the degree of deformation can calculate the distance from the object to the camera. The Light Coding technology l38 used by Kinect is a kind of structured light technology, but the difference from the traditional structured light technology is that Light Coding is realized through a technology called "laser speckle" [11].

Laser speckle has high randomness and is very sensitive to the change of distance. In other words, in the scene to be tested, the shape or size of speckle pattern in any two regions is different. If you want to obtain the "speckle pattern" of the whole scene to be tested, that is, some mark is made on the distance information of the whole scene to be tested, so that the distance information of an object in the scene to be tested can be converted into the "speckle pattern" of the object. Through the "speckle pattern" you can get the specific location information of the object in the scene. Therefore, the light

source needs to be calibrated to obtain all speckle patterns of the scene to be measured, as shown in Fig. 2.
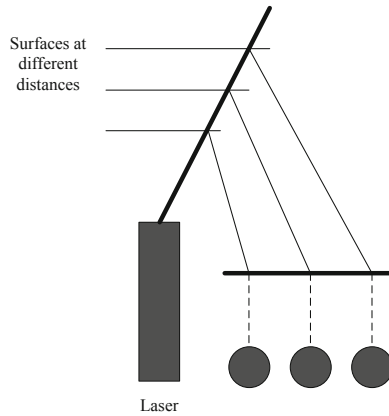

Fig. 2. Principle of speckle pattern formation

It can be seen from Fig. 2 that the used light waves formed by the interference between multiple light waves have different frequencies and have different phases and amplitudes. The speckle pattern is actually a random diffraction pattern. Laser speckle uses the higher coherence of laser light to make the diffraction result more prominent. The information recorded by the speckle image contains information about the light and the objects that the light can pass through, and the speckle images of any two regions in space are different. Therefore, people use this feature to measure the roughness of the reflective surface, the movement speed of the object, and the electronic speckle photography technology and electronic speckle interferometry used in the industrial field.

The speckle pattern can reduce the optical complexity of the projection system to a large extent, the reason is that its measuring depth is very large, generally about a few meters. Assuming that the distance between the object and the camera is Z, the average vertical and horizontal dimensions of the speckle here are as follows:

$$S_L = \lambda \left(\frac{Z}{\phi} \right)^2 \tag{1}$$

$$S_T = \frac{\lambda Z}{\phi} \tag{2}$$

In the formula, λ represents the wavelength of the emitted light, and ϕ represents the size of the diffuser. When Kinect works, the COMS sensor receives the infrared light reflected from the object. Some areas in the depth image appear black, and the depth value is zero. In fact, the sensor does not receive the reflected speckle pattern.

2.2 Independently Adjustable Design Point Cloud Registration Algorithm Based on Sparse Coefficient

The super-resolution reconstruction of a single image can be defined as the process of generating a corresponding high-resolution image by adding more details and resolution on the basis of the original low-resolution single input image, and generating a high-resolution image from the low-resolution input image through a predetermined mathematical model. The pixel intensity value obtained by the weighted average interpolation method is locally similar to that of adjacent pixels, which can generate a better smooth region, but it does not perform well when high-frequency regions such as edges produce large gradients [12]. The high score image is reconstructed by using the image prior information learned from the edge features. Edge is an important original image structure, which plays a decisive role in visual perception. Because the prior information is learned from the edge, the reconstructed image has high edge quality, appropriate sharpness and less artifacts. However, the effectiveness of edge a priori in modeling other high-frequency structures is poor.

With the introduction of machine learning and compressed sensing theory, learning-based super-resolution algorithms are increasingly used in single-image super-resolution applications [13]. This paper proposes a sparse coding super-division method based on sparse coding and dictionary learning. The edge sharpness of the super-division image is better but the artifacts are more obvious. The joint dictionary training generates a complete dictionary and proposes a method to build a sample library based on the redundant information of the image itself. The sample library is generated by interpolation, but the adaptive ability is poor. The training of the high-score dictionary and the low-score dictionary are carried out separately, the dimensionality of the image block is reduced, the neural network approximation is introduced and the training of the high-score dictionary is simplified with the pseudo-inverse, which improves the over-score speed. The dictionary learning is no longer limited to small-scale, definite training samples, and the accuracy of dictionary training is improved, and the application effect of image restoration is better. The super-resolution reconstruction based on ODL is compared with other algorithms, but the improvement effect is limited. Double sparse coding in the wavelet domain, the reconstruction effect is better. Combining the advantages of reconstruction and learning methods, a multi-level dictionary super-resolution reconstruction algorithm based on image pyramid is proposed. The learning-based super-score algorithm represented by the SCSR method can achieve a better single-image super-score effect. If it can effectively suppress the edge artifacts and increase the texture details, the super-score effect will be greatly improved.

In view of the problems in obtaining depth images by Kinect, many researchers have proposed methods to repair depth data, but many methods benefit from the traditional depth data repair algorithms. In view of the noise phenomenon and low resolution of traditional depth data, the traditional depth data repair algorithm has a good enhancement effect. However, for Kinect's strong noise and large area loss, the traditional depth data enhancement algorithms need to be improved. Therefore, these traditional algorithms have great reference value for the research of depth data restoration based on Kinect. The following briefly introduces the main two types of algorithm principles and their

representative algorithms, namely, depth data enhancement algorithm based on local information and depth data enhancement algorithm based on global optimization.

The basic idea of depth image enhancement algorithm based on local information is to calculate it through the similar pixel with depth data [14]. Most of these algorithms are based on linear transformation. Let p and q denote two pixels and their positions in the image, and h denotes the weight coefficient of the pixels in the neighborhood, then the output image is as follows:

$$g(p) = \sum_{q \in N(P)} h(p-q) f(q) \tag{3}$$

Among them, h in the linear transformation is generally constant in the entire image. In order to facilitate the calculation of the center of gravity of the neighborhood, a matrix with an odd number of rows and columns is generally used to represent h.

In practical applications, many methods only use the information of one image, and use the information as a guide to filter the secondary image, that is, for different pixels, the guided image will form different filter cores. Bilateral filtering is composed of two sub-filter cores, one is related to spatial distance, called spatial domain weight, and the other is related to pixel value difference, called range weight. The output J of bilateral filtering at point p is as follows:

$$Jp = \frac{\sum_{q \in N(P)} s(p-q) r(I_P - I_q) I_q}{g(p)} \tag{4}$$

$$p = \frac{s(p-q)}{g(p)} \tag{5}$$

$$r(I_P - I_q) = \exp\left(-\frac{\|I_P - I_q\|^2}{2\sigma_r^2}\right) \tag{6}$$

In the formula, I_P, I_q represents the color value of the pixel, and σ_r represents the range parameter of bilateral filtering. The weight coefficient of the bilateral filter is inversely proportional to the spatial distance or the difference in pixel value. This feature makes the bilateral filter have the advantage of keeping the edges better. Since the filtering kernel of bilateral filtering is calculated based on the guided image, it is not constant during the filtering process of the entire image. In traditional bilateral filtering, the leading image is the original image. Subsequent researchers replaced the guided image with another image that was different from the original image, thereby obtaining a joint bilateral filtering algorithm. For the repair of the black hole in the depth image, the color image is usually used as the pixel value difference degree for calculation, and the black hole in the depth image is used as the filtering object, thus the following formula is obtained:

$$E = \exp\left(-\frac{I_P - I_q}{\sigma_r}\right) \tag{7}$$

$$E(x) = \frac{1}{E} = \frac{1}{\exp\left(-\frac{I_P - I_q}{\sigma_r}\right)} \tag{8}$$

$$J(p) = \frac{\sum\limits_{q \in N(P)} s(p-q)r(I_P - I_q)}{\sum\limits_{q \in N(P)} f(q)} \tag{9}$$

$$J(X) = \frac{\frac{I_P - I_q}{\sigma_r}}{\sum\limits_{q \in N(P)} s(p-q)f(q)} \tag{10}$$

Compared with the depth image enhancement algorithm based on local information, the depth image enhancement algorithm based on independently adjustable coefficients can retain more global information. The depth image enhancement algorithm based on local information converts the input image into the expected result through the local image repair operator. In order to obtain the ideal image processing effect, the depth image enhancement algorithm based on global optimization constructs the global energy term according to the image information and some limited conditions. One of the most popular ideas in this kind of methods is based on Markov random field model. The method based on the model constructs the maximum likelihood term based on the difference between the input image and the estimated image and the prior constraints that the latter needs to meet, so as to construct the final probability expression of the model and obtain the maximum value of the probability.

In fact, in the methods based on global optimization, it is a common method to solve the image processing problem by introducing regular terms. The method based on this idea has a great relationship with the image enhancement method of the above model. Its data item corresponds to the maximum likelihood item in the above model, and its smoothing item corresponds to the a priori item in the above model. Generally, researchers equate the above two. As mentioned above, in solving problems such as sparse image interpolation, image noise reduction, filling invalid pixels of the image, giving priority to the image enhancement method based on global optimization can obtain a good repair effect.

2.3 Color 3D Image Surface Reconstruction

Surface representation is the most basic method to represent the shape of three-dimensional material. It can provide comprehensive information of three-dimensional objects. There are two specific forms: boundary contour representation and surface surface representation.

The initial surface reconstruction method adopts the description method based on contour line, that is, in the sectional image, the deterministic segmentation of the target contour is realized manually or automatically, and then the contour lines of each layer are "stacked" to represent the boundary of the object of interest. This contour line representation method is simple and the amount of data is small, but it is not very intuitive. In addition to representing the object by contour, it can also be represented by the surface of the contour reconstructed object. The earliest method is based on polygon technology, mainly using the triangle algorithm of plane contour, fitting the surface of this group of surface contours with triangles, and solving the problem of three-dimensional

connectivity of a series of surface contours. The surface of the object is formed by filling the small planes of triangles or polygons between the adjacent boundary contours, and the result is only a piecewise smooth surface. Firstly, determine a surface threshold, calculate the gradient value in each voxel, compare it with the surface threshold, find out those cubes containing surfaces, and use the interpolation method to calculate these surfaces, which is actually the process of extracting isosurfaces.

The main advantage of the surface-based method is that it can be displayed using more mature computer graphics methods. The amount of calculation is small and the running speed is fast. With the help of dedicated hardware support, drawing on a high-performance PC can completely realize real-time interactive display.

Because volume rendering directly studies the relationship between light passing through the volume data field and voxels, there is no need to construct an intermediate plane, and many detailed information of the voxels are retained, and the fidelity of the results is greatly improved. In terms of the quality of the resulting image, volume rendering is better than surface rendering. But in terms of interactive performance and algorithm efficiency, at least on the current hardware platform, surface rendering is better than volume rendering.

The surface reconstruction of tomographic data is to deduce the spatial structure of the corresponding entities from the contour lines on a series of sections. In order to ensure the correctness and uniqueness of the derivation, the boundary of the entity is required to be composed of two-dimensional points, and the intersection of the entity and the plane is required to be two-dimensional. These assumptions are reasonable. The entire surface regroove process can be divided into two steps, topology reconstruction and geometric reconstruction. The former derives the topological representation of the entity, and the latter establishes the geometric representation of the entity.

The purpose of topology reconstruction is to classify the contour lines on each fault in the three-dimensional tomographic data set, confirm the entity to which each contour line belongs, and ensure the correctness of the reconstruction. Therefore, topological reconstruction is the basis of tomographic data reconstruction. When there are multiple contour lines, it means that there will be entity intersections, the problem is more complicated, and it is more necessary to reconstruct the topology first.

The classification of contour lines is described by a classification map. Each vertex of the classification map corresponds to a contour line, and its edges are connected to two contour lines belonging to adjacent layers. If there is an entity described by the classification map, the classification map is considered valid. In other words, a valid classification map corresponds to entities that meet the conditions.

Linear smoothing filter is suitable for many situations and its design is simple, which makes it an important method in signal processing, especially for signal spectrum and noise spectrum with obvious differences. However, for the signal with a wide spectrum, that is, the steep edge in the general sense, although the linear smoothing filter can smooth the noise, it will also blur the steep edge. At the same time, for impulse noise, the linear filter can not be completely smooth. Therefore, in most cases, median filter is used to solve the above problems. However, due to the lack of large-area depth pixels in the depth image, the repair effect of median filter is very poor. Therefore, the iterative idea is added to median filter to make the improved median filter suitable for the lack

of large-area depth pixels. On this basis, guided by image edge information and color image information, the detailed characteristics of image edge are guaranteed.

In the depth image obtained by Kinect, most of the positions of black holes are located at the edge of the object. It is difficult to accurately find the depth value only by using the color image information. To solve this problem, a depth image enhancement algorithm based on edge information guided filtering is proposed in this paper. Firstly, the edge of color image and depth image is extracted to obtain color edge image and depth edge image. The color edge image and depth edge image are enhanced respectively, and the enhanced edge image is fused to obtain the edge image as a guide; Then, guided by the fused edge image, the iterative median filter is used to fill the black hole; Finally, the adaptive median filter is used to smooth the noise of the image.

There are real edges and wrong edges in depth edge image. The pixels in the black hole area in the depth image are called invalid pixels. The invalid pixels are zero, and the effective pixels are greater than zero. In order to determine these two edges, it is necessary to detect the neighborhood pixel of an edge pixel in the depth edge image. If there is no invalid pixel, this pixel is the real edge pixel. If there is an invalid pixel, this pixel is the pixel of the wrong edge. Although the wrong edge is not the real edge of the object, it can provide a guide to find the real edge of the object in the black hole part.

When looking for the real edge in the black hole area, for each invalid pixel, calculate its vector gradient, the calculation method is the same as the above calculation of the color edge image pixel vector gradient. The boundary of the object in the black hole area always has two edges, so the pixel can be found on the other wrong edge. Through this method, the two boundaries around the black hole in the depth edge image are enhanced, thereby enhancing the depth edge image.

Through the above steps, an enhanced color edge image and an enhanced depth edge image are obtained respectively. In order to obtain more accurate image edge information, the enhanced edge image is fused. Since the edges of color images are more reliable, color edge images are the main ones. If a certain part of the color edge image is unreliable, then the edges of the depth edge image are used. The specific method is as follows: For the enhanced depth edge image, first calculate the vector direction of each pixel. If the edge of the color edge image is close to this direction, the edge of the color edge image is used, and the edge of the depth edge image is discarded, otherwise it will be The edge of the deep edge image replaces the edge of the color edge image. By fusing the enhanced color edge image and depth edge image, more accurate image edge information is obtained, which provides favorable guiding conditions for subsequent image processing.

3   Experiment and Analysis

In order to test the application effect of the color 3D image surface reconstruction method based on the independent adjustable sparse coefficient designed in this paper, it is compared with the traditional 3D image surface reconstruction method, and the following experiments are carried out.

3.1   Experiment Preparation

Using the ray projection algorithm, in the Windows environment, combined with the MFC and MITK platform to realize the volume rendering reconstruction of the image, and the reconstructed image can be observed in various directions. After setting different opacity, the volume reconstruction image with different transparency is obtained. At this time, the experiment process is shown in Fig. 3.
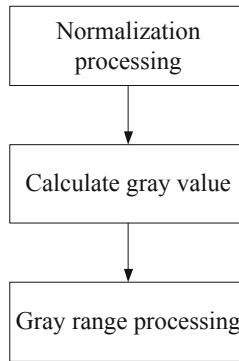
```
┌─────────────────────┐
│   Normalization     │
│    processing       │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│ Calculate gray value│
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│ Gray range processing│
└─────────────────────┘
```

Fig. 3. Image reconstruction process

Relevant experimental parameters are shown in Table 2.

Table 2. Experimental parameters

Project	Parameter
Training data set	DIV2K data set
Number of images	500
Rebuild times	80
Operating environment	Windows 10

3.2   Experimental Results and Discussion

The color three-dimensional image surface reconstruction method designed in this paper and the traditional color three-dimensional image surface reconstruction method are used for three-dimensional reconstruction respectively, and the three-dimensional reconstruction indexes of the two methods are calculated using formulas (3)–(8).

The calculation results are shown in Table 3.

Table 3. Experimental results

Rebuild times/time	Reconstruction index	Traditional method reconstruction index
10	0.954	0.456
20	0.964	0.534
30	0.978	0.521
40	0.995	0.498
50	0.934	0.561
60	0.934	0.663
70	0.948	0.574
80	0.957	0.642

By analyzing the results shown in Table 3, it can be seen that with the increase of the number of image reconstruction, the reconstruction index of both the proposed method and the traditional method presents a trend of constant change. The reconstruction index of the traditional method varies between 0.456–0.663, while the reconstruction index of this method varies between 0.934–0.985. In contrast, the reconstruction index of the 3d image surface reconstruction method designed in this paper is higher, indicating that the reconstruction effect of the method proposed in this paper is good and has certain application value.

4   Conclusion

With the continuous development of scientific research technology and the continuous improvement of people's quality of life, 3D reconstruction technology has become the focus of current research. How to achieve high precision scene reconstruction has become a hot topic in the field of computer vision. There are several key steps in the process of 3D reconstruction. In this paper, the problems of depth image enhancement and point cloud registration are studied deeply, and relevant improvement methods are proposed.

In this paper, based on the idea of global optimization and guided by edge information, iterative median filter is used to estimate the depth pixel value. Finally, adaptive median filter is used to reduce the noise, so that the obtained restoration results have more complete global structured information and clearer edges.

The experimental results show that the proposed method achieves good reconstruction effect and is more suitable for practical application.

References

1. Paoli, A., Neri, P., Razionale, A.V., et al.: Sensor architectures and technologies for upper limb 3D surface reconstruction: a review. Sensors **20**(22), 6584 (2020)

2. Raudonienea, J., Skaudziusa, R., Zarkova, A., et al.: Wet-chemistry synthesis of shape-controlled Ag3PO4 crystals and their 3D surface reconstruction from SEM imagery - ScienceDirect. Powder Technol. **345**, 26–34 (2019)

3. Wang, S., Wu, T., Wang, K., et al.: 3D Particle surface reconstruction from multi-view 2D images with structure from motion and shape from shading. IEEE Trans. Indust. Electron. **99**, 1–1 (2020)

4. Alvarez, J., Saudino, G., Musteata, V., et al.: 3D analysis of ordered porous polymeric particles using complementary electron microscopy methods. Sci. Rep. **9**(1), 1–10 (2019)

5. Meng, Y., Ma, S., Zhang, Z., et al.: 3D nanoscale chemical imaging of core-shell microspheres via microlensed fiber laser desorption postionization mass spectrometry. Anal. Chem. **92**(14), 9916–9921 (2020)

6. Perraud, J.B., Guillet, J.P., Redon, O., et al.: Shape-from-focus for real-time terahertz 3D imaging. Opt. Lett. **44**(3), 483–486 (2019)

7. Hepp, B., Niessner, M., Hilliges, O.: Plan3D: viewpoint and trajectory optimization for aerial multi-view stereo reconstruction. ACM Trans. Graph. **38**(1), 4.1–4.17 (2019)

8. Gao, P., Li, J., Liu, S.: an introduction to key technology in artificial intelligence and big data driven e-learning and e-education. Mobile Networks Appl. **26**, 2123–126 (2021)

9. Shuai, L., Shuai, W., Xinyu, L., et al.: Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**(1), 90–102 (2021)

10. Lee, S.A., Lee, B.G.: Accurate 3D surface reconstruction for smart farming application with an inexpensive shape from focus system. J. Sens. **2020**, 1–7 (2020)

11. Shuai, L., Dongye, L., Khan, M., Ding, W.: Effective template update mechanism in visual tracking with background clutter. Neurocomputing **458**, 615–625 (2021)

12. Lacher, R.M., Vasconcelos, F., Williams, N.R., et al.: Nonrigid reconstruction of 3D breast surfaces with a low-cost RGBD camera for surgical planning and aesthetic evaluation. Med. Image Anal. **53**, 11–25 (2019)

13. Méndez-Manjón, I., Luiz, H., Raquel, G.M., et al.: Semi-automated three-dimensional condylar reconstruction. J. Craniofacial Surg. **30**(8), 2555–2559 (2019)

14. Wu, H., Zhang, Z.: Three-dimensional image reconstruction method based on absolute conic image. Comput. Simul. **38**(8), 203–206+211 (2021)

Personalized Dialogue Generation Method of Chat Robot Based on Topic Perception

Junmei Li[(✉)]

School of Computer Engineering, Jingchu University of Technology, Jingmen 448000, China
`chenweiliang7895@163.com`

Abstract. Human-Computer interaction system is a significant research direction in the field of human-computer interaction, and the research of open domain chat robot has received extensive attention. There are many problems in the existing chat robot: lack of personalized features, resulting in the process of the same chat, and the conversation has nothing to do with the topic. Therefore, a method of creating personalized conversation based on topic perception is proposed, and a personalized conversation model based on topic perception is designed. Semantic analysis and text similarity calculation are needed to build a conversation model. Based on the dialogue model, the training robot collects the corpus data related to the subject, convolves the corpus data related to the subject, and carries out the topic perception training. Finally, a personalized dialogue mechanism is established to generate personalized dialogue. Through experimental comparison, it is proved that the dialogue generated by this method is more suitable for the chat topic.

Keywords: Chatbot · Personalization · Dialogue generation

1 Introduction

Man-machine dialogue system is a significant research direction in the field of human-machine interaction of chat robot, and various dialogue systems are developing vigorously. Text generation, also known as natural language generation, is a key technology to realize dialogue system. Various types of information, such as text and image, can be used to automatically generate smooth and clear natural language text. BENGIO et al. call neural network language model applied to the task of text generation [1], using neural network language modeling. In order to solve the long term dependency problem in natural language, MIKOLOV uses RNN to build language model, and puts forward RNNLM, which improves the accuracy of language model. Since then, RNN and its variants such as the long short termmemory (LSTM) have become the most commonly used method in natural language processing. However, the recently proposed Transformer model has successfully solved some problems in the RNN model, which has triggered a wave of research. Reference [2] method applies the LSTM algorithm to chatbots. The method extracts the fictional dialogue content in the chatbot film and television database. Taking into account the target program model factors, the fusion of LSTM and BiLSTM

models is used to provide accurate dialogue texts. This method improves the accuracy of the session. The reference [3] approach integrates the contextual bandit algorithm into MathBot to personalize the pacing of the conversation, allowing bandits to insert additional practice questions or skip explanations. Provides valuable experience for teaching course dialogue.

Although the conversation rationality of chatting robot is controlled, it depends on the natural sequence structure of RNN. Although the natural sequence structure of RNN is suitable for the task of natural language processing, its strict linear structure will lead to the problem of gradient disappearance or explosion during the training process, and it is difficult to carry out parallel training, which is a serious problem in large-scale application scenarios. To address these issues, Google introduced a new sequence modeling model, the Transformer Model, in 2017. As soon as this model was put forward, it aroused great repercussions in the field of NLP and abandoned the sequence structure in RNN. The whole model is made up of Attention module, which effectively solves the problems of long distance dependence and poor parallel computing abili. Transformer model can capture the semantic information of text sequence efficiently, and its semantic feature extraction ability, long distance feature capture ability and task comprehensive feature extraction ability are much better than RNN model. Recent popular large-scale pre-training models such as GPT model, BERT model of the basic structure of Transformer, in a variety of natural language processing tasks to create excellent results, its superior ability is obvious.

One of the longest research goals in the field of artificial intelligence is the social chat robot, which is a human-computer dialogue system capable of empathy with human beings. If the chat robot wants to establish emotional contact with the user, it must have several abilities, first of all, the ability to integrate context and context, in the process of chat.

At the same time chat robot must have a consistent personality, such as age, gender, etc., if these features change, it is easy to make users feel stripped. Finally, conversation content must be diverse, not always produce "I don't know", "yes, that's right" such generic replies, otherwise users are very easy to produce boredom. The design of personalized dialogue generation method for chat robot is particularly important [4]. In order to solve the problem of single robot dialogue, this paper constructs a topic-aware-based personalized dialogue content generation model. In the process of dialogue, contextual information and personalized feature information are considered, which can better perceive the dialogue topic and improve the accuracy of human-computer conversation responses. And adopt a variety of optimization methods to increase the diversity of the generated response content, so that the generated context is coherent and consistent with high-quality dialogue content.

2   Personalized Dialogue Model Based on Topic Awareness

2.1   Semantic Analysis

The chatting system of chatting robot is mainly composed of speech synthesis module and speech recognition module, which transforms text into speech and speech into text respectively. The natural language understanding module uses NLP technology to

express user's intention in the form of data, which is converted into specific semantic data and then handed over to the next dialog management module. The function of the dialogue management module is to coordinate the work of several modules and to maintain the current system state. The natural language generation module is the most important one [5], which is the focus of this paper. The natural language generation framework is shown in Fig. 1:
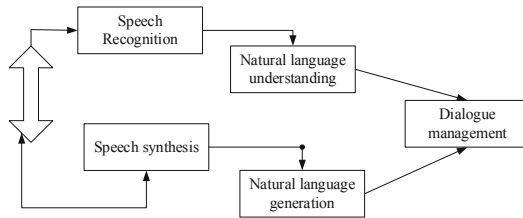

Fig. 1. Natural language generation framework

After generating the natural language, the chat robot should analyze and judge the aim of the syntactic structure analysis, judge whether a text accords with the grammar of the corresponding language, and then analyze the syntactic structure in line with the grammar norms. There are mainly three tasks, one is to perceive the topic, to judge whether the text belongs to a certain language category, the other is to disambiguate the word meaning, and the third is to analyze the sentence structure, context and syntactic relationship [6].

How to get a powerful parser, usually need to solve the following two problems, one is the formal expression of syntax, the other is the description of the entry information. Based on the semantic analysis, we can use the coding model and the Transformer model to construct a non-target-driven dialog system, namely chat robot. And realized with the user to carry on in the open domain the dialog. The Transformer model is essentially a codec architecture, and the overall structure of the model is shown in Fig. 2:
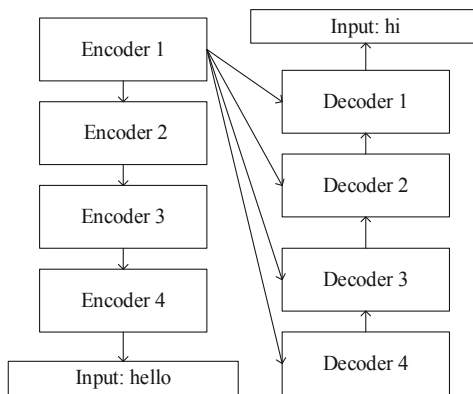
Fig. 2. Overall structure of model

The encoder and decoder model is composed of 4 encoders and 4 decoders. The structure of each encoder and decoder is shown in Fig. 2, respectively. During the encoding process, the data is first weighted by the eigenvector obtained from the Attention module, and then outputted by the feedforward neural network. The decoder has more code-decoding attention module than the encoder, which is used to get the relation between the current time output of the decoding stage and the input of the encoding stage.

2.2 Text Similarity Calculation

At present, most of the corpus-based short text similarity studies use the statistical description method based on context, because the context can provide sufficient semantic information for the definition of words. Lexical Vector Space Model (VSM) is a widely used statistics-based lexical similarity calculation strategy with relatively low algorithm complexity and easy implementation. The Lexical Vector Space Model (VSM) pre-selects a set of feature words [7] and then calculates the relevance of the set of feature words to each word (usually measured by the frequency of the words appearing in the context of the actual large corpus. So each word gets a word vector with the same dimension, and then the similarity between the words is calculated using a formula like this:

$$NDG(x, y) = \frac{\max(\log f(x), \log f(y)) - \log f(x, y)}{\log N - \min(\log f(x), \log f(y))} \qquad (1)$$

Among them, *NDG* stands for standard Google distance, and the larger *NDG* stands for higher similarity, ranging from 0 to 1. $f(x)$ and $f(y)$ represent the number of pages containing the words x and y, $f(x, y)$ represents the number of pages containing the words x and y, and N represents the total number of pages referenced.

Distance is used to calculate the similarity of text sequences. It works as follows: there are currently two short text sequences A and B, in which B is the reference sequence [8]. There are three main steps: the short text sequence A deletes a word. Short text sequence A adds a word; short text sequence A replaces a word.

This repeats until Short Text Sequence A is converted to Short Text Sequence B, then the middle number of operations is recorded as $ED(A, B)$, and 2-$ED[i][j]$ array is the minimum operand, indicating that the first $[i]$ words of Short Text Sequence A are converted to the first $[j]$ characters of Short Text Sequence B. The recursion formula for $ED[i][j]$ is:

$$ED[i][j] = \begin{cases} ED[i-1][j-1] \\ \min\big(ED[i-1][j-1], ED[i][j-1], ED[i-1][j]\big) + NDG(x, y) \end{cases} \qquad (2)$$

3 Training Robot to Discourse Based on Dialogue Model

3.1 Collecting Corpus Data Related to Subjects

Personalized training of robot conversations requires a large amount of corpus data, so Scrapy is used to fetch the data first. Scrapy is a Web content crawling system developed

in Python that is fast, crawls content with less noise, and can be used to crawl a Web site and get data from those pages. When crawling data, due to crawlers and the site itself [9], a considerable amount of garbage data is collected, especially when crawlers crawl through URLs, some of which have no valuable content. Therefore, this article will remove garbage data in the following ways:

(1) Some URLs have no valuable content, based on filtering of worthless URLs, so regular expressions are used to match the URLs, thereby deleting the data.
(2) According to some sensitive words processing corpus data, some sensitive words are fetched in the data, and regular expressions are used to match such statements, thereby filtering the garbage data.
(3) Filter according to the length of the text sequence, some spam data length is very small, text less than 6 bytes will be filtered out.

As for word segmentation, we've already covered it, but we won't go further here. This paper uses the jieba word segmentation tool based on Python to process the corpus word segmentation. There are three patterns of Chinese word segmentation, among which the accurate pattern is the most accurate, mostly used in emotional analysis, syntactic analysis, etc. The full pattern is the fastest, but the accuracy is not high enough.

3.2 Convolution Processing of Corpus Data Related to Topic

On the basis of the corpus collected above, short text information related to the topic needs to be extracted. And construct short text topic graphs through word co-occurrences and document word relationships. In order to improve the accuracy of topic information extraction, it is necessary to use convolutional neural network to convolve the data, and make full use of document node and word node representation to improve text classification results. Convolutional Neural Network (CNN) is currently an advanced technology for subject-related corpus data processing. There is not only one network layer. There will be many different network layers appearing in turn, and the order is not fixed. There are Pooling Layer, Fully Connected Layer, Convolutional Layer, ReLU Layer, etc. The purpose of the convolution operation of the convolutional layer is to classify the picture by discovering the characteristics of a certain part of the picture. The maximum pooling operation is shown in Fig. 4 below, and the average pooling is shown in Fig. 3 below.
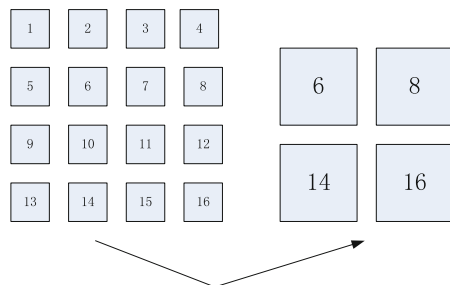

Fig. 3. Average pooling

The last layer in a convolutional neural network is generally a fully connected layer, and the fully connected layer can accept input from the remaining layers, which can be a pooling layer, a convolutional layer, and so on. The input of the fully connected layer is a multi-dimensional vector, and the dimension represents the categories included in the system. This output means that the system contains 8 categories, which are identified as the first. The probability of class I is 0.1, the probability of identifying the second class is 0.3, the probability of identifying the sixth class is 0.1, the probability of identifying the seventh class is 0.4, and the probability of identifying the eighth class is 0.1. On the basis of this convolutional processing, perception training is performed on topic data to improve the accuracy of the robot's response to questions.

3.3   Topic Awareness Training Based on Processed Data

The subject perception training can use the cyclic neural network training method. On the basis of the ordinary multi-layer BP neural network, the horizontal connection between the units of the hidden layer is increased. Through a weight matrix, the value of the neural unit in the previous time series can be transferred to the current neural unit [10]. As a result, the neural network has a memory function, and has good applicability for processing contextual NLP or time series machine learning problems. In the hierarchical expansion of Hidden Layer, $t-1$, t, and $t+1$ represent time series. X represents the input sample. S_t represents the memory of the sample at time t:

$$S_t = f(W * S_{t-1} + U * X_t) \tag{3}$$

where W represents the weight of the input, U represents the weight of the input sample at the moment, and V represents the weight of the output sample. At $t = 1$, the input $S_o = 0$ is generally initialized, and W,U,V are initialized randomly, where W,U,V are equal at each time (weight sharing). Carry out the following formula calculation:

$$h = U * X_1 + W * S_o \tag{4}$$

$$S = f(h) = f(U * X_1 + W * S_o) \tag{5}$$

$$O = g(V * S) \tag{6}$$

Among them, f and g are both activation functions. The state S at this time is used as the memory state of the previous time to participate in the calculation of the next time, and so on, as shown in the following formula.

$$h_t = U * X_t + W * S_{t-1} \tag{7}$$

$$S_t = f(h_t) = f(U * X_t + W * S_{t-1}) \tag{8}$$

$$O_t = g(V * S_t) \tag{9}$$

However, the training process will force the model to make non-zero or one predictions to distinguish between real data and generated content, reducing the generalization performance of the model. Topic perception solves this problem by acting like a regular term to reduce the model's confidence in its prediction results. Use a prior distribution that is not related to the current input parameters to smoothly predict the distribution function of the target, usually using a uniform prior distribution of all words. Label smoothing is equivalent to adding a divergence term on the basis of the negative log-likelihood function, that is, calculating the distance between the prior distribution and the predicted output probability of the model, that is, by preventing the model from over-concentrating the predicted value on the higher probability In terms of categories, it reduces the probability of general replies and increases the diversity of generated replies.

The cluster search algorithm is an algorithm commonly used in the decoding stage of the Seq2seq model. Its parameters are that the word with the highest probability is selected as the output at each moment in the decoding process. By selecting the word with the highest probability at each moment, the algorithm finally generates the sequence of sentences with the highest probability. And maximize it to generate more reasonable results and improve the quality of generated sentences. Since the probability of a sentence sequence is obtained by multiplying the probabilities of multiple words, the longer the generated sentence sequence, the smaller the probability value obtained by the multiplication, so the cluster search algorithm tends to generate shorter sentences. Google proposed a length penalty method to solve this problem. By reducing the probability value of short sequences and increasing the probability value of long sequences, the model has more opportunities to generate a longer sequence P, namely:

$$P = \frac{f(U * X_1 + W * S_o)}{(5 + g(V * S))} \tag{10}$$

Another problem in the cluster search algorithm is that the generated sentences have little difference and low diversity. By grouping the generated results, a similarity penalty is added between the groups to reduce the similarity of multiple results, forcing the model to generate more diversified content, and reducing the appearance of general responses.

4 Establishing Personalized Dialogue Mechanism of Robot

The robot's personalized dialogue mechanism is to imitate the process of humans observing a certain thing. When humans observe a thing, they must only pay attention to a part of the thing, and their attention moves with the movement of the focus. In other words, when human beings observe a thing, the attention given to various parts of the thing is inconsistent. It must be that one part gets more and the rest gets less. Therefore, the robot's personalized dialogue mechanism is very suitable for natural language processing.

The Encoder and Decoder in the traditional model exchange data through an intermediate semantic vector, and the length of this semantic vector is fixed, which will bring the problem of long-distance dependence to the model, that is, for long sequences of text. As the input progresses, the information in the latter part of the sequence will overwrite the information in the former part of the sequence. Therefore, Attention, by

retaining part of the output results of the Encoder on the input sequence, the training model selectively learns these output results, and will eventually be associated with the corresponding output sequence. In other words, the output and input will be selectively associated together.
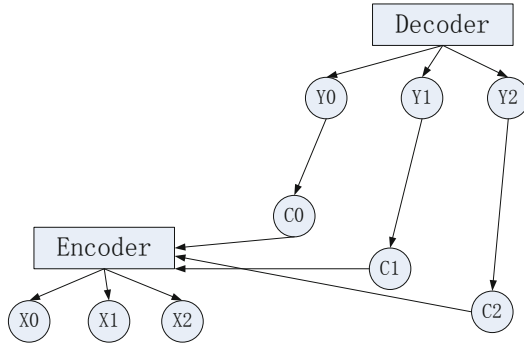

Fig. 4. Mechanism abstract

In the traditional robot personalized dialogue mechanism, the output of the decoder is directly based on the output with the highest probability, but sometimes the output with the highest probability is often the most common words in the corpus, usually "I don't know", "Hello", "Haha" and other meaningless replies, we call them safe replies. Therefore, how to avoid this kind of safe response is also a key issue of the seq2seq model. In order to solve this problem, the Beam Search algorithm can be used. Beam Search (cluster search) reduces the search space and improves time efficiency through "pruning" and multi-layer search. The so-called "pruning" is to filter out some nodes with a small probability every time you search downwards, and will not use this node as the parent node to continue the downward search. This reduces the number of searches and improves space utilization and time efficiency. Suppose that machine translation is used as an example to illustrate. The task is to translate "I am Chinese" in Chinese and "I am Chinese" in English. Assuming that the vocabulary is only three words in size, it is "I", "am", and "Chinese". So if the beam size is 2. In the decoder process, after having the beam search method, in the first output, we select the two words "I" and "am" with the highest probability, instead of selecting only the word with the highest probability.

On the basis of the above selection mechanism and the self-attention mechanism, the multi-head attention mechanism is proposed in the model. Each self-attention module is called a attention head. In the calculation of multi-attention mechanism, the input sequence is first computed through h different attention heads, then the h different feature matrices are assembled into a feature matrix by column, and then compressed into a matrix with the same dimension as a single attention head through a fully connected layer to obtain the output results of multi-attention modules. The general conversation model can carry out several rounds of conversations without considering personalized information. Chatbots without personalized features may have semantic inconsistencies in their conversations, such as being a student and working at work. So that its users have a sense of stripping, easy to make users aware of their own with a fake robot chat, it

is difficult to establish long-term emotional contact. In order to solve this problem, this paper designs a personalized dialogue model, which can generate appropriate replies according to its own characteristics and improve users' interactive chat experience.

This model is based on the codec framework in Fig. 2 above. The encoder and decoder are composed of encoder and decoder of Transformer model respectively, and the number of layers is the same as the general model. In contrast, personalized dialogue model needs to consider personalized information, and add personalized information in the input part of the model to encode. In this paper, personalized features refer to a set of sentence information that describes the character, and the personalized feature vector after coding together with the current input and the historical dialogue content guides the reply generation process in decoding stage. In the decoding process, the code-decoding attention module can determine the impact of user input, historical conversations and personalized information on the output of the current moment. By integrating contextual information and personalized information, personalized dialogue model can generate replies that are consistent with context and accord with specific personalized features.

5   Test Experiment

5.1   Set Up an Experimental Environment

This article uses Google's open source deep learning framework, TensorFlow. Tensor-Flow is an open source software library for high performance numerical computing. TensorFlow has a flexible architecture that allows users to easily deploy computing tasks to multiple platforms, devices, and even mobile devices. Tensorflow is a processing framework based on data flow graph. The nodes in the graph represent numerical computation, and the edges represent the data interaction between computing nodes. The hardware and software environment for the test is shown in Table 1.

Table 1. Experimental environment

Serial number	Software/hardware	Parameter
1	Operating system	Linux
2	Framework platform	TensorF1ow
3	Development language	python
	Hardware	Parameter
1	Graphics card	GTX-1060
2	CPU	Xeon-E5
3	RAM	8G

From the perspective of data transmission and processing, the data flow diagram graphically expresses the logical function of the system, the logical flow of data within the system and the logical transformation process. It is the main expression tool for structured

system analysis methods and a graphical method for representing software models. Data flow diagram can be more intuitive representations of TensorFlow. Data flow graphs represent mathematical computations through directed graphs of "nodes" and "lines". The "nodes" in the diagram are used to represent the mathematical calculations being performed. They can also represent the start or end of the data input or output, or the end of reading in and writing out persistent variables. Lines represent data interactions between Nodes. A "line" can transmit a multidimensional array of data, known as a "tensor," that is "resizable." The tensor flows through the graph, which is why this tool is called "Tensorflow." As long as all the tensors at the input are ready, the "nodes" are allocated to each computing device to perform asynchronous parallel operations.

5.2 Experimental Data

In order to realize the model of chat robot, we need a lot of Chinese conversation corpus. So the most appropriate one is the dialogue in movies and TV plays. In contrast, there are more dialogues in foreign films and TV series, so this paper selects the open subtitles corpus, because this article is to verify the Chinese chat robot, but the open subtitles is English dialogue, so the use of translation tools to translate into Chinese. The corpus size is shown in Table 2.

Table 2. Corpus size

Serial number	Corpus	Open subtitles corpus
1	Development set	2000
2	Test set	2000
3	Training set	44063050
4	Data set	44067050

Using the model to construct the chat robot model, the model is simply a translation model, translating one language sequence into another language sequence. The whole process is to map one sequence as input to another output sequence 62 by using long and short memory network or recursive neural network.

5.3 Experimental Result

In order to judge the practicality of the method, this paper compares the method based on LSTM with the method designed in this paper. It should be noted that the hidden layer of the model has 512 nodes, the word vector dimension is 64, and both Encoder and Decoder use the LSTM model, the size of the batch is 128. Table 3 shows some experimental results.

Table 3. Experimental result

Number of experiments	Question	Answer (LSTM)	Answer (context bandit algorithm)	Answers (in this article)
1	What do you like to eat?	Thank you!	I do not know	There are many favorite foods
2	What do you like to eat?	Sorry!	Thank you!	I don't know
3	What do you like to eat?	Thank you!	There is no favorite food	There is no favorite food
4	What do you like to eat?	Thank you!	There is no favorite food	What kind of food do you like?
5	What do you like to eat?	I have no idea	I do not know	No favorite food
6	What do you like to eat?	I do not understand you	Glad you like it, too	The food is very nutritious

As can be seen from Table 3, compared with the traditional method based on LSTM and context bandit algorithm, the dialogue answer of the robot based on topic perception is more personalized and diversified, but the traditional method has a single answer. And there are many repeated sentences in 6 experiments. And based on the "What do you like to eat?" Obviously, the method designed in this paper to generate personalized conversations between chat robots is more in line with the topic of chat - favorite food.

In order to compare the number of dialogue rounds of the dialogue generation model, we use the test set to simulate the dialogue process, and count the number of dialogue rounds before the answer with no clear meaning like "I don't know". The results are shown in Table 4:

Table 4. Comparison of different model dialogue turns

Method name	Number of dialogue rounds
LSTM	3.7
Context bandit algorithm	4.2
The method of this paper	4.6

It can be seen from Table 4 that this method has more dialogue rounds than the other two methods, probably because the context management of this method records the historical dialogue information, which can be answered according to the context topic. This method improves the diversity of dialogues, reduces the probability of meaningless answers, and increases the number of dialogue rounds to some extent.

6 Conclusion

The chatbot dialogue model proposed in this paper improves the problems of traditional models to a certain extent. This paper innovatively applies a topic-aware approach to generate personalized bot chat conversations. On the basis of semantic analysis, a human-machine dialogue model is constructed, and then the text similarity of chat machine short texts is calculated. Based on the dialogue model, the robot is trained on the dialogue data, the corpus data related to the subject is collected, and the convolution training is performed on the corpus data related to the topic. Finally, the robot personalized dialogue is generated through the robot personalized dialogue mechanism.

But the current dialogue model of chat robot still faces many problems. At present, the conversation model based on topic perception needs a large number of standard Chinese pairs. The more the number of question and answer pairs, the less the noise in the data, the better the model will be in theory. However, there are few open source corpus for Chinese dialogues, so how to collect large scale standardized Chinese corpus is an urgent problem to be solved. From the point of view of chat robot on the market at present, the development of chat robot is still in its infancy, and breakthrough is needed in technology and Chinese corpus. So I hope that in the near future, to find a better technology to achieve a breakthrough and development of chat robot.

References

1. 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
2. Anki, P., Bustamam, A.: Measuring the accuracy of LSTM and BiLSTM models in the application of artificial intelligence by applying chatbot programme. Indones. J. Electr. Eng. Computer Sci. **23**(1), 197–205 (2021)
3. Cai, W., et al.: Bandit algorithms to personalize educational chatbots. Mach. Learn. **110**(9), 2389–2418 (2021). https://doi.org/10.1007/s10994-021-05983-y
4. Zhao, L.N., Li, W., Kang, B., Zhang, K.: Python-based intelligent robot multi-channel knowledge base push simulation. Comput. Simul. **37**(3), 328–332 (2022)
5. Chen, W., Chen, X., Sun, X.: Emotional dialog generation via multiple classifiers based on a generative adversarial network. Virtual Real. Intell. Hardw. **3**(1), 18–32 (2021)
6. Yang, M., Huang, W., Tu, W., Qu, Q., Lei, K.: Multitask learning and reinforcement learning for personalized dialog generation: an empirical study. IEEE Trans. Neural Netw. Learn. Syst. **32**(1), 49–62 (2020)
7. Lee, C.H.: Dual policy network for speaker-specific dialog generation in deep reinforcement learning. J. Inst. Electron. Inf. Eng. **56**(4), 44–49 (2019)
8. Stein, B.: Zum titelbild: intergenerationeller dialog gezeichnet von leonard von der stein. Psychotherapie im Alter **16**(2), 217–219 (2019)
9. Liu, S., Wang, S., Liu, X., Lin, C.T., Lv, Z.: Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**(1), 90–102 (2020)
10. Liu, S., He, T., Dai, J.: A survey of CRF algorithm based knowledge extraction of elementary mathematics in Chinese. Mob. Netw. Appl. **26**, 1891–1903 (2021). https://doi.org/10.1007/s11036-021-01777-7

Multi-round Dialogue Intention Recognition Method for a Chatbot Baed on Deep Learning

Junmei Li[✉]

School of Computer Engineering, Jingchu University of Technology, Jingmen 448000, China
chenweiliang7895@163.com

Abstract. With the continuous development of human-computer dialogue system, more and more dialogue robot products come into people's lives. However, when human beings use short sentences and omit words, and in the process of identification often face problems such as more text noise, sparse characteristics, polysemy, backward and backward dialogue information. In order to solve the above problem, a deep learning based chatbot multi-round dialogue intention recognition method, according to the fit of deep learning algorithm and chatbot multi-round dialogue intention recognition model, by transforming the problem into a mathematical model, and obtain the final dialogue intention through the calculation of the model. First, the chat dialogue text was preprocessed, and the BERT model was established based on the processing results, the BERT model fused the deep learning model in the B E R T model, established a joint model, and data vectorized the short text of the human-computer dialogue. Finally, the multi-round dialogue intention identification similarity is calculated through the robot, realizing the dialogue intention recognition, and experiments show that the highest accuracy of the recognition method can reach 0.9912, the highest recall rate can reach 0.9914, and the highest f price is 0.9914, which can prove the superiority of the design method.

Keywords: Deep learning · Chatbots · Multiple rounds of dialogue · Dialogue intent · Intent recognition

1 Introduction

With the advent of the AI era, more and more intelligent products have been widely used in everyday life, such as the emotional escort robot personal mobile assistant Siri, voice assistant Google Now, and Cortana. XiaoBing, an intelligent chatbot launched by Microsoft Asia Research Institute, and a Xiaodu robot launched by Baidu. These intelligent dialogue systems cannot only communicate with normal information with users, but also bring a lot of convenience to users' lives [1]. The dialogue system consists mainly of five parts: Automatic Speech Rec-ognition (ASR), Spoken Language Understanding (SLU), dialogue management (Dialog Management, DM), dialogue generation (Dialogue Generation, DG), and Textto Speech (TTS), as shown in Fig. 1. In order to let the machine better understand the expression of users, and then feedback the correct

information, spoken understanding plays an extremely important role. IntentDetection (ID), as a submodule of oral understanding, is also the key to the human-computer dialogue system. Traditional spoken understanding is mainly divided into two subtasks: intent recognition and semantic slot filling. Since the early research is limited by application scenarios, data and computing power, most oral comprehension is limited to some fields. However, with the innovation of technology and the emergence of multi domain dialogue system, today's oral understanding is often divided into three tasks: domain recognition, intention recognition and semantic slot filling [2].

In dialogue systems, intention recognition is crucial. The intention is the user's intention, namely what the user wants to do. Intent is sometimes referred to as "dialogue behavior" (Dialog Act) I', the behavior where the information status or context shared in the conversation changes and is constantly updated. The intention is generally named after the "verb + noun", such as weather inquiry, hotel booking, etc. intent recognition, also known as intent classification, is classified into a previously defined category of intent based on the areas and intent involved in the user's utterance.

With the widespread use of the human-computer dialogue system, users may have different intentions in different occasions, so they will involve multiple fields in the human-computer dialogue system, including the task-type vertical field and small chat [3]. The purpose of the task text is clear and easy to retrieve, such as flight tickets, weather, hotels, etc. The chat intention text generally has the characteristics of unclear theme, semantic width and short statements, paying attention to the communication with humans in the open domain. In the dialogue system only clear the user's topic field, to correctly analyze the specific needs of the user's intention, otherwise will cause later intention error identification when the user input a query, first need to clarify the user input text topic field is "train" "flight", because the intention category more granular than the topic field, so need to determine the user's specific semantic information is refund or query time, and semantic slot filling is also helpful to the user intention judgment. Therefore, in the intention recognition module of the human-computer dialogue system, it is first necessary to identify the user topic field, and then the specific intention needs of the users should be defined, and finally express the form of the semantic framework [4]. However, previous methods often face some problems in the process of recognition, such as text noise, sparse features, polysemy, backward dialogue information and so on. In order to solve the above problems, this paper proposes a multi round dialogue intention recognition method of chat robot based on deep learning. According to the deep learning algorithm and the multi round dialogue intention recognition model of chat robot, the problem is transformed into a mathematical model, and the final dialogue intention is obtained through the calculation of the model, in order to provide some help to improve the accuracy of multi round dialogue intention recognition of chat robot.

2 Design a Multi-round Dialogue Intent Recognition Model Based on Deep Learning

Intent recognition has become a new research hotspot in academia and industry, and to correctly understand user intentions in human-computer dialogue systems, intention identification can be solved as a short text classification problem, where a category is

automatically determined for text in certain specific categories according to pre-defined topic categories. Intent to identify the user intention from the user short text of the dialogue process through the short text classification. The process of intention recognition is expressed as mathematically symbolic, as shown in Fig. 1, which can be viewed as a mapping relationship $f : U \rightarrow I$, where $I = 1, 2....$, Where the set U is the type of intent preset by the natural language statement i entered by the user, and I is the list of intentions resolved from the natural statement i.
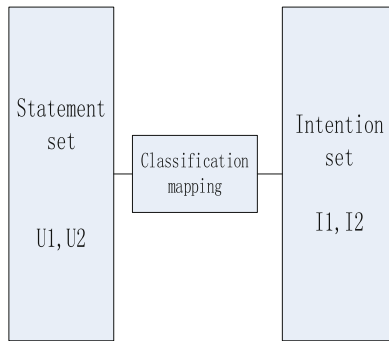
Fig. 1. Intent to identify the mathematical model

The process of intention identification can be roughly described as follows: models with parameters conduct parameter learning and model optimization on the training set with intention markers, and then use the trained optimal model to identify each data on the test set of hidden intention labels, calculate the identification results and compare them with the real label.

2.1 Chat Conversation Text Preprocessing

Text preprocessing is at the beginning of the entire intent recognition process, and many anomaly data or values in text data will directly or indirectly affect the results of downstream tasks, so preprocessing has great significance and necessity for conventional anomaly data or values [5–7], With important implications for the results of intent recognition. Because the short text format is not standard and the number of words is small, if it is not reasonably standardized processing, it will affect the results of intention recognition, so in the process of intention recognition, the preprocessing of short text text is a step that cannot be ignored. Data preprocessing can avoid many non logical exceptions before the algorithm starts, such as denoising data, processing outliers, missing values and so on. Text preprocessing mainly includes noise removal, parti, and removal of stop words.

Usually, the short text data crawled from the Internet platform contains not only the text and characters that identify the semantic, but often has a large number of additional structures with low connection to text labels and content, such as hyperlinks, emojis, description symbols, HTML markers, XML tags, pictures, etc., etc., which have no significance to build intention recognition, they usually have no clear semantic information,

but only express the text information more intuitively. However, in the intention recognition task, these data do not only contribute to the semantic expression of short text data, but will increase the processing time and operation scale of short text data, and will also adversely affect the operation process such as word segmentation and text vectorization, ultimately resulting in reducing the accuracy and reliability of text processing methods. Therefore, these abnormal information and noise data should be cleaned and sorted out before the short text data is formally processed.

In our daily life, English is the most widely used language, involving major countries in various regions, and Chinese is the most widely used language, but the two most representative languages differ greatly in terms of processing. It is relatively simple to divide English, only to divide according to the space characters between words. There is no clear separation between Chinese words, and the words in Chinese are separated by semantic and context. Therefore, compared with English segmentation segmentation, it is difficult to handle Chinese segmentation, and more rules and restrictions should be considered. Word segmentation is the basic step of Chinese text preprocessing and the premise of text representation. Under certain syntactic semantic rules, word segmentation is the process of dividing the original continuous expression into single words or words. In this process, many components with little correlation to the semantic expression of the original text will be ignored, and only the key and core words or words are retained. The effect of participle will directly affect the effect of words, semantic representation, we can choose the appropriate participle tool according to the different use scenarios and requirements.

There are three text segmentation methods in natural language processing: one is based on grammar and rules, one is based on dictionaries, and the other is based on statistics. This paper will adopt the method of stuttering segmentation, which is based on statistical segmentation methods. The rationale of a statistical-based partitioning method is to determine whether a string constitutes a word based on the statistical frequency of its occurrence in the corpus. Words are a combination of words, and the more times adjacent words appear simultaneously, the more likely it is to form a word [8]. Therefore, the frequency or probability of co-occurrence adjacent to words can better reflect their credibility to becoming words.

Stop words refer to be frequent in text, but from the perspective of semantic understanding and expression, it has little influence on tasks such as text representation or intention recognition. Stop words mainly include public stop words and professional stop words. Public stop words usually have commonly used prepositions, crowns, aids, pronouns, conjunctions, etc. Based on the empirical summary of the numerous research work, Stop words are roughly divided into two categories: one refers to some words that are very widely used, Words like "I", "just" that appear in almost every text, However, its association with the intent labels is very low, Not only did there have any positive impact on the identification task, Instead, because of the excessive number of appearances, consumption of time, Also reduces the efficiency of identification; The other category refers to the high frequency of both appearing and being used in the text, But these words are not substantive or decisive in semantic expression, Its role in the text is only to ensure the standardization and integrity of the text in the grammatical structure, This type of words usually contain tone aids, adverbs, prepositions, conjunctions, etc., They have no

practical meaning of themselves, Nor decisive determine the intention and emotional tendency of the text, Only put into the complete sentence can show a certain auxiliary effect, Such as the common "of", "in", "and", "then", etc.

2.2 Establish the BERT Model

This chapter presents a joint model BERT _ word2vec based on BERT and word2vec to quantify the short text of human-computer dialogue, whose model structure diagram is shown in Fig. 3-1. The vectorization representation method first trains the word vector in the word2vec model, and calculates each word vector to the sentence hierarchy vector pre-trained by BERT, then transforms the resulting similarity value into weights assigned to the corresponding word vector, and finally combines the weighted word vectors into the sentence vector and the sentence vector of BERT [9] (Fig. 2):
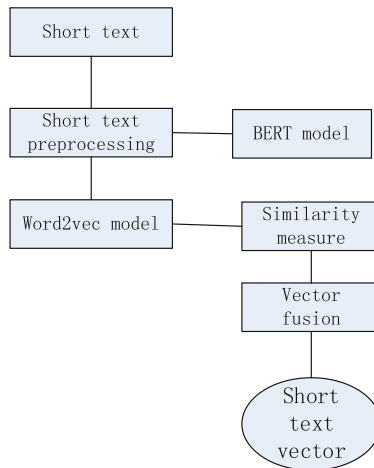
Fig. 2. Structural diagram of the BERT_word2vec model

The BERT (Bidirectional Encoder Representations from Transformers) model, is essentially implemented on the basis of a two-way Transformer encoder, where E, E, …, Ex is the input vector of the model, and bidirectional Transformer coding yields a vectorized representation of the text T. Transformer is a self-attention (Self-attention)-based seq2seq model, a Encoder-Deocder-structured neural network whose input and output are a sequence [10]. In the model, Encoder transforms the input sequence of variable length into a fixed-length vector expression, and then decodes this fixed-length vector into a variable-length target signal sequence by Decoder, and Figs. 3-3 are the structural diagram of the traditional model, where C is the state vector between Encoder and Deocder. In fact, the basic Encoder-Decoder structure is implemented based on RNN, and its core module is composed of RNN units, but with the increase of sequence length, RNN itself has some unavoidable problems, such as unable to parallel, slow operation, etc. At the same time, because the state vector size of the connecting Encoder

and Decoder is fixed, the Decoder is unable to directly follow more details of the input information [11]. To improve on the above deficiencies, Transformer uses self-attention to replace the RNN. Since the Encoder part of the Transformer is mainly used in the BERT model, the Encoder structure in the Transformer model is highlighted below, as shown in Fig. 3:
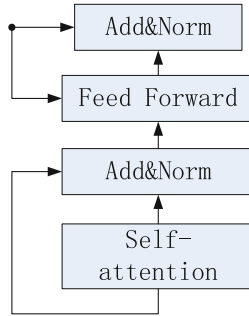
Fig. 3. The Encoder structure

As can be seen from Fig. 3, the input of Encoder consists of a vectorial representation of the short text and the location information of each word in the text, and then through the Self-attention layer gives Encoder the information to view the word before and after the word when encoding each word. Another Add & Norm layer, Add adds the input and output of the self-attention layer, Norm normalizes the added output, so that the output of the self-attention layer has a fixed mean 0 and standard deviation 1, and the normalized vector representation will be processed through a fully connected feedforward neural network (Feed Forward). Similarly, the Feed Forward layer also contains the Add & Norm layer [12]. As mentioned above, a new list of word vectors will be output. The core module in part encoder is Self-attention, whose main idea is to calculate the mutual relationship between each word in the short text and all the words in the short text, and then adjust the weight of each word to obtain a new expression of each word [13]. This new expression not only contains the semantic meaning of the word itself, but also contains the relationship between other words and the word, so it is a more global expression compared with the traditional word vector. The Self-attention procedure is calculated as follows:

Suppose the input short text is expressed as:

$$X = (x_1, x_2, ..., x_n)^T \tag{1}$$

where, x_i is the i word in the short text, and now it is expressed as a_i by one hot vector. There are n words in total, and the vector matrix is obtained:

$$A = (a_1, a_2, ...a_n)^T \tag{2}$$

Then multiply the vector matrix by three different weight matrices W_g, W_k and W_v to obtain query, key and value matrices Q, K and V. The importance s of each word is

calculated as follows:

$$s = Q \bullet K \tag{3}$$

The smoothed result of s is multiplied by the softmax function to obtain the value of attention. Each line represents the attention vector of the corresponding word in the input short text. The vector has been integrated with the information of other position words, which is a new vector representation. The specific calculation formula is shown in 4:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where: d_k is the penalty factor to ensure that the inner product of Q and K is not too large. As can be seen from the calculation formula above, the whole computational process is a series of matrix multiplication that can be parallelized and superior to RNN. In practice, Transformer uses Multi-headSelf-attention, or multiple Self-attention in parallel, to enhance the attention power of the model.

2.3   Robot Multiple Rounds of Dialogue Intent Recognition Similarity Calculation

The similarity metric adopted is cosine similarity, which measures the cosine value of the two vector clip angles in vector space as the size of the difference between two individuals, which pays more attention to the difference between the two vectors in direction than distance or length, so the method is more suitable for the vectized data in this chapter [14]. Assuming that the output vector of the Bert model is expressed as A, the word vector trained by word2vec is B_i, $i = 1, 2, ..., n$, which represents the word vector of the i-th word in the short text, and n is the total number of words in the short text, the calculation formula of the similarity S between the i-th word in the short text and the short text is shown in 5, and the weight w corresponding to the th word can be obtained from the similarity S, as shown in formula 6:

$$S_i = \frac{A \bullet Bi}{\|A\|\|B\|} \tag{5}$$

$$w_i = \frac{S_i}{\sum_{i=1}^{n} S_i} \tag{6}$$

Then multiply each word vector by its corresponding weight, splice it into a vector, and then add it to the short text preprocessed by Bert to obtain the final vectorized representation of the short text. The vector contains both semantic features at the sentence level and highlights words closer to the sentence meaning, which not only compensates for the disadvantage that word2vec cannot reflect the word polysemy, but the presence of the sentence vector generated by BERT also complements the semantic information lost during word vector splicing. Finally, the scaling point product attention results of h times were spliced from left to right, and the attention matrix X obtained from a second linear transformation was used as the result of multi-head attention. The specific calculation formula is as follows:

$$head_i = Attention(Qw_i, Kw_i, Vw_i) \tag{7}$$

$$MultiHead(Q, K, V) = (head_1 + head_2 + head_3 + ... + head_i) \tag{8}$$

In Eqs. (7) and (8): Q, K and V represent query matrix, key matrix and value matrix respectively, with equal values, which are vectorized output E; $\sqrt{d_k}$ refers to the square root of the bond vector dimension, which plays a regulatory role and controls that the inner product of Q and K will not be too large; W is the parameter of linear transformation, and W is different every time Q, K and V are linearly transformed; h represents the number of heads and i represents the attention head.

3 Test Experiments

In order to verify the effectiveness of the deep learning-based multi-round dialogue intent recognition method designed in this paper, compare the method designed in this paper with the traditional dialogue intent identification method (literature [11] method), highlighting the advantages of the method designed in this paper.

3.1 Dataset

The experimental data are selected from the corpus of the booked restaurant, but because the corpus only contains the data of a single round of dialogue, considering the specific requirements of the experiment in this chapter, the dialogue of the data set is required for this chapter. Multiple rounds of dialogue in the dataset are shown in Table 1:

Table 1. Multiple-round dialogue data

Conversation object	Dialog box text	Intention
p1	Is there any restaurant near here?	find_restaurant
b1	Please provide your current location	The dialogue robot identified the key word "nearby"
p2	I am on XX Road, XX No	ask_ location
b2	Recommend XX restaurants to you	The dialogue robot identified the questioner

A list of intent labels involved in the conversation and data examples are shown in Table 2:

The multi-round dialogue dataset annotated in this paper contains 10,432 dialogue data and 100 sets of dialogue, ranging from 1 to 8 rounds, with 11 intention labels appearing. It can be seen from the data distribution that the intention categories of data are unbalanced, especially the number of data of greet, thanks and deny, which has a great impact on the accuracy and is easy to bring large errors to the experiment. Therefore, so

Table 2. Schematic labels and data examples

Order number	Intent name	The intent label	Example data
1	Query the restaurant	find restaurant	Is there any restaurant near here?
2	Number of information	info people	About 5 people had their meals
3	Location information	ask_location	I am on XX Road, XX No
4	Telephone information	phone_number	My cell phone is 99999999999999
5	Price information	info _price	How is your meal price here
6	Confirm information	confirm	Yes, that is what I am looking for
7	Denial information	deny	This restaurant does not meet my requirements
8	send one's respects to	greet	shalom
9	thanks	thanks	Thank you very much
10	good-bye	goodbye	a form of greeting by women
11	other	others	Recommend XX restaurants to you

these three intentions are not included when calculating the comprehensive performance of each model in the subsequent experiments. During the experiment, the datasets were divided into training, test, and validation set in a ratio of 6:2:2. To ensure the contrast of the experiment, the datasets in the experiment used the same division method, and the data used the same short text vector representation method.

3.2 Experimental Result

Due to the data imbalance, the amount of data in the intention categories greet, thanks and d e n y does not affect the evaluation of the overall performance of the model. The identification results of these three intentions are not included in the calculation of the model indicators: Precision (accuracy) and RecallK (recall) and F-measure are used to evaluate the classifier performance, as shown in (9), (10), (11):

$$\Pr ecision = \frac{TP}{TP + FN} \tag{9}$$

$$\mathrm{Re} call = \frac{TP}{TP + FN} \tag{10}$$

$$F - measure = \frac{2 \times P \times R}{P + R} \tag{11}$$

where: if sample *A* belongs to a category *B* and is recognized as such by the model, it is recorded as *TP*; If sample *A* does not belong to category *B* but is identified as category *B*

by the model, it is recorded as *FN*; If sample *A* belongs to a category *B* but is identified by the model as a category other than category *B*, it is recorded as *TN*; If sample *A* does not belong to category *B* and is identified by the model as a category other than category *B*, it is recorded as *TN*.

For the multi-classification problem of this task, the evaluation index of each category is first calculated, and then the macro average value (Macro-average) is used as the final evaluation index.

In the algorithm text extension method, the minimum support is set to 0.01 and the minimum confidence is 0.5; In the LDA subject extension method, the number of topics is set to 10. After obtaining the subject distribution of the chat text, the word with the maximum probability under the subject is added to the text. During the feature selection process, this paper starts from 100 dimensions and calculates the performance effect of the recognition model with 100 as a step length, finally retaining the most relevant 3,000 words as the final feature. Different recognition models were also tested for different parameters.

To reduce the contingency of the experimental results, this chapter uses a 5-fold cross-validation method to obtain the final identification results. Where "original" refers to the word feature + feature selection method; "+ apriori" means the content extension on the original base; "+ lda" means the theme extension on the original base; "+ apriori, 1da" means merging original, apriori and 1da3 methods.

The final results are shown in Table 3:

Table 3. Comparison results of different methods

Numerical classification	Naive Bayesian multi-round dialogue intent recognition methods			Random Forest multi-round dialogue intent identification method			The multi-round dialogue intent identification method designed in this paper		
	Precision	Recall	F price	Precision	Recall	F price	Precision	Recall	F price
Original	0.7445	0.7412	0.7541	0.7412	0.7544	0.7445	0.9135	0.9912	0.9178
+apriori	0.7845	0.7415	0.7153	0.7415		0.7845	0.9912	0.9145	0.9914
+lda	0.7746	0.7745	0.7544	0.7745		0.7746	0.9541	0.9914	0.9145
+apriori, lda	0.7256	0.7523	0.7826	0.7523		0.7256	0.9416	0.9514	0.9914

From the experimental results, it can be seen that in three different intention recognition models, the content and theme are expanded through Apriori and LDA, and the effect is improved. However, compared with the other two methods, it can be seen that the multi round dialogue intention recognition method of the design method in this paper has better effect, with the highest accuracy of 0.9912, the highest recall rate of 0.9914 and the highest f price of 0.9914, which can prove the superiority of the design method. The reason for this result is that this design method first calculates the correlation on the corpus, then finds out the co-occurrence relationship of each word in the chat text, and completes the limited relationship with the text. Because there are some connections between common words, after completing the words, it can enrich the current text content and supplement the synonyms not mentioned in the original text. After the description information is added to the supplementary text, the recognition effect is

improved compared with the original text. Moreover, because deep learning needs to set word support, which is the minimum frequency, a priori mining common associations in the corpus and display content supplement is the premise, that is, the user's expression of high-frequency words also limits the user's need for relative expression norms, and try not to appear random rather than standard low-frequency words. However, in the actual process, users may have more irregular expressions, and the expression patterns are very diverse, resulting in a large number of repetitions of diversified spoken words, which can not be found according to the association mining algorithm. Based on this situation, this chapter uses LDA to mine the topic information of the text, which can bypass the word level supplement. The overall semantics of the text is very helpful to express non-standard text.

4   Conclusion

In order to improve the accuracy of short text intention recognition in man-machine dialogue system, this paper proposes a short text vectorization method and two intention recognition methods. Firstly, the text of chat dialogue is preprocessed, and the Bert model is established according to the processing results. The Bert model integrates the deep learning model in the Bert model, establishes a joint model, and quantifies the short text of man-machine dialogue. Finally, the similarity of multi round dialogue intention recognition is calculated by the robot, and the dialogue intention recognition is realized. The experimental results show that the recognition accuracy of the design method can reach 0.9912, the recall rate can reach 0.9914, and the f price is 0.9914, which can prove the superiority of the design method.

References

1. Al-Mayyahi, A., Aldair, D., Chatwin, C.R.: Control of a 3-RRR planar parallel robot using fractional order PID controller. Int. J. Autom. Comput. **17**(6), 822–836 (2020)
2. Rutschi, C., Dibbern, J.: Towards a framework of implementing software robots: transforming human-executed routines into machines. Data Base Adv. Inf. Syst. **51**(1), 104–128 (2020)
3. Travagnin, S.: From online Buddha halls to robot-monks: new developments in the long-term interaction between Buddhism, media, and technology in contemporary China. Rev. Relig. Chin. Soc. **7**(1), 120–148 (2020)
4. Perugia, G., Paetzel-Prüsmann, M., Alanenp, M., Castellano, G.: I can see it in your eyes: gaze as an implicit cue of uncanniness and task performance in repeated interactions with robots. Front. Robot. AI **8**, 1–18 (2021)
5. Zhang, L., Yang, Y., Zhou, J., Chen, C.C., He, L.: Retrieval-polished response generation for chatbot. IEEE Access **8**, 123882–123890 (2020)
6. Ren, F., Xue, S.: Intention detection based on Siamese neural network with triplet loss. IEEE Access **8**, 82242–82254 (2020)
7. Saha, T., Gupta, D., Saha, S., Bhattacharyya, P.: Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. Cogn. Comput. **13**(3), 277–289 (2020)
8. Gupta, D., Bansal, P., Kavita: Emotion recognition: differences between spontaneous dialogue and active dialogue. J. Shanghai Jiaotong Univ. (Sci.) **16**(9), 633–644 (2021)

9. Li, J., Guo, H., Chen, S., Yang, D., Zhao, L.: A novel semantic inference model with a hierarchical act labels embedded for dialogue act recognition. IEEE Access **7**, 167401–167408 (2019)
10. Yang, W., Wan, B., Qu, X.: A forward collision warning system using driving intention recognition of the front vehicle and V2V communication. IEEE Access **8**, 11268–11278 (2020)
11. Chen, Y., Li, C.: Simulation of target tactical intention recognition based on knowledge map. Comput. Simul. **36**(8), 5 (2019)
12. Liu, S., Liu, D., Muhammad, K., Ding, W.: Effective template update mechanism in visual tracking with background clutter. Neurocomputing **458**, 615–625 (2021)
13. Liu, S., et al.: Human memory update strategy: a multi-layer template update mechanism for remote visual monitoring. IEEE Trans. Multimedia **23**, 2188–2198 (2021)
14. Liu, S., Wang, S., Liu, X., et al.: Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**(1), 90–102 (2021)

Vibration Failure Analysis of Civil Aircraft Engine Blades Based on Virtual Reality

Mingfei Qu[✉] and Yaping Li

College of Aeronautical Engineering, Beijing Polytechnic, Beijing 100176, China
qmf4528@163.com

Abstract. With the development of new aircraft, the flight speed of aircraft is faster and faster, and its requirements for engine performance are higher and higher. If the vibration accuracy of the blade is reduced, it will lead to its vibration failure. Therefore, this study designs a vibration failure analysis method of civil aircraft engine blades based on virtual reality. Firstly, the vibration state of single blade of civil aircraft engine is analyzed by finite element method, and then combined with the analysis results of virtual reality technology, the vibration failure analysis equation of blade is designed. The experimental results show that the analysis effect of this method is good, and can be used as a reference for subsequent aircraft engine research.

Keywords: Virtual reality · Aircraft engine · Blade vibration · Failure analysis

1 Introduction

Blade structure is widely used in various modern machinery and daily electrical appliances, such as gas turbine, steam turbine, water turbine, various booster pumps, electric fan, air conditioner, refrigerator and so on [1]. In the typical power plant, except the piston engine, most of them adopt the structural form of blade for design. Blades are one of the main components of various aviation engines, including turbojet engine, turbofan engine, turbine and turboplasma engine [2].

The research on Aeroengine Blades belongs to a multidisciplinary comprehensive field, involving vibration mechanics, fluid mechanics, material mechanics, solid mechanics, structural strength, mechanical design, manufacturing technology and so on [3]. These disciplines cooperate with each other, but restrict the research of engine blades, which makes the development of Aeroengine Blades a complex system engineering.

Since the birth and operation of turbojet engine in 1930s, blade failure has been one of the main factors hindering its development. In the late 1960s and early 1970s, turbofan engines began to develop. The turbofan engine looks similar to the double rotor turbojet engine. The main difference is that the blades of the low-pressure compressor of the turbofan engine are lengthened into a fan, and an outer culvert is added behind it [4].

With the development of new aircraft, the requirements for engine driving force are higher and higher, the rotating speed of engine blades is faster and faster, and there are more and more compressor stages. The total boost ratio of the engine is the multiplication of boost ratios at all levels. When the boost ratio at all levels is certain, more compressor stages are required to improve the total boost ratio, and the weight of the engine is large, Therefore, only increasing the compressor stage cannot solve the problem of engine efficiency, but also improve the single-stage boost ratio [5]. The improvement of total boost ratio and average single-stage boost ratio, on the one hand, greatly improves the thrust of the engine, but on the other hand, it also brings a series of new problems, such as smaller surge margin, lower working efficiency, increased engine weight and so on. With the increase of the number of engine blades, the designed structure tends to be lighter and thinner in order to pursue high efficiency. In addition, turbine blades are often designed into complex cooling structures to meet the working needs in the environment of high temperature and high load. These factors cause a high probability of blade failure in the working process.

In fact, in order to meet the design requirements of the engine, the safety factor of blade design is close to "1" at present, and the potential of materials has been applied to the limit. In order to further develop the use value of blades, researchers put forward the method of adjusting the amount of blade cover [6]. The cover value of the blade, that is, the offset of the blade, is an important parameter to balance the aerodynamic bending stress. The blade is subjected to centrifugal force generated by rotating speed, bending moment generated by pneumatic pressure, resonance force generated by various excitation factors, etc. Under the action of aerodynamic pressure and centrifugal force, the blade root produces relatively large bending stress, which often exceeds the yield strength of the material after combined with tensile stress, which is not allowed for the design requirements of infinite life of the blade.

Based on the above analysis, this study designed a method for vibration failure analysis of civil aircraft engine blade based on virtual reality. On the basis of analyzing the vibration state of single blade of civil aircraft engine by using finite element method, this study combined virtual reality technology and finite element software to design the vibration failure analysis equation of blade, providing reference for the subsequent research of aircraft engine.

2 Method Design

2.1 Finite Element Analysis of the Vibration State of a Single Blade of a Civil Aviation Aircraft Engine

The finite element method can be understood as dividing the structure to be solved into a series of elements connected by nodes. The shape of these elements is very simple, such as triangles, rectangles, etc., so for each element, it is easy to establish equations according to the balance relationship or energy relationship, and then combine the equations of each element to obtain the total equation system of the structure [7]. The basic idea of finite element analysis of real structural systems is to simulate complex continuous real structures with a simple finite number of interacting elements.

With the rapid development of computer technology, a large number of finite element calculation software have been developed. Now the application of finite element method can not only analyze plane problems, but also spatial problems, plate and shell problems, not only static analysis, but also stability analysis and dynamic analysis.

The three basic elements of the finite element system are nodes, elements and degrees of freedom [8]. The finer the units, the more accurate the results. The basic task of element analysis is to establish the relation between element joint force and node displacement, namely the element stiffness equation, so as to determine the element stiffness matrix and transform the external load into the element equivalent node load. The rectangular element shown in Fig. 1 is taken as an example to establish the finite element equilibrium equation.

For the blade, which is similar to a rectangular thin plate, it can be easily discretized by rectangular elements. If the four corners of the rectangle are taken as nodes, a simple rectangular plate element is obtained, and its structure is shown in Fig. 1.
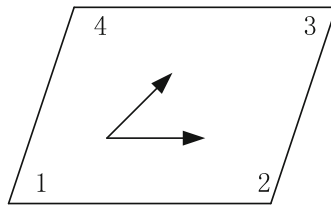
Fig. 1. Schematic diagram of finite element element

It can be seen from Fig. 1 that when the plate is bent, there is moment transmission between the elements, and the nodes are rigidly connected. A point on the element surface actually represents a normal line segment with the length of the plate thickness. According to the hypothesis, the length of the normal line segment is constant, and the points on the midplane of the thin plate do not produce displacements in the x and y directions. Therefore, the possible displacements of the thin plate nodes are only the deflection in the z direction and the rotation angle of the normal around the x and y axes. Taking the vector marked according to the right-hand spiral rule to be positive along the positive direction of the coordinate axis, the two corners θ_x, θ_y at this time are as follows:

$$\theta_x = \frac{\partial w}{\partial y} \tag{1}$$

$$\theta_y = \frac{\partial w}{\partial x} \tag{2}$$

In formulas (1) and (2), w represents deflection, and the element displacement vector at this time is as follows:

$$\{\theta_x, \theta_y\} \tag{3}$$

Substituting the nodal displacement component and nodal coordinates into the above formula, the expression of the undetermined coefficient can be obtained as follows:

$$g = Na \tag{4}$$

In formula (4), N represents the shape function, and a represents the displacement component. In the finite element analysis of the plate, since the deflection and rotation angle have been used as nodal parameters, the known deflection and rotation angle on the boundary can be used as mandatory boundary conditions. Considering the equivalent load caused by the surface force, assuming that there is a lateral distributed load q acting on the surface of the thin plate, the virtual work of the external force generated in the case of the lateral shear force V and the upper bending moment M is as follows:

$$V = \int g \frac{N}{VM} \tag{5}$$

Set the element balance equations according to degrees of freedom to obtain the overall balance equation of blade finite element analysis as follows:

$$Ka = P \tag{6}$$

In formula (6), K is the overall stiffness matrix of the structure according to the node group, a is the overall node displacement of the structure, and P is the overall node load of the structure.

Any deformed body has natural frequency and vibration mode. When there is external excitation, it will produce a series of responses. In addition to structural static analysis, structural vibration analysis is also an important aspect of structural evaluation, which is of great significance to the working state and functional control of the structure [9]. The structure in the vibration problem is also a deformed body, which also needs three kinds of mechanical variables to describe. The time-dependent inertial force and damping force can be considered in a static way by using the darumbel principle.

In the solution domain Q, if the field function u is an exact solution, then any point in the domain Q satisfies the control differential equation, and at the same time any point on the boundary I satisfies the boundary conditions, then the equivalent integral form must be strictly satisfied. But for complex practical problems, such exact solutions are often difficult to find, so people need to try to find approximate solutions with a certain degree of accuracy.

Assuming that in an engineering problem, the general forms of the governing equations and boundary conditions of the system are shown in Eqs. (7) and (8) respectively:

$$A(u) - f = 0 \tag{7}$$

$$B(u) - g = 0 \tag{8}$$

In the formula, u represents the function to be solved, A and B represent the differential operator of the boundary, and g represents the known function. By selecting the parameters to be determined, the approximate residual value is considered to be zero. The system

of equations obtained in this way can be solved to obtain the undetermined parameters, and then the approximate solution of the problem can be obtained by solving. Properly choose the trial function w so that it satisfies all the conditions on the boundary, so that the boundary residual value R is zero, and the approximate solution has a high precision. The more approximate functions should be selected, as the number tends to At infinity, the approximate solution is infinitely close to the exact solution.

The method of using the weighted integral of the residual value to zero to obtain the approximate solution of the differential equation is called the weighted residual value method, which was first proposed by Crandall. Weighted residual method is an effective method to find approximate solutions of differential equations. Obviously, any independent complete function set can be used as a weight function [10].

The residual equation is orthogonal to each basis function of the trial function, which ensures the convergence of Galerkin method. In many cases, the coefficient matrix obtained by Galerkin method is symmetrical. Therefore, Galerkin finite element method is used almost without exception when using weighted residual method to establish finite element scheme. At the same time, it should be pointed out that Galerkin method and variational method are often equivalent when there are corresponding functionals for differential equations and their boundary conditions.

2.2 Design Failure Analysis Equation Based on Virtual Reality

Based on the above finite element analysis of the vibration state of a single blade of a civil aircraft engine, this study then combined virtual reality technology and finite element software to design a blade vibration failure analysis equation.

In theory, virtual reality technology is a computer simulation system that can create and experience the virtual world. It uses the computer to generate a simulation environment and immerse users in the environment. Virtual reality technology is to use the data in real life and the electronic signals generated by computer technology to combine them with various output devices to transform them into phenomena that people can feel. These phenomena can be real objects in reality or substances that can not be seen by our naked eyes, which are expressed through three-dimensional models [11]. Because these phenomena are not what we can see directly, but the real world simulated by computer technology, they are called virtual reality. Therefore, we can use virtual reality technology combined with finite element software to design vibration failure analysis equations.

ANSYS software is a large-scale general-purpose finite element analysis software that integrates structural mechanics, fluid mechanics, electromagnetics, acoustics, and thermodynamics. ANSYS provides a complete modeling and analysis module, capable of pre- and post-processing. The analysis that ANSYS can carry out includes: structural linear static analysis, structural dynamic analysis, structural nonlinear analysis, fatigue and fracture analysis, and structural optimization design. ANSYS itself contains modeling functions, which can build structural solid models. There are mainly top-down and bottom-up modeling schemes. The former is to directly build the structure of the body, and then operate on the body through Boolean operations to obtain the structure. The model is suitable for simple and regular structures; the latter is a solid model for building structures in the order of points, lines, areas, and bodies, and is suitable for building more

complex structures [12]. You can also build a model in a specialized modeling software, and then import it into ANSYS. ANSYS has powerful meshing functions, including free meshing and mapped meshing. The quality of the grid determines the accuracy of the finite element calculation results. Calculation and analysis ability is the most powerful function of ANSYS. For each analysis type, ANSYS provides a variety of solution methods, which can be selected for different structures. The general post-processing program and the event-history post-processing program can separately view the results of structural analysis such as deformation displacement, stress, and time-varying curve.

The plate member in the structure has a remarkable geometric feature, that is, the size in one direction is much smaller than that in the other two directions [13]. Flat plates are usually divided into thin plates and medium plates. Quantitatively speaking, when the thickness is less than one fifteenth of the minimum size in the other two directions, it can be called thin plate. The bending deformation of thin plate is very small compared with its thickness. According to the particularity of thin plate, there are some reasonable thin plate theoretical assumptions. First, the extrusion deformation in the thickness direction of thin plate can be ignored. Transverse fibers similar to beams are assumed to have no extrusion. Second, in the plate bending deformation, the normal of the middle plane remains a straight line and still the normal of the elastic surface, which is the famous Kirchhoff straight normal hypothesis, which is similar to the plane section hypothesis in the beam bending theory. At this time, the vibration failure analysis equation of the aircraft engine blade is designed as follows:

$$G = \{M\}\{\delta\} \tag{9}$$

$$D = [C]\{\delta\} + [K]\{\delta\} \tag{10}$$

In the formula, $\{M\}$ represents the matrix at this time, $\{\delta\}$ represents the nodal matrix, $[C]$ represents the vibration set, $[K]$ represents the blade stiffness, and $\{p\}$ represents the engine load.

2.3 Achieve Vibration Failure Analysis of Aircraft Engine Blades

Rotating stall is another aerodynamic excitation source that produces high frequency response. Stall is caused by the limited ability of the boundary layer to bear the inverse pressure gradient. Experiments have confirmed that there are two types of stall in the compressor. One is simple blade stall, that is, the boundary layer separation on the blade surface; The other is rotating stall, which is a unique phenomenon in compressor. If all blades in the blade row are identical and feel the same stall inlet angle at the same time, the blades in the whole annular channel shall stall at the same time, that is, simple blade stall, such as local separation of blade surface of compressor under normal working condition. When some blades reach the stall condition before other blades in the blade row, a stall mass is formed. The stall mass rotates at a speed less than the rotor speed. When the blade rotates, it will alternately pass through the stall area and non stall area and be periodically excited, which is called rotating stall.

When a cascade works in its critical stall state and the air flow attack angle at the blade increases due to some local disturbance, resulting in air flow separation. Due to

the blocking effect of stall on the cascade channel, the inlet air deflects to the channels on both sides of the blade, reducing the inlet attack angle of the front blade and the rear blade (in the opposite direction of rotor rotation) the inlet attack angle of the blade increases, so the air flow also separates and becomes the next stall blade. In the case of blade air flow separation, as above, the air flow flows to the front blade at a smaller attack angle, so that the separation phenomenon of blade air flow disappears and exits the stall state. Such a continuous reaction forms a stall mass continuously moving towards the blade back, i.e. It is transmitted in the direction opposite to the rotor rotation, but because its transmission speed is less than the rotor rotation speed, it is observed in the absolute coordinate system that the stall mass still moves along the rotor rotation direction, but the speed is much lower [14].

The characteristic parameters that usually represent the rotating stall phenomenon include the rotation speed of the stall group, the number of the stall group, the width of the stall group, the strength of the stall group, and the type of the rotating stall. Most of the content about the rotating stall comes from the axial compressor test; the stall group observed in the experiment generally has a changing geometric shape and is rarely stable, but it is generally considered that the fully developed rotating stall group is stable and no longer changes with time; The rotation speed of the stall group is generally 10%–70% of the rotor speed; the number can be as high as 9 or as low as 1; the width can only occupy part of the blade height, or occupy the entire blade height, all of which are progressive stalls with the rotating stall type, it's still related to the sudden stall.

The type of rotating stall is divided into progressive stall and sudden stall according to the characteristics of the characteristic line change after the compressor enters the stall state: For progressive stall, the pressure ratio characteristic of the compressor is gradually reduced during the change from the stable working state to the rotating stall state; and the pressure ratio of an abrupt stall has obvious discontinuity or a sudden decrease. In the results presented in the experiment, progressive stalls generally have multiple stall clusters, and abrupt stalls seem to always produce only a single rotating stall cluster; full-leaf high-rotating stalls are usually abrupt stalls, but in some cases it can also be a gradual stall.

The magnitude and direction of the load exerted on the blade surface by rotating stall are affected by time and space changes, so it is difficult to obtain an accurate load model. In order to simplify the model, this paper only considers the sudden rotating stall with single stall group of whole blade height, and sets up a simplified load calculation model to simulate the load applied to the blade surface during the rotating stall. The dynamic pressure of rotating stall with or without distortion was measured by placing a dynamic pressure measuring sensor on the pneumatic interface near the compressor inlet.

Compressor inlet distortion is an important factor for stall boundary degradation and steady-state characteristic attenuation of aviation fan/compressor. Geometric asymmetry of inlet, gust, strong crosswind, change of yaw angle and pitch angle caused by high angle of attack flight, change of inlet flow field during hovering take-off and landing, airborne weapon Factors such as transient flow field changes during launch (such as missiles) will lead to uneven compressor inlet velocity field, temperature field and pressure field, resulting in degradation of aircraft stall boundary and reduction of stability margin, which is usually called compressor inlet distortion.

The influence of inlet distortion on the engine includes aerodynamic response and mechanical response. It not only affects the performance and stability of the engine and causes the compressor to surge, but also the non-uniform inlet conditions will produce exciting force on the rotating blade, produce a variety of complex harmonic exciting components, and increase the vibration level of the blade (for example, in most compressor blades, high-intensity and small angle inlet distortion can produce high-intensity high-order modal response), cause blade flutter, forced vibration and resonance, and may lead to blade vibration fatigue failure.

There are many examples of blade vibration fatigue and fracture damage caused by distortion and therefore modifying the design of relevant parts. Serious blade vibration occurred during the modification of a jet engine. According to the analysis, it is known that the blade resonance is caused by the flow field distortion caused by the inlet support plate. The failure of the first stage compressor blade of a certain engine is also due to the flow field mismatch between the inlet and the engine It is caused by serious inlet distortion.

No matter how complex the aerodynamic pressure and excitation factors borne by the compressor blades are, these different excitation factors form excitation force through one channel and act on the rotor blades. Therefore, the above aerodynamic excitation factors will also affect each other. There is interaction between wake, distortion and rotating stall.

Wake and intake distortion have an impact on the stall. Under certain circumstances, the interaction between the upstream stator wake and the unsteady flow inside the rotor channel may improve the aerodynamic performance near the stall point, thereby increasing the stable operating range of the rotor; but At the current row of blades where the downstream wake is generated, if the wake area is sharply expanded to cause severe separation, it will cause a rotating stall. Therefore, the change of wake can be used as a way to predict rotating stall. Intake distortion will reduce the duration of the stall prelude wave and induce the generation of rotating stall, which will lead to the reduction of compressor stall margin, and the influence of circumferential inlet distortion on the decline of stall margin is stronger than radial distortion. When the total intake pressure is distorted, after the system enters the rotating stall state, its flow value is less than the flow when the intake is uniform. However, intake distortion does not affect the propagation frequency of the rotating stall.

Intake distortion will change the original design flow conditions of the compressor, make the intake angle of attack of the compressor cascade deviate from the design value, increase the intensity of vortex disturbance, flow loss and the separation of the rotor blade exhaust flow, thereby affecting the blade wake speed and direction Influence. Rotating stall will result in increased turbulence in the wake area. And when it is close to rotating stall, it will cause a significant change in the wake waveform behind the rotor.

The rotor structure includes not only the rotor blades and the disc, but also the blades and disc connections such as tenons and collars, as well as the connections between the disc and disc, such as connecting rods and radial pins. Due to the influence of stiffness, the analysis model of the rotor is not just a simple solid model of the rotor and blisk, but a three-dimensional coupled model including connecting rods, radial pins, tenons and collars. Since the focus of this article is the study of the dynamic response of the

blade under aerodynamic loading, if the rotor structure is accurately modeled and the friction of the connected parts is considered, the mathematical solution will be extremely complicated and the solution time will be increased. Therefore, this article grasps the main contradiction and simplifies the rotor model.

When modeling, only the first-level roulette is considered, and the connection between the roulette and the roulette is ignored, and the axial and radial displacements of the roulette are constrained to simulate the constraints between the roulettes. There are fixed parts such as tenon and clamp ring between the blade and the wheel. The model does not consider its function, but uses the bonding of the blade and the wheel to simulate. Ignore some rounded corners and bosses on the roulette. Because if there are rounded corners and bosses in the solid model, it will make the meshing complicated, reduce the speed of simulation, and make it difficult for the accuracy of the analysis to reach the expected results.

Since the rotor is a rotationally symmetric structure, the ANSYS rotationally symmetrical entity modeling function is used when modeling, and the rotor disc is regarded as a rotationally symmetric entity for analysis, which reduces the amount of calculation for analysis. ANSYS program provides two solid modeling methods: top-down and bottom-up. When performing solid modeling from top to bottom, first define the most advanced volume primitives of the model. The program will automatically define the relevant surfaces, lines and key points, and use these advanced primitives to directly construct the geometric model. Bottom-up is to construct the model from the lowest-level primitives upwards, first define key points, and then use these key points to define high-level primitives, namely line, area, and volume primitives, until the entire model.

In order to facilitate structural improvement, this paper adopts a bottom-up solid modeling method, using the coordinate data of 70 points of the axial section of the rotor blade parallel to the y-axis as the key points, and inputting it into the ANSYS preprocessor to establish More advanced primitive lines, due to the complex characteristics of the leaf shape, especially the large curvature of the leaf shape at the leading and trailing edges of the leaf, the polyline fitted by cubic spline fitting is used to fit the leaf shape, and then pass through the boundary The line generates a leaf-shaped section, and then stretches in the y direction to form a solid model of the leaf.

Twenty-node hexahedral element SOLID95 is used to mesh the blade, because this is a high-precision element with intermediate nodes, and to adapt to boundary conditions, it can be degenerated into a pentahedral or tetrahedral element. However, there are many problems in the actual division, and there are many failure units, and even the mesh cannot be divided. After many times of analysis, it is believed that because the thickness of the rotor blade is thin and the edge angle is too small, the lines and surfaces generated by the hexahedral element cannot be recognized in ANSYS. Therefore, SOLID95's degenerate element-ten-node tetrahedral element SOLID92 was used for meshing in the later stage, and denser meshes were used for parts with large data gradients, such as the leading edge and trailing edge of rotor blades., And in the parts where the data change gradient is small, such as the wheel and the middle of the blade, in order to reduce the scale of the model, a relatively sparse grid is divided, and finally the finite element grid of the model is free of failure elements.

3   Experiment and Analysis

In order to test the analysis effect of the engine blade vibration failure analysis method designed in this paper, it is compared with the traditional blade vibration analysis method, and the following experiments are carried out.

3.1   Experiment Preparation

The data source used in this paper is that the front stator of the compressor rotor in the experiment has 30 blades. Although the change of the pressure in the front half of the suction with time is chaotic, which shows that there are multiple oscillations in one wake cycle, the pressure pulsation at the rear of the whole pressure surface and the suction surface shows good periodicity, which can be approximated as a high-frequency large amplitude sine wave whose mean value is not at zero, Moreover, the mean pressure on the pressure surface and suction surface first decreases and then increases along the leading edge and rear edge of the blade, while the amplitude of sinusoidal pulsation gradually decreases. The standard table of pulsation values at this time is shown in Table 1.

Table 1.   Standard table of pulsation value

Measuring point	Standard value/Pa
1	712.55
2	590.54
3	585.16
4	535.16
5	370.46

It can be seen from Table 1 that the blade pulsation value at this time is the standard blade vibration pulsation. Based on this value, the subsequent vibration failure analysis can be carried out.

3.2   Experimental Results and Discussion

The engine blade vibration failure analysis method designed in this paper and the traditional blade vibration failure analysis method were used to analyze the blade pulsation value, and compared with the standard value in Table 1. The comparison results are shown in Table 2.

It can be seen from Table 2 that the numerical value of the engine blade vibration failure analysis method designed in this paper is the closest to the standard value, which proves that its analysis effect is good and has certain application value. The reason for this result is that the vibration state of a single blade of a civil aircraft engine is analyzed by the method in the finite element environment. Based on this, in the virtual reality environment, the engine blade is meshed by SOLID92, which is the degradation unit

Table 2. Experimental results

Measuring point	The analysis value of the vibration analysis method designed in this paper/Pa	The traditional vibration analysis method analyzes the value/Pa
1	712.54	711.54
2	590.44	589.24
3	585.26	584.46
4	535.18	536.49
5	370.49	371.46

of SOLID95, so that it can better adapt to the gradient of data change and obtain more accurate analysis effect.

In order to further verify the effectiveness of the proposed method, the accuracy of the analysis results is taken as an indicator to validate the application performance of the proposed method and the traditional method. The results are shown in Table 3.

Table 3. Comparison of accuracy of analysis results

Test times	The vibration analysis method is designed in this paper	Traditional vibration analysis method
10	95.2%	88.1%
20	92.6%	86.9%
30	97.7%	83.5%
40	93.8%	84.4%
50	94.7%	85.0%

As can be seen from Table 3, in many experiments, the accuracy of analysis results of the vibration analysis method designed in this paper has always been above 90%, while the accuracy of analysis results of the traditional vibration analysis method is between 80% and 90%. By contrast, the method presented in this paper can more accurately analyze the vibration failure state of civil aircraft engine blades.

4   Conclusion

With people's attention and investment in aviation technology, various high and new technologies have been applied to the development of aero-engine blades. At present, engine blade research has become a multidisciplinary technical field, involving theoretical mechanics, vibration mechanics, material mechanics, thermodynamics, casting technology, coating technology, finite element analysis theory and so on. On the one

hand, these disciplines promote each other and make the engine blade technology more and more perfect, on the other hand, they restrict each other, and the optimal scheme of each discipline cannot be realized on the blade at the same time. After decades of development, engine blade research is not only to meet the performance requirements of the engine, but also to the comprehensive pursuit of reliability, usability, durability and maintainability.

Engine failure caused by engine blade fracture has always been one of the main causes of aero-engine failure, and blade surge and flutter are the main factors of blade crack and even fracture. Research on blade vibration has a very important role in preventing blade flutter and other phenomena. Therefore, in this study, the vibration failure analysis method of civil aircraft engine blade was designed, and the vibration failure analysis equation of blade was designed after the vibration state of a single blade of civil aircraft engine was analyzed by making full use of the advantages of virtual reality, so as to provide reference for subsequent aircraft engine research.

References

1. Aust, J., Mitrovic, A., Pons, D.: Assessment of the effect of cleanliness on the visual inspection of aircraft engine blades: an eye tracking study. Sensors **21**(18), 6135 (2021)
2. Wang, Y., Tang, B., Qin, Y., et al.: Rolling bearing fault detection of civil aircraft engine based on adaptive estimation of instantaneous angular speed. IEEE Trans. Ind. Inf. **16**(7), 4938–4948 (2019)
3. Vucetic, N., Jovii, G., Krstic, B., et al.: Further investigation of the repetitive failure in an aircraft engine cylinder head - mechanical properties of Aluminum alloy 242.0. Mechanika **26**(4), 285–292 (2020)
4. Huang, W.Q., Yang, X.G., Li, S.L.: Evaluation of service-induced microstructural damage for directionally solidified turbine blade of aircraft engine. Rare Met. **38**(02), 65–72 (2019)
5. Atilgan, R., Turan, O., Aydin, H.: Dynamic exergo-environmental analysis of a turboprop aircraft engine at various torques. Energy **186**(01), 1–9 (2019)
6. Li, C., She, H., Tang, Q., et al.: The coupling vibration characteristics of a flexible shaft-disk-blades system with mistuned features. Appl. Math. Model. **67**, 557–572 (2019)
7. Sun, H., Ren, A., Wang, Y., et al.: Deformation and vibration analysis of compressor rotor blades based on fluid-structure coupling. Eng. Fail. Anal. **122**, 1–18 (2021)
8. 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)
9. Arkadiusz, S., Jacek, C., Piotr, J.: Detection of cylinder misfire in an aircraft engine using linear and non-linear signal analysis. Measurement **174**(1), 108–112 (2021)
10. Torres-Carrillo, S., Siller, H.R., Vila, C., et al.: Environmental analysis of selective laser melting in the manufacturing of aeronautical turbine blades. J. Clean. Prod. **246**(10), 1–14 (2020)
11. Liu, S., et al.: Human memory update strategy: a multi-layer template update mechanism for remote visual monitoring. IEEE Trans. Multimedia **23**, 2188–2198 (2021)
12. Liu, L., Yang, Z., Chen, W., et al.: The whole engine dynamic response and security analysis during aero-engine blade out event. Comput. Simul. **37**(02), 47–52+124 (2020)

13. Liu, S., Wang, S., Liu, X., et al.: Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**(1), 90–102 (2021)
14. Chen, X., Hong, J., Wang, Y., et al.: Fatigue failure analysis of the central-driven bevel gear in a turboshaft engine arising from multi-source excitation. Eng. Fail. Anal. **119**(1), 104–111 (2021)

Intelligent Regulation Method of University Heating Water Flow Based on Adaptive Control Algorithm

Shengzuo Lin[✉]

Guangdong Polytechnic of Environmental Protection Engineering, Foshan 528000, China
serndror@sohu.com

Abstract. With the increasing requirements of energy saving and indoor environment, variable flow technology is widely used in heating system, and the adjustment of resistance is the throttle adjustment of valve. However, the economic loss of this adjustment method is too great. Therefore, a new intelligent adjustment method of university heating water flow is designed based on adaptive control algorithm. First, a single loop network is connected. Secondly, an intelligent adjustment system of university heating water flow is designed based on adaptive control algorithm. Experiments show that the designed intelligent adjustment of heating water flow is achieved.

Keywords: Adaptive control algorithm · Heating in colleges and universities · Water flow · Intelligent regulation

1 Introduction

In the traditional double-pipe heating pipe network system, regardless of the branch pipe network or the ring pipe network, the hydraulic conditions of the parallel branch lines in the pipe network system are mutually influenced, that is, if the hydraulic conditions of a branch line change, it will lead to the change of the hydraulic conditions of the parallel branch lines, thus changing the hydraulic conditions of the whole heating system [1]. In actual engineering, if the hydraulic condition of a branch changes greatly and the system itself cannot meet the requirements [2–4], in order to ensure the operating pressure of a branch line, a booster pump [5] should be installed on the branch line without increasing the head of the heat source circulating pump. However, this way will change the flow of the trunk line, and then change the hydraulic condition of the trunk line, which will lead to changes in the operating flow and pressure of other branches.

Because the water supply and backwater share the same pipe, the hot water in the pipe flows out of the heat source and flows through each branch trunk line in turn along the water flow direction, and finally flows back to the heat source [7]. In the process, the hot water in the ring trunk line flows into the branch trunk line and then flows back to the ring trunk line, so the circulating flow of each point of the ring trunk line is equal,

and its value is equal to the flow of the heat source outlet pipe section [8]. It can be seen that the hot water flow of each branch trunk line is independent and unaffected. The hydraulic condition of the ring trunk line of the unidirectional ring heating system affects the hydraulic condition of the branch line [9], while the hydraulic condition of the branch line does not affect the hydraulic condition of the ring trunk line.

Because the traditional double-pipe heating system needs to lay two water supply and return pipes at the same time, the initial investment and operation cost of the pipe network are high, and the hydraulic condition of the system is complex [10], and the stability is poor. Reasonable design and planning of the heating pipe network system can ensure the high operating efficiency of the heating system and the heat medium transmission and distribution capacity under network expansion or accident conditions, which is of great benefit to reducing the initial investment and operating cost of the heating pipe network system and improving the heating quality of the heating network. One-way loop network has the greatest advantages of simple and stable hydraulic conditions and convenient adjustment of hydraulic conditions. However, this kind of system has been in existence for a short time. In order to ensure the heating quality of the system, meet the heat demand of users, make the heat energy production, transmission and distribution reasonable and economical, and reduce the operating cost of the heat supply network, it is necessary to study the operation regulation mode of the unidirectional ring network heating.

Intelligent regulation of university heating water needs to consider the increase of heating area, and solve the above problems such as large initial investment, high operating cost and complicated hydraulic conditions. It has become a trend to use multi-heat source combined heating. Compared with single heat source heating system, multi-heat source combined heating system has outstanding advantages, which disperses the supply position of heat sources, reduces the pipe diameter and flow rate of heating system, and increases the investment of pipe network and the energy consumption of system operation. The operation cost of heating system is very high, and its operation consumes huge power. Therefore, the research on heating regulation mode and operation regulation of multi-heat source combined heating system is of great significance to reduce the operation energy consumption of heating system.

2 Design of Intelligent Regulation Method of University Heating Water Flow Based on Adaptive Control Algorithm

2.1 Connect the Single Ring Network

One-way loop network system is composed of heat source, heat network and heat users, which is a heating system based on loop pipe network, while the core of double-pipe heating pipe network is heat source, which is different from each other. The unidirectional ring network system mainly includes unidirectional ring trunk line, ring trunk line boosting point, ring trunk line heating point and ring trunk line cooling point; The circulation schematic diagram of branch trunk water supply pipe, branch trunk return pipe, heating station and single ring network is shown in Fig. 1 below.

As can be seen from Fig. 1, the unidirectional ring trunk line: the hot water flows circularly along the unidirectional ring trunk line after flowing out from the heating point,
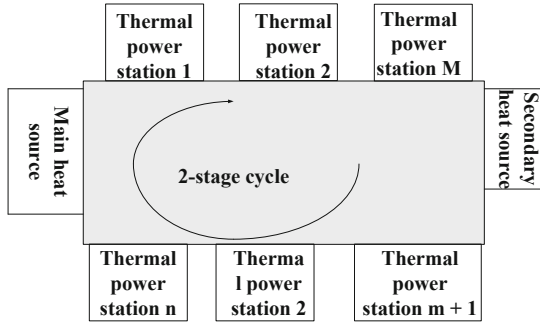
Fig. 1. Schematic diagram of single ring network connection

and some of the hot water flows out of the ring trunk line through the cooling point during the circulation process, and then flows back to the ring trunk line after cooling down, so that the temperature of the hot water in the ring trunk line gradually decreases along the flow direction, but the circulation flow of the ring trunk line is constant along the way, and the pressure of the pipe network gradually decreases along the circulation flow direction. Finally, the hot water in the one-way ring flows back to the heating point for heating, thus continuing to circulate.

Ring heating point of trunk line: In the one-way loop heating system, if the heat source heats the hot water in the loop, the heat source is the heating point of the one-way loop network. Cooling point of ring trunk line: When there is a branch trunk line connected to the ring trunk line, part of hot water flows into the branch trunk line from the ring trunk line to deliver heat to the heat users, and then flows back to the ring trunk line after the temperature drops. Ring-shaped trunk line boosting point: A circulating water pump is set on the ring-shaped trunk line to provide circulating power to eliminate the along-way resistance and local resistance caused by the hot water flowing along the one-way ring network, and then the device circulating water pump of the ring-shaped trunk line is the ring-shaped trunk line boosting point. One-way ring network system consists of three cycles-heat source cycle, heat network cycle and heat user cycle.

The operating temperature level of heat medium in the three-stage cycle is different, the heat source cycle temperature level is the highest, the loop trunk cycle temperature level is the second, and the heat user cycle temperature level is the lowest. Increasing the water supply temperature of heat source can increase the temperature difference between supply and return water of heat source circulation, thus reducing the power consumption of heat source circulation; The mixed temperature of the circulating water supply of heat source and the hot water of the ring trunk line decreases through the heating point of the ring trunk line, thus lowering the operating temperature of the ring trunk line and reducing the insulation investment of the ring trunk line network to a certain extent; As the heat is gradually transferred from the ring trunk line to the heat users along the flow direction of heat medium, the temperature of hot water on the ring trunk line gradually decreases, while the water supply temperature obtained by the heat users located in different branch trunk lines gradually decreases along the flow direction of hot water on the ring trunk line, and the temperature of hot water on the ring trunk line presents

a stepwise distribution. Therefore, it can be seen from the form of unidirectional ring network system that the energy in unidirectional ring network system is in the form of cascade utilization.

The circulating water pump located at the heat source of the common double-pipe heating pipe network system provides the transmission power of the whole heating system. However, when the heating scale of the heating system is large, it is necessary to make the head of the circulating water pump large enough to meet the pressure requirements of the most unfavorable users at the end of the system, resulting in high operating pressure of the pipe network and heat users near the heat source. Excessive operating pressure of the system increases the investment cost of pipe network and equipment, and increases the damage probability of heating equipment. Moreover, the heat users close to the heat source need to reduce the operating pressure of the secondary network system by throttling and depressurizing, which leads to the waste of energy consumption and increases the operating cost of the system.

In the traditional double-pipe heating system, the change of hydraulic condition of any heat user will cause the change of hydraulic condition of the main line. Compared with the traditional double-pipe heating system [11], the trunk line of the unidirectional ring network system adopts the form of a single pipe, and the system has only one water supply pipe. The resistances of the ring trunk line and the branch trunk line of the heating system are powered by the pumps located on the ring trunk line and the branch trunk line respectively, and the required lift of the circulating pump on the ring trunk line is reduced, which reduces the pressure bearing capacity of the pipe network. This arrangement of power transmission equipment theoretically eliminates the throttling loss on the branch lines and reduces the operating cost. The hydraulic conditions of branch trunk lines in unidirectional ring network system do not affect the hydraulic conditions of trunk lines, and the hydraulic conditions of branch trunk lines do not affect each other.

In the traditional double-pipe heating system, because of the complex series-parallel control between pumps and the mutual influence of hydraulic conditions, it is not conducive to the application of distributed variable frequency pump system. However, in the unidirectional ring network system, the system operates in different levels, and each level of operation is relatively independent. The hydraulic condition of the secondary side of the branch trunk line does not affect the hydraulic condition of the ring trunk line. And the pressure loss of each branch line is overcome by the circulating pump on the branch line, which avoids the throttling loss and reduces the operation energy consumption compared with the double-pipe system. In the unidirectional loop network system, the pressure difference between the water supply and return branch lines from a branch node is small, and each heat user of the secondary network at the branch line side can select the appropriate distributed variable frequency pump according to the hydraulic calculation data. In the traditional double-pipe heating system, when the heat load in the area under its jurisdiction increases, the demand of load increase can only be met by increasing the circulating water volume of the pipe network. However, the increase of circulating water volume will lead to the increase of the hydraulic loss of the system, which will lead to the situation that the pressure difference for return pipe in the system becomes smaller or even negative.

The heat transfer capacity of single pipe system can be improved not only by increasing the circulating water flow, but also by raising the hot water temperature, so that the heat transfer and heat source configuration of the system are no longer limited by the circulating water. According to the distribution of heat load in the heating area, the pipe network and heat sources are scientifically arranged. When the heating load increases in the future, new heat sources can be installed at appropriate locations to provide the heating capacity of the high-heat network, thus improving the heat load carrying capacity of the heating system. Compared with the traditional double-pipe heating system, when new users are added, the supply and return water pressure of some branch end users does not meet the requirements and the heating effect becomes worse. Because the unidirectional ring network system is a cascade heating system, the hydraulic conditions of branch lines do not affect the hydraulic conditions of ring trunk lines. When new users are added to branch lines, the circulating water pump head of ring trunk lines remains unchanged, and the resistance of new users to branch lines is borne by the internal water pump of branch lines. For the increase of heating load of the system caused by new user branch lines, it is only necessary to increase the water supply temperature of the heat source or add new heat sources in the heating system to meet the demand of increasing heating load. The connection mode between heat network and heat users has different effects on the initial investment, hydraulic and thermal conditions, and operating costs of the heating system. Therefore, we should try our best to choose the appropriate heating connection mode. There are three traditional connection modes: direct connection, mixed water connection and indirect connection. The latter two connection modes are also suitable for one-way loop system. The schematic diagram of heating station connection at this time is shown in Fig. 2 below.
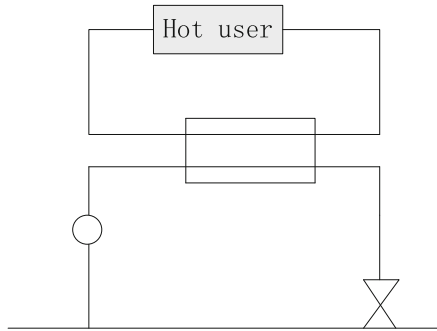
Fig. 2. Schematic diagram of thermal station connection

As can be seen from Fig. 2, the indirect connection form in the one-way loop network heating station is that the primary and secondary sides of the heating station are two independent circulation systems, and the calculation of the user heating heat load Q_i' of the heating station at this time is shown in the following (1).

$$Q_i' = q_f \cdot F \times 10^{-6}$$

(1)

In the formula (1), q_f represents the thermal index of heating buildings of heat users, and F represents the heating area of users of heating stations. For the primary and secondary sides of the thermal station, ignoring the heat loss along the heat network and the heat exchange efficiency of the heat exchanger, the heat balance equation for the primary side of the thermal station is shown in (2).

$$Q_i = cG_{i1} \cdot (\tau_i - \tau_{ih}) \tag{2}$$

In formula (2), G_{i1} represents the water supply flow of the branch line, τ_i represents the water supply temperature of the branch line on one side, and τ_{ih} represents the backwater temperature of the branch line. At this time, the thermal balance equation of the branch line will be shown in the following (3).

$$(G_c - G_{i1})c\tau_i + G_{i1}c\tau_{i1} = G_c c\tau_{i+1} \tag{3}$$

In formula (3), G_c represents the hot water circulating flow of the ring trunk line, and τ_{i+1} represents the water supply temperature at the cooling point. By integrating the above formulas, we can get the temperature of the cooling node of the ring trunk line and the primary side flow of the thermal station as shown in (4) and (5) below.

$$\tau_{i+1} = \tau_i - \frac{Q_i}{cG_c} \tag{4}$$

$$G_{i1} = \frac{Q_i}{c(\tau_i - \tau_{ih})} \tag{5}$$

When the mixed water pressurizing pump is installed in the heating station in the system, when the mixed water direct connection is adopted, compared with the direct connection without mixed water, the mixed water connection can increase the temperature difference between the supply and return water of the ring trunk line and the branch trunk line, reduce the flow in the ring network, reduce the pipe network diameter, reduce the initial investment cost and also reduce the operation power consumption. Compared with indirect connection, hybrid connection has stronger adaptability to the temperature of heat medium on the primary side of thermal power station.

However, when the mixed water direct connection mode is adopted, the hydraulic condition of the secondary side of the thermal power station will be affected by the primary side. When the traditional double-pipe heating system adopts mixed water connection, there are three forms of mixed water pump installed on the water supply pipe of the secondary network, return pipe and bypass pipe respectively. Because there is no return pipe in the unidirectional ring network system, the pressure head provided by the ring trunk line to the heating station is only the loss of hot water along the way between the connection point of the ring trunk line and the branch line for return pipe section. The pressure head is very small and negligible, so it cannot be used as the pressure head for hot water flow in the secondary side of the heating station. The mixed water pressurizing pump is needed to overcome the resistance loss in the secondary side of the heating station. However, when the water mixing pump is installed on the bypass pipe, it cannot overcome the resistance of the user side, so the water mixing connection of the circulating pump installed on the water mixing pipe is not suitable for the single-pipe

water mixing direct connection system. As shown in the figure below, the water mixing connection of the water pump installed on the secondary water supply pipe is suitable for the middle and lower reaches of the heating system. In this area, the pressure of the pipe network is low, and the water supply of the water pump is pressurized, which not only meets the pressure requirements of the hydrostatic pressure line of the users on the secondary side, but also does not cause overpressure on the users' side. The mixed water connection of the water pump installed on the return pipe on the secondary side is suitable for the upper and middle reaches of the heating system. In this area, the pipe network pressure is high, and the backwater of the water pump is pressurized, so that the backwater on the secondary side can meet the pressure of the pipe network on the primary side. At this time, the circulating water flowmeter of the users on the secondary side of the thermal station is shown in the following formula (6). Among them, the economic ratio of hydraulic calculation of indoor mechanical circulating hot water heating system is set to 0.86.

$$G_{i2} = 0.86 \frac{Q_i}{t_g - t_h} \qquad (6)$$

In formula (6), t_g represents the water supply temperature of heat users and t_h represents the backwater temperature of heat users. At this time, assuming that G_{ih} represents the backwater quantity drawn from the backwater of the secondary side of the heat station, the mixing water of the heat station at this time is as shown in (7) and (8) below.

$$G_{i1} = \frac{0.86Q_i}{\tau_i - h} \qquad (7)$$

$$u_i = \frac{G_{ih}}{G_{i1}} \qquad (8)$$

From the above formula, we can know the relationship between water mixing ratio and temperature, and the relationship between water supply from the primary side of the thermal power station, water withdrawal from the secondary side and circulating water from the secondary side of the thermal power station.

Because the user system of the secondary side of the heating station is a traditional double-pipe heating system, the traditional double-pipe user system regulation mode is also suitable for the operation regulation of the secondary side of the heating station of the unidirectional ring network system, that is, the user system. When the user adopts quality control, the circulating water flow of the user system is a fixed value, and only the water supply temperature of the system can be adjusted. When the secondary network adopts quality control of changing flow by stages, the operating flow of the system is different in different operation stages. In each operation stage, the flow of the system is a fixed value. The water supply and return temperature formulas of quality control of changing flow by stages are shown in (9) and (10) below.

$$G_{i1} = \frac{G_{i2}}{G_{i1} + 1} \qquad (9)$$

$$j = \frac{G_{i2} - 1}{G_{i1} + 1} \qquad (10)$$

When the user system adopts the quality regulation of changing flow by stages, the larger the temperature difference between the supply and return water at the user side and the smaller the relative operation flow, the smaller the power of the water pump at the user side, the smaller the operation energy consumption of the system and the better the energy-saving effect. However, when the flow of the user system is too small, it will affect the heating effect of the user system. For the double-pipe user heating system, the user circulation flow is too small, which will lead to an increase in the proportion difference of gravity circulation effect among users at all levels, resulting in the vertical imbalance of the system. For a single-pipe user heating system, the circulation flow is too small, which leads to different degrees of change of heat transfer coefficient values of radiators of users in each floor, and also causes vertical misalignment of the user system.

2.2 Design an Intelligent Water Flow Regulation System Based on Adaptive Control Algorithm.

The adaptive process is a process of approaching the target continuously. The path it follows is expressed by mathematical model, which is called adaptive algorithm. Gradient-based algorithm is usually used, especially the least mean square error algorithm. Adaptive algorithm can be implemented by hardware or software. The former designs the circuit according to the mathematical model of the algorithm, while the latter compiles the mathematical model of the algorithm into a program and realizes it with a computer. There are many kinds of algorithms, and its selection is very important, which determines the performance, quality and feasibility of the processing system. In order to avoid vertical misalignment of user system, an intelligent water flow regulation system is designed based on adaptive control algorithm. The temperature of hot water on the ring trunk line of unidirectional ring network system gradually decreases with the direction of hot water flow, which is different from that of the ring trunk line.

The flow of branch trunk lines can be adjusted synchronously or asynchronously. The regulation of one-way loop network system includes hierarchical regulation of loop trunk line, primary side regulation of thermal station and secondary side regulation of thermal station. Therefore, when choosing the adjustment mode of unidirectional loop network system, the adjustment of primary and secondary sides of loop trunk line and thermal station must be considered at the same time. When the heating range of the heating system is relatively small, the single heat source and single loop ring pipe network system can meet the heating requirements of heat users. When the heating system adopts different regulation modes, the energy consumption of the system pumps is different. Because the study of heating regulation mode is of great significance to the economic and reliable operation of heating system, it is necessary to study the heating regulation mode of single heat source unidirectional ring network system as the basic model of multi-heat source unidirectional ring network system.

Under the condition that all the heating stations in the unidirectional loop network system are connected by mixed water, the regulation rules of the loop trunk line, branch trunk line and user side with different regulation modes are obtained, and the energy consumption of the system pumps combined by various regulation modes and the stability of the user system hydraulic conditions are analyzed, as well as the thermal accessibility

of the heat users in the unidirectional loop network system with different regulation modes of the heat source water supply temperature.

When the heating stations in the unidirectional ring network system are connected by mixed water, in some cases (the user's geographical location is higher or the pressure level of the user's node is lower), the user-side circulating pump should not only overcome the system resistance but also pressurize the system. At this time, for the normal operation of the user-side, the head change of the user-side circulating pump should not be too large, thus limiting the change of the user-side operating flow. At this time, the user side can only use quality adjustment. Because each branch trunk line adopts quality control, and the temperature of the cooling point of the ring trunk line is equal to the water supply temperature of the primary side of the water mixing station, the temperature control of each cooling point of the ring trunk line also meets the quality control law, so when the ring trunk line keeps constant flow operation, the regulation law of the ring trunk line is also quality control, so it is necessary to design the design working condition parameters of the ring trunk line, as shown in Table 1 below.

Table 1. Working condition parameters

Cooling point	Temperature (°C)	Flow rate t/h	Water mixing ratio
1	130	54	3
2	124	58	2.7
3	118	63	2.4
4	112	69	2.1
5	106	77	1.8
6	100	86	1.5
7	94	98	1.2
8	88	113	0.9
9	82	134	0.6
10	76	165	0.3

It can be seen from Table 1 that the unidirectional ring network is also a hot water heating pipe network, and its heat medium parameters have little difference compared with the traditional double pipe system, so the diameter of the ring trunk line is selected according to the specific friction recommended by the design specification of heating pipe network.

On the premise that the secondary side of the heating station is subject to quality control, when the loop trunk line and branch trunk line adopt quality control at the same time, the temperature of heat source supply and return water and the temperature of each cooling point gradually decrease with the increase of outdoor temperature. The loop trunk line and the secondary side of the heating station both operate at a constant flow rate, and the water mixing ratio of each water mixing station remains unchanged with the outdoor temperature. However, the lower the water supply temperature at the cooling

point along the direction of hot water circulating flow, the smaller the water mixing ratio of the water mixing station, and the larger the hot water flow from the primary side is required. When this regulation method is adopted, the system only needs to adjust the water supply temperature of the heat source of the unidirectional ring network system to meet the heat load demand of heat users with the change of outdoor temperature. This adjustment method has the advantages of small adjustment amount and simple operation. At the same time, the water pressure diagram of the ring trunk line has been kept at the design working condition, and the pressure at each cooling point remains unchanged, which does not affect the normal operation of the users on the secondary side, and the heating network runs stably. However, the circulating flow rates at all levels of the unidirectional ring network system are all constant, which makes the pump operation energy consumption of the system very high.

When the annular trunk line runs at a constant flow rate and the water supply temperature of the heat source remains unchanged, with the increase of outdoor temperature, the heat load of heat users decreases, the backwater temperature of the heat source increases, and the water supply temperature of the primary side of each cooling point also increases correspondingly, so the hot water flow from the annular trunk line into the water mixing station through the branch trunk line decreases, while the backwater flow of the secondary side pumped by the bypass pipe in the water mixing station increases, and the water mixing ratio of the water mixing pumps in each heat station increases with the increase of outdoor temperature. Under this regulation mode,

Under the operation mode of constant flow and constant water supply temperature of the ring trunk line, only the hot water flow supplied by the primary side of each mixing station and the return water flow of the secondary side pumped by the bypass pipe need to be adjusted, with few adjustment parameters. However, during the operation, the mixing ratio of each mixing station is constantly changing, and the above two parameters need to be constantly adjusted, which requires high operation requirements. In addition, when this regulation mode is adopted, the operation flow of the ring trunk line of the unidirectional ring network and the secondary side of the water mixing station are all designed flow, which leads to high operation energy consumption. Because the heat source water supply keeps the design temperature all the time, it is not conducive to the utilization of low-grade energy, and the heat source efficiency is low. The temperature of water supply at each cooling point of the ring trunk line increases with the increase of outdoor temperature, which requires higher pipeline insulation materials for the ring trunk line and larger initial investment of pipeline system. Although there are some disadvantages in the operation regulation mode of constant flow rate and constant water supply temperature of the ring trunk line for the unidirectional ring network heating system with single heat source, it has great applicability for the operation mode of the main heat source in the unidirectional ring network system with multiple heat sources, because the main heat source in the system runs at full load for a long time, and with the increase of outdoor temperature, the change law of water supply temperature of each cooling point within the heating range of the main heat source is the same as this.

2.3   Realize Intelligent Regulation of Heating Water Flow in Colleges and Universities

The temperature change of the supply and return water in the one-way loop network system with the quality control mode of changing the flow by stages is different from that of the double-pipe system. The increase of the supply water temperature in the two-pipe system is equal to the decrease of the return water temperature, while the temperature change of the heat source supply water in the one-way loop network system is larger, the change range of the return water temperature is smaller than that of the supply water temperature, and the supply and return water temperature of the heat source all show temperature increase. Along the circulating water flow direction of the ring trunk line, the increasing range of the primary water supply temperature of the heating station at each cooling point at the flow staging point gradually decreases.

On the premise that the secondary side of the heating station is of quality control, when the loop trunk line and branch trunk line simultaneously adopt the quality control of changing the flow by stages, the loop trunk line and the primary side of the heating station both operate at constant flow by stages, and the water mixing ratio of each water mixing station remains unchanged at each stage. In the low flow operation stage, the water mixing ratio of each heating station is greater than that of the design flow operation stage. In each operation stage, the system only needs to adjust the water supply temperature of heat source to meet the heat load demand of heat users with the change of outdoor temperature. When switching between different stages, it is necessary to adjust the water mixing ratio of each water mixing station. This combination of adjustment modes also has small adjustment amount and simple operation.

When the outdoor temperature rises, when the flow rate of the ring trunk line drops to 75% of the designed flow rate, the backwater temperature of the heat source decreases to expand the temperature difference between the supply and the backwater to ensure the heat supply; At the same time, the water supply temperature of each cooling point also decreases, and the cooling range of the water supply temperature of each cooling point is different, and the temperature of the cooling point decreases more along the direction of hot water circulating in the ring trunk line; On the contrary, the water supply flow of the primary side of each cooling point thermal station increases more along the hot water circulation direction of the ring trunk line.

During the operation of the system, it is only necessary to adjust the hot water flow in the primary side of each thermal station and the backwater flow in the secondary side of pumping, with few adjustment parameters. However, the mixing ratio of each mixing pump is constantly changing, so the above two parameters need to be constantly adjusted, and the operation requirements are high. The water supply temperature of the heat source in this regulation mode is always kept at the design temperature, which reduces the thermal efficiency of the heat source and wastes high-grade energy.

When equal temperature difference adjustment is adopted for the ring trunk line and branch trunk line, the temperature difference between the supply and return water of the heat source and the water supply temperature difference between the cooling points remain unchanged. With the increase of outdoor temperature, the flow rate of the primary side of the ring trunk line and branch line gradually decreases, and because of the constant flow rate of the secondary side of the heat station, the amount of backwater

pumped by the heat station through the bypass pipe increases, and the water mixing ratio of the water mixing pump of each heat station gradually increases. During the operation stages of the heating system with different flow rates, the loop trunk line, branch trunk line and the three-stage pipe network on the secondary side of the heating station are all regulated by constant flow rate, and the water mixing ratio of each heating station keeps constant. When the system is switched from the high flow rate stage to the low flow rate stage, the temperature of the supply and return water of the heat source and the temperature of each cooling point increase, which is different from that of the ordinary double-pipe system, and the temperature increase degree of each cooling point decreases along the circulation direction.

When the secondary side of the thermal station is changed from the high-flow stage to the low-flow stage, the temperature difference between the supply and return water temperature of the heat source and the water temperature of each cooling water supply decreases by the same temperature difference, but at the same time, the flow rate of the primary side of the thermal station increases significantly along the direction of hot water circulation along the ring trunk line, especially at the last cooling point, and the water mixing ratio becomes negative, which shows that the water temperature of the cooling point at the end is lower than that of the secondary side of the thermal station. Under this combination mode, when the secondary side of the thermal station runs at low flow rate, the water supply temperature of the heat source decreases too much, which directly leads to the water supply temperature of the lower cooling point on the ring trunk line being lower than that of the users on the secondary side of the thermal station, resulting in poor heating effect of the end users.

3 Experiment

In order to test the conditional effect of the intelligent regulation method of heating water flow in colleges and universities designed in this paper, it is compared with the traditional intelligent regulation method of water flow. Ensure the environment of the control group and the experimental group is consistent. Methods Each experiment was repeated ten times. The experiment is as follows.

3.1 Experimental Preparation

The operation requirements of the adjustment mode combination when the primary side flow of the thermal station and the return water flow of the pumping secondary side continuously change are relatively high. The combination of variable flow regulation has a higher energy-saving rate. When the loop trunk line and branch trunk line are regulated by equal temperature difference, the combination of quality regulation by changing flow in stages at the user side has the highest energy-saving rate. The connection table of intelligent water supply regulation at this time is shown in Table 2 below. Among them, the energy saving rate exceeding 8% is regarded as a higher level, and the hydraulic condition of users is stable the stability of users' hydraulic state means that when other users' flow changes, it is a higher level to keep their own flow changes within 2%.

Table 2. Intelligent connection table of water supply

Operation requirement	1	2	3
Energy saving rate	Low	High	low
Insulation investment	1	1	26%
The hydraulic condition of users is stable	Low	High	low
Meet the water supply needs of end users	Yes	Yes	no
Avoid energy waste	Yes	Yes	yes

Table 3. Results of traditional intelligent regulation method of hot water flow

Operation requirement	1	2	3
Energy saving rate	Low	Low	Low
Insulation investment	67%	54%	23%
The hydraulic condition of users is stable	Low	Low	Low
Meet the water supply needs of end users	No	Yes	No
Avoid energy waste	No	No	No

It can be seen from Table 2 and Table 3 that the water flow adjusted by the automatic adjustment method designed in this paper is more stable, which proves that the adjustment effect of the designed method is better and has certain application value.

4   Discussion

Through the analysis of the hydraulic and thermal conditions of the unidirectional ring network, it can be seen that the hydraulic and thermal conditions of the unidirectional ring network system are more independent than those of the traditional double-pipe system. According to the analysis of thermal conditions, the temperature of the ring trunk line cooling point of the unidirectional ring network is calculated and determined, and the temperature characteristics of the unidirectional ring network system are obtained. It is concluded that the determination of the inlet temperature of heating point of trunk line of unidirectional ring network should meet the heating demand of end users.

For the single heat source unidirectional loop network system, under the mixed water connection or indirect connection mode, the operation regulation modes of the system are solved with different regulation modes combinations for the loop trunk line, branch trunk line and the secondary side of the thermal station, and the regulation rules of various regulation modes combinations are obtained. When the ring trunk line and branch trunk line are regulated by equal temperature difference, and the secondary side of the thermal station is regulated by changing the flow by stages, the combined system has the lowest energy consumption. When the same regulation mode combination is adopted, the power consumption of water pump in mixed water connection mode of the

system is lower than that in indirect connection mode. The indirect connection system has more obvious energy-saving effect than the mixed water connection system by adopting the operation mode of constant water supply temperature of the ring trunk line, but this mode wastes high-grade energy. Moreover, the operation mode of constant water supply temperature and constant flow of the ring trunk line requires higher pipeline insulation materials, which increases the initial investment of the pipeline. When mixed water connection is used in the heating system, the equipment investment of the system is low, and the adaptability to the water supply temperature of the primary side of the heating station is strong. However, the operation regulation mode of the ring trunk line with variable flow will affect the normal operation of the user system, and the operating pressure of the ring trunk line should not be too high.

5   Conclusion

In this paper, the intelligent regulation method of university heating water flow is studied, and the unidirectional loop network system is introduced. By virtue of its superior hydraulic and thermal conditions compared with the traditional double pipe network system, the operation regulation modes such as trunk line, branch line and thermal station of heating water flow regulation loop network are improved. Experiments show that this method has the performance of more stable flow regulation, lower energy consumption and lower investment cost, and has certain application value.

References

1. Gu, L., Zeng, D., Li, W., et al.: Intelligent VNF orchestration and flow scheduling via model-assisted deep reinforcement learning. IEEE J. Sel. Areas Commun. **38**(2), 279–291 (2020)
2. Liu, Y., Liu, X., Duan, B., et al.: Polymer-water interaction enabled intelligent moisture regulation in hydrogels. J. Phys. Chem. Lett. **12**(10), 2587–2592 (2021)
3. Yan, X., Gong, J., Wu, Q.: Pollution source intelligent location algorithm in water quality sensor networks. Neural Comput. Appl. **33**(1), 209–222 (2020). https://doi.org/10.1007/s00521-020-05000-8
4. Cerone, A.: Model mining: integrating data analytics, modelling and verification. J. Intell. Inf. Syst. **52**(3), 501–532 (2019)
5. Ma, Y., Fan, X., Cai, J., et al.: Application of sensor data information cognitive computing algorithm in adaptive control of wheeled robot. IEEE Sen. J. **22**(18), 17343–17350 (2021)
6. Yuan, D., Wang, Y.: Data driven model-free adaptive control method for quadrotor formation trajectory tracking based on RISE and ISMC algorithm. Sensors **21**(4), 1289 (2021)
7. Shukl, P., Singh, B.: Grid integration of three-phase single-stage PV system using adaptive Laguerre filter based control algorithm under nonideal distribution system. IEEE Trans. Ind. Appl. **55**(6), 6193–6202 (2019)
8. Ke, X., Zhang, D.: Fuzzy control algorithm for adaptive optical systems. Appl. Opt. **58**(36), 9967 (2019)

9. Landau, I.D., Melendez, R., Dugard, L., et al.: Robust and adaptive feedback noise attenuation in ducts. IEEE Trans. Control Syst. Technol. **27**(2), 872–879 (2019)
10. Mostafa, E., Li, F.U., Arafa, I.I., et al.: A solution of UAV localization problem using an interacting multiple nonlinear fuzzy adaptive H_∞ models filter algorithm. Chin. J. Aeronaut. **32**(04), 212–224 (2019)
11. Li, N., Gao, B., Fiona, F.: Design and simulation of bio-oxidation tank heating system in Alpine mining area. Comput. Simul. **38**(07), 214–218+260 (2021)
12. Liu, S., Liu, D., Muhammad, K., Ding, W.: Effective template update mechanism in visual tracking with background clutter. Neurocomputing **458**, 615–625 (2021)
13. Liu, S., et al.: Human memory update strategy: a multi-layer template update mechanism for remote visual monitoring. IEEE Trans. Multimedia **23**, 2188–2198 (2021)
14. Liu, S., Wang, S., Liu, X., et al.: Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**(1), 90–102 (2021)

Binaural Spatialization: Comparing Head Related Transfer Function Models for Use in Virtual and Augmented Reality Applications

Simone Angelucci, Fabio Franchi, Fabio Graziosi, and Claudia Rinaldi$^{(\boxtimes)}$

Department of Information Engineering, Computer Science and Mathematics (DISIM), Università degli Studi dell'Aquila, via Vetoio 1, 67100 L'Aquila, Italy
simone.angelucci@student.univaq.it,
{fabio.franchi,fabio.graziosi,claudia.rinaldi}@univaq.it
https://www.univaq.it/

Abstract. This work is focused on the binaural spatialization, presenting an analysis of the most common solutions to understand and classify their advantages and drawbacks and to find the one that results in a better virtual/augmented experience on the basis of subjective tests. This work is a preliminary step toward the implementation of an augmented reality system for cultural heritage enjoyment exploiting spatial audio through low-cost devices. Two different models are implemented in order to avoid the use of non individualized HRTFs and results show promising opportunities that must be further exploited.

Keywords: Audio Signal Processing · Binaural spatialization · Head Related Transfer Function · Augmented Reality (AR) · Virtual Reality (VR)

1 Introduction

The recent widespreading of multimedia applications for learning and entertainment, even pushed by the recent 5G paradigm, has also to take into account the proper treating of audio signals. In this context many research topics are involved as efficient coding solutions, innovative effects development, artificial reverberation advancements, audio synthesis improvements and audio spatialization issues. The latter aspect is particularly important while conceiving with different kind of Augmented and Virtual Reality (AR/VR) applications, [8,13].

This paper focuses on the exploitation of audio spatialization for improving AR/VR experiences for Cultural Heritage (CH). Indeed, the progress of digital information has significantly affected the evolution of CH dissemination [23], offering new technological possibilities for developing, e.g., the market of tourist

services [18] and CH organisations have to address new users needs by creating innovative applications [1], i.e. AR/VR based. Indeed, AR technology gives a different perception of reality, as it enriches reality with a computer-generated layer containing visual, audio, and tactile information while using a "virtual" representation of a classic museum allows access to aspects of the artwork that may otherwise be hidden [19].

However, in order to build a complete and complex cultural representation via digital heritage technologies, developers must also understand how users interact with the system or interface [29]. When designing AR applications, it is important to choose the best combination of techniques for presenting the appropriate digital information to the users. During the past years, the main aim of AR/VR applications for CH changed from a mere virtual recreation of object to display, to the development of an entire virtual environment able to disseminate and teach culture. The idea is the opposite of a "dead museum": users don't have to see an accumulation of 3D heritage objects, but feel and understand another culture through those items. An important aspect is the relation between AR/VR and education. This new way to present Culture enhances the learning process, encouraging students and researchers through stimulating methods of presentation of archival materials and historical events. Users can therefore travel through space and time without moving from their home [27]. Numerous AR/VR applications exist for CH or tourists enjoyment of places with a rich past, allowing a realistic navigation of environments that no longer exist or that may be inaccessible, [2,6,11,14,25].

The most of the previously cited experiences do not take into account the advantages that may arise by a proper exploitation of sounds together with visual effects, except for audio content presentation purposes. On the contrary, an acoustic guide, properly placed in the virtual space, may drive the user toward a certain direction or the acoustic landscape of a specific historical period could be reproduced to improve the virtual experience. To the authors' knowledge only a few experimentations have been carried on in this context. One of the few exceptions is given by the work in [17], where authors present a signal processing method for fast real-time binaural synthesis, whose main target application is the fruition of cultural heritage, and the work in [8] where a smart headphones set is presented that remotely takes the orientation of the listener's head and properly generates an audio output to attract the tourists'attention toward specific points of interest in the 3D space. An interesting analysis of hardware and software requirements for this purpose is presented in [20] without references to real applications.

In this paper we present a preliminary study on the possible models for binaural spatialization, already available in literature, that could be exploited for AR/VR applications for CH, taking into account the kind of devices involved in this context which do not offer a high definition audio experience. We are aware of existing solutions offered by Google (i.e. Resonance Audio, [10]) or the Oculus Unity Spatializer, [21] by Facebook and the Steam Audio Unity Plugin [28], but in this initial part of our research we decided to build our own solution in order

to precisely control all the involved parameters given the specific application and devices we are targeting. The paper is organized as in the following. The analyzed methods are summarized in Sect. 2, Sect. 3 reports the implementation of the previously described models, subsequently tests and results are discussed in Sect. 4, while conclusions are drawn in Sect. 5.

2   Analyzed Models

This section briefly describes two models that have been analyzed and tested to achieve binaural spatialization for AR/VR applications. The two models have been chosen due to their different characteristics to achieve the spatialization. The first one, i.e. the anthropometric model, is able to achieve the spatialization by adapting some parameters directly to the listener, so it is able to take into account the differences between human parts. The second one, the minimum phase representation model which is basically an approximation of the HRTFs, is useful to investigate which characteristics of the HRTFs are most relevant to locate a sound source in the space.

2.1   Anthropometric Models

During the previous years researches have focused on how human body parts can affect the acoustic experience and which of them are the most influencing ones. Models coming from these studies are called *anthropometric models*. They are based on several geometric approximations, e.g. a sphere for the head, and they also exploit particular manikins, known as KEMAR (Knowles Electronic Manikin for Acoustic Research), because they allow the use of in-ear microphones to measure the Head Related Transfer Functions (HRTF), i.e. the representation of the ways in which the positions of the head, trunk, and ears filter sounds, altering the way they are perceived. In this paper we mainly focus on the analysis and implementation of models related to the head and the pinna, namely the outer part of the ear.

It is well known that the variation in magnitude of the audio signal due to the head can be represented by a low-pass filter, as mentioned by Lord Rayleigh in [26].

One of the parameter that the human auditory system (HAS) uses to locate sounds in the space is the Interaural Time Difference (ITD). This parameter is given by the difference of the arrival times of a sound wave to the left and the right ear. By simply manipulating this parameter, it is possible to place a sound source in a virtual environment.

However, such parameter allows to locate sound sources only on a horizontal plane.

The ITD is included in the developed anthropometric model by using an all-pass filter with group delay given by the Woodworth-Schlosberg formula:

$$\tau_h(\theta) = \begin{cases} -\frac{a}{c}\cos\theta & \text{if} \quad 0 \le |\theta| < \pi/2 \\ \frac{a}{c}\big(|\theta| - \pi/2\big) & \text{if} \quad \pi/2 \le |\theta| < \pi \end{cases} \tag{1}$$

where θ represents the azimuth coordinate, a the head radius and c the speed of sound.

Considering human ears as two "observation points", the formula has to be used once for the right ear and once for the left one in order to take into account the Interaural Time Difference (ITD).

As previously said, the ITD allows sound sources positioning on a horizontal plane. It has been shown that the influence of the pinnas affects the positioning of sound sources on a vertical plane.

The pinna is mainly responsible for multiple reflections of the incoming wave [31] and, according to the procedure described in [4], the pinna effects can be modeled with five reflections by means of a tapped delay line.

2.2   Minimum Phase Representation

Several studies have focused on relevant characteristics of the HRTF for the localization task, [16,30].

In this paper we refer to the work in [16], regarding the sensitivity of humans to the variations of the phase spectra of the HRTFs where HRTFs were approximated by a minimum phase transfer function including also a factor capable of incorporating the ITD.

This approximation is justified from the fact that every system can be decomposed in a product of a minimum-phase system and an all-pass system [22]:

$$H(e^{j\omega}) = H_{min}(e^{j\omega})H_{ap}(e^{j\omega}) \tag{2}$$

The HRTF can thus be expressed as follows:

$$H(e^{j\omega}) = |H_{min}(e^{j\omega})|e^{j\phi(\omega)}e^{-j\tau} \tag{3}$$

where $|H_{min}(e^{j\omega})|$ is the magnitude of the HRTF, $e^{j\phi(\omega)}$ represents the phase response of the minimum phase transfer function (TF) and $e^{-j\tau}$ is the all-pass function able to model the ITD, as explained below. The ITD is thus a delay line, since it is an all-pass function, to be applied to the lagging ear, [22].

Since in every causal system the real and the imaginary parts of the frequency response are related to each other by the Hilbert transform, for minimum-phase systems, it is possible to state that the *cepstrum*, the sequence associated to the complex logarithm of the frequency response, is causal, and so the real and the imaginary part of the logarithm of the frequency response, which correspond to the magnitude and the phase respectively, are related to each other [22].

Due to the minimum-phase representation, the ITD cue has to be reintroduced as a constant delay. In this study the ITD has been extracted from a set of HRTFs taken by the ARI (Acoustics Research Institute) database through the Interaural Cross-Correlation (IACC) method, [9,15], where the ITD is interpreted as the delay which maximizes the similitude between the Head Related Impulse Responses (HRIRs) of the right and the left ears, i.e. the delay at which the cross-correlation is maximum.

3   Models Implementation

In this section details on the implementation of the previously described models are given.

3.1   Anthropometric Model

In order to model the low-pass effect of the head, we referred to the single pole-single zero function modelled by [4]. The transfer function of the IIR filter representing the head behaviour and taking into account both the angle of incidence θ of the acoustic wave and the head radius (a), is given in [31]:

$$H_{\mathrm{HS}}(z) = \frac{(\omega_0 + F_s\alpha) + (\omega_0 - F_s\alpha)z^{-1}}{(\omega_0 + F_s) + (\omega_0 - F_s)z^{-1}} \qquad (4)$$

where $\omega_0 = c/a$, c is the speed of sound ($343\,\mathrm{m/s}$ circa), α is a function of the azimuth angle given by, [4]:

$$\alpha(\theta) = \left(1 + \frac{\alpha_{min}}{2}\right) + \left(1 - \frac{\alpha_{min}}{2}\right)\cos\left(180°\frac{\theta}{\theta_{min}}\right) \qquad (5)$$

The ITD is represented as a constant delay on all frequencies through an allpass filter given in 6, where τ is the group delay defined in Eq. 1, [31]:

$$a = \frac{1 - \tau}{1 + \tau} \qquad (6)$$

Concerning the model of the pinna, the choice of coefficients has been done empirically. Indeed from previous studies, pinna effects can be found in the first 0.7 ms of the HRIRs, [4], thus, using a sampling frequency of 44.1 KHz, in the first 32 samples. This implies that a 32 taps FIR filter is enough for pinna modeling. Moreover, as previously said, five reflections are enough representative of the pinna effect, therefore parameters describing these reflections have to be properly chosen. First of all the intensity of the reflections is assumed to be independent on the source direction, thus the corresponding parameters are constant. For each reflection the characterizing parameters are reported in Table 1 where ρ is the reflection coefficient which is different for each reflection and A_n, B_n, D_n are experimentally derived for 3 different persons as described in [31].

It is worth noting that two different columns characterize the parameter D_n since it allows the adjustment of the model to the individual characteristics of the pinna and it indeed results to be different for one person of the 3 involved for the derivation of this table. The delay of each path is instead dependent on the audio source direction as follows, [4]:

$$\tau_{pn}(\phi,\theta) = A_n \cos(\theta/2)\sin[D_n(90 - \phi)] + B_n \qquad (7)$$

where coefficients values can be obtained from Table 1.

Table 1. Coefficients table for Eq. 7

n	ρ_{pn}	A_n	B_n	D_n for PB & NH	D_n for RD
2	0.5	1	2	1	0.85
3	−1	5	4	0.5	0.35
4	0.5	5	7	0.5	0.35
5	−0.25	5	11	0.5	0.35
6	0.25	5	13	0.5	0.35

Table 2. Values of τ_{pn} for every event of the pinna with θ equals to $0°, 15°, 30°, 45°$ and $60°$. The generic value it has been calculated using the values of Table 1

n	0°	15°	30°	45°	60°
2	3	7.53	10.53	14.53	16.53
3	2.99	7.50	10.50	14.50	16.50
4	2.96	7.41	10.41	14.41	16.41
5	2.92	7.26	10.26	14.26	16.26
6	2.86	7.06	10.06	14.06	16.06

In the current paper only one column for D_n has been taken into account because there are no information related to the "original" listeners and the persons involved in the experimentation reported in Sect. 4 are physically very different.

The last aspect to be discussed is the relation between the FIR filter coefficients and the coefficients of the pinna anthropometric model. By the analysis of τ_{pn} it has been shown that the delay associated to the different pinna events were almost constant as the incident angle of the horizontal plane was varying, as it is possible to see in Table 2. It is thus reasonable to assume a mean value for each delay associated to a generic event. The delays obtained with this assumption were approximated to the upper and lower integer numbers of samples.

The values of the reflection coefficients were also interpolated since the number of reflections was less than the number of delays. Finally, due to these assumptions, also the extreme values of the reflection coefficients were adjusted since they were not defined.

3.2   Minimum Phase Representation

For the derivation of the minimum phase response, the real cepstrum solution has been exploited, as described in [24]. Then, in order to reintroduce the ITD, the cross correlation method has been used. This method returns the number of samples by which the sequence, representing the audio signal, must be shifted in order to take into account the ITD.

4 Tests and Results

The previously discussed models have been tested only on 4 adult individuals with different anthropometric characteristics. The subjects were 2 males and 2 females with no experience in listening tests and with no hearing problems. The age varied within 25 up to 70 years.

The audio source for both administered tests was a 4000 ms periodic pseudo-random sequence. The generation of the sequence and all the processings have been done by using MATLAB. All the listeners have used a pair of Marshall Major headphones to hear the processed sequences through a PC with a Realtek High Definition Audio on-board. The type of administered tests has required a differentiation, as described below.

4.1 Anthropometric Model Test

The anthropometric model has been tested over 27 different audio source positions on each person. In order to compare the performances of the tested model, the same tests were performed by using also 2 sets of non-individualized HRTFs per person, taken from the ARI database.

So, the tested positions were the ones in which the HRTFs were measured. It is worth noting that the original discretization between audio source positions was equal to $15°$ but preliminary tests showed that users were not sensitive to this tight distances, so a distance of $30°$ was chosen. The positions involved are reported in Fig. 1.

Since the anthropometric model can be adapted to anthropometric features of the listeners, for each of them the model was tested by exploiting both an average head radius of 8 cm and a personalized head radius (approximated on each listener's head). This is due to the fact that one of the main purposes of this experimentation is to understand how much a personalized model influences the correct perception of the source location when medium quality devices are used.

Each listener were asked to point on a grid pattern the perceived audio source position. This procedure was repeated for each position testing both the two HRTFs and the two anthropometric models. Results are discussed in Sect. 4.3.

4.2 Minimum Phase Model Test

Concerning the minimum phase model, the Four Interval - Two Alternative Forced Choice (4I-2AFC) has been chosen, [3], which consists of administrating to each individual 4 sequences, within which only 1 is different from the others and asking them to point out the different sequence [16].

In our case, the sequences were the audio source, a pseudo-random noise, filtered by a HRTF taken from the ARI database or its minimum-phase approximation. In particular the different sequence was the one filtered with the approximation, while the other 3 identical sequences were given by filtering the noise through a non personalized HRTF.

With this approach, the experimentation is successful if the 50% circa of the responses is correct because this would mean that the two models are actually indistinguishable and thus the minimum phase solution is properly approximating the HRTF.

4.3   Anthropometric Models Results

Results for the anthropometric model are reported in Table 3 where in a single cell the average percentage of error for the model involved and for a specific direction is reported, as it is evaluated only for the azimuth coordinate. Having a look at these results there is not a particular model that behaves better than others.

Results show a not optimal outcome since the typical problems associated to the use of non personalized HRTF, i.e. front/back reversal, arise also in the case of the anthropometric model. The most of the problems arise when θ is equal to 0°, 30° and 60° that are often perceived as 180°, 150° and 120° degrees respectively.

The less problematic directions were 90° and 270°. Indeed these are those directions for which the binaural parameters do not show ambiguities [7], thus rarely reporting "big" mistakes. The elevation perception which is not here reported, has shown very good results. A particular trend that is worth to discuss is that performance improve as the elevation increases. This result show a proper behaviour of this model for localization of the height of an audio source. While this behaviour of the elevation perception is in line with the results reported in [4], the same cannot be stated for the horizontal perception, indeed the authors in [4] state that the horizontal angle perception did not show any problem thus avoiding to report results. It is possible that these errors are due to the non personalized coefficients for the pinna model (i.e. D_n in Eq. 7), future works will be devoted also to understand this aspect.

Other data can be derived from Figs. 2 and 3, where the abscissa represents the horizontal angle of the emitted audio source, while the ordinate axis reports the angle perceived by the listener. Each set of plots is referred to the antropometric model with a minimum radius for the head and the anthropometric model with personalized head radius. The angles on each plot are parametrized with respect to a certain elevation angle, pointed as ϕ. For some elevation values a few points have been tested given the points chosen for the HRTFs. This point of view allows to notice that there is not a particular model that behaves better than others in terms of perceived localization, moreover, the diversity of choices made by the listeners is even more evident.

4.4   Minimum Phase Representation Results

As hinted before, these tests were based on the 4I-2AFC paradigm. Results have reported that equally perceived sequences were the ones generated by the non personalized HRTF, while the one perceived to be different was generated through minimum phase HRTF with external ITD. Results have thus shown a

100% percentage of correct responses, which is the opposite of the 50% expected for stating the correctness of the solutions. This is not a failure because the differences perceived were mainly related to the quality of the sound and no significant differences in the sources localization were reported.

Both the previously used HRTFs and the 27 positions have been exploited to exploit the minimum phase model through the 4I-2AFC paradigm. Each person was asked to explain which sound was different and why. The most of the reasons were related to a slightly different perceived position. In particular the sound filtered through the minimum phase representation appeared to be higher than the one reproduced through the reference HRTF. A reversal was perceived in only one case i.e. when the correct position was placed at $\theta = 90°$ and $\phi = 80°$ and was instead perceived at $\theta = 240°$. Nevertheless, with respect to the previous results it is possible to state that the minimum phase model introduces improvements.

Table 3. Test results for the anthropometric model. The first column represents the position tested indicated as (ϕ,θ), where ϕ is the elevation and θ is the azimuth.

Positions	HRTF n1	HRTF n2	Average radius	Adapted radius
(0, 0)	58%	31%	46%	48%
(0, 30)	19%	15%	23%	23%
(0, 60)	21%	6%	8%	19%
(0, 90)	4%	4%	8%	10%
(0, 120)	4%	10%	17%	8%
(0, 150)	0%	27%	27%	23%
(0, 180)	6%	21%	17%	17%
(0, 210)	10%	8%	8%	6%
(0, 240)	6%	2%	2%	2%
(0, 270)	6%	6%	8%	10%
(0, 300)	10%	15%	6%	8%
(0, 330)	12%	19%	17%	8%
(30, 0)	13%	25%	35%	65%
(30, 30)	17%	13%	15%	19%
(30, 90)	8%	10%	4%	4%
(30, 150)	17%	15%	4%	15%
(30, 180)	6%	38%	23%	25%
(30, 270)	8%	6%	19%	15%
(60, 0)	25%	23%	50%	27%
(60, 30)	17%	21%	10%	8%
(60, 90)	10%	8%	4%	13%
(60, 180)	15%	17%	0%	4%
(60, 270)	15%	8%	6%	8%
(80, 30)	19%	23%	25%	13%
(80, 90)	10%	8%	4%	6%
(80, 180)	13%	27%	0%	2%
(80, 270)	8%	8%	2%	8%
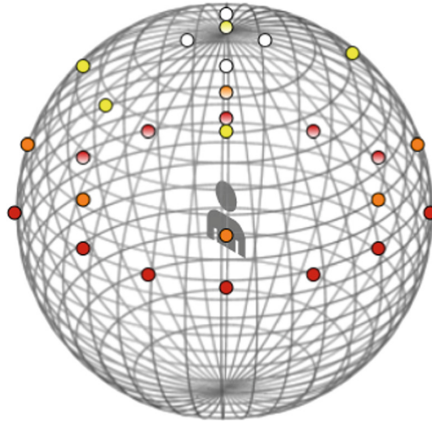
Fig. 1. Each colour represents a specific elevation, red is for 0, orange is for 30, yellow is for 60 and white is for 80. (Color figure online)
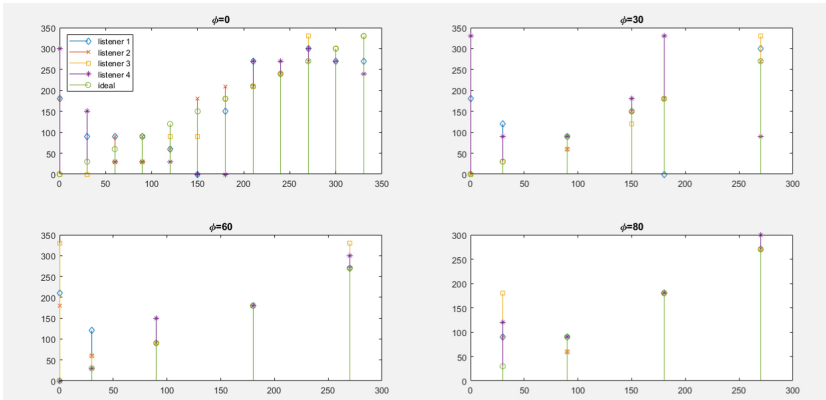
Fig. 2. Results related to the anthropometric model with average head radius

The validity of such approximation was already tested by Kulkarni et al., [16], in a bigger work regarding the sensitivity of humans to variations of the phase spectra of HRTFs, that is why the 4I-2AFC paradigm has been taken into consideration in this work. They obtained the minimum phase reconstruction by a set of non individualized HRTFs, just as presented here, so it seems reasonable to conclude that this model is not valid over these HRTFs. The problems related to the use of non individualized HRTFs are well-known so it would be interesting to repeat this study with individualized ones.
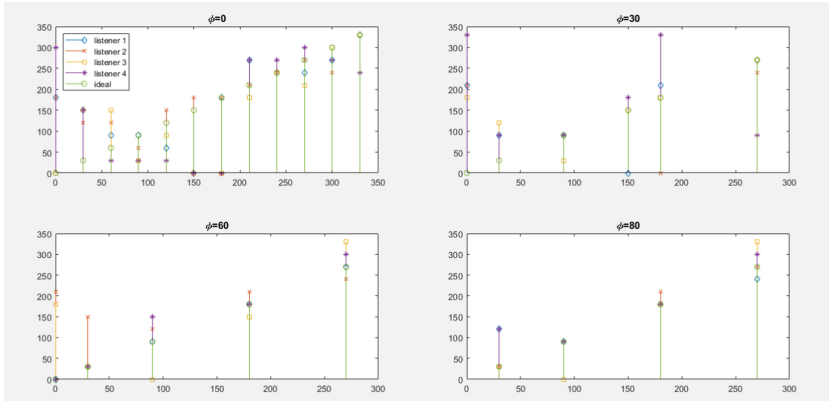
Fig. 3. Results related to the anthropometric model with listener adapted head radius

5   Conclusions and Future Works

This work dealt with the analysis of binaural spatialization in order to support AR/VR applications with particular reference to Cultural Heritage Enhancement services. This work presented the techniques able to support those kind of applications in terms of audio experience, taking into account the need of adapting to low-cost, low-complexity devices that may be exploited in the described scenario. Two models have been studied, implemented and tested in order to avoid the use of non individualized HRTFs in order to obtain binaural spatialization. Results obtained for the antropometric model demonstrate a proper behaviour of the solution when conceiving with the elevation perception, while the azimuth information perception suffered from the typical problems associated to the exploitation of non personalized HRTF such has front/back reversal. This is in contrast with previous studies, thus suggesting a possible change of used coefficients to be adapted to the listener.

Concerning the minimum phase solution, listeners were always able to distinguish the different sound source, but only in terms of its timbre, while the spatial position of the source was not significant for discrimination. Reference parameters properly derived by a set of HRTF directly obtained from the listener, would help in obtaining more reliable solutions and this is part of future works.

Future works are also devoted to solve the problem of front/back reversal [12], and to take into account the issue of the externalization. It is well-known that the only use of the HRTFs lacks on information about the external environment, so the sound is perceived as inside the head. The use of some reverberation in this cases is very helpful.

We have only considered sound sources at a fixed distance without varying it, so the problem of sources closer than 1 m, where the binaural parameters have a fundamental role [5], has to be taken into account as well. A sample demon-

stration system is under development in order to setup a validation scenario for audio spatialization supporting AR/VR applications using Google's Resonance Audio open-source library [10].

References

1. Addison, A.C.: Emerging trends in virtual heritage. IEEE Multimedia **7**(2), 22–25 (2000)
2. Bernardini, F., Rushmeier, H., Martin, I.M., Mittleman, J., Taubin, G.: Building a digital model of Michelangelo's Florentine Pieta. IEEE Comput. Graphics Appl. **22**(1), 59–67 (2002)
3. Bi, J., Kuesten, C.: The four-interval, two-alternative forced-choice (4I2AFC): a powerful sensory discrimination method to detect small, directional changes particularly suitable for visual or manual evaluations. Food Qual. Prefer. **73**, 202–209 (2019)
4. Brown, C.P., Duda, R.O.: A structural model for binaural sound synthesis. IEEE Trans. Speech Audio Process. **6**(5), 476–488 (1998)
5. Brungart, D.S., Rabinowitz, W.M.: Auditory localization of nearby sources. Head-related transfer functions. J. Acoust. Soc. Am. **106**(3), 1465–1479 (1999)
6. Brusaporci, S., Graziosi, F., Franchi, F., Maiezza, P., Tata, A.: Mixed reality experiences for the historical storytelling of cultural heritage. In: Bolognesi, C., Villa, D. (eds.) From Building Information Modelling to Mixed Reality. STCE, pp. 33–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-49278-6_3
7. Carlile, S.: The physical and psychophysical basis of sound localization. In: Virtual Auditory Space: Generation and Applications, pp. 27–78. Springer (1996). https://doi.org/10.1007/978-3-662-22594-3_2
8. D'Auria, D., Di Mauro, D., Calandra, D.M., Cutugno, F.: A 3D audio augmented reality system for a cultural heritage management and fruition. J. Digit. Inf. Manage. **13**(4) (2015)
9. Estrella, J.: On the extraction of interaural time differences from binaural room impulse responses. Master's thesis, Technische Universität Berlin (2010)
10. Gorzel, M., et al.: Efficient encoding and decoding of binaural sound with resonance audio. In: Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio. Audio Engineering Society (2019)
11. Guidi, G., et al.: 3D digitization of a large model of imperial Rome. In: Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM 2005), pp. 565–572. IEEE (2005)
12. Gupta, N., Barreto, A., Ordonez, C.: Improving sound spatialization by modifying head-related transfer functions to emulate protruding pinnae. In: Proceedings IEEE SoutheastCon 2002 (Cat. No. 02CH37283), pp. 446–450. IEEE (2002)
13. Härmä, A., et al.: Augmented reality audio for mobile and wearable appliances. J. Audio Eng. Soc. **52**(6), 618–639 (2004)
14. Ikeuchi, K., Nakazawa, A., Hasegawa, K., Oishi, T.: The great buddha project: modeling cultural heritage for VR systems through observation. In: ISMAR, vol. 3, pp. 7–16 (2003)
15. Katz, B.F., Noisternig, M.: A comparative study of interaural time delay estimation methods. J. Acoust. Soc. Am. **135**(6), 3530–3540 (2014)
16. Kulkarni, A., Isabelle, S., Colburn, H.: Sensitivity of human subjects to head-related transfer-function phase spectra. J. Acoust. Soc. Am. **105**(5), 2821–2840 (1999)

17. Lapini, A., Calamai, G., Argenti, F., Carfagni, M.: Application of binaural audio techniques for immersive fruition of cultural heritage. IOP Conf. Ser. Mater. Sci. Eng. **364**, 012099 (2018). https://doi.org/10.1088/1757-899X/364/1/012099

18. Madirov, E., Absalyamova, S.: The influence of information technologies on the availability of cultural heritage. Procedia. Soc. Behav. Sci. **188**, 255–258 (2015)

19. Malpas, J.: Cultural heritage in the age of new media. In: Kalay, Y.E., Kvan, T., Affleck, J. (eds.) New Heritage: New Media and Cultural Heritage, Abingdon (2008)

20. Murphy, D., Neff, F.: Spatial sound for computer games and virtual reality. In: Game Sound Technology and Player Interaction: Concepts and Developments, pp. 287–312. IGI Global (2011)

21. Oculus: Explore the oculus unity spatializer with the sample scene. https:// developer.oculus.com/documentation/unity/audio-osp-unity-scene/. Accessed 15 Jan 2021

22. Oppenheim, A.V., Buck, J.R., Schafer, R.W.: Discrete-Time Signal Processing, vol. 2. Prentice Hall, Upper Saddle River (2001)

23. Palombini, A.: Storytelling and telling history. Towards a grammar of narratives for cultural heritage dissemination in the digital era. J. Cult. Heritage **24**, 134–139 (2017)

24. Pei, S.C., Lin, H.S.: Minimum-phase FIR filter design using real cepstrum. IEEE Trans. Circuits Syst. II Express Briefs **53**(10), 1113–1117 (2006)

25. Pietroni, E., Pagano, A., Rufa, C.: The Etruscanning project: gesture-based interaction and user experience in the virtual reconstruction of the Regolini-Galassi tomb. In: 2013 Digital Heritage International Congress (DigitalHeritage), vol. 2, pp. 653–660. IEEE (2013)

26. Rayleigh, L., Lodge, A.: Iv. on the acoustic shadow of a sphere. Philos. Trans. R. Soc. London Ser. A, Contain. Papers Math. Phys. Charact. **203**(359–371), 87–110 (1904)

27. Roussou, M., Efraimoglou, D.: High-end interactive media in the museum. In: International Conference on Computer Graphics and Interactive Techniques: ACM SIGGRAPH 1999 Conference Abstracts and Applications, vol. 8, pp. 59–62 (1999)

28. Software, V.: Steam audio unity plugin. https://valvesoftware.github.io/steam-audio/doc/phonon_unity.html. Accessed 15 Jan 2021

29. Thwaites, H.: Digital heritage: what happens when we digitize everything? In: Ch'ng, E., Gaffney, V., Chapman, H. (eds.) Visual Heritage in the Digital Age. SSCC, pp. 327–348. Springer, London (2013). https://doi.org/10.1007/978-1-4471-5535-5_17

30. Zagala, F., Noisternig, M., Katz, B.F.: Comparison of direct and indirect perceptual head-related transfer function selection methods. J. Acoust. Soc. Am. **147**(5), 3376–3389 (2020)

31. Zölzer, U.: DAFX: Digital Audio Effects. Wiley, New York (2011)

Enhancement of Gravity Centrality Measure Based on Local Clustering Method by Identifying Influential Nodes in Social Networks

Pham Van Duong[1], Xuan Truong Dinh[2], Le Hoang Son[3], and Pham Van Hai[1(✉)]

[1] School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, Vietnam
`haipv@soict.edu.vn`
[2] Posts and Telecommunications Institute of Technology, Hanoi, Vietnam
[3] Information Technology Institute, Vietnam National University, Hanoi, Vietnam
`sonlh@vnu.edu.vn`

Abstract. Identifying influential nodes has great theoretical and practical implications in real-world scenarios such as search engines, social networks, and recommendation systems. Among the most essential issues in the field of complicated networks. Many approaches have been developed and deployed that have proven to be as effective as the gravity model. However, these models only focus on the local information of the node and ignore the information about the node neighbors or the global information of the network, leading to the gravity model is not really effective. This study focuses on improving the gravity model by considering the position information of the node based on the improvement of the k-shell decomposition algorithm. In addition, the article also uses the link of the node's neighbors by the local neighbor coefficient to increase the rigor for the local information of the node. The paper applies the SIR model to simulate the propagation effect of the node, then uses the Kendall Tau coefficient to evaluate the efficiency between the list of influence rankings. This research applies the monotonicity ratio to evaluate the resolution of the proposed ranking list. The efficiency of the recommended method is proven to outperform other methods on 5 social network datasets.

Keywords: Influential node · Node ranking · Gravity model · Local clustering · K-shell · SIR model · Kendall's Tau

1 Introduction

The ideas of complex networks have attracted people's attention due to their wide application in the real-world [1], as well as their infinity to biology systems social systems, and multi-agent systems, etc. Network science plays an extremely key role in many fields. As one of the topics of network science research, the identification of influential nodes has been widely discussed in recent years. Influential spreaders identification may be widely

employed in several fields and applications, like disease analysis, rumor analysis, knowledge graph [2], social computing [3], information propagation, community detection [4] and so on. A typical real-life example is in a product promotion campaign on social networks, identifying the most influential users will help effect of helping the promoted product reach the most people. There are many proposed and implemented studies for determining the influence node. There are two main approaches: node centrality method and network embedding method.

Network embedding is an approach for the progressively large network information, the first problem is the way to represent the network structure information a lot of efficiently. The traditional network illustration methodology rep-resented by the adjacency matrix has the characteristics of high dimensions and information scantness, which typically means the procedure complexness and the computing cost square measure high in large networks. With the event and wide application of network illustration learning technology in language process and alternative fields, researchers have turned to explore a sort of low-dimensional and dense vector representation methodology to represent a high-dimensional network, which is termed network embedding. As one of the pioneers of the network embedding algorithms, DeepWalk [5] uses Word2vec [6] to generate low-dimensional representation vectors of nodes using random walk sequences. Additionally, another algorithm based on network embedding can be mentioned as Node2vec, CANE and SDNE.

Node centrality is the most intuitive method for identifying a node's influence in the network. To solve the matter of a way to effectively determine the influence of nodes within the network, researchers have proposed several methods, usually in step with the central score of nodes within the network to rank the nodes, the higher the score, the larger the influence. Degree centrality (DC) is the simplest method to determine node influence, this metric focuses only on the number of neighbors of the node without considering the global information of the network, which leads to incomplete results [7]. Closeness centrality (CC) and Betweenness centrality (BC), these two measures take into account the global information of the network, BC measures the influence of a node by calculating the number of shortest paths through it while CC computes the average of the short distances between that node and other nodes in the network [8]. However, these two measures are quite complicated to calculate, BC and CC are not suitable for large networks [8]. An existing commonly used approach to identify the influential nodes is PageRank (PR) which not only considers the number of neighbors, but also considers how influential that neighborhood is. Hence, it works well in directed graphs but does not suit for undirected graphs [9]. The K-shell decomposition algorithm considers the node's position in the network by determining the node ratio to find out the influence of the node [10]. Because of its low algorithmic complexity, it is suitable for large networks but this algorithm does not distinguish between different influences of nodes in the same shell layer. In addition, researched and proposed algorithms such as eigenvector centrality (EC) [1], HITS [11] to improve node ranking, however, these algorithms have not been appreciated for networks with tight links. Recently, a few more effective approaches for determining the typical influence node are the gravity model, which considers both the influence information of the neighborhood and the information about the connectivity in the graph [12]. Even so, this algorithm is not suitable for large

networks due to the algorithm's complexity as well as being only interested in local information of the node. Therefore, for the problem of determining the affected node, it is necessary to consider both local and global information of the network.

In this paper, to solve the above problem, we propose the algorithm DKGM_CLC based on the improvement of the degree centrality and the k-shell decomposition method. This model considers both the influence information of the node using the improved K-shell metric combined with the Coefficient Local Centrality (CLC) metric and the restriction on the path information. To analyze the efficiency of the algorithm, two types of models that we use to calculate the ranking results of the node influence are SIR and the correlation Kendall Tau coefficient.

The rest of this paper is organized as follows. We present in Sect. 2 the overviews of the related work. In Sect. 2.1, we discuss the centrality measures algorithm and introduce performances metrics in Sect. 2.2, and Sect. 3 is about the proposal method whereas its performance evolution is discussed in Sect. 4. Finally, we elaborate conclusion and future recommendations in Sect. 5.

2   Backgrounds

Given a network $G = <V, E>$ with G is an undirected and unweighted network, where V and E represent nodes and edges. We denote $A = (a_{ij})_{N \times N}$ is the adjacency matrix of G. If there is an edge between node i and node j then $a_{ij} = 1$, otherwise $a_{ij} = 0$.

2.1   Centrality Measures

Degree Centrality (DC). [7] is defined as the number of edges occurring on a node is known as the number of edges node. $DC(i)$ of node i can be calculated by:

$$DC(i) = k(i) = \sum_j a_{ij} \tag{1}$$

where $k(i)$ is the degree of node i.

Closeness Centrality (CC). [8] is defined as the average length of the shortest path from it to the other nodes in the network. $CC(i)$ of node i can be calculated as follows:

$$CC(i) = \frac{1}{\sum_j^N d(i, j)} \tag{2}$$

where $d(i, j)$ represents the shortest path from node i to node j.

Betweenness Centrality (BC). [8] is defined as the number of shortest paths through it. $BC(i)$ of node i can be calculated by the following equation:

$$BC(i) = \sum_{s \neq i, s \neq t, i \neq t} \frac{g_{st}(i)}{g_{st}} \tag{3}$$

where g_{st} represents the number of shortest paths from node s to node t and $g_{st}(i)$ is the number of shortest path through node i from node s to node t.

K-shell Decomposition: One of the limits of centrality-based methodologies is the obliviousness of node position. Even if the degree of effect is small, a node at the network's core position has a lot of influences. Consider this point of view, [10] uses k-shell decomposition to determine where the nodes in the network are. The peripheral nodes on the outside layers are stripped away, whereas nodes in the core layer have a lot of influence. This approach can be considered as a node degree-based coarse graining sorting algorithm. The following is the specific decomposition procedure: The initial stage in KS is to remove all nodes in the network with a degree of 1 from the network. Then, after one round of removal, it removes nodes with a degree of $k \leq 1$ since this step may cause the degree values to be reduced throughout the removal process. All nodes deleted in this stage generate 1-shell and their k-shell values are equal to one until there are no nodes in the network with degree $k \leq 1$. Then repeat the process to get two shells, three shells, and so on. Finally, all nodes are separated into distinct shells, and each node's k-shell value may be calculated.

Gravity Centrality (GC). The gravity centrality model [12] has a similar structure to Isaac Newton's universal gravitation formula. The results of the k-shell values divided by the shortest path lengths between the two nodes is the gravity metrics.

$$GC(i) = \sum_{j \in \omega_i} \frac{k_s(i)k_s(j)}{d^2(i,j)} \tag{4}$$

where $k_s(i)$ is the degree of node i, $d(i,j)$ is the shortest path distance between node i and node j; ω_i is the local nodes whose distance to node i is less than or equal to a given value r.

PageRank (PR). [9] is an iterative method for determining the importance of a node. $PC(i)$ of node i can be determined as follows:

$$PR(i)^q = \sum_{j=1}^{n} \left(a_{ij} \frac{PC(j)^{q-1}}{k_j} \right) \tag{5}$$

where $PR(i)^q$ represents centrality score of node i in step q.

GGM. Based on the gravity model, the generalized gravity model (GGM) [13] measures local information from both the local clustering coefficient and the degree of each node. $GGM(i)$ of node i can be determined using the equation below:

$$GGM(i) = \sum_{d(i,j) \leq R} \frac{Sp(i) \times Sp(j)}{d^2(i,j)} \tag{6}$$

$$Sp(v) = e^{-\propto C_v} \times DC(v)$$

where $Sp(v)$ is the spreading ability of node v, where $\alpha \geq 0$ and $k(v)$ is the degree of node v. α is a free parameter that can be modified flexibly real application. In this experiment, we choose $\alpha = 2$ and $R =< d > /2$.

DKGM. is a high-resolution index combining both degree centrality and the k-shell decomposition method. DKGM [14] is based on this proposed index and the well-known gravity law. *DKGM* (*i*) of node *i* can be calculated as follows:

$$DKGM(i) = \sum_{d(i,j)\leq R, i\neq j} \frac{DK(i) \times DK(j)}{d^2(i,j)} \tag{7}$$

$$DK(v) = DC(v) + KS(v) + \frac{p(v)}{maxq(k) + 1} \tag{8}$$

where during the process of k-shell decomposition for the k-degree iteration, the total number of stages is $q(k)$, and node v is removed in the $p(v)$ stage and $R = < d > /2$.

Local Clustering Coefficient: Other measurements can be computed for each node and used to create a distribution function, but the most common is degree. A node's clustering coefficient [15] estimates how many vertices will more often than not be grouped together. There are two types of clustering coefficients, global and local. In this paper, we use the local clustering coefficient. Specifically, given a vertex v and và $d_v = |N_v|$ its degree. The local clustering coefficient C_v of node v can be calculated as follows:

$$C_v = \frac{2.\left|\left\{v`, v''\right\} \in E(G) : v`, v'' \in N_v\right|}{d_v(d_v - 1)} \tag{9}$$

2.2 Performances Metrics

Susceptible Infected Recovered (SIR) Model and Susceptible Infected (SI) Model
The susceptible infected recovered (SIR) model is a classic infectious disease model that can also be used to abstractly represent information transmission. The SIR model categorizes the population into three groups: the susceptible (S), the infected (I), and the recovered (R). The SIR model assumes all nodes to be susceptible (S) at first, with the exception of the source node, which is infected (I). Each infected node has a possibility of infecting its susceptible neighbors with probability β. Furthermore, at each time step, each infected node recovers to become a recovered node with the probability μ. The infection process will continue until there are no more infected nodes. [16] Among them, $S(t)$ is the number of nodes, which are susceptible to the disease at time t, $I(t)$ denotes the number of those infected at time t, and $R(t)$ denotes the number of the recovered at time t. The influence of node I could be calculated using the formula:

$$P(i) = \frac{R(t^*)}{N} \tag{10}$$

where $R(t^*)$ is the number of recovered nodes when the dynamic process achieves steady-state and N is the total number of nodes in the network. For simplicity, μ chose to 1 and the corresponding epidemic threshold is:

$$\beta_c \approx \frac{< k >}{< k^2 > - < k >} \tag{11}$$

The susceptible infected (SI) model is a special case of the susceptible infected (SIR) model. The SI model is based on the SIR model and assumes that the recovery rate $\mu = 0$, meaning that once a node has been infected, it cannot be recovered. The number of experiments is K. $F(t)$ represents the average number of infected nodes in the SI model at time t.

KENdall's Tau or Kendall correlation coefficient [17] is useful and significant metric for determining the linear correlation between two sequences. The Kendall coefficient is a number that ranges from 0 to 1. The closer the Kendall coefficient's absolute value is to 1, the higher the linear correlation between two sequences. Two sequences having a Kendall coefficient of 0 can be considered non-linear. $X = (x_1, x_2, \ldots, x_N)$ and $Y = (y_1, y_2, \ldots, y_N)$ are two sequences with N elements. For any pair of two-tuples (x_i, x_j) and (y_i, y_j) $(i \neq j)$, if $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i > y_j$, the pair is concordant. If xi > xj and yi < yj or xi < xj and yi > yj, the pair is inconsistent. If xi = xj or yi = yj. Kendall's Tau of X and Y can be defined as:

$$\tau = \frac{2(n_+ - n_-)}{N(N-1)} \tag{12}$$

where n_+ and n_- are the number of consistent pairs and inconsistent pairs, respectively, and N is the total number of pairs.

Monotonicity Relation. A monotonicity index $M(R)$[18] for a ranking list R is used to quantify the resolution of different ranking algorithms. Monotonicity relation could be calculated as follows:

$$M(R) = \left[1 - \frac{\sum_{r \in R} N_r(N_r - 1)}{N(N-1)} \right]^2 \tag{13}$$

where N is the size of the ranking vector R, and N_r denotes the number of nodes that have the same rank index value r. This metric measures the percentage of rank nodes in the ranking list that are the same. $M(R) \in [0, 1]$. If $M(R) = 1$, the ranking algorithm is perfectly monotonic, and each node is categorized using a different index value. Otherwise, $M(R) = 0$ puts all nodes in the same rank. For rank list R, a higher M value indicates more variation and uniformity.

3   Methodologies

The proposed algorithm will consider combining measures from the local clustering coefficient, the vertex of each node, and the k-shell decomposition improvement to measure the importance of each node in the network and this index considers as mass in GC model.

In terms of the basic k-shell decomposition approach, take node 2 and node 3 as an example in this Fig. 1 compared with node 2, node 3 is nearer to the center of the network, therefore node 3 could also be additional conducive to propagation. However, we can not distinguish the two nodes by the on top of proposed methodology. Though
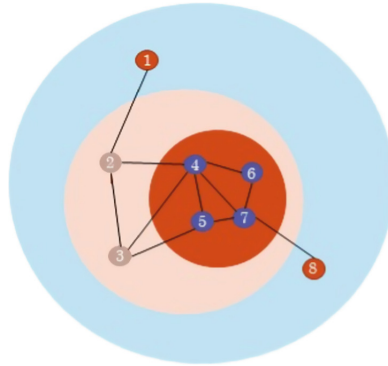
Fig. 1. An eight nodes network to illustrate the resolution for K-shell decomposition

each node a pair of and node 3 square measure within the 2-shell, node 3 is removed later than node 2, that is, the 2-shell decomposition process includes two stages, node 2 is removed within the initial stage and node 3 is removed within the second stage. Therefore, we tend to introduce the stage number at that the node is removed from the network whereas performing the k-shell decomposition. Based on this point of view, [26] proposed he improved k-shell index of node i, denoted by $k_s^*(i)$, can be calculated by:

$$k_s^*(i) = k_s(i) + \frac{p(i)}{q(k) + 1} \tag{14}$$

where $p(i)$ is the stage in a k-shell layer decay for which node i is removed from the graph, $q(k)$ is the set of the number of stages in the k-shell decay.

To increase the "tightness" of the local information of the node, we will consider the information of the neighboring nodes, and we will apply the node-local center measure with the local neighborhood coefficient [19]. This index is calculated using the formula:

$$CLC(i) = f(c(i)) \times LC(i) \tag{15}$$

$$LC(i) = \sum_{u \in \Gamma(i)} Q(u) \tag{16}$$

$$Q(u) = \sum_{w \in \Gamma(u)} N(w) \tag{17}$$

where $\Gamma(i)$ is the set of nearest neighbors to node i, $N(w)$ is the number of nearest and nearest neighbors of node w. The function $f(c(i))$ is proved by the formula:

$$f(c(i)) = e^{-c(i)} \tag{18}$$

where $c(i)$ is the local clustering coefficient of node i.

Combining the above two measures, we build a new measure using the formula:

$$DK\_CLC(i) = k_s^*(i) + CLC(i) \tag{19}$$

Use this index DK_CLC as mass in GC, Hence the influence of node i can be estimated as follows:

$$DKGM\_CLC(i) = \sum_{j \neq i, d(i,j) \leq R} \frac{DK\_CLC(i)DK\_CLC(j)}{d^2(i,j)} \tag{20}$$

where $R = \, <d> \, /2$.

In addition, to evaluate the proposed algorithm more effectively than other methods, we will change the node location information by changing the CLC index with the PageRank, DC indexes combined with the local neighborhood index:

$$DKGM\_PR(i) = \sum_{j \neq i, d(i,j) \leq R} \frac{DK\_PR(i)DK\_PR(j)}{d^2(i,j)} \tag{21}$$

where $DK\_PR(i) = k_s^*(i) + PR(i)$

$$DKGM\_DKC(i) = \sum_{j \neq i, d(i,j) \leq R} \frac{DK\_DKC(i)DK\_DKC(j)}{d^2(i,j)} \tag{22}$$

where α is a coefficient that can be arbitrarily changed depending on the problem. We choose $\alpha = 2$ and $R = \, <d> \, /2$.

4   Experimental Results

In this paper, we use 5 data networks to evaluate the model and select a few evaluation indicators to evaluate the effectiveness of our proposed model. Data networks for experiment include: PB, Facebook, Jazz, NS, USAir, Email. Table 1 describes some basic characteristics of the above 5 datasets.

Table 1. Basic features of 5 experimental networks

Networks	N	M	$<k>$	$<d>$	C	r	H	β_c
PB	1222	16,714	27.3552	2.7375	0.3600	− 0.2213	2.9707	0.0125
Facebook	4039	88,234	46.6910	3.6925	0.6170	0.0636	2.4392	0.0095
Jazz	198	2742	27.6970	2.2350	0.6334	0.0202	1.3951	0.0266
NS	379	914	8.8232	4.0419	0.7981	− 0.0817	1.6630	0.1424
Email	1133	5451	9.6222	3.6060	0.2540	0.0782	1.9421	0.0565

To evaluate the ability and effectiveness of the proposed model, we use the SIR model and the Kendall's Tau index to evaluate the similarity between the list of influence node proposed by the algorithm and the simulation. For each experimental network, the paper will use 1000 independent tests with infection probability β in the SIR model to simulate and average the results obtained. Then use this result to come up with a standard ranking

Table 2. The accuracy of the algorithms is calculated by the Kendall Tau index with $\beta = \beta_c$

Networks	DC	BC	CC	KS	GC	GGM	DKGM	DKGM_DKC	DKGM_PR	DKGM_CLC
PB	0.87040	0.68725	0.78647	0.88418	0.89865	0.81829	0.89994	0.87495	**0.90208**	0.90080
Facebook	0.68478	0.45659	0.39622	0.71886	0.78829	0.70466	0.78539	0.77730	0.78826	**0.81766**
Jazz	0.8247	0.46501	0.71140	0.79951	0.86897	0.78381	0.89475	0.89092	0.87492	**0.89854**
NS	0.61644	0.39688	0.34435	0.56266	0.80952	0.58582	0.78790	0.77913	0.77384	**0.81802**
Email	0.79012	0.63486	0.81300	0.81424	0.87474	0.77602	0.86270	0.84169	0.87546	**0.88489**

list of the nodes in the graph. The accuracy of the algorithm is determined by Kendall's Tau algorithm between the list of standard ratings and the list of ratings suggested by the algorithm. Table 2 describes the accuracy of the proposed model and the studied models with $\beta = \beta_c$.

With the results obtained in the table above, compared with other algorithms, the proposed model has a relatively high efficiency. For the Facebook, Jazz, NS, Email dataset, the accuracy of DKGM_CLC algorithm is the largest. For the remaining dataset, PB, the accuracy is also relatively high, just behind the DKGM_PR algorithm. It can be seen that the proposed model DKGM_CLC is more effective than the dataset with many links and closely related neighbors.
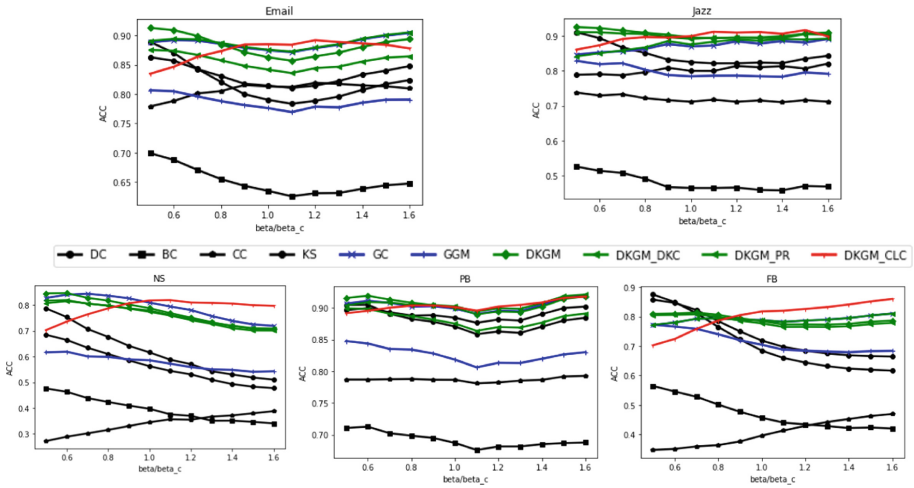
Fig. 2. Accuracy of algorithms by Kendall's Tau model with different β values

With each different network will have different strong structure. Therefore, when we choose the probability of infection β to determine the list of standard influence nodes with the SIR model, the value of β of each different network may be different. Therefore, to generalize the model and increase its accuracy, the paper examines the results of random $\beta(\beta = k\beta_c)$ with k from 0.5 to 1.5. Figure 2 shows the Kendall's Tau index of the proposed model and other models with random $\beta(\beta = k\beta_c)$. The performance results show that the proposed model has the greatest efficiency (red line) with β in the near vicinity of β_c. This effect shows that the proposed method is really effective. The black line represents the accuracy of the algorithms using the basic scale (DC, KS, CC, BC), the blue line represents the accuracy of the improved algorithms based on the GM model, the blue line The construction leaves show the algorithms that improve from the paper and the red line shows the accuracy of the proposed algorithm.
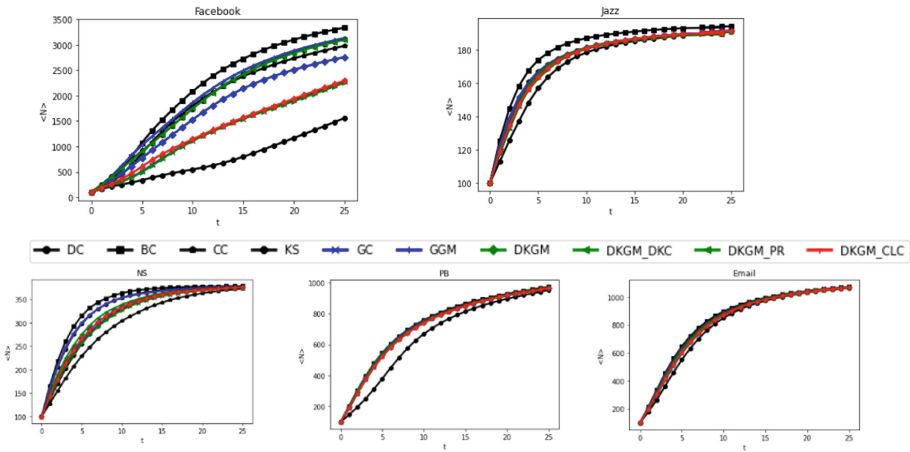
Fig. 3. Comparison of the spreading scale S(t) as a function of infected time t of nine methods on eight networks

Figure 3 describes the number of affected nodes in each time point with different models using the SI model. At baseline, the probability of infection for the SIR model was $\beta = \beta_c$, and $\mu = 0$. A list of 100 most influential nodes is selected for each method and SI model is applied to determine the number of influential nodes at each time point. It shows the influence of 100 proposed influence nodes with the proposed model being similar to other models, even with the Email and PB dataset at time $15 < t < 25$, a number of nodes affected with the proposed model is the largest. It is found that, because this dataset has a large number of links, the nodes are closely linked, so the proposed model is quite effective.

Table 3 shows the monotonicity relation for the ranking list proposed by different algorithms. The algorithm proposed by this paper has better results than the other methods.

Table 3. The accuracy of the algorithm is calculated by the monotonicity index

Networks	DC	BC	CC	KS	GC	GGM	DKGM	DKGM_DKC	DKGM_PR	DKGM_CLC
PB	0.9328	0.9064	0.9489	0.9980	0.9993	0.9993	0.9993	**0.9996**	0.9993	0.9994
Facebook	0.9739	0.9855	0.9967	0.9419	0.9998	**0.9999**	**0.9999**	**0.9999**	**0.9999**	**0.9999**
Jazz	0.9659	0.9885	0.9878	0.7944	0.9994	0.9993	**0.9995**	0.9994	0.9994	**0.9995**
NS	0.7642	0.3389	0.9928	0.6421	0.9949	0.9955	0.9955	0.9955	**0.9957**	0.9952
Email	0.8874	0.9400	0.9988	0.8088	**0.9999**	**0.9999**	**0.9999**	**0.9999**	**0.9999**	**0.9999**

5   Conclusions and Future Work

This paper has presented an DKGM_CLC model based on the improvement of the degree centrality and the k-shell decomposition method. In complex networks, this approach efficiently ranks and quantifies the influential spreaders. We employed three evaluation metrics (SIR, Kendall τ and monotonicity relation) to verify its efficiency. The proposed DKGM_CLC identifies the difference in node influence better than standard centrality approaches such as DC, BC, CC, KS, GC and some the other improved algorithms on GC like DKGM, GGM, DKGM_DKC, DKGM_PR, according to experimental results conducted on five real networks. Experimental results demonstrated that our algorithm performed better than the benchmarks. The proposed DKGM_CLC algorithm can be developed to several forms for better outcomes in the future. For instance, our future work will consider the edge weight, properties of node in the network and base on networking embeded to improve the performance. Additionally, we improve DKGM_CLC to identify the node spreading influence in large-scale dynamic networks.

References

1. Gao, Y.-C., Fu, C.-J., Cai, S.-M., Yang, C., Eugene Stanley, H.: Repulsive synchronization in complex networks. Chaos **29**, 053130 (2019)
2. Van Pham, H., Tien, D.N.: Hybrid louvain-clustering model using knowledge graph for improvement of clustering user's behavior on social networks. In: Tran, D.-T., Jeon, G., Nguyen, T.D.L., Lu, J., Xuan, T.-D. (eds.) ICISN 2021. LNNS, vol. 243, pp. 126–133. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-2094-2_16
3. Dinh, X.T., Van Pham, H.: Social network analysis based on combining probabilistic models with graph deep learning. In: Sharma, H., Gupta, M.K., Tomar, G.S., Lipo, W. (eds.) Communication and Intelligent Systems. LNNS, vol. 204, pp. 975–986. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-1089-9_76
4. Zheng, Z., Ye, F., Li, R.H., Ling, G., Jin, T.: Finding weighted k-truss communities in large networks. Inf. Sci. **417**, 344–360 (2017)
5. Dinh, X.T., Van Pham, H.: A proposal of deep learning model for classifying user interests on social networks. In: Proceedings of the 4th International Conference on Machine Learning and Soft Computing, pp. 10–14 (2020)
6. Perozzi, B., Al-Rfou, R., Skiena, S.S.: Deepwalk: online learning of social representations. In: Proceedings of the ACM Sigkdd International Conference on Knowledge Discovery & Data Mining (2014)
7. Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. **26**, 3111–3119 (2013)
8. Grover, A., Leskovec, J.: Node2vec: Scalable feature learning for networks. In: Proceedings of the ACM Sigkdd International Conference on Knowledge Discovery & Data Mining (2016)
9. Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. J. Math. Sociol. **2**(1), 113–120 (1972)
10. Sabidussi, G.: Thee centrality index of a graph. Psychometrika **31**(4), 581–603 (1966)

11. Brin, S., Page, L.: Reprint of: the anatomy of a large-scale hypertextual web search engine. Comput. Netw. **56**(18), 3825–3833 (2012)
12. Li, Z., Ren, T., Ma, X., Liu, S., Zhang, Y., Zhou, T.: Identifying influential spreaders by gravity model. Sci. Rep. **9**(1), 1–7 (2019)
13. Li, H., Shang, Q., Deng, Y.: A generalized gravity model for influential spreaders identification in complex networks. Chaos, Solitons Fractals **143**, 110456 (2021)
14. Li, Z., Huang, X.: Identifying influential spreaders in complex networks by an improved gravity model. Sci. Rep. **11**(1), 1–10 (2021). https://doi.org/10.1038/s41598-021-01218-1
15. Eguiluz, V.M., Klemm, K.: Epidemic threshold in structured scale-free networks. Phys. Rev. Lett. **89**(10), 108701 (2002)
16. Shulgin, B., Stone, L., Agur, Z.: Pulse vaccination strategy in the SIR epidemic model. Bull. Math. Biol. **60**(6), 1123–1148 (1998)
17. Xiao-Ping, S., Yu-Rong, S.: Leveraging neighborhood "structural holes" to identifying key spreaders in social networks. Acta Physica Sinica **64**(2) (2015)
18. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry, 35–41 (1977)
19. Zhao, X., et al.: Evaluating influential nodes in social networks by local centrality with a coefficient. ISPRS Int. J. Geoinf. **6**(2), 35 (2017)

Deep Factorized Multi-view Hashing
for Image Retrieval

Chenyang Zhu, Wenjue He, and Zheng Zhang$^{(\boxtimes)}$

Harbin Institute of Technology, Shenzhen 518055, China
`darrenzz219@gmail.com`

Abstract. Multi-view hashing has been paid much attention to due to its computational efficiency and lower memory overhead in similarity measurement between instances. However, a common drawback of these multi-view hashing methods is the lack of ability to fully explore the underlying correlations between different views, which hinders them from producing more discriminative hash codes. In our work, we propose the principled Deep Factorized Multi-view Hashing (DFMH) framework, including interpretable robust representation learning, multi-view fusion learning, and flexible semantic feature learning, to deal with the challenging multi-view hashing problem. Specifically, instead of directly projecting the features to a common representation space, we construct an adaptively weighted deep factorized structure to preserve the heterogeneity between different views. Furthermore, the visual space and semantic space are interactively learned to form a reliable hamming space. Particularly, the flexible semantic representation is obtained by learning regressively from semantic labels. Importantly, a well-designed learning strategy is developed to optimize the objective function efficiently. DFMH as well as compared methods is tested on benchmark datasets to validate the efficiency and effectiveness of our proposed method. The source codes of this paper are released at: https://github.com/chenyangzhu1/DFMH.

Keywords: Multi-view hashing · Deep factorization · Image retrieval · Learning to hash

1 Introduction

With the increasing usage of big data and multimedia, there has been rising need for large-scale data analytic methods, especially methods measuring the similarities or distances between items, with low computational complexity and memory usage. Hashing provides an elegant solution to this challenging problem by representing the original high-dimensional features as binary codes which preserve the original similarity of features.

While some early methods [4,5,7] use fixed hash functions to obtain the binary codes, most of the recent methods [1,3,10,22] implement a learning-to-hash strategy, which borrows the ideas from machine learning and deep learning to obtain more flexible hash codes. These methods are no longer independent of data, and are expected to yield better results. More specifically, there are two types of learning-to-hash methods, *i.e.*, unsupervised and supervised, according to whether semantic information is used. Among them, supervised methods [12, 16] are usually preferred to unsupervised ones [15,24] since they explore the semantic information and thus can provide a more promising result.

However, all of the aforementioned methods are designed for single-view situations, which cannot solve the real-world problems that usually appear in a multi-view form. For instance, multiple features, such as HOG, GIST, and LBP, can be extracted from a picture, every feature is taken as a view. An event can be reported in different forms including texts, photos, and videos, and each of the forms is considered a view. The multi-view circumstances usually contain more information than the summation of the information within every single view, since connections between views provide additional information. Unfortunately, existing single-view algorithms are not able to utilize such complementary information, resulting in unsatisfying clustering results. Therefore, a number of multi-view hashing methods are put forward to explore the heterogeneous information under the multiple views. Utilizing non-negative matrix factorization (NMF), [8] fuses information from multiple views and drops the useless and even noisy ones. [20] solves the near-duplicate video retrieval (NDVR) problems by exploring information from multiple feature types extracted from video keyframes, and learns hash codes and the hash functions simultaneously to deal with the large-scale data. [17] measures the local similarities as well as the semantic similarities by a Locally Linear Embedding (LLE) based method.

Although there has been some encouraging progress in multi-view hashing, most existing methods still encounter many challenges, which prohibit them from providing accurate and discriminative hash codes. First, the conventional multi-view hashing always learns the common representations from multiple views based on feature projecting from the original feature space, which lacks enough interpretability for robust representation learning. Moreover, the underlying characteristics across multiple views are missing in common feature learning. Furthermore, the complementary property of multi-view learning is under-explored to generate a unified representation. Additionally, most of the existing works employ fixed discrete labels for classification or regression in hash code learning, which lacks enough flexibility in semantic knowledge discovery.

To deal with the drawbacks of existing methods mentioned above, a novel Deep Factorized Multi-view Hashing (DFMH) method is proposed to produce hash codes which not only preserve the latent characteristics between views, but also capture the categorical information from the regressive semantic labels (Fig. 1). Specifically speaking, we implement a novel deep factorized structure to learn a consensus representation space instead of using the traditional factorize method, and employ an orthogonal rotation strategy to convert the real values

into the hash codes. In this way, more characteristics, especially the complementary information between views, are incorporated in the learned hash codes. Furthermore, adaptively learned weights are applied to the real-value learning process to automatically make the trade-off between different views, making sure that informative views contribute more to the representation space and the noisier ones contribute less. Additionally, learning from semantic labels, instead of directly using fixed labels, ensures that the flexibility is preserved and underlying information is fully explored.

The main contributions of the DFMH method are listed as following:

1. We propose a novel multi-view hashing method to capture the hidden characteristics from multi-view data for discriminative multi-view information retrieval. In the method, deep factorized common representation learning, adaptively-weighted multi-view fusion, and flexible semantic feature learning are jointly considered in one unified learning framework.
2. An adaptively weighted deep factorized learning scheme is designed to explore the heterogeneous properties of multi-view data, and the consensus representations are further transformed to the discrete hash space based on an orthogonal rotation strategy.
3. The flexible semantic representation is learned based on the regressive semantic labels, which can guide the intrinsic semantic transfer in joint binary code learning.
4. Extensive experiments validate the effectiveness of the proposed method in multi-view fusion and semantic-preserving multi-view hashing compared with multiple state-of-the-art multi-view hashing methods.

2   Related Work

Multi-view hashing is an efficient method for multimedia retrieval. Unlike single-view methods which focus on one view only, multi-view hashing not only utilizes information within each individual view but also employs the complementary information between views to improve its hashing performance. Existing multi-view hashing methods can be classified into two types: those do not use labelled information, *i.e.*, unsupervised methods, and those take advantage of labelled information *i.e.*, supervised methods.

Unsupervised methods are intended to generate hash codes and hash functions independent of semantic labels. For example, [20] generates hash codes under individual structure as well as global structure. In [18], Shen *et al.* generate hash codes using matrix factorization methods with an adaptively learned kernel space. Going one step further, Lu *et al.* [13] propose a method that has no additional parameter for the regularization term to shorten training time, making it possible to deal with the retrieval of large-scale multi-view data. Although encouraging improvement has been made, hash codes generated without the supervision of semantic labels are usually not discriminative enough.

Contrarily, information underlying in the semantic labels has been proven very useful in discovering similarities between views, and thus many supervised

multi-view methods are proposed to make full use of it. For instance, [11] designs a model which combines different kernels of different features to generate hash codes. In [23], Yang *et al.* propose a novel method to explore the local as well as semantic similarity and optimizes the resulting objective function without relaxing the constraint. However, in most of the existing methods, different views share the same weight, while the real-world case is that some views usually contain more important information than others.
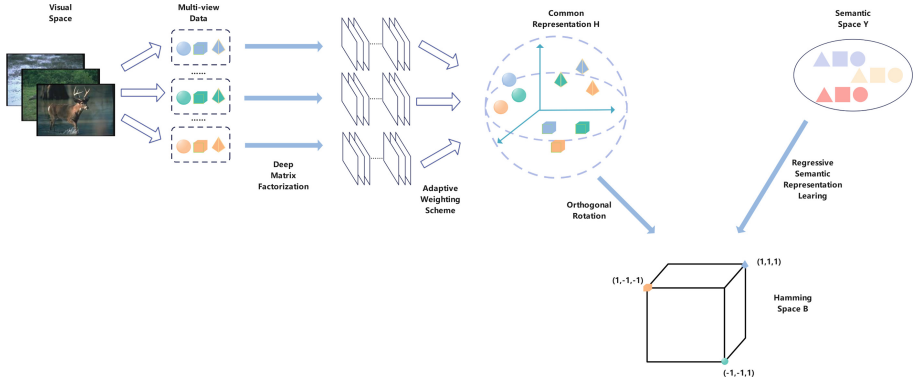
Fig. 1. The general structure of our DFMH method. The learning framework is composed of a deep matrix factorization scheme to extract the features from complex structures and an adaptive weighting scheme to balance each view. Moreover, our framework adopts orthogonal rotation strategy and regressive semantic representation learning to fuse visual and semantic information together in Hamming space.

3   The Proposed Method

3.1   Notation and Problem Definition

In this paper, matrices are represented by upper case letters such as $\boldsymbol{X}$ and vectors are represented by lowercase letters, such as $\boldsymbol{\alpha}$. The multi-view data training set is denoted as $\mathcal{X} = \left\{ \boldsymbol{X}^{(v)} \right\}_{v=1}^{m}$, where $\boldsymbol{X}^{(v)} = [x_1^{(v)}, \cdots, x_n^{(v)}] \in \Re^{d_v \times n}$ is the feature matrix of the v-th view, d_v is the feature dimension of v-th view, and n is the number of instances. Based on this, each x_i has a corresponding label vector $y_i \in \{0,1\}^c$ indicating its class, where c is the class number and

$$y_{ij} = \begin{cases} 1 & \text{if } x_i \text{ is an instance of the j-th class} \\ 0 & \text{otherwise} \end{cases}. \quad \boldsymbol{Y} = [y_1, \cdots, y_n] \in \Re^{c \times n} \text{ is the}$$

class label matrix. $\boldsymbol{B} = [b_1, \cdots, b_n] \in \{-1, 1\}^{l \times n}$ is the learned hash code, where l is the hash code length. DFMH aims at learning the hash function $\mathcal{H} : \mathcal{X} \rightarrow \boldsymbol{B}$, which enables the proposed method to return a series of binary codes which preserves the similarity of original instances, *i.e.*, the more similarity instances share, the closer the distance between the learned hash codes of these instances is.

3.2   The Proposed Deep Factorized Multi-view Hashing Method

Unlike most of the aforementioned methods which usually use a shallow learning method, deep matrix factorization is employed to learn a consensus multi-view representation $\boldsymbol{H} \in \Re^{l \times n}$, where l is the length of the hash code.

Inspired by [21], we conduct matrix factorization on $\boldsymbol{X}^{(v)}$, which writes $\boldsymbol{X}^{(v)} \approx \boldsymbol{U}_1^{(v)} \boldsymbol{H}_1^{(v)}$, so that the latent representation of the first layer is obtained. After that, the first layer is represented as $\boldsymbol{H}_1^{(v)} = \boldsymbol{U}_2^{(v)} \boldsymbol{H}_2^{(v)}$ in the same manner. Such process is repeated for $(m-1)$ times to get $\boldsymbol{H}_{m-1}^{(v)}$.

The factorization process can be listed as following:

$$
\begin{aligned}
\boldsymbol{X}^{(v)} &\approx \boldsymbol{U}_1^{(v)} \boldsymbol{H}_1^{(v)} \\
\boldsymbol{X}^{(v)} &\approx \boldsymbol{U}_1^{(v)} \boldsymbol{U}_2^{(v)} \boldsymbol{H}_2^{(v)} \\
&\cdots \\
\boldsymbol{X}^{(v)} &\approx \boldsymbol{U}_1^{(v)} \boldsymbol{U}_2^{(v)} \cdots \boldsymbol{U}_{m-1}^{(v)} \boldsymbol{H}_{m-1}^{(v)} \\
&\text{s.t. } \boldsymbol{H}_i^{(v)} \geq 0.
\end{aligned}
\tag{1}
$$

The last layer is specially treated in order to learn the final consensus multi-view representation $\boldsymbol{H}$:

$$
\begin{aligned}
\boldsymbol{H}_{m-1}^{(v)} &= \boldsymbol{U}_m^{(v)} \boldsymbol{H}, v = 1, 2, \cdots, m \\
&\text{s.t. } \boldsymbol{H} \geq 0.
\end{aligned}
\tag{2}
$$

As such, the common representation learning process can be expressed in the following equation:

$$
\min_{\boldsymbol{U}_i^{(v)}, \boldsymbol{H}} \sum_{v=1}^m \left\| \boldsymbol{X}^{(v)} - \boldsymbol{U}_1^{(v)} \boldsymbol{U}_2^{(v)} \cdots \boldsymbol{U}_m^{(v)} \boldsymbol{H} \right\|_F^2 \quad \text{s.t. } \boldsymbol{H} \geq 0, \tag{3}
$$

where $\boldsymbol{U}_i^{(v)} \in \Re^{p_{i-1} \times p_i}$ is the layer basis matrix of the i-th layer and v-th view, $\boldsymbol{H}_m^{(v)}$ denotes the latent representation of the m-th layer and v-th view, and p_i represents the dimension of the i-th layer with $p_0 = d_v$, $p_m = r$.

Since the physical meanings of features vary from one to another, different weights should be assigned to different views to explore as much information in provided data as possible. To balance the information between views, an adaptive weighting scheme is proposed, which writes:

$$
\min_{\boldsymbol{U}_i^{(v)}, \boldsymbol{H}, \alpha^{(v)}} \sum_{v=1}^m \left(\alpha^{(v)} \right)^r \left\| \boldsymbol{X}^{(v)} - \boldsymbol{U}_1^{(v)} \boldsymbol{U}_2^{(v)} \cdots \boldsymbol{U}_m^{(v)} \boldsymbol{H} \right\|_F^2
$$
$$
\text{s.t. } \sum_v \alpha^{(v)} = 1, 0 < \alpha^{(v)} < 1, \boldsymbol{H} \geq 0, \tag{4}
$$

where $\alpha^{(v)}$ is the adaptive weighting parameter of the v-th view. r controls the distribution of the weight, and is usually set to 3 or 5.

The space rotation strategy is introduced to make sure that the real-valued common representation space $\boldsymbol{H}$ and the binary code space $\boldsymbol{B}$ are of enough similarity:

$$
\min_{\boldsymbol{U}_i^{(v)},\boldsymbol{H},\alpha^{(v)},\boldsymbol{B},\boldsymbol{R}} \sum_{v=1}^{m} \left(\alpha^{(v)}\right)^r \left\|\boldsymbol{X}^{(v)} - \boldsymbol{U}_1^{(v)}\boldsymbol{U}_2^{(v)}...\boldsymbol{U}_m^{(v)}\boldsymbol{H}\right\|_F^2
$$
$$
+ \lambda \left\|\boldsymbol{B} - \boldsymbol{R}\boldsymbol{H}\right\|_F^2 \tag{5}
$$
$$
\text{s.t. } \sum_v \alpha^{(v)} = 1, 0 < \alpha^{(v)} < 1, \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}, \boldsymbol{H} \geq 0,
$$

where λ is a balance parameter and $\boldsymbol{R}$ is the orthogonal rotation matrix.

Notably, hash code generated from the Eq. (5) only contain the visual information, while the semantic information which is also significant is ignored. Therefore, a regressive semantic representation learning step is introduced to extract the flexible semantics from the fixed one-hot binary class label matrix $\boldsymbol{Y}$:

$$
\min_{\boldsymbol{U}_i^{(v)},\boldsymbol{H},\alpha^{(v)},\boldsymbol{B},\boldsymbol{R},\boldsymbol{W}} \sum_{v=1}^{m} \left(\alpha^{(v)}\right)^r \left\|\boldsymbol{X}^{(v)} - \boldsymbol{U}_1^{(v)}\boldsymbol{U}_2^{(v)}...\boldsymbol{U}_m^{(v)}\boldsymbol{H}\right\|_F^2
$$
$$
+ \lambda \left\|\boldsymbol{B} - \boldsymbol{R}\boldsymbol{H}\right\|_F^2 + \beta \left\|\boldsymbol{B} - \boldsymbol{W}\boldsymbol{Y}\right\|_F^2 \tag{6}
$$
$$
\text{s.t. } \sum_v \alpha^{(v)} = 1, 0 < \alpha^{(v)} < 1, \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}, \boldsymbol{H} \geq 0,
$$

where β is a trade-off parameter.

Adding a fourth term to avoid trivial solutions of regressive semantic encoding, the overall objective function in DFMH is shown below:

$$
\min_{\boldsymbol{U}_i^{(v)},\boldsymbol{H},\alpha^{(v)},\boldsymbol{B},\boldsymbol{R},\boldsymbol{W}} \sum_{v=1}^{m} \left(\alpha^{(v)}\right)^r \left\|\boldsymbol{X}^{(v)} - \boldsymbol{U}_1^{(v)}\boldsymbol{U}_2^{(v)}...\boldsymbol{U}_m^{(v)}\boldsymbol{H}\right\|_F^2
$$
$$
+ \lambda \left\|\boldsymbol{B} - \boldsymbol{R}\boldsymbol{H}\right\|_F^2 + \beta \left\|\boldsymbol{B} - \boldsymbol{W}\boldsymbol{Y}\right\|_F^2
$$
$$
+ \gamma \left\|\boldsymbol{W}\right\|_F^2 \tag{7}
$$
$$
\text{s.t. } \sum_v \alpha^{(v)} = 1, 0 < \alpha^{(v)} < 1, \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}, \boldsymbol{H} \geq 0.
$$

The first term of this function learns a common representation $\boldsymbol{H}$, the second term minimizes the gap between $\boldsymbol{H}$ and the binary code space $\boldsymbol{B}$, and the third and forth term obtain flexible semantics from the class label matrix $\boldsymbol{Y}$.

4 Optimization

Optimizing problem Eq. (7) is a non-convex problem whose closed-form solution can hardly be achieved. Therefore, a feasible iterative optimization strategy is employed to address the problem Eq. (7).

Step 1: Set other variables as constant values to update $U_i^{(v)}$. The subproblem for $U_i^{(v)}$ refers to:

$$C = \left\| X^{(v)} - \phi U_i^{(v)} H_i^{(v)} \right\|_F^2 \quad \text{s.t. } H_i^{(v)} \geq 0. \tag{8}$$

where $\phi = U_1^{(v)} U_2^{(v)} \cdots U_{i-1}^{(v)}$. Setting $\frac{\partial C}{\partial U_i^{(v)}} = 0$, $U_{(i)}^v$ can be expressed in the following form:

$$U_i^{(v)} = \phi^\dagger X^{(v)} H_i^{(v)\dagger} \tag{9}$$

where $X^\dagger$ means the pseudo inverse of matrix X.

Step 2: Set other variables as constant values to update $H_i^{(v)}$: The optimization of $U_i^{(v)}$ requires an up-to-date $H_i^{(v)}$, so here we update $H_i^{(v)}$ though it does not appear in the final objective function. The subproblem for $H_i^{(v)}$ refers to:

$$C = \| X^{(v)} - \Phi H_i^{(v)} \|_F^2 \quad \text{s.t. } H_i^{(v)} \geq 0. \tag{10}$$

where $\Phi = U_1^{(v)} U_2^{(v)} \cdots U_i^{(v)}$. Following [2], $H_i^{(v)} (i < m)$ is updated with the following equation:

$$H_i^{(v)} = H_i^{(v)} \odot \sqrt{\frac{[\Phi^\top X^{(v)}]^+ + [\Phi^\top \Phi H_i^{(v)}]^-}{[\Phi^\top X^{(v)}]^- + [\Phi^\top \Phi H_i^{(v)}]^+}}, \tag{11}$$

where $[A]^+ = (|A| + A)/2, [A]^- = (|A| - A)/2$.

Step 3: Set other variables as constant values to update W. The subproblem for W refers to:

$$C = \beta \| B - WY \|_F^2 + \gamma \| W \|_F^2 \tag{12}$$

Setting $\frac{\partial C}{\partial W} = 0$, the following solution for W is obtained:

$$W = \beta B Y^\top (\beta Y Y^\top + \gamma I)^{-1} \tag{13}$$

Step 4: Set other variables as constant values to update R. The subproblem for R refers to:

$$\max_{R^\top R = I} tr(R^\top B H^\top) \\ \text{s.t. } R^\top R = I, H \geq 0. \tag{14}$$

According to [9], the solution for R is:

$$R = P_R Q_R^\top \tag{15}$$

where P_R is a matrix constructed of the left-singular vectors of $BH^\top$, and Q_R is a matrix constructed of the right-singluar vectors of $BH^\top$.

Step 5: Set other variables as constant values to update H. The subproblem for H refers to:

$$
\begin{aligned}
C = \sum_{v=1}^{m} \left(\alpha^{(v)} \right)^r & \left\| X^{(v)} - U_1^{(v)} U_2^{(v)} \cdots U_m^{(v)} H \right\|_F^2 \\
& + \lambda \left\| B - RH \right\|_F^2 \\
\text{s.t.} \sum_v \alpha^{(v)} = & 1, 0 < \alpha^{(v)} < 1, R^\top R = I, H \ge 0.
\end{aligned}
\tag{16}
$$

By setting $\frac{\partial C}{\partial H} = 0$, the following solution can be obtained:

$$
H = (\sum_{v=1}^{m} (\alpha^{(v)})^r Z^\top Z + \lambda R^\top R)^{-1} (\sum_{v=1}^{m} (\alpha^{(v)})^r Z^\top X + \lambda R^\top B)
\tag{17}
$$

where $Z = U_1^{(v)} U_2^{(v)} \cdots U_m^{(v)}$.

Step 6: Set other variables as constant values to update B. The subproblem for B refers to:

$$
\max_{B \in \{-1,1\}} tr(B^\top (2\lambda RH + 2\beta WY))
\tag{18}
$$
$$
\text{s.t.} \ R^\top R = I, H \ge 0.
$$

The solution for B is:

$$
B = sgn(2\lambda RH + 2\beta WY).
\tag{19}
$$

Step 7: Set other variables as constant values to update $\alpha^{(v)}$. For convenience, we set $h_v = \left\| X^{(v)} - U_1^{(v)} U_2^{(v)} \cdots U_m^{(v)} H \right\|_F^2$. The subproblem for α refers to:

$$
\min_{\alpha} \sum_{v=1}^{m} (\alpha^{(v)})^r h_v
\tag{20}
$$
$$
\text{s.t.} \sum_v \alpha^{(v)} = 1, 0 < \alpha^{(v)} < 1.
$$

We can then construct the Lagrange function as following:

$$
L(\alpha, \varepsilon) = \sum_{v=1}^{m} (\alpha^{(v)})^r h_v - \varepsilon (\sum_{v=1}^{m} \alpha^{(v)} - 1),
\tag{21}
$$

then we need

$$
\nabla_{\alpha^{(v)}} L = r(\alpha^{(v)})^{r-1} h_v - \varepsilon = 0
\tag{22}
$$

$$
\nabla_\varepsilon L = \sum_{v=1}^{m} \alpha^{(v)} - 1 = 0,
\tag{23}
$$

where $(v = 1, 2 \cdots m)$. Combining Eq. (22) and Eq. (23), we have:

$$\sum_{v=1}^{m} (\frac{\varepsilon}{r h_v})^{\frac{1}{r-1}} = 1, \tag{24}$$

which can be reformulated as

$$(\frac{\varepsilon}{r})^{\frac{1}{r-1}} = \frac{1}{\sum_{v=1}^{m}(\frac{1}{h_v})^{\frac{1}{r-1}}}. \tag{25}$$

Thus, Eq. (25) can be transformed into:

$$\alpha^{(v)} = \frac{(\frac{1}{h_v})^{\frac{1}{r-1}}}{\sum_{v=1}^{m}(\frac{1}{h_v})^{\frac{1}{r-1}}}. \tag{26}$$

Notably, the adaptive weight $\alpha^{(v)}$ is determined by different multi-view data $\mathcal{X}$.

The proposed optimization algorithm is run for several times until its convergence. According to our experiments, the number of iterations needed is usually less than 10. The whole learning process of DFMH is concluded in Algorithm 1.

Algorithm 1. DFMH

Input: $\mathcal{X}$: set of multi-view data; $\boldsymbol{Y}$: Label set; λ, β, γ: parameters; h: length of binary codes; T: maximum number of iterations.
Output: $\alpha^{(v)}$: Adaptive weight parameters; $\boldsymbol{U}_1^{(v)}\boldsymbol{U}_2^{(v)} \cdots \boldsymbol{U}_m^{(v)}$: layer basis matrices; $\boldsymbol{B}$: ideal binary code.
 1: Initialize $\boldsymbol{H}$, $\boldsymbol{U}_i^{(v)}$, $\boldsymbol{B}$, $\boldsymbol{R}$ by random matrix.
 2: **for** Not Convergence or k=1:T **do**
 3:    Update $\boldsymbol{U}_i^{(v)}$ via solving Eq. (9)
 4:    Update $\boldsymbol{H}_i^{(v)}(i < m)$ via solving Eq. (11)
 5:    Update $\boldsymbol{W}$ via solving Eq. (13)
 6:    Update $\boldsymbol{R}$ via solving Eq. (15)
 7:    Update $\boldsymbol{H}$ via solving Eq. (17)
 8:    Update $\boldsymbol{B}$ via solving Eq. (19)
 9:    Update $\alpha^{(v)}$ via solving Eq. (26)
10: **end for**

---

5   Nonlinear Feature Embedding and Out-of-Sample Cases

5.1   Nonlinear Feature Embeddings

High-quality multi-view features are indispensable in the generation of accountable hash codes. However, to avoid redundancies and noise in the handcraft features is of great difficulty, in this work we employ a kernelized method to solve

this problem. According to [15], Gaussian Mixed Model (GMM) makes great contribution to the production of nonlinear feature embeddings. The function writes:

$$\varphi(x_i^{(v)}) = [exp(-\left\|x_I^{(v)} - a_1^{(v)}\right\|^2/\sigma_v), \cdots, exp(-\left\|x_i^{(v)} - a_k^{(v)}\right\|^2/\sigma_v)]^\top, \quad (27)$$

where $\varphi(x_i^{(v)}) \in \Re^{k \times 1}$ is a feature of the i-th sample from v-th view learned using the Radial Basis Function (RBF) kernel. $a_i^{(v)}(i = 1, 2, \cdots, k)$ is k randomly selected anchor samples in view v. σ_m is the m-th view's kernel width which means the median of the distance between $a^{(v)}$ and $\boldsymbol{X}^{(v)}$. Replacing $\boldsymbol{X}^{(v)}$ by nonlinear feature $\varphi(\boldsymbol{X}^{(v)}) = [\varphi(x_1^{(v)}), \cdots, \varphi(x_n^{(v)})] \in \Re^{k \times n}$, the above optimization methods can be applied to this learning algorithm.

5.2   Out-of-Sample Extension

In this section, we extend the proposed DFMH method to out-of-sample cases, in which data are not included in the training stage. Generally, a two-step method is adopted to generate the final hash function [9,25]. Typically, for the given training samples, we can obtain their common latent representation $\boldsymbol{H}$ by using Eq. (17). Based on this, we can formulate the hash function from the latent representation $\boldsymbol{H}$ to the objective binary code matrix $\boldsymbol{B}$ by using the following simple linear regression function:

$$\min_{\boldsymbol{W}} \left\|\boldsymbol{W}^\top \boldsymbol{H} - \boldsymbol{B}\right\|_F^2 + \|\boldsymbol{W}\|_F^2, \quad (28)$$

which has closed-form solution:

$$\boldsymbol{W} = (\boldsymbol{H}\boldsymbol{H}^\top + \boldsymbol{I})^{-1}\boldsymbol{H}\boldsymbol{B}^\top. \quad (29)$$

Similarly, for any query sample $\boldsymbol{x}_q^{(v)}$, we first embed it into the nonlinear feature vector $\varphi(\boldsymbol{x}_q^{(v)})$ by using Eq. (27). Subsequently, we learn the shared latent representation $\hat{h}$ of the multi-view query samples based on the following equation:

$$\hat{h} = \sum_{v=1}^m (\alpha^{(v)})^r \boldsymbol{U}_1^{(v)} \boldsymbol{U}_2^{(v)} \cdots \boldsymbol{U}_m^{(v)} \varphi(\boldsymbol{x}_q^{(v)}), \quad (30)$$

Hence, the corresponding hash codes can be learned by using:

$$b = sgn(\boldsymbol{W}^\top \hat{h}). \quad (31)$$

6   Experiments

6.1   Datasets

To evaluate our model, we conduct experiments on two publicly available benchmark datasets, *i.e.*, CIFAR-10 [6] and MIRFlickr [8]. CIFAR-10[1] is consisted of

---

[1] https://www.cs.toronto.edu/~kriz/cifar.html.

Table 1. MAP values with different code length on **CIFAR-10** and **MIRFlickr**

CIFAR-10					
Methods	8	16	32	64	128
MFKH [11]	0.1943	0.1892	0.1654	0.1467	0.1330
MFH [20]	0.1235	0.1349	0.1443	0.1649	0.1798
MAH [8]	0.1051	0.1117	0.1104	0.1089	0.1509
MVLH [19]	0.1103	0.1189	0.1238	0.1146	0.1103
DMVH [23]	0.1162	0.1510	0.1484	0.1549	0.1599
MVDH [17]	0.1524	0.1691	0.1796	0.1877	0.1952
DFMH (ours)	**0.3590**	**0.4265**	**0.4820**	**0.5892**	**0.6362**
MIRFlickr					
Methods	8	16	32	64	128
MFKH [11]	0.5763	0.5805	0.5827	0.5773	0.5745
MFH [20]	0.5596	0.5643	0.5673	0.5692	0.5703
MAH [8]	0.5626	0.5622	0.5629	0.5653	0.5714
MVLH [19]	0.5638	0.5670	0.5744	0.5729	0.5700
DMVH [23]	0.5673	0.5623	0.5619	0.5631	0.5739
MVDH [17]	0.5660	0.5693	0.5723	0.5745	0.5766
DFMH (ours)	**0.6176**	**0.6414**	**0.6701**	**0.6753**	**0.6936**

10 classes with 6000 images each, *i.e.*, 60000 images in total. Three features, *i.e.*, LBP, Gist, and HOG, with dimensions of 1450, 1024, and 1152 respectively, are extracted as 3 views to evaluate the performances of the methods. Following [9], 12500 labeled images from 38 categories are extracted from MIRFlickr dataset[2] in our experiment. Gist (512 dimensions), Hue histogram (100 dimensions) and SIFT bag of visual words (BoVW) (1000 dimensions) serve as three views.

6.2 Experimental Settings and Evaluation Matrics

Our DFMH method is compared with six different state-of-the-art multi-view hashing methods, *i.e.*, MFKH [11], MFH [20], MAH [8], MVLH [19], DMVH [23], and MVDH [17]. All these experiments are conducted on MATLEB R2021a on a computer with Intel (R) Core (TM) i5-9300HF CPU @ 2.40 GHz and 16 GB RAM. The original code of the compared methods are from the corresponding authors. The parameters in each compared method are set to the recommended values if it is mentioned in corresponding papers, or in other cases, default values. For the CIFAR-10 dataset, 1000 samples are randomly selected from 10 views with 100 samples from each view as a query set. Similarly, MIRFlickr is separated into a query set with 1000 images and a retrieval set containing the rest of the images. Two pictures are considered relevant if at least one label exists in both

---

[2] http://lear.inrialpes.fr/people/guillaumin/data.php.

(a) 16 bits   (b) 32 bits   (c) 64 bits   (d) 128 bits

(e) 16 bits   (f) 32 bits   (g) 64 bits   (h) 128 bits

(i) 16 bits   (j) 32 bits   (k) 64 bits   (l) 128 bits
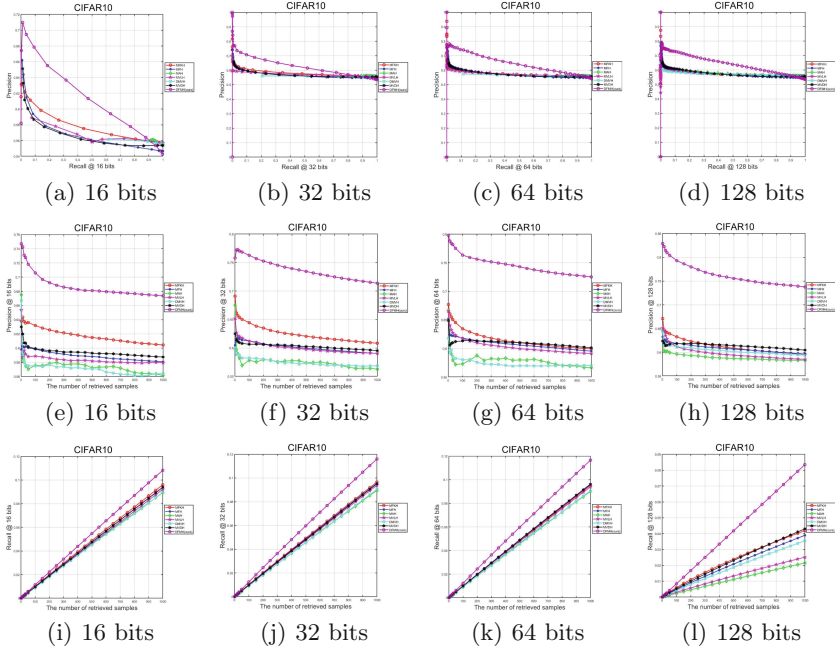
Fig. 2. Precision-recall curves, Precision @TopN curves, and Recall @TopN curves of DFMH as well as compared methods on the CIFAR-10 with different code lengths

of their label sets. We randomly select 1000 images as anchor graphs for anchor-based methods. Gaussian kernel is used for the kernel-based methods [9].

Following previous works [14, 26], we use metrics including mean average precision (mAP), precision-recall curves, recall @TopN and precision @TopN to evaluate the performances of tested methods.

6.3   Experimental Results

Table 1 reports the mAPs of DFMH as well as compared algorithms on CIFAR-10 and MIRFlickr when code length is 8, 16, 32, 64, and 128, respectively. The following observations can be made from Table 1. 1) With the increment of code length, the mAPs generally rise, which is intuitive since longer hash codes usually contain more information. 2) Due to the lack of semantic information, unsupervised methods normally have inferior performances than the supervised ones. 3) MFKH performs the best among all the compared methods. 4) Our method defeats all the other methods dramatically on both datasets. For CIFAR-10, when the hash code is consisted of 8 bits, our method has an mAP 16.47% higher than that of the second-placed method MFKH, and the mAP of our method is more than two times as high as that of the second highest method in the 16, 32, 64 or 128 bits situations. On the MIRFlickr dataset, DFMH defeats all the compared methods, and as the hash code length increases, the advantage of our

method over the compared methods becomes even larger. For instance, when the hash codes is only 8 bits, the mAP of our method is 4.13% higher than the second highest method, while when the hash code is lengthened to 128 bits, our method outperforms the second-placed method by 11.7%. This indicates that with the same code length increment, our method incorporates more additional information.

The precision-recall relationships of all the tested methods with different length of hash code on CIFAR-10 and MIRFlickr are reported in figure (a)–(d) in Fig. 2 and Fig. 3, respectively. The P-R curve of our method lies on the upper right side of the other curves in almost all situations, which strongly justifies the effectiveness of our proposed method. The relationship between precision and the number of retrieved images on both datasets are illustrated in figure (e)–(h) in Fig. 2 and Fig. 3, and figure (i)–(l) demonstrate how recall change with the number of retrieved images. It can be concluded from the graphs that with the number of image retrieved fixed, our method always yields a better precision and recall.

(a) 16 bits   (b) 32 bits   (c) 64 bits   (d) 128 bits

(e) 16 bits   (f) 32 bits   (g) 64 bits   (h) 128 bits

(i) 16 bits   (j) 32 bits   (k) 64 bits   (l) 128 bits
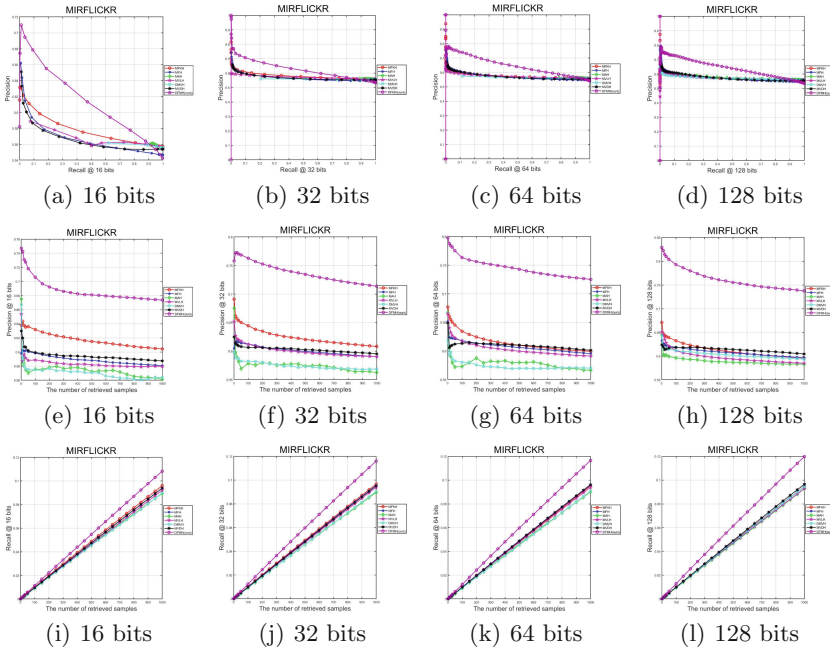
Fig. 3. Precision-recall curves, Precision @TopN curves, and Recall @TopN curves of DFMH as well as compared methods on the MIRFlickr with different code lengths
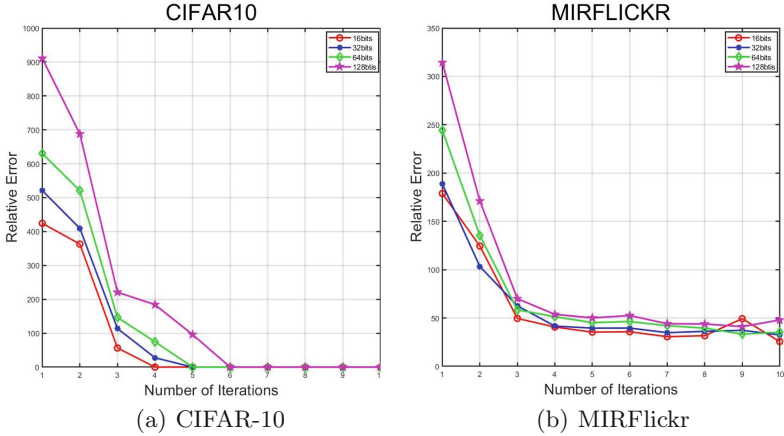
Fig. 4. How retrieval error of DFMH change with iteration times on CIFAR-10 and MIRFlickr when the code lengths are different.

6.4   Convergence and Computational Complexity Analysis

Figure 4 shows how $\|\boldsymbol{B}_t - \boldsymbol{B}_{t-1}\|_F$ of different code length changes with the number of iterations, where $\boldsymbol{B}_t$ denotes $\boldsymbol{B}$ gained from the t-th iteration. It can be observed that the curve has a downward trend, while in the first two iterations the difference drops sharply, and in the following iterations the curve flattens down until the final convergence. These figures also show that our objective function converges within a reasonable number of iterations, which validates the proposed optimization method.

Table 2 compares the running time of different methods on MIRFlickr learning a 32-bit hash code. While outperforming other methods on mAP, our method is also among the fastest methods which are able to finish training within 10 seconds. This shows that our method keeps a balance between efficiency and effectiveness.

Table 2. Comparison of training time on **MIRFlickr** dataset with 32-bit hash code.

Methods	Training time (s)
MFKH [11]	2.03
MVLH [19]	5.13
MFH [20]	6.42
MVDH [17]	13.14
MAH [8]	62.15
DMVH [23]	408.97
DFMH (ours)	9.69

7   Conclusion

We proposed a novel *Deep Factorized Multi-view Hash* (DFMH) method for efficient and effective multi-view image retrieval in this work. DFMH first learned a common representation space through an adaptively weighted deep factorization scheme and then fused it with the semantic information space to learn a hamming space. Furthermore, a feasible algorithm was put forward to optimize the non-convex problem. The superiority of the proposed DFMH method in both accuracy and computational complexity was validated by extensive experiments on benchmark datasets.

References

1. Chen, Y., Wang, S., Lu, J., Chen, Z., Zhang, Z., Huang, Z.: Local graph convolutional networks for cross-modal hashing. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1921–1928 (2021)
2. Ding, C.H.Q., Li, T., Jordan, M.I.: Convex and semi-nonnegative matrix factorizations. IEEE Trans. Pattern Anal. Mach. Intell. **32**(1), 45–55 (2008)
3. Liong, V.E., Lu, J., Wang, G., Moulin, P., Zhou, J.: Deep hashing for compact binary codes learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2475–2483 (2015)
4. Gorisse, D., Cord, M., Precioso, F.: Locality-sensitive hashing for Chi2 distance. IEEE Trans. Pattern Anal. Mach. Intell. **34**(2), 402–409 (2011)
5. Ji, J., Li, J., Yan, S., Zhang, B., Tian, Q.: Super-bit locality-sensitive hashing. In: Advances in Neural Information Processing Systems, vol. 25 (2012)
6. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
7. Kulis, B., Grauman, K.: Kernelized locality-sensitive hashing. IEEE Trans. Pattern Anal. Mach. Intell. **34**(6), 1092–1104 (2011)
8. Liu, L., Mengyang, Yu., Shao, L.: Multiview alignment hashing for efficient image search. IEEE Trans. Image Process. **24**(3), 956–966 (2015)
9. Liu, L., Zhang, Z., Huang, Z.: Flexible discrete multi-view hashing with collective latent feature learning. Neural Process. Lett. **52**(3), 1765–1791 (2020)
10. Liu, W., Wang, J., Kumar, S., Chang, S.-F.: Hashing with graphs. In: International Conference on Machine Learning (2011)
11. Liu, X., He, J., Liu, D., Lang, B.: Compact kernel hashing with multiple features. In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 881–884 (2012)
12. Liu, X., Yadong, M., Zhang, D., Lang, B., Li, X.: Large-scale unsupervised hashing with shared structure learning. IEEE Trans. Cybern. **45**(9), 1811–1822 (2014)
13. Lu, X., Zhu, L., Li, J., Zhang, H., Shen, H.T.: Efficient supervised discrete multi-view hashing for large-scale multimedia search. IEEE Trans. Multimedia **22**(8), 2048–2060 (2019)
14. Schütze, H., Manning, C.D., Raghavan, P.: Introduction to Information Retrieval, vol. 39. Cambridge University Press, Cambridge (2008)
15. Shen, F., Shen, C., Liu, W., Shen, H.T.: Supervised discrete hashing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 37–45 (2015)

16. Shen, F., Xu, Y., Liu, L., Yang, Y., Huang, Z., Shen, H.T.: Unsupervised deep hashing with similarity-adaptive and discrete optimization. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 3034–3044 (2018)
17. Shen, X., Shen, F., Liu, L., Yuan, Y.-H., Liu, W., Sun, Q.-S.: Multiview discrete hashing for scalable multimedia search. ACM Trans. Intell. Syst. Technol. **9**(5), 1–21 (2018)
18. Shen, X., Shen, F., Sun, Q.-S., Yuan, Y.-H.: Multi-view latent hashing for efficient multimedia search. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 831–834 (2015)
19. Shen, X., Shen, F., Sun, Q.-S., Yuan, Y.H.: Multi-view latent hashing for efficient multimedia search. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 831–834 (2015)
20. Song, J., Yang, Y., Huang, Z., Shen, H.T., Luo, J.: Effective multiple feature hashing for large-scale near-duplicate video retrieval. IEEE Trans. Multimedia **15**(8), 1997–2008 (2013)
21. Trigeorgis, G., Bousmalis, K., Zafeiriou, S., Schuller, B.W.: A deep matrix factorization method for learning attribute representations. IEEE Trans. Pattern Anal. Mach. Intell. **39**(3), 417–429 (2016)
22. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: Advances in Neural Information Processing Systems, vol. 21 (2008)
23. Yang, R., Shi, Y., Xu, X.-S.: Discrete multi-view hashing for effective image retrieval. In: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 175–183 (2017)
24. Zhang, Z., et al.: Scalable supervised asymmetric hashing with semantic and latent factor embedding. IEEE Trans. Image Process. **28**(10), 4803–4818 (2019)
25. Zhang, Z., Luo, H., Zhu, L., Lu, G., Shen, H.T.: Modality-Invariant asymmetric networks for cross-modal hashing. IEEE Trans. Knowl. Data Eng. (2022). https://doi.org/10.1109/TKDE.2022.3144352
26. Zhang, Z., Xie, G., Li, Y., Li, S., Huang, Z.: SADIH: semantic-aware discrete hashing. Proc. AAAI Conf. Artif. Intell. **33**, 5853–5860 (2019)

Covid-19 Detection by Wavelet Entropy and Artificial Bee Colony

Jia-Ji Wang[1], Yangrong Pei[2], Liam O'Donnell[3], and Dimas Lima[4(✉)]

[1] School of Math and Information Technology, Jiangsu Second Normal University, Nanjing 210016, Jiangsu, People's Republic of China
[2] Huai'an Tongji Hospital, Huai'an 223000, Jiangsu, China
[3] School of Engineering, University of Limerick, Limerick, Ireland
[4] Department of Electrical Engineering, Federal University of Santa Catarina, Florianópolis 88040-900, Brazil
dimaslima@ieee.org

Abstract. Computer analysis of patients' lung CT images has become a popular and effective way to diagnose COVID-19 patients amid repeated and evolving outbreaks. In this paper, wavelet entropy is used to extract features from CT images and integrate the information of various scales, including the characteristic signals of signals with transient components. Combined with the artificial bee colony optimization algorithm, we used the advantages of fewer parameters and simpler calculation to find the optimal solution and confirm COVID-19 positive. The use of K-fold cross validation allows the data set to avoid overfitting and unbalanced data set partition in small cases. The experimental results were compared with those of WE + BBO, GLCM-SVM, GLCM-ELM and WE-Jaya. Experimental data show that this method achieves our initial expectation.

Keywords: Covid-19 detection · Wavelet entropy · Artificial bee colony

1 Introduction

With the spread of the SARS COV-2, the number of confirmed patients with COVID-19 is increasing rapidly and the number of deaths from severe cases is still on the rise [1]. Under these circumstances, hospitals have to take on more pressure and call on more resources, including drugs, manpower and equipment [2], to screen patients. COVID-19 is a respiratory disease caused by the new coronavirus SARS COV-2 [3].

In February 2020, WHO announced that the official name of this disease is COVID-19. In this name, CO stands for corona, VI for the virus, D for disease, and 19 is for the first outbreak year of that disease [4]. The majority of patients have the following respiratory tract symptoms, common clinical manifestations include fever, cough, expectoration, weakness of limbs, headache and other symptoms, and there are also patients without obvious symptoms [5–7]. However, it is not yet clear what symptoms and complications

---

J.-J. Wang and Y. Pei — Contributed equally to this paper

can occur long after contracting COVID-19 [8, 9]. To improve the efficiency of screening, we use Wavelet Entropy and Artificial Bee Colony model to diagnose COVID-19 images.

Some researchers have proposed some classical classification methods based on CT images of COVID-19 patients' lung. Research [10] uses CNN to detect COVID-19. By testing the performance of different pre-training models on CT tests [11], they found that a larger field data set could improve the testing capabilities of the models.

In this study [12], VGG16 and ResNet50 models are improved and optimized by using data expansion and fine-tuning techniques. The robustness and validity of the model were verified and tested using hierarchical 5-fold cross-validation. Finally, the model performs much better in binary classification than in multi-classification.

The research [13] proposed two deep learning structures including AlexNet architecture are proposed in this study. The input images are pre-segmented by ANN [14]. In one structure, BiLSTM layer was added, and time attribute was added, and the accuracy reached 98.7%.

In this study, we attempted to use Wavelet Entropy [15] and Artificial Bee Colony [16] to diagnose lung CT images of COVID-19 [17] patients. The following article is organized as follows. Section 2 shows the dataset we used. Section 3 describes our proposed approach. Section 4 discusses the experimental results. In the Sect. 5, we summarize the research of this paper.

2 Dataset

The data set we used consisted of three different types of patients and a healthy control group. All images have a resolution of $1024 \times 1024 \times 3$. A total of 148 COVID-19 images and 148 healthy control (HC) images are obtained. Figure 1 shows two figures in the dataset we used [18].

(a) COVID-19   (b) Healthy
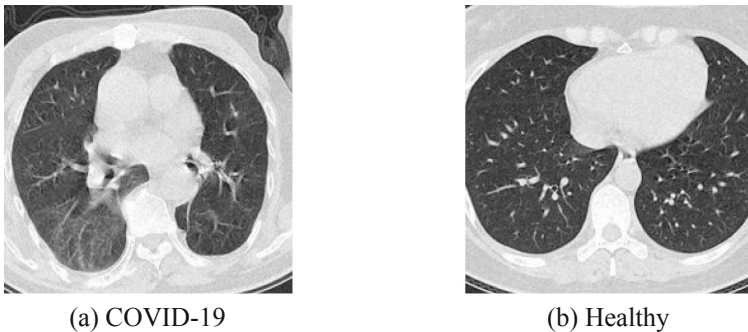
Fig. 1. These two graphs are samples from the dataset

3 Methodology

3.1 Wavelet Entropy

The main factors to be considered in signal analysis are frequency, amplitude and phase. The main function of the Fourier transform [19] is to take a signal and decompose it into its constituent frequencies, converting the function into the form of multiple sinusoidal combinations, or e exponents [20, 21]. The signal is still the original signal essentially, but in a different way so that we can more analyze the frequency, amplitude, and phase components of a function intuitively. Therefore, after analyzing a complex signal through Fourier transform, its frequency [22], phase and amplitude components can be easily seen. It is a general analysis of the signal. The formula is:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt \tag{1}$$

where ω represents frequency, t represents time.

But the Fourier transform has its own limitations. It can be seen from the Fourier transform formula that it is based on sine wave and its higher harmonics [23, 24] as the standard basis. If you need a good localization in the time domain, you need to analyze all the signals in the frequency domain [25–27]. If we want good results in one aspect, we have to give up the other one. Wavelet entropy changes the basis for the Fourier transform directly, replacing an infinite trig basis with a finite decaying wavelet basis [28, 29]. Thus, not only can we get the frequency, but also locate the time. The formula is as follows:

$$W(a, \tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t)\psi\left(\frac{t - \tau}{a}\right)dt \tag{2}$$

where a is scale, τ is translation, ψ is parent wavelet function, and t is time.

That's obvious from the formula that the number of variables of wavelet transform is more than that of Fourier transform. In the wavelet transform, variable scale a controls the scaling of the wavelet function and variable translation τ controls the translation of the wavelet function. The scale a corresponds to frequency (inverse ratio) and the translation τ corresponds to time. As a more ideal tool, wavelet transform can provide a "time-frequency" window that changes with frequency, and transform the original fixed window size of Fourier transform into an adaptive window size for signal processing.

3.2 Feedforward Neural Network

Feedforward neural network [30] has a one-way multi-layer structure [31–33], and each neuron is arranged hierarchically, and each neuron is only connected with the previous neuron. The output of the previous layer is received and sent to the next layer [34–36], with no feedback between layers. A typical multilayer feedforward neural network is shown in Fig. 2.
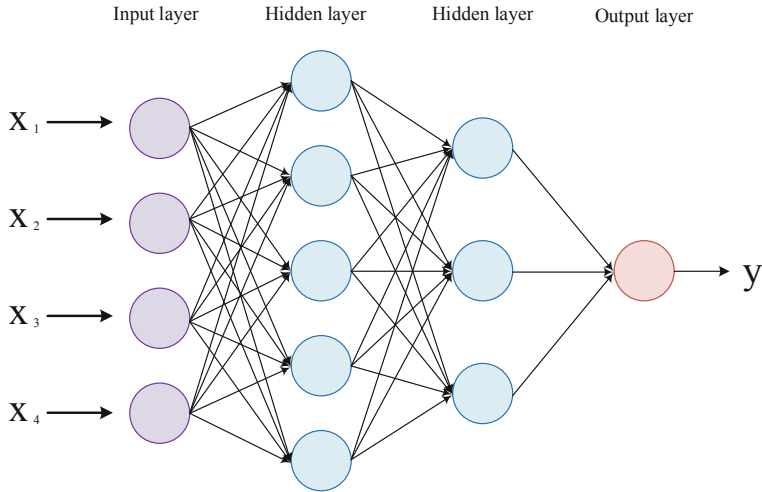
Fig. 2. A typical multilayer feedforward neural network

Feedforward neural network can be divided into single layer feedforward neural network [37] and multilayer feedforward neural network [38]. The single-layer feedforward neural network contains only one output layer, and the output value can be obtained directly by multiplying the input value by the weight value [39–41]. The multilayer feedforward neural network includes a input layer, one or more hidden layers and a output layer, which is relatively complex. Each layer of neurons is completely connected to the next layer of neurons, and there is no same-layer connection or cross-layer connection between neurons. Common feedforward neural networks include perceptron network [42], BP network [43] and so on.

3.3 Artificial Bee Validation

Artificial bee colony algorithm is an optimization method proposed by Karaboga in 2005 [44], which imitated bee behavior and has fast convergence speed. This method is similar to that after each worker bee finds the optimal solution, the colony is also displayed as the global optimal solution [45, 46]. In the basic ABC algorithm, artificial bee colony consists of three types of individuals: employed bees, onlooker bees and scout bees. Randomly generate feasible solution equal to the number of employed bees during initialization. Since each food source x_m is a feasible solution, each feasible solution containing n variables. It's a vector.$(x_{mi}, m = 1, 2 \ldots SN, i = 1, 2 \ldots n)$.

The initialization formula is as follows:

$$x_{mi} = l_i + rand(0, 1) * (u_i - l_i), \tag{3}$$

where u_i means the maximum boundary values of x_{mi} and l_i means the minimum boundary values of x_{mi}.

Each employed bees corresponds to a certain feasible solution that already exists, and the field of nearby feasible solution is searched in the iteration. Employed bees continue to search for new feasible solutions near existing ones. The formula for finding new food sources is as follows:

$$v_{mi} = x_{mi} + \phi_{mi}(x_{mi} - x_{ki}) \tag{4}$$

The x_k is a random food source, i also is random, ϕ is a random number between $[-a, a]$. When a new feasible solutions is found, the fitness of the feasible solution is estimated [47–49] and the greedy selection is used to select between old and new food sources. The fitness calculation formula is as follows:

$$fit_m(x_m) = \begin{cases} \frac{1}{1+f_{m(x_m)}} & \text{if } f_{m(x_m)} \geq 0 \\ 1 + abs\left(f_{m(x_m)}\right) & \text{if } f_{m(x_m)} < 0 \end{cases} \tag{5}$$

where f_m is the objective function of the feasible solution x_m.

Scout bees use roulette to choose possible solutions based on the information brought back to the hive by the employed bees. The p_m formula for the probability of following bees choosing a food source is as follows:

$$p_m = \frac{fit_m(x_m)}{\sum\limits_{m=1}^{SN} fit_m(x_m)} \tag{6}$$

If the feasible solution is still not improved after several updates, the feasible solution will be gave up, and the employed bee is turned into the scout bee to continue to randomly search for new feasible solutions.

3.4 K-fold Cross Validation

The total data set can be divided into training set and test set. K-fold cross Validation, a replacement-free resampling technique, can be used when the sample size is insufficient. In order to test the effectiveness of the algorithm, K-fold cross Validation randomly divides the total data set into K parts to make the best use of every piece of data in the data set, and takes out one package as the test set each time, and the remaining K-1 packages for training. The Fig. 3 shows the workflow of the K-fold method.

After k times, and these k models and performance evaluation were averaged to obtain the average performance. In many cases, K is 10. If the training set is relatively small, you can increase the K value. If the data set is large, reducing k value can reduce the refitting on different folds and the computational burden of model evaluation.
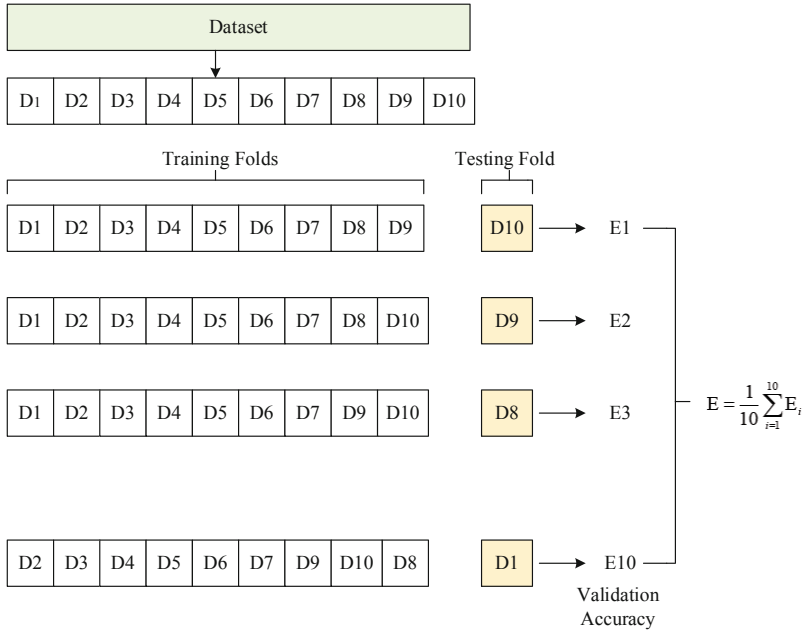
Fig. 3. The workflow of the K-fold method

4 Experience Results and Discussion

4.1 WE Results

Figure 4 shows an example of a level 4 decomposition of a biorthogonal decomposition. Wavelet transform divides the image into low-frequency part which changes slowly and high-frequency part which changes quickly. The upper left corner of Fig. 4(a) is the low-frequency information image of the input image obtained after the first-level wavelet transform, which is the frame and contour of the image. The other three images in Fig. 4(a) (located in the upper right corner, lower left corner and lower right corner) are the high-frequency information of the input image, reflecting the details of the input image. The upper sampling in the wavelet is the interval zero insertion, which aims to reconstruct the signal. The lower sampling introduced in the wavelet is the interval sampling of the signal. The edge lengths of the four self-band maps are universal for the input, and the purpose is to compress and store the information. The input image in Fig. 4(b) is the low-frequency information image in Fig. 4(a). Three images (upper right corner, lower left corner and lower right corner of Fig. 4(b)) are obtained after transformation similar to first-level wavelet transform. Figure 4(c) and Fig. 4(d) are obtained after the recursive operations.

(a) 1-level (b) 2-level

(c) 3-level (d) 4-level
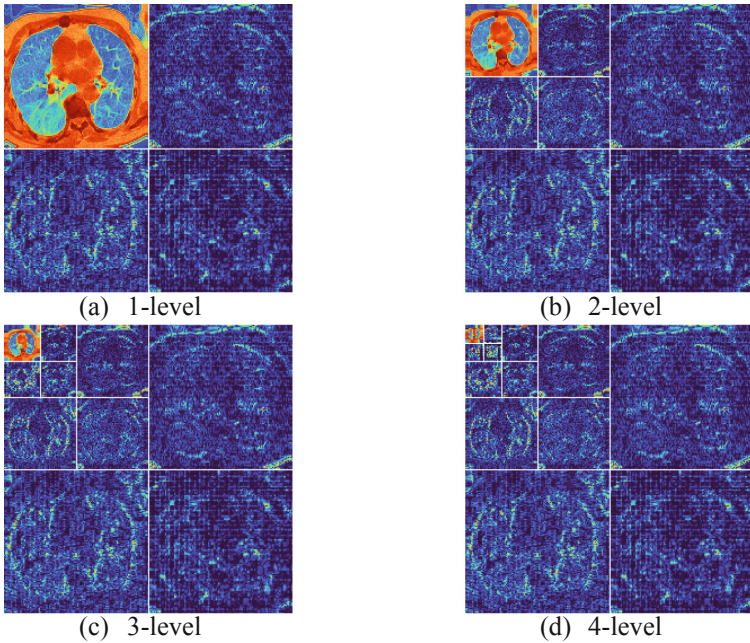
Fig. 4. An example of a level 4 decomposition of a biorthogonal decomposition

4.2 Statistical Results

The feature extraction method used in Wavelet Entropy and Artificial Bee Colony is wavelet entropy, the classifier is FNN, and 10 times 10-fold cross validation reports unbiased performance. The data show that the experiment achieves good performance (Table 1). The average sensitivity was 75.14 ± 1.93, the specificity was 77.30 ± 2.45, the precision was 76.85 ± 1.72, the accuracy was 76.22 ± 0.96, the F1 score was 75.95 ± 0.95, the Matthews correlation coefficient was 52.48 ± 1.95, and the feature mutual information of 75.97 ± 0.95.

4.3 Comparison to State-of the Art Approaches

To understand the level of this method proposed by us, we consulted relevant literature and compared our experimental data with the experimental data of the current more advanced method, as shown in Table 2. As the results in the Table 2, the Sen, Acc, F1, MCC and FMI data of this method are the best among these methods. ABC algorithm has strong ability to find the optimal solution globally and fast convergence speed. However, similar to the gradient descent method, when approaching the global optimal solution, it may fall into the local optimal solution, and the later search speed slows down in the subsequent search. According to these characteristics, ABC algorithm is applicable to solving the multivariable function optimization problems.

Table 1. Results of 10-fold cross-validation

Run	Sen	Spc	Prc	Acc	F1	MCC	FMI
1	75.68	77.70	77.24	76.69	76.45	53.39	76.45
2	72.97	82.43	80.60	77.70	76.60	55.65	76.69
3	76.35	75.00	75.33	75.68	75.84	51.36	75.84
4	77.70	75.00	75.66	76.35	76.67	52.72	76.67
5	74.32	77.03	76.39	75.68	75.34	51.37	75.35
6	75.00	76.35	76.03	75.68	75.51	51.36	75.51
7	71.62	78.38	76.81	75.00	74.13	50.11	74.17
8	75.68	74.32	74.67	75.00	75.17	50.00	75.17
9	74.32	79.73	78.57	77.03	76.39	54.13	76.42
10	77.70	77.03	77.18	77.36	77.44	54.73	77.44
Mean + SD	75.14 ± 1.93	77.30 ± 2.45	76.85 ± 1.72	76.22 ± 0.96	75.95 ± 0.95	52.48 ± 1.95	75.97 ± 0.95

Table 2. The results are compared with the existing advanced methods

Method	Sen	Spc	Prc	Acc	F1	MCC	FMI
WE + BBO [50]	72.84 ± 3.00	75.00 ± 1.99	74.47 ± 1.20	73.92 ± 1.18	73.61 ± 1.57	47.89 ± 2.34	73.63 ± 1.56
GLCM-SVM [51]	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
GLCM-ELM [52]	74.19 ± 2.74	77.81 ± 2.03	77.01 ± 1.29	76.00 ± 0.98	75.54 ± 1.31	52.08 ± 1.95	75.57 ± 1.28
WE-Jaya [18]	73.31 ± 2.26	78.11 ± 1.92	77.03 ± 1.35	75.71 ± 1.04	75.10 ± 1.23	51.51 ± 2.07	75.14 ± 1.22
WE-ABC (Ours)	75.14 ± 1.93	77.30 ± 2.45	76.85 ± 1.72	76.22 ± 0.96	75.95 ± 0.95	52.48 ± 1.95	75.97 ± 0.95

5 Conclusions

The use of computer analysis of CT images for classification is attracting more and more attention. People realized that this was a way to save a lot of medical and human resources, and to reduce unnecessary contact with patients and reduce the risk of infection. Combined with wavelet entropy and artificial bee colony algorithm, this study proves that the model has improved the classification of CT images. The data show that the results are relatively ideal. This method has not been applied in other medical image classification tasks so far. In the future, we will apply this method to different types of medical images to continuously verify, optimize and improve the performance of this method. We believe that after continuous experiments, the application scope of

this method will be more extensive, suitable for the classification and diagnosis of various diseases, and improve the efficiency and accuracy of diagnosis, so as to face more challenges in the future.

References

1. Lufler, R.S., et al.: The glass ceiling thickens: the impact of COVID-19 on academic medicine faculty in the United States. Med. Educ. Online **27**(1), 2058314 (2022)
2. Jerzak, M., Szafarowska, M.: Preliminary results for personalized therapy in pregnant women with polycystic ovary syndrome during the COVID-19 pandemic. Arch. Immunol. Ther. Exp. **70**(1), 1–7 (2022). https://doi.org/10.1007/s00005-022-00650-z
3. Santana, J.C., et al.: Complicações em testes para COVID-19 com swab nasal: relatos de caso. Revista de Medicina **101**(2) (2022)
4. Acter, T., et al.: Evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as coronavirus disease 2019 (COVID-19) pandemic: a global health emergency. Sci. Total Environ. **730**, 138996 (2020)
5. Zhang, X.: Diagnosis of COVID-19 pneumonia via a novel deep learning architecture. J. Comput. Sci. Technol. **37**(2), 330–343 (2022)
6. Yang, L.: EDNC: ensemble deep neural network for Covid-19 recognition. Tomography **8**(2), 869–890 (2022)
7. Guo, X.: A survey on machine learning in COVID-19 diagnosis. Comput. Model. Eng. Sci. **130**(1), 23–71 (2022)
8. Balasubramanian, P., et al.: Outcomes in patients with mild COVID-19 treated with casirivimab and imdevimab or bamlanivimab-a single-center retrospective cohort study in the bronx. Infect. Dis. Clin. Pract. **30**(3), e1128 (2022)
9. Hashimoto, K., et al.: Severe infectious acute respiratory failure mimicking COVID-19 in a healthy adolescent. Respirology Case Rep. **10**(4), e0933 (2022)
10. Zhao, W., et al.: Deep learning for COVID-19 detection based on CT images. Sci. Rep. **11**(1), 14353 (2021)
11. Khan, M.A.: VISPNN: VGG-inspired stochastic pooling neural network. Comput. Mater. Continua **70**, 3081–3097 (2022)
12. Mishra, N.K., et al.: Automated detection of COVID-19 from CT scan using convolutional neural network. Biocybern. Biomed. Eng. **41**(2), 572–588 (2021)
13. Aslan, M.F., et al.: CNN-based transfer learning-BiLSTM network: a novel approach for COVID-19 infection detection. Appl. Soft Comput. **98**, 106912 (2021)
14. Shanbehzadeh, M., et al.: Developing an artificial neural network for detecting COVID-19 disease. J. Educ. Health Prom. **11** (2022)
15. Wang, W., et al.: Covid-19 diagnosis by WE-SAJ. Syst. Sci. Control Eng. **10**, 325–335 (2022)
16. Jacob, I.J., et al.: Artificial bee colony optimization algorithm for enhancing routing in wireless networks. J. Artif. Intell. **3**(01), 62–71 (2021)
17. Govindaraj, V.: Deep rank-based average pooling network for Covid-19 recognition. Comput. Mater. Continua **70**, 2797–2813 (2022)
18. Wang, W.: Covid-19 detection by wavelet entropy and jaya. Lecture Notes in Computer Science, vol. 12836, pp. 499–508 (2021)
19. Anaya-Isaza, A., et al.: Fourier transform-based data augmentation in deep learning for diabetic foot thermograph classification. Biocybern. Biomed. Eng. **42**, 437–452 (2022)
20. Sahabuddin, M., et al.: Co-movement and causality dynamics linkages between conventional and Islamic stock indexes in Bangladesh: a wavelet analysis. Cogent Bus. Manag. **9**(1), 2034233 (2022)

21. Meenpal, A., Majumder, S.: Image content based secure reversible data hiding scheme using block scrambling and integer wavelet transform. Sādhanā **47**(2), 1–17 (2022). https://doi.org/10.1007/s12046-022-01828-z

22. Khani, M.E., et al.: Translation-invariant zero-phase wavelet methods for feature extraction in terahertz time-domain spectroscopy. Sensors **22**(6), 2305 (2022)

23. Fang, Y., et al.: Optimal control over high-order-harmonic ellipticity in two-color cross-linearly-polarized laser fields. Phys. Rev. A **103**(3), 033116 (2021)

24. Wu, X.: Diagnosis of COVID-19 by wavelet renyi entropy and three-segment biogeography-based optimization. Int. J. Comput. Intell. Syst. **13**(1), 1332–1344 (2020)

25. Biswas, A., et al.: Revisiting OD-stretching dynamics of methanol-d(4), ethanol-d(6) and dilute HOD/H2O mixture with predefined potentials and wavelet transform spectra. Chem. Phys. **553**, 111385 (2022)

26. Messer, P.K., Henß, A.-K., Lamb, D.C., Wintterlin, J.: A multiscale wavelet algorithm for atom tracking in STM movies. New J. Phys. **24**(3), 033016 (2022). https://doi.org/10.1088/1367-2630/ac4ad5

27. Chia, C., et al.: Interpretable classification of bacterial Raman spectra with Knockoff wavelets. IEEE J. Biomed. Health Inform. **26**(2), 740–748 (2022)

28. Utudee, S., Maleewong, M.: Multi-resolution wavelet basis for solving steady forced Korteweg–de Vries model. J. Inequalities Appl. **2021**(1), 1–14 (2021). https://doi.org/10.1186/s13660-021-02696-7

29. Jiang, X.: Multiple sclerosis recognition by biorthogonal wavelet features and fitness-scaled adaptive genetic algorithm. Front. Neurosci. **15**(1098), 737785 (2021)

30. Yee, J., et al.: Image features of a splashing drop on a solid surface extracted using a feedforward neural network. Phys. Fluids **34**(1), 013317 (2022)

31. Nimmanterdwong, P., et al.: Artificial neural network prediction of transport properties of novel MPDL-based solvents for post combustion carbon capture. Energy Rep. **8**, 88–94 (2022)

32. Ullah, W., et al.: Artificial intelligence of things-assisted two-stream neural network for anomaly detection in surveillance big video data. Future Gener. Comput. Syst. Int. J. Escience **129**, 286–297 (2022)

33. Christensen, O., et al.: A neural network approach for property determination of molecular solar cell candidates. J. Phys. Chem. A **126**(10), 1681–1688 (2022)

34. Yan, Y., Yao, X.-J., Wang, S.-H., Zhang, Y.-D.: A survey of computer-aided tumor diagnosis based on convolutional neural network. Biology **10**(11), 1084 (2021). https://doi.org/10.3390/biology10111084

35. Wang, S.-H., Satapathy, S.C., Anderson, D., Chen, S.-X., Zhang, Y.-D., Deep fractional max pooling neural network for COVID-19 recognition. Front. Pub. Health **9** (2021). https://doi.org/10.3389/fpubh.2021.726144

36. Zhang, Y.-D., Satapathy, S.C., Wu, D., Guttery, D.S., Górriz, J.M., Wang, S.-H.: Improving ductal carcinoma in situ classification by convolutional neural network with exponential linear unit and rank-based weighted pooling. Complex Intell. Syst. **7**(3), 1295–1310 (2020). https://doi.org/10.1007/s40747-020-00218-4

37. Koçak, Y., et al.: New activation functions for single layer feedforward neural network. Expert Syst. Appl. **164**, 113977 (2021)

38. Rizk-Allah, R.M., Hassanien, A.E.: COVID-19 forecasting based on an improved interior search algorithm and multilayer feed-forward neural network. In: Hassanien, A.E., Bhatnagar, R., Snášel, V., Yasin Shams, M. (eds.) Medical Informatics and Bioimaging Using Artificial Intelligence. SCI, vol. 1005, pp. 129–152. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-91103-4_8

—

39. Hajiahmadi, M., Zarei, M., Khataee, A.: An effective natural mineral-catalyzed heterogeneous electro-Fenton method for degradation of an antineoplastic drug: modeling by a neural network. Chemosphere **291**, 132810 (2022). https://doi.org/10.1016/j.chemosphere.2021.132810
40. Ankobea-Ansah, K., et al.: A hybrid physics-based and stochastic neural network model structure for diesel engine combustion events. Vehicles **4**(1), 259–296 (2022)
41. Goudarzi, F., Hedayatiaghmashhadi, A., Kazemi, A., Fürst, C.: Optimal location of water quality monitoring stations using an artificial neural network modeling in the Qarah-Chay River Basin, Iran. Water **14**(6), 870 (2022). https://doi.org/10.3390/w14060870
42. Radhakrishnan, S., et al.: Multilayer perceptron neural network model development for mechanical ventilator parameters prediction by real time system learning. Biomed. Sign. Process. Control **71**, 103170 (2022)
43. Chen, L., Jagota, V., Kumar, A.: Research on optimization of scientific research performance management based on BP neural network. Int. J. Syst. Assurance Eng. Manag., 1–102021). https://doi.org/10.1007/s13198-021-01263-z
44. Karaboga, D.: An idea based on honey bee swarm for numerical optimization. Technical report-tr06, Erciyes university, engineering faculty, computer (2005)
45. Wu, L.: Magnetic resonance brain image classification by an improved artificial bee colony algorithm. Prog. Electromagn. Res. **116**, 65–79 (2011)
46. Wu, L.: Optimal multi-level thresholding based on maximum tsallis entropy via an artificial bee colony approach. Entropy **13**(4), 841–859 (2011)
47. Khababa, G., et al.: An extended artificial bee colony with local search for solving the skyline-based web services composition under interval QoS properties. J. Intell. Fuzzy Syst. **42**(4), 3855–3870 (2022)
48. Mahmoodabadi, M.J., et al.: Optimal design of an adaptive robust controller using a multi-objective artificial bee colony algorithm for an inverted pendulum system. Trans. Can. Soc. Mech. Eng. **46**(1), 89–102 (2022)
49. Rahimi, A.M., et al.: Artificial bee colony algorithm with proposed discrete nearest neighborhood algorithm for discrete optimization problems. Jurnal Kejuruteraan **33**(4), 1087–1095 (2021)
50. 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
51. 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
52. Pi, P.: Gray level co-occurrence matrix and extreme learning machine for Covid-19 diagnosis. Int. J. Cognitive Comput. Eng. **2**, 93–103 (2021)

Heat-Map Algorithm Based Multi-robots Path Planning Method

Shuhui Bi[1], Zhihao Li[1], Lei Wang[2], and Yuan Xu[1(✉)]

[1] School of Electrical Engineering, University of Jinan, Jinan 250022, China
xy_abric@126.com
[2] HRG (Shandong) Intelligent Equipment Research Institute,
No. 1268 Gongye 4 Road, Jinan 250000, Shandong, China

Abstract. Robot conflicts elimination and efficiency improvement from a global perspective are important issues in intelligent storage systems. In order to improve the operation efficiency of intelligent storage system more effectively, a heat-map Algorithm by combing the reservation table in this paper. Firstly, a small storage grid model applicable to multiple storage modes is established. Secondly, considering the frontal collision problem of multiple robots, an improved reservation table is established, which greatly reduces the storage space occupied and improves the operation efficiency at the same time. In addition, the heat map algorithm is added to reasonably allocate the tasks, avoid the congested area and realize the dynamic assignment of tasks. Finally, the effectiveness of the proposed design scheme is demonstrated by simulation.

Keywords: Improved A* algorithm · Improved reservation form · Heat map algorithm

1 Introduction

In recent years, intelligent storage has developed rapidly, and the sales of multi-robot intelligent storage systems have been growing. Conflict-free scheduling of multiple robots is the core of intelligent storage systems. The existing warehouse is expanding, the number of robots is gradually increasing, the system operation process is complex, and the problems are closely connected, and the best route for a single robot and the optimal strategy for a single problem are not optimal when the system is running globally. Since there is more than one robot in the warehouse, conflicts are caused between multiple robots, and the shortest path may have other robots passing through frequently, creating serious conflict problems, and conflicts are strongly dynamic and interleaved, easily causing serial conflicts and congestion, and even deadlock phenomena.

The conflicts between robots can be grouped into the following categories: catch-up conflicts, intersection conflicts, and phase conflicts. For these conflicts, there are various heuristic rules that can be employed, such as improved particle

swarm algorithm [1], or adopting various approaches such as leaving, detouring and waiting before starting to avoid different conflicts [1]. The literature [1] proposed a Firefly-based Approach (FA) for robot cluster path planning, where firefly social behavior is used to optimize group behavior. In the literature [1], a Combined Roadmaps and Potentials for Swarms (CRoPS) path planning algorithm is proposed by combining probabilistic roadmaps and potential field methods to enable clusters to move efficiently to desired destinations while avoiding collisions with each other and with static obstacles. The literature [1] investigates an improved bidirectional A* algorithm to reduce the path length and the time required to plan the path for search and rescue UAVs. The literature [1] uses an improved Q-Learning algorithm to plan the shortest path for each robot to complete the mission goal and form a reservation table to reduce the standby state of robots without a task and balance the workload among robots. The literature [1] proposed a particle swarm optimization-based path planning algorithm for UAV clusters in dynamic environments. In route planning, it is easy to fall into the local optimum problem by using a particular algorithm alone, and the literature [1] proposed a hierarchical path planning method based on a hybrid genetic particle swarm optimization algorithm, which can find the optimal path quickly and efficiently by avoiding obstacles in a complex environment.

In the intelligent storage system, the existing methods mostly minimize the conflict range from the path planning, but can not completely optimize the conflict from the global perspective, and can only qualitatively reduce the conflict but not quantitatively eliminate it.

To address this problem, the multi-robots vertex and edge collision problem is considered by using an improved A* algorithm for path finding for each robot based on the establishment of a warehouse raster map with an improved reservation table. Moreover, for the purpose of achieving an efficient multi-robots conflict-free scheduling scheme, a heat map is added by reflecting the congestion level of the lanes and determining the matching rules between robots and order and picking tables.

2   Algorithm Design

2.1   Establishment of the Environment

Raster Map Building. This paper proposes a more standard small storage model applicable to a variety of storage modes, which can cope with the storage requirements of a variety of scenarios. A reasonable environment representation facilitates the establishment of planning methods and the selection of suitable search algorithms to finally achieve a more satisfactory path with less time overhead. There are many ways to build environment maps, such as visual map method, free space method, topology method, raster method, etc. This thesis intends to use the raster method to model the warehouse environment.

A small warehouse model with a length of 25 m and a width of 26 m was constructed, and the map was partitioned using the raster method. Each raster was 1 m long and 1 m wide, and the map was partitioned into 650 rasters, as

shown in Fig. 1. The blue area is the picking table location. The picking tables were arranged in the leftmost two columns of the warehouse, and each picking table occupied two grids. The black area is the placement of warehouse shelves. The warehouse shelves are arranged in the form of shelf groups, each group occupies 8 grids, arranged in the form of 2 rows and 4 columns, and the shelf groups are kept at an interval of one grid wide as a transportation path for the robot, the location of the white area in Fig. 1 is the transport aisle. As shown in the figure, the robot represented by the green icon delivers the goods to the designated shelf through the transportation lane. In this paper, we ignore the problem of different sizes of pallets caused by different goods, and set each pallet to be a standard module of 1 m in length and 1 m in width, and the robot transports only one pallet per task.
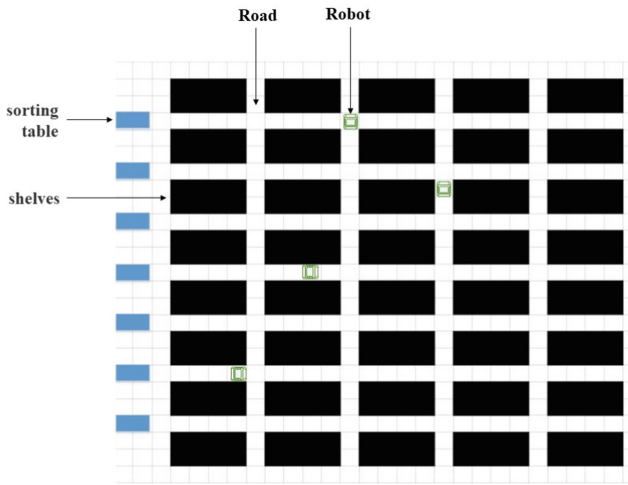
Fig. 1. Raster map and warehouse structure. (Color figure online)

Improved Appointment Form. In intelligent storage systems, multiple robots are involved in transporting goods in and out of the warehouse at the same time, and the environment is dynamic and variable, which can lead to congestion, collision and even deadlock. To address this problem, this thesis uses a reservation table to monitor and record the motion paths of robots, which facilitates the subsequent path optimization algorithm to query the historical paths of robots by calling the reservation table to avoid path conflicts.

The reservation table created is shown in Fig. 2. Variable k indicates the number of reservation tables corresponding to the current moment, the reservation table is arranged in time order, and the time interval is the time consumed by the robot to move from the center position of the grid to the center position of the adjacent grid Δt. The system records the appointment table every Δt. The kth reservation table records the position states of all robots at the moment of

$k \cdot \Delta t$, and the (k-c)th reservation table records the information at the moment of $(k-c) \cdot \Delta t$. The $r_i x$ in the kth reservation table denotes the horizontal coordinate of the warehouse location where robot i is at the moment of $k * \cdot \Delta t$, and $r_i y$ denotes the vertical coordinate of the location where robot i is at this moment. When the robot updates its path, the position information in the reservation table will be updated simultaneously.
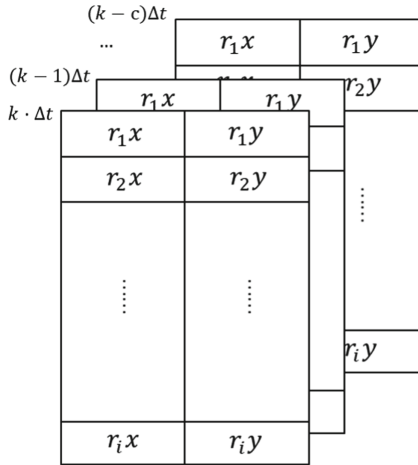
Fig. 2. Improved reservation form.

The number of rows and columns of the matrix of the unimproved reservation table is equal to the number of rows and columns of the raster map, and the route is circumvented by querying the "0" and "1" status of a location. This method occupies a lot of useless matrix space, causing matrix redundancy, and the query step is tedious. In the improved reservation table, the number of rows in each reservation table is equal to the number of robots, the number of columns is 2, and only the X and Y coordinates of the grid location of each robot at the current moment are stored. When the algorithm calls the improved reservation table, it only needs to compare whether the same coordinates appear at the same moment to determine whether there is a conflict and achieve the purpose of collision avoidance.

2.2   Improved A* Algorithm

A* algorithm, as a heuristic algorithm, prioritizes the extended nodes according to the estimated cost function, selects the best node, and repeats the above steps at this node until it reaches the target point, which is the most effective direct search method for solving the shortest path in static maps. The warehouse map built in this thesis is simple and regular, which is suitable for calculating the shortest path using A* algorithm.

The robot starts from the starting raster point and expands the surrounding raster at the current raster point. In this paper, the robot in the warehouse only moves up, down, left and right, so the four-neighborhood search method is chosen. The current position is called the parent node, and the estimated cost of the surrounding four directional grids is calculated and put into the candidate table for storing the extended nodes. After all the surrounding nodes finish estimating the cost, the grids with the smallest estimated cost are selected as the new moving position, and this position becomes the new parent node. The new parent node is used as the center to expand the grid, and this step is repeated until the robot reaches the target point. Finally, an optimal path is obtained from the starting point to the target point with the minimum cost.

The expression of the improved cost estimation function of the A* algorithm is:

$$f(n) = g(n) + h(n) + \sum_{j=1}^{m} t_{j(turn)} + \sum_{j=1}^{m} t_{j(wait)} \tag{1}$$

where $\sum_{j=1}^{m} t_{j(turn)}$ is the sum of the extra time spent by the robot in steering from the starting grid to the current grid, and $\sum_{j=1}^{m} t_{j(wait)}$ is the extra time spent by the robot in waiting in place due to path conflicts in the process from the starting grid to the current grid. $g(n)$ denotes the actual cost of moving from the starting grid to the current grid n. The actual cost is generally expressed by distance or time, and this paper uses time as a uniform scalar to compare the size of the cost of the function $f(n)$. $g(n)$ is expressed as:

$$g(n) = \frac{d}{v} \tag{2}$$

d is the actual distance traveled by the robot from the starting grid to the current grid n, and v is the speed at which the robot travels at a uniform speed. $h(n)$ denotes the heuristic estimation cost from the current grid n to the target grid point, and the expression is:

$$h(n) = \frac{d_n}{v} \tag{3}$$

d_n is the estimated shortest distance of the robot from the current raster n to the target raster point, where the estimated distance is calculated using the Harmattan distance:

$$d_n = abs(n.x - goal.x) + abs(n.y - goal.y) \tag{4}$$

The Harmattan distance is the sum of the horizontal and vertical distances of the current node n and the target point.

However, when the A* algorithm expands nodes in the parent node, there will be a situation that the expanded nodes are already occupied by other robots, which will cause vertex conflict among robots if not avoided. Adding a query step to the reservation table in the A* algorithm can achieve the purpose of avoiding the vertex conflict.

Meanwhile, since the expansion interval of the A* algorithm is a fixed value, it generally takes a raster as a unit and expands according to the raster. This will lead to the situation of edge conflict, i.e., the reservation table is not queried to have robots in the same position at the same time, but the run will collide at the raster junction, as shown in Fig. 3.

Time t

Time t+1
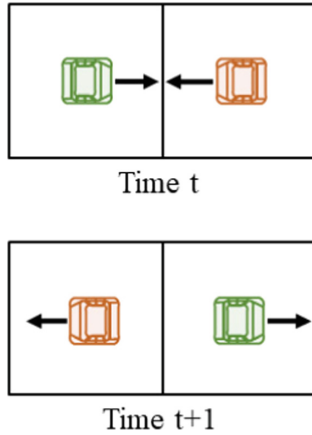
Fig. 3. Situations of edge conflict.

In order to avoid edge conflict, the reservation table is cross-called in the A* algorithm and brought into the decision formula. If the decision formula is satisfied, it is determined that this extended node will have edge conflict and the node needs to be reselected. The determination formula is as follows:

$$F_{point}(x,y) \in R_{(k+1)} \bigcap N_{point}(x,y) \in R_{(k)} \tag{5}$$

where $F_{point}(x,y)$ denotes the coordinates of the parent node, $N_{point}(x,y)$ denotes the coordinates of the extended node, R_k denotes the reservation table matrix at the moment of the parent node, and $R_{(k+1)}$ denotes the reservation table matrix at the moment of the extended node.

2.3   Heat-Map Algorithm

By improving the A* algorithm and the control of the reservation table, the collision problem is avoided, but only local conflicts can be solved, and the operation of the whole raster map path cannot be observed globally. As the number of robots increases, there may be local abnormal congestion while other locations in the warehouse are free, which greatly reduces the efficiency of the robot system in delivering goods.

For this reason, this thesis incorporates a heat map algorithm that reacts to the congestion level of the aisles, and takes the shelf groups and picking tables as

units, and reacts the congestion level of the surrounding aisles to the corresponding shelf groups and picking tables as the basis for selecting the target points and picking tables for new tasks, as shown in Fig. 4. By calling the reservation table data for a certain time period, the number of robots passing through each aisle during this time period is calculated and used as the congestion level value for each aisle, which is involved in the calculation of the heat value for each shelf group.
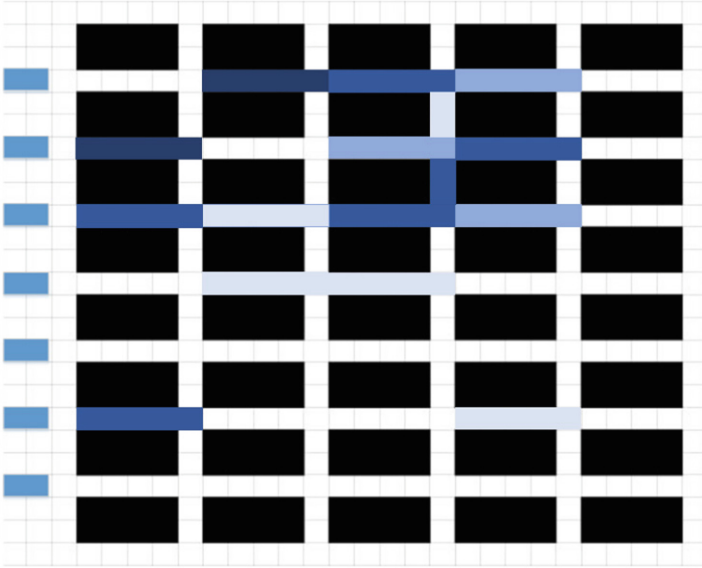
Fig. 4. Diagram of the heat map reflecting the level of congestion.

The congestion level of each lane is calculated by the following formula:

$$Congestion = \frac{N(k - RI \cdot \Delta t, k)}{RI} \qquad (6)$$

where $Congestion$ is the congestion level of the lane at the recording range of RI, and $N(k - RI \cdot \Delta t, k)$ is the number of robots passing through this lane from the moment of $k - RI \cdot \Delta t$ to the current moment of k.

To facilitate the calculation of the thermal value, the coordinates of the shelf groups and the lanes are numbered, both using the top-down and left-to-right numbering, with the coordinates of the first shelf at the top left being $(1, 1)$, and so on, to obtain the following formula for calculating the thermal value of each shelf group.

$$Wh(i,j) = C_w(i,j) + C_w(i,j+1) + C_l(i,j) + C_l(i+1,j) \qquad (7)$$

where $Wh(i, j)$ is the heat value of the shelf group with coordinate (i, j), $C_w(i, j)$ is the congestion degree value of the short aisle to the left of this shelf group, and $C_l(i, j)$ is the congestion degree value of the long aisle in front of this shelf group.

The unassigned tasks correspond to the shelf group where their target points are located one by one, and are sorted in ascending order based on the heat value of the shelf group where they are located; the smaller the heat value is, the more forward the task point is, and if the heat values are the same, the task point with the closest distance to the matching transport robot takes precedence. The reordered task list avoids orders for goods in shelves within the congestion range, as well as choosing to avoid picking stations with long waiting times, to achieve dynamic allocation.

3   The Simulation Results

For demonstrating the effectiveness of the proposed algorithm, the reservation table-based A* algorithm was compared with the improved A* algorithm based on the heat map algorithm (hereafter referred to as the optimized algorithm) designed in this paper in the warehouse raster map model created in this thesis. Simulation experiments are conducted using MATLAB, and the simulation experiments are as follows.

(1) Compare the situation of transporting different number of task orders with the same number of robots, the number of robots is set to 8, and the number of orders is 50, 100, 150, 200, 250 in order. Simulation results are shown in Fig. 5, compared with the A* algorithm, the efficiency of order completion under the optimization algorithm are greatly improved, the more the number of orders, the more obvious the efficiency improvement of the optimization algorithm, from 50 groups of tasks 7.91% improvement, gradually increasing to 19.36% improvement for 250 groups of tasks.

(2) Comparing different numbers of robots delivering goods under the same order, the number of goods is set to 100, and the number of robots is 2, 5, 8 and 12. simulation results are shown in Fig. 6, compared with the A* algorithm, the order completion time under the optimization algorithm are shortened, and the effect of the optimization algorithm gradually decreases with the increasing of robots number, from 38.28% improvement of 2 robots to 15.06% improvement of 12 robots, but it still has a significant lifting effect.
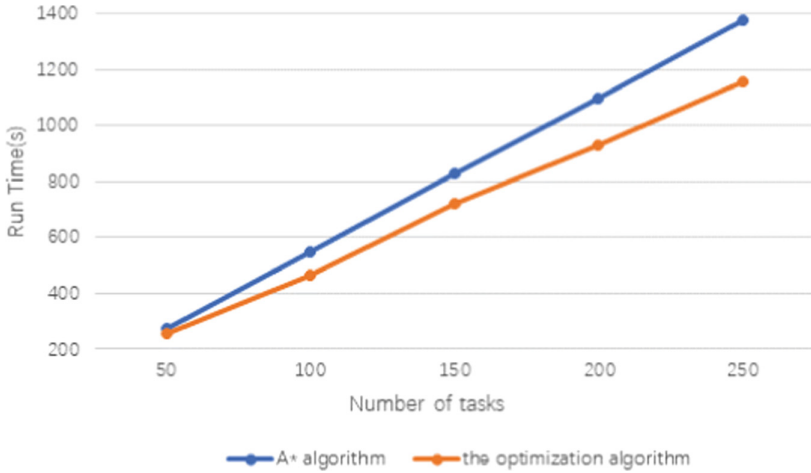
Fig. 5. Simulation results with different number of orders.
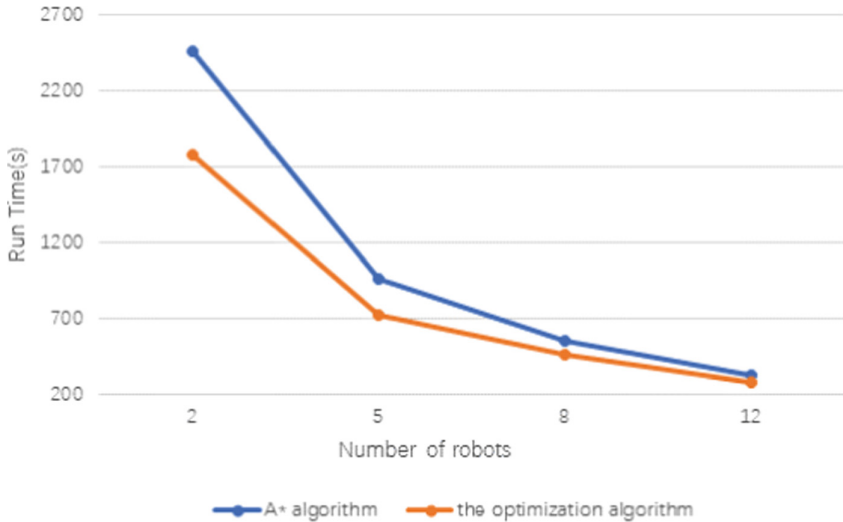
Fig. 6. Simulation results with different number of robots.

4 Conclusion

In this paper, a heat map based improved A* algorithm is proposed for the multi-vehicle scheduling problem in the intelligent storage system. The following work is accomplished: the warehouse raster map is established; the reservation table is improved, which greatly reduces the storage space occupied by the reservation table while improving the operation efficiency. The A* algorithm was improved to avoid vertex conflicts and edge conflicts at the same time. The heat map

algorithm is added to reasonably allocate tasks, avoiding congested areas and realizing the dynamic assignment of tasks. Finally, the feasibility and effectiveness of the scheme are verified by simulation.

References

1. Zhang, Y., Jun, Y., Wei, G., Lenan, W.: Find multi-objective paths in stochastic networks via chaotic immune PSO. Expert Syst. Appl. **37**(03), 1911–1919 (2010)
2. Zhang, Y., Lenan, W., Wei, G., Wang, S.: A novel algorithm for all pairs shortest path problem based on matrix multiplication and pulse coupled neural network. Digit. Sig. Process. **21**(04), 517–521 (2011)
3. Hidalgo-paniagua, A., Vega-rodrguez, M.A., Ferruz, J.: Solving the multi-objective path planning problem in mobile robotics with a firefly-based approach. Soft. Comput. **21**(04), 949–964 (2017)
4. Uzol, O., Yavrucuk, I., Sezer-Uzol, N.: Panel-method based path planning and collaborative target tracking for swarming micro air vehicles. J. Aircr. **47**(02), 544–550 (2010)
5. Li, S., Weihua, S., Guo, P., Zhang, S., Xie, P.: Research on SAR drone global path planning based on improved A* algorithm. Chin. Med. Equip. J. **41**(12), 16–20 (2020)
6. Chen, M., Qian, T., Zhang, S., Wang, J.: Obstacle avoidance and cooperative path planning method of warehouse logistics robot cluster. Mod. Electron. Tech. **42**(22), 174–177+182 (2019)
7. Zhang, Y., Lenan, W., Wang, S., Balochian, S.: UCAV path planning by fitness-scaling adaptive chaotic particle swarm optimization. Math. Probl. Eng. **2013**(03), 1–9 (2013)
8. Ouyang, H., Quan, Y., Gao, L., Zou, D.: Hierarchical path planning method for mobile robots based on hybrid genetic particle swarm optimization algorithm. J. Zhengzhou Univ. (Eng. Sci.) **41**(04), 35–40 (2020)

Workshop of Data Fusion
for Positioning and Navigation

Distributed Quaternion Kalman Filter for Human Tracking Using IMU and UWB Measurement

Jing Cao[1], Yunguang Wang[2], Wanjie Ren[3], Mingran Li[1], and Shuhui Bi[1(✉)]

[1] School of Electrical Engineering, University of Jinan, Jinan 250022, China
cse_bish@ujn.edu.cn
[2] State Grid Intelligent Technology Company Limited, Jinan 250000, China
[3] Shandong Institute of Non-Metallic Materials, Jinan 250000, Shandong, China

Abstract. For improving the accuracy of the indoor navigation system, a Kalman filter (KF) will be proposed to filter the quaternion and to obtain the acceleration. Then, the human body is tracked based on the distributed quaternion extended Kalman filter (DQEKF) by combing with inertial navigation system (INS)/ultra-wide band (UWB) technology. In the proposed algorithm, the used local data filter is composed by four sub-filters, the position information is used as the state vector, which is more effective in dealing with the noise. In the following, the outputs of the local filters are the inputs to the main filters for fusion and provide the best estimate. Finally, experimental results show that the proposed scheme can reduce the positioning error effectively.

Keywords: Quaternion · Inertial measurement unit · Ultra wide band · Indoor human localization

1 Introduction

With the development of science and technology, people's demand of accurate location information becomes more and more intense. For example, in a large store, customers can use positioning technology to find the location where they need to buy items and get the best route more easily [1]. Parents can use positioning information to locate their children in real time. Positioning technology is the basis of providing for various location services [6]. However, with the existing positioning technology, it is difficult to achieve the expected results of indoor positioning, and indoor metal components, electrical signals, obstacles, etc., will interfere with the positioning signal, resulting in inaccurate indoor positioning. For the positioning technology in the indoor environment, how to use the limited sensor information obtained to eliminate the influence of the complex indoor navigation environment on the obtained pedestrian position information has become a hot topic of research.

The Kalman filter (KF) was proposed in the last century and has made a remarkable contribution to the connection between cybernetics and information theory. Unlike traditional frequency-domain filtering, Kalman filtering is a time-domain state predictor. Since there will be some errors due to internal noise and external interference when collecting data, Filters can correct the data, and this process can also be seen as a filtering process. In recent years, KF has developed rapidly. When dealing with some nonlinear systems, the extended Kalman filter (EKF) comes into play, which is a KF that linearizes expectation and variance [5].

This paper studies a method that can effectively improve the positioning accuracy. First, KF filters the quaternion to calculate acceleration [3]. Then, distributed quaternion EKF (DQEKF) filtering is performed on the position and velocity position of the pedestrian. Then the output of the local filter is input to the main filter for fusion to eliminate the influence of the complicated indoor navigation environment on the obtained pedestrian position information and ensure the continuous stability of the pedestrian navigation information [4].

The remainder of this paper is arranged as follows. Section 2 discusses the pedestrian positioning scheme based on quaternion. The performance of the proposed algorithm is verified through simulation results in Sect. 3, and Sect. 4 summarizes this article.

2   Fusion Model

In this section, the indoor pedestrian integrated navigation scheme based on quaternion distributed filter will be designed in details. Figure 1 shows the block diagram of the distributed filter. For distributed filters, it includes four sub-filters and a main filter. In the structure, the sub-filter is used to estimate the system parameters. Among them, d_1, d_2, d_3, d_4 are the distance information collected by ultra-wide band (UWB). The output of the sub-filter is used as the input of the main filter to perform data fusion to obtain more accurate pedestrian position [2]. When the signal is collected, it will be affected by the interference of external signals and the internal noise of the device, and hence error and randomness happens in the received signal. In order to obtain the desired signal, the signal needs to be filtered to eliminate the disturbance.

This procedure includes two steps: KF is used to filter the quaternion, and DQEKF is designed to filter the position information of pedestrians.

2.1   Quaternion Filtering

A quaternion is used as state vector and the state equation is:

$$\mathbf{Q}_k^q = \mathbf{F}_{k-1}^q \mathbf{Q}_{k-1}^q + \mathbf{M}_{k-1}^q, \tag{1}$$

where $\mathbf{Q}_k^q = \begin{bmatrix} Q_1 & Q_2 & Q_3 & Q_4 \end{bmatrix}^T$, $\mathbf{M}_k^q$ is the noise of $\mathbf{Q}_k^q$ at time k.

The state matrix is:

$$\mathbf{F}_k^q = \left[\mathbf{I}_{4\times4} + \frac{\Delta t}{2} \mathbf{\Omega} \left(\mathbf{c}_k - \mathbf{v}_k \right) \right], \tag{2}$$

where $\mathbf{c}_k$ represents the current gyroscope measurement, and $\mathbf{v}_k$ is the average value of the gyroscope measurement in an attitude phase. The definition of $\mathbf{\Omega}\left(\mathbf{n}\right)$ is as follows:

$$\mathbf{\Omega}\left(\mathbf{n}\right) = \begin{bmatrix} 0 & -\mathbf{n}^T \\ \mathbf{n} & -\left[\mathbf{n}\times\right] \end{bmatrix}, \forall \mathbf{n} = \begin{bmatrix} n_1\ n_2\ n_3 \end{bmatrix}^T \in R^3, \tag{3}$$

$$\left[\mathbf{n}\times\right] = \begin{bmatrix} 0 & -n_3 & n_2 \\ n_3 & 0 & -n_1 \\ -n_2 & n_1 & 0 \end{bmatrix}. \tag{4}$$

The observation matrix is:

$$\mathbf{H}_k = \begin{bmatrix} 0 & -\left(\mathbf{g}_k - \left[\mathbf{A}\left(\mathbf{Q}_k^q\right)\right]^{-1}\mathbf{g}_k\right)^T \\ \mathbf{g}_k - \left[\mathbf{A}\left(\mathbf{Q}_k^q\right)\right]^{-1}\mathbf{g}_k & \left(\mathbf{g}_k - \left[\mathbf{A}\left(\mathbf{Q}_k^q\right)\right]^{-1}\mathbf{g}_k\right)\times \end{bmatrix}, \tag{5}$$

In the formula, $\mathbf{g}_k = \begin{bmatrix} 0\ 0\ G \end{bmatrix}$, G is the gravitational acceleration at the corresponding location. $\mathbf{A}\left(\mathbf{Q}_k^q\right)$ is the direction cosine matrix:

$$\mathbf{A}\left(\mathbf{Q}_k^q\right) = \left(r^2 - \mathbf{e}^T\mathbf{e}\right)\mathbf{I}_{3\times3} + 2\mathbf{e}\mathbf{e}^T - 2r\left[\mathbf{e}\times\right], \tag{6}$$

where r and $\boldsymbol{e}$ are the scalar part and vector part of the quaternion $\boldsymbol{Q}_k^q$ respectively, $\boldsymbol{I}_{3\times3}$ is the identity matrix.

Based on this model, the next step is to design a KF filter. $\boldsymbol{a}_k^b$ represents the acceleration in the carrier coordinate system, converted to the geographic coordinate system using the filtered quaternion:

$$\boldsymbol{a}_k^n = \boldsymbol{A}\left(\boldsymbol{Q}_k^q\right)\boldsymbol{a}_k^b. \tag{7}$$

2.2 Position Filtering

The DQEKF is used to filter the position and estimate the position. The state vector consists of position and velocity. The state equation is:

$$\mathbf{S}_k^l = \underbrace{\begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{A}_{k-1}^l}\mathbf{S}_{k-1}^l + \underbrace{\begin{bmatrix} \frac{\Delta t^2}{2} \\ \frac{\Delta t^2}{2} \\ \Delta t \\ \Delta t \end{bmatrix}}_{\mathbf{C}_{k-1}^l}\mathbf{a}_{k-1}^{l,n} + \mathbf{w}_k^l, \tag{8}$$

where $\mathbf{S}_k^l = \left[x_k, y_k, v_{x,k}, v_{y,k}\right]^T$ is the state vector of the l_{th} local filter at the time index k. At time k, $\left[x_k, y_k\right]$ represents the current pedestrian position, $\left[v_{x,k}, v_{y,k}\right]$ represents the current pedestrian speed vector, $\mathbf{a}_k$ represents the acceleration vector , the $\mathbf{w}_k^l \sim \mathcal{N}\left(0, \mathbf{B}^l\right)$ is the noise of the $\mathbf{S}_k^l$.

The observation equation is:

$$d_{i,k}^l = \sqrt{\left(\left(x_k\right) - \left(x_i\right)\right)^2 + \left(\left(y_k\right) - \left(y_i\right)\right)^2} = h\left(\mathbf{S}_k^l\right) + \nu_k^l, i = 1, 2, \dots \tag{9}$$
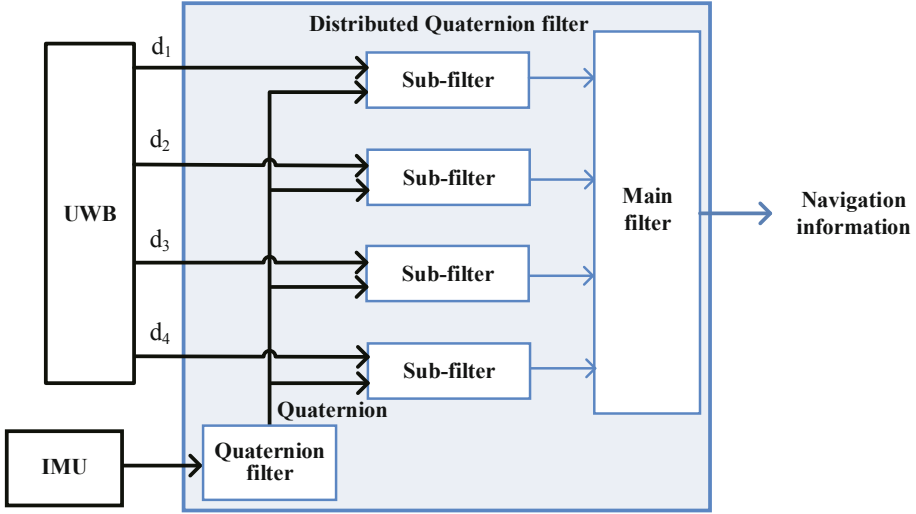
Fig. 1. Structure diagram of distributed filter.

Among them, (x_i, y_i) is the coordinate of the i-th reference node (RN), (x_k, y_k) are the location information of the target pedestrian in the east and north directions, $\nu_k^l \sim \mathcal{N}\left(0, \mathbf{R}^l\right)$ is the measurement noise.

Firstly, with the initial value $\mathbf{S}_0^l$ and $\mathbf{P}_0^l$, the $\mathbf{S}_{k+1}^l$ and $\mathbf{P}_{k+1}^l$ can be predicted at the time index k via Eqs. (10) and (11).

$$\mathbf{S}_k^l = \mathbf{A}_{k-1}^l \mathbf{S}_{k-1}^l + \mathbf{C}_{k-1}^l \mathbf{a}_{k-1}^{l,n} + \mathbf{w}_k^l, \tag{10}$$

$$\mathbf{P}_k^l = \mathbf{A}_{k-1}^l \mathbf{P}_{k-1}^l \left(\mathbf{A}_{k-1}^l\right)^T + \mathbf{B}_{k-1}^l, \tag{11}$$

where $\mathbf{P}_k^l$ represents the covariance matrix of $\mathbf{S}_k^l$. Then, the $\mathbf{S}_k^l$ and $\mathbf{P}_k^l$ can be updated at the time index k via Eqs. 12, 13 and 14.

$$\mathbf{K}_k^l = \mathbf{P}_k^l \left(\mathbf{H}_k^l\right)^T [\mathbf{R}_k^l + \mathbf{H}_k^l \mathbf{P}_k^l \left(\mathbf{H}_k^l\right)^T]^{-1}, \tag{12}$$

$$\mathbf{S}_k^l = \mathbf{S}_k^l + \mathbf{K}_k^l [\mathbf{d}_{i,k}^l - h(\mathbf{S}_k^l)], \tag{13}$$

$$\mathbf{P}_k^l = [\mathbf{I} - \mathbf{K}_k^l \mathbf{H}_k^l]\mathbf{P}_k^l, \tag{14}$$

where $\mathbf{H}_k^l = \frac{\partial \mathbf{h}(\mathbf{s}_k^l)}{\partial \mathbf{s}_k^l}$.

With the local EKF's output $\mathbf{S}_k^l$ and $\mathbf{P}_k^l$, the main filter works to provide the optimal output by fusing $\mathbf{S}_k^l$ and $\mathbf{P}_k^l$ by Eqs. (15) and (16).

$$\mathbf{S}_k = \mathbf{P}_k \left(\left(\mathbf{P}_k^1\right)^{-1}\mathbf{S}_k^1 + \left(\mathbf{P}_k^2\right)^{-1}\mathbf{S}_k^2 + \left(\mathbf{P}_k^3\right)^{-1}\mathbf{S}_k^3 + \left(\mathbf{P}_k^4\right)^{-1}\mathbf{S}_k^4\right), \tag{15}$$

$$\mathbf{P}_k = \left(\left(\mathbf{P}_k^1\right)^{-1} + \left(\mathbf{P}_k^2\right)^{-1} + \left(\mathbf{P}_k^3\right)^{-1} + \left(\mathbf{P}_k^4\right)^{-1}\right)^{-1}. \tag{16}$$

3 Experimental Testing

In this section, the performance of the above mentioned dual filters is verified through tests.

3.1 Experimental Environment

The indoor environment selected for this experiment is the lobby on the first floor of the Machinery Building of the West Campus University of Jinan, as shown in the Fig. 2. In which, four UWB reference nodes (UWB RNS) are used to make the measured data more accurate, so that the height of the UWB blind node (UWB BN) is consistent with the height of RNS. The UWB BN on the shoulder is used to receive RNS signals to measure the distance between them, and then process them to obtain the UWB trajectory. The target pedestrian is shown in Fig. 3, the data measured by the UWB and the encoder on the wheel are processed as a reference trajectory, the inertial measurement unit (IMU) on the foot is used to measure the quaternion data, and the computer is used to collect and process data.

Fig. 2. Test environment.
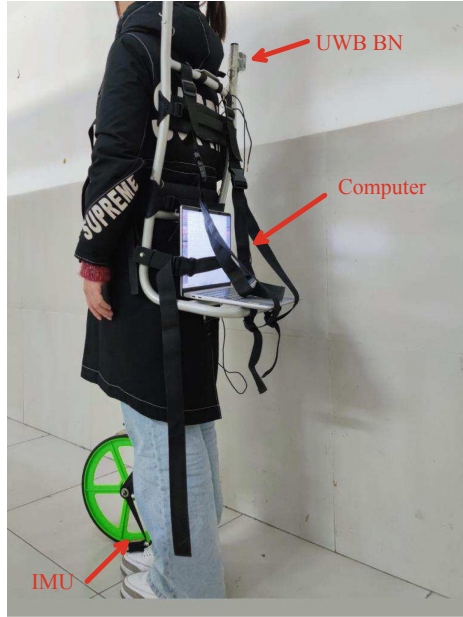
Fig. 3. Experimental equipment and target personnel.

3.2 Performance Analysis of the Proposed Algorithm

In this chapter, to verify the effectiveness of the proposed algorithm, it is compared with UWB and distributed EKF (DEKF) trajectories. The reference path and trajectory estimated by UWB, DEKF and DQEKF are shown in the Fig. 4, which shows that DQEKF can provide accurate path. In addition, the purple circle in the Fig. 4 represents the position of the UWB RNS, the triangle represents the starting position, and the square represents the end point. Figure 5 is given to depict the cumulative distribution function (CDF), and the Fig. 5 shows that the DQEKF has good estimation performance. To be more convincing, Table 1 lists the root mean square error (RMSE) of the estimated positions for UWB, DEKF, and DQEKF. It can be seen from Table 1 that the RMSE is larger than that of UWB due to the large error in the intermediate position of the DEKF experiment. The RMSE estimated by DQEKF is the smallest at 0.358 m and 0.297 m in the east and north positions, respectively.
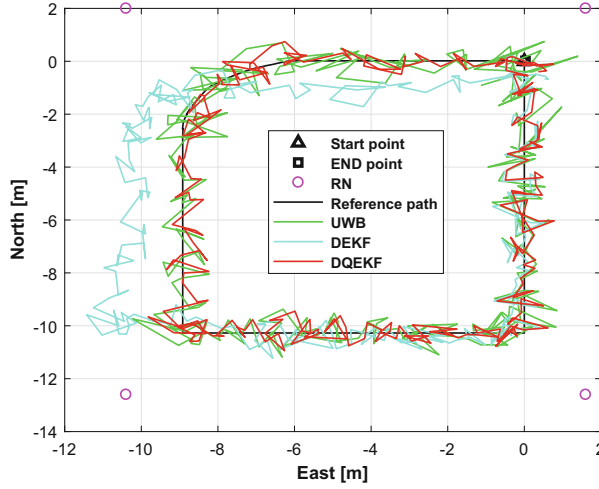
Fig. 4. The estimated path is provided by DQEKF, DEKF and UWB.
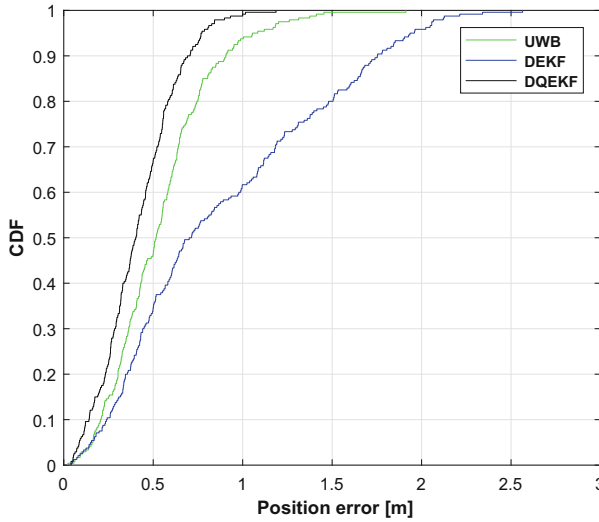
Fig. 5. The CDFs of the DQEKF, DEKF, and UWB.

Table 1. The RMSE (m) generated by UWB, DEKF and DQEKF

Filter	RMSE	
	East	North
UWB	0.474	0.385
DEKF	0.930	0.5058
DQEKF	0.358	0.297

4   Conclusion

This article studies how to perform distributed filtering on quaternion and location information. The acceleration is calculated from the filtered quaternion, and then the pedestrian's position and velocity information is used as the state vector, and the distributed EKF is used for filtering. In this algorithm, four EKFs are used as local data fusion filters, and the output of the local filters is used as the input of the main filter. The fusion obtains the optimal estimation, thereby obtaining more accurate pedestrian position information. The experiment test results show that comparing with the traditional model, the presented design model can effectively reduce the error.

References

1. Dardari, D., Closas, P., Djuric, P.M.: Indoor tracking: theory, methods, and technologies. IEEE Trans. Veh. Technol. **64**(4), 1263–1278 (2015)
2. 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)
3. Wen, K., Yu, K., Li, Y., Zhang, S., Zhang, W.: A new quaternion Kalman filter based foot-mounted IMU and UWB tightly-coupled method for indoor pedestrian navigation. IEEE Trans. Veh. Technol. **69**(4), 4340–4352 (2020)
4. Xu, Y., Chen, X.: Online cubature Kalman filter Rauch-Tung-Striebel smoothing for indoor inertial navigation system/ultrawideband integrated pedestrian navigation. Proc. Inst. Mech. Eng. **232**(4), 390–398 (2018)
5. 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)
6. Zhang, Z., Meng, X.: Use of an inertial/magnetic sensor module for pedestrian tracking during normal walking. IEEE Trans. Instrum. Meas. **64**(3), 776–783 (2015)

Robot Navigation in Crowds Environment Base Deep Reinforcement Learning with POMDP

Qinghua Li[1,3], Haiming Li[1,3], Jiahui Wang[1,3], and Chao Feng[2,3]([✉])

[1] School of Information and Automation, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, People's Republic of China
[2] International School for Optoelectronic Engineering, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, People's Republic of China
cfeng@qlu.edu.cn
[3] Jinan Engineering Laboratory of Human-Machine Intelligent Cooperation, Jinan 250353, People's Republic of China

Abstract. With the development of deep learning technology, the navigation technology of mobile robot based on deep reinforcement learning is developing rapidly. But, navigation policy based on deep reinforcement learning still needs to be improved in crowds environment. The motion intention of pedestrians in crowds environment is variable, and the current motion intention information of pedestrian cannot be judged by only relying on a single frame of sensor sensing information. Therefore, in the case of only one frame of input, the pedestrian motion state information is partially observable. To dealing with this problem, we present the P-RL algorithm in this paper. The algorithm replaces traditional deep reinforcement learning Markov Decision Process model with a Partially Observable Markov Decision Process model, and introduces the LSTM neural network into the deep reinforcement learning algorithm. The LSTM neural network has the ability to process time series information, so that makes the algorithm has the ability to perceive the relationship between the observation data of each frame, which enhances the robustness of the algorithm. Experimental results show our algorithm is superior to other algorithms in time overhead and navigation success rate in crowds environment.

Keywords: Deep reinforcement learning · Robot navigation · Partially observable Markov decision process

1 Introduction

With the development of technology that automatic driving and artificial intelligence, the application scene of the robot has been expanded from the industrial environment to the social environment of sharing activity space with human. Mobile robot navigation is used in factories, hospitals and shopping malls. These

tasks are still a challenging problem, because they require mobile robots to navigate safely and effectively in crowds environment [7,9,14].

Collision avoidance is an essential ability which the mobile robots navigate in crowds environment. Early works have proposed many methods which based on pedestrian motion model to dealing with the problem of mobile robot navigation in crowds environment. The pedestrian motion models have three main categories: social force model, data-driven and geometric approaches. The social force model [6,8] proposes a model of crowds interaction based on Gaussian process. These methods perform well in a crowds simulation, but they usually do not predict the movement of individual pedestrian. The data-driven approaches [1,13] can learn pedestrian dynamics from past trajectories. But these methods usually hard to obtain the required training data and the learning model may be not well generalized to different scenarios. The geometric approaches include the Reciprocal Velocity Obstacle(RVO) [2] and the Optimal Reciprocal Collision Avoidance(ORCA) [11]. This kind of methods through optimization the geometric feasible space to calculates the obstacle avoidance paths for multi-agents. But, these methods can not understand the diversity of human behavior, the movement trajectory of robot is short-sighted in time which lead to unnatural robot behaviors and create movement oscillatory in crowds environment [3].

With the development of technology which artificial intelligence and automatic driving. Robots encode features related to the interaction between the crowds or robots in the navigation policy and use neural network learning experience to understand crowds environment which produces paths that are close to human behavior through learning. Many researches have proposed the methods which motion planner base on deep reinforcement learning [10,15]. These methods learn policies from raw sensor input of the environment by reinforcement learning methods. However, it is difficult to extract the richer high-level representation of pedestrian intention in raw sensor information which makes them difficult navigation results in crowds environments. In order to dealing with this problem, some people proposed the deep reinforcement learning navigation algorithm based on the representation of pedestrian state which integrates pedestrian motion prediction into the decision making process to generate a path close to human behavior [3–5]. The algorithm extracts pedestrian state information from the original data as the input of reinforcement learning. The relevant features of the interaction between the crowds and robot are encoded into a fixed-length vector, which processes the state of each pedestrian in descending order according to the distance which between the robot and pedestrians. Although these methods have proved well results when working in a crowds environment, but there are still some limitations for robot navigation. These methods are based on single frame data and do not process the time series information, which will lead to short-sighted and make a detour of robots in crowds, because of they do not consider the future movement state of pedestrians.

In this work, we propose a new algorithm which can dealing with these previous shortcomings. Inspired by DRQN algorithm [12], our algorithm will consider: First, we incorporated the interactions between the observation data of adjacent

frames into the reinforcement learning network to overcome the short-sighted problem of robot trajectory time, and without using multi frame data class to predict the future pedestrian trajectory. Second, we use of the POMDP model to replace with the MDP model in deep reinforcement learning, which enhance performance whith robustness of the algorithm. Thirdly, we add the attention network into the neural network, infer the relative importance of the adjacent frame data relative to its future state through the attention network, so as to focus on the key frame data and improve the learning efficiency.

This paper is structured as following. In the second section, we introduces the related work of robot navigation algorithm. In the third section, we introduces the robot navigation base on deep reinforcement learning. Introduces the details of the P-RL algorithm in the fourth section. Introduces the experiments and result in the fifth section. Finally we concluded the algorithm.

2   Related Work

Early works have proposed many methods which based on pedestrian motion model to dealing with the problem of robot navigation in crowds environment. The Optimal Reciprocal Collision Avoidance(ORCA) [11] is the best performance algorithm in these algorithms. In this algorithm, the robot calculates the velocity space of other agents to avoid collisions with them, the robot can selects the optimal velocity in the intersection of all permitted geometric feasible velocity space. However, since the velocity space-based method does not consider the change of future state with agents, they will create oscillatory and unnatural behaviors in crowds environment.

The method of using pedestrian motion model to dealing with the problem of mobile robot in dense crowded is too depend on the human-engineered hyper parameters and rules which the effect is often poor. In order to dealing with these problems, deep reinforcement learning method has been widely studied in the field of robot navigation. In Deep reinforcement learning method, the robot use neural network learning experience to understand crowds environment which produces paths that are close to human behavior through learning.

There are a number of recent studies proposed the collision avoidance algorithm using deep reinforcement learning which integrates pedestrian motion prediction into the decision-making process to generate a path close to human behavior [3–5]. Chen proposed a collision avoidance algorithm [4,5], this methods extracts the pedestrian motion state information from the original data and takes the pedestrian motion state information as the input of reinforcement learning. The relevant features of the interaction between the crowds and robot are encoded into a fixed-length vector, which processes the state of each pedestrian in descending order according to the distance between the pedestrian and the robot. However, it is not reasonable to allocate importance according to distance, the pedestrian following the robot may not be as important as the farther pedestrian in front of it. A recent work proposed a approaches named SARL [3]. This method improved on previous work which uses the self-attention module

to allocate different importance weights to different parts of the crowds, so as to further improve the navigation performance.

Although these methods have proved well results when working, but there are still some shortcomings for robot navigation. Firstly, these methods are based on an idealized assumption that the motion states of robots and pedestrians are globally known and absolutely real. And, these methods are based on single frame data and do not process the time series information. It will lead to the robots short-sighted and make a detour in the dense crowds because of they do not consider the future movement state of pedestrians.

3 Robot Navigation Base Deep Reinforcement Learning

3.1 Problem Formulation

We designed a task of robot navigation that robot through the crowds and move to a random goal position. The navigation task of mobile robot in the crowds can be regarded as a decision process problem in deep reinforcement learning. We take crowds environment as the multiple agent problem to modeling between the robot and the crowds, where only robot agents' policy π is trainable, and the crowds agents' policy $\tilde{\pi}$ is designed to a unknown function that is modeled as a part of the environment.

In the past research, it is usually assumed that there is the robot and crowds in the X-Y plane of the 2D-workspace. They suppose each agent state can be observed, it include the position p, velocity v, orientation θ, goal position g, preferred speed v_{pref} and radius r, which are expressed as $p = [p_x, p_y]$, $v = [v_x, v_y]$, θ, r, $g = [g_x, g_y]$ and v_{pref}. The robot state at time t can be defined as $S_t = [p_x, p_y, v_x, v_y, \theta, g_x, g_y, r, v_{pref}]$, and the crowds state at time t of each human state can be defined as $O_t^i = [p_x^i, p_y^i, v_x^i, v_y^i, \theta, g_x^i, g_y^i, r^i, v_{pref}^i]$. In the real-world, it is difficult for sensors to perceive the absolute real motion state of pedestrian based on single frame data. So as to simulate the real-world environment, we remove the speed, goal position and preferred speed information of each crowds agent, and assume that the robot only observe the position, direction and radius of each crowds agent. After modify, the robot data of each frame become that states $S_t = [p_x, p_y, v_x, v_y, \theta, g_x, g_y, r, v_{pref}]$ and observation state $O_t^i = [p_x^i, p_y^i, \theta, r^i]$.

At each time step t, the robot observes a state of crowds which is include the robot state and the each humans motion states. It can be defined as $S_t^{jn} = [S_t, O_t^1, O_t^2, ..., O_t^n]$.

The reward $R(s_t^{jn}, a_t)$ is designed to excitation the robot when reach goal position and punish the robot close to humans or collisions humans.

$$
R(s_t^{jn}, a_t) = \begin{cases} -0.25 & d_{min} < 0, \\ -0.1 + d_t/2 & 0 < d_{min} < 0.2, \\ 1 & p_t = p_g, \\ 0 & \text{otherwise} \end{cases} \tag{1}
$$

where d_{min} is the minimum safe distance between humans and robot for a cycle, $p_t = [p_x, p_y]$ is the robot position at each time step t, $p_g = [g_x, g_y]$ is the goal position.

We expect to obtain an optimal policy π^* by deep reinforcement learning that maximizes the expectation of discounted total rewards, the optimal value function V^* of the state S_t^{jn} can be formulated as:

$$V^*(S_t^{jn}) = \sum_{t=0}^{T} \gamma^t \cdot R\left(s_t^{jn}, \pi^*\left(s_t^{jn}\right)\right), \tag{2}$$

where $\gamma \in [0, 1)$ is a discount factor.

Base on Bellman Equation, the optimal value function can get a optimal policy $\pi^*\left(s_t^{jn}\right)$ by the value iteration method. It can be derived as:

$$\pi^*(S_t^{jn}) = \underset{a_t \in A}{argmax} R(s_t^{jn}, a_t) +$$

$$\gamma^{\triangle t} \int_{S_{t+\triangle t}^{jn}} P(S_{t+\triangle t}^{jn}|S_t^{jn}, a_t) V^*(S_{t+\triangle t}^{jn}) \, dS_{t+\triangle t}^{jn} \tag{3}$$

where $\triangle t$ is the time of decision interval between two actions, A is the action space, $P(S_{t+\triangle t}^{jn}|S_t^{jn}, a_t)$ is a transition probability from S_t^{jn} to $S_{t+\triangle t}^{jn}$ when the action a_t is executed.

3.2   Partially Observable Markov Decision Processes

In the real environment, Markov property is difficult to hold. The Partially Observable Markov Decision Process model is an extension of Markov Decision Process in partially observable environment, it can better capture the dynamics of many real-world environments. Typically, the POMDP model can be defined as the tuple (S, A, T, R, Z, O).

S is a collection of real state s_t in the real environment.

A is a collection of all available actions of the robot, and $a_t \in A$ stand for the action which the robot take in time t.

T is the probability distribution of the agent transferring to other states s_{t+1} after executing action a_t at the state s_t, $T(s_t, a_t, s_{t+1}) = p(s_{t+1}|s_t, a_t)$.

R is the reward which represents the reward after the agent takes an action a_t, and it can be defined as $R(s_t, a_t) = r_t$.

Z is a collection of observation results, which is the environmental data obtained by the robot's sensors.

O is the probability distribution of receiving observation z_t after the agent take action a_t in the state s_t, $O(s_t, a_t, z_t) = p(z_t|s_t, a_t)$.

In the partially observable markov decision processes system, the system state is not completely known. We maintain a belief over possible states. It defined a

belief update function τ to estimate the current state.

$$b_t(s_{t+1}) = \eta O(s_t, a_t, z_t) \sum_{s_t \in S} T(s_t, a_t, s_{t+1}) b_{t-1}(s_t) \tag{4}$$

$$b_t = \tau(b_0, a_1, z_1, a_2, z_2, ..., a_t, z_t) \tag{5}$$

where η is a normalizing constant, b_0 is an initial belief, $s_t \in S$, $b(s_t) \geqslant 0$, and $\sum_{s_t \in S} b(s_t) = 1$.

4   P-RL

4.1   Overall Framework

In this paper, we inspired by the DRQN algorithm [12] and proposed a new algorithm. In our algorithm, we use of POMDP model to replace the MDP model in reinforcement learning, and introduces the LSTM neural network into the Value-Network. It deal with the problem of the robot navigation unsafe and detour in the crowds environments. This algorithm as outlined in Algorithm 1.

Algorithm 1: P-RL algorithms

1 Initialize Value-Network Q;
2 Initialize the observation space S, action space A;
3 Set a random goal position g;
4 **while** *Goal not reached do* **do**
5   Get a observation state of robot and crowds S_t^{jn};
6   Select $a_t = \underset{a_t \in A}{argmax} R(s_t^{jn}, a_t) + \gamma^{\triangle t \cdot v_{pref}} \cdot Q(s_t^{jn})$, Execute action a_t;
7   Obtain reward r_t and result observation s_{t+1}^{jn};
8 **end**

4.2   Value-Network

In order to make the algorithm have the ability to process time series, we introduce the LSTM neural network and attention module into the Value-Network. We divide the Value-Network into feature module, belief module and decision module, as show in Fig. 1.

Feature Module. We used the multi-layer perceptron neural networks(MLP1) to encode state of robot and crowd into a fixed length vector, and obtain the high-dimensional feature information of human-robot interaction ε_t.

$$\varepsilon_t = \varphi_\varepsilon(s_t^{jn}; W_\varepsilon) \tag{6}$$

where $\varphi_\varepsilon(\cdot)$ is an encoded function composed of a multi-layer perceptron with relu activations function, W_ε is the encode weights.
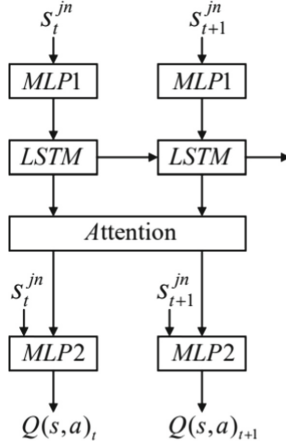
Fig. 1. The Value-Network is made up of the multi-layer perceptron neural networks with activations function Relu and long-short memory networks and attention module. It can be divided into feature module, belief module and decision module. The inputs is the robot and crowds motion state, and the outputs is a optimal value of the value function.

Belief Module. The module is made up of the LSTM neural network and the attention module. In this module, the LSTM neural network infers the relationship between each frame's feature information of human-robot interaction h_t. The attention module scored each frame, and weighted by the attention score α_t to find the keyframe. Since the module has strong inference ability to time series, we use it as the belief update function of POMDP model. The module input is the feature information of human-robot interaction for each frame, and output the corresponding belief b_t.

$$h_t = \psi_h(\varepsilon_t) \tag{7}$$

where $\psi_h(\cdot)$ is a time series inference function composed of a long-short memory network.

$$\alpha_t = \phi_\alpha(h_t; W_\alpha) \tag{8}$$

where $\phi_\alpha(\cdot)$ is an attention function composed of a multi-layer perceptron.

$$b_t = \sum_{t=0}^{T} softmax(\alpha_t)h_t \tag{9}$$

Decision Module. The module a multi-layer perceptron(MLP2). This module is a regression network, which is composed of multiple fully connected layers. It can be predicted the Q-value through observation state S_t^{jn} and belief value b_t.

$$Q_t = \mathcal{V}_Q(S_t^{jn}, b_t; W_Q) \tag{10}$$

where $\mathcal{V}_Q(\cdot)$ is a decision function composed of a multi-layer perceptron, W_Q is the decision weights.

5   Experiments

5.1   Simulation Setup

We built a simulation environment based on Gym which can simulation robot navigation in crowds environment. In this simulation environment, we used circles with the radius of 0.25 m to behalf of robots and crowds. Among them, robots for crowds are controlled by Optimal Reciprocal Collision Avoidance algorithm [11]. We assume that the robot and the crowd move freely in a space, in which the crowd will not actively avoid the robot. The robot and the crowd are randomly placed on a circular side length with a diameter of 4 m, and the opposite point on the side length over the center of the circle is set as the goal position. The goal of robots and crowd was to move from the start position to the goal position. So as to improve the training efficiency, we set the upper limit of robot navigation time to 25 s. It is defined as navigation failure, when the navigation time exceeds this upper limit.

5.2   Quantitative Evaluation

There are three most advanced methods, CADRL [5], LSTM-RL [4] and SARL [3], are implemented as base-line methods for this experiments. It is difficult to obtain the information of pedestrian speed in the simulated real world, so we delete the information of speed in the population state based on the baseline method. So as to assure the equity of the experiment, we run all algorithms in this environment.

We use the method with Temporal-Difference Learning to train the deep reinforcement learning algorithm. We create some data sets by ORCA [11] algorithm to pretreatment train the neural network before deep reinforcement learning training. We generate 2000 pretrain data with ORCA [11] to initialization of deep reinforcement learning. After pretraining, we train the algorithm for 4000 times of reinforcement learning. We use the method of dynamic greedy coefficient ε to improve the learning efficiency of deep reinforcement learning. At the initial stage of training, the greedy coefficient e is set to 0.5, which makes the algorithm more exploratory to generate more training data. With the increase of training times, we gradually reduce the greedy coefficient ε to 0.01, making the algorithm trust the trained neural network to promote the convergence of the algorithm.

Figure 2 shows the curves of success rate, collision rate, reward and cost time to reach goal in the simulated environment after 4000 rounds of training for CADRL [5], LSTM-RL [4], SARL [3] and P-RL.

Analyzing Fig. 2, we can conclude that in the training environment without pedestrian speed information, all base-line methods have been affected to a certain extent. Among them, CADRL [5] can not converge without pedestrian speed
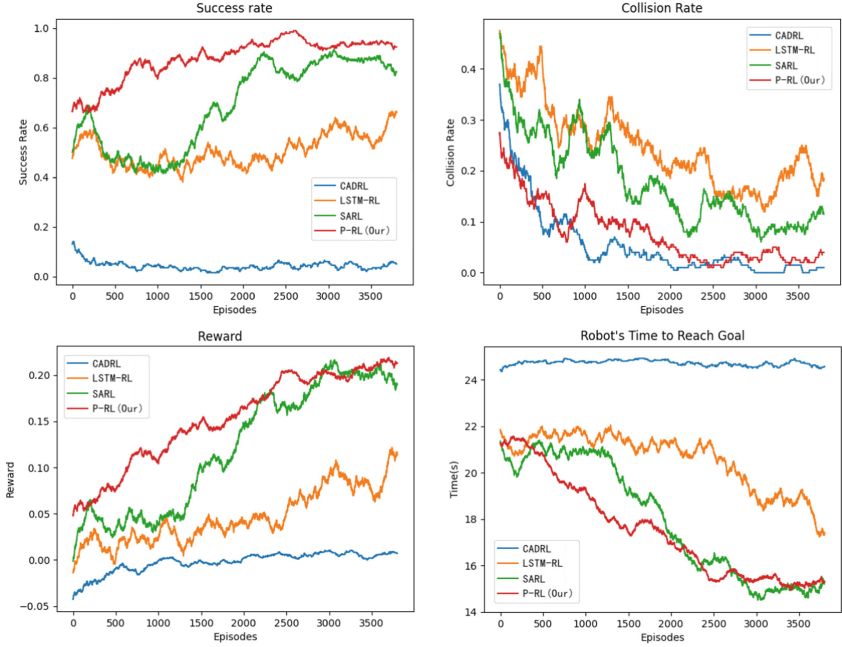
Fig. 2. The curves of success rate, collision rate, reward and cost time to reach goal in the simulated environment after 4000 rounds.

information. In collision rate, LSTM-RL [4] and SARL [3] can not achieve good convergence, and the collision rate remains at a high level. Although CADRL [5] has the lowest collision rate, but the success rate and cost time is basically zero. It can be seen that CADRL [5] maintains excellent collision avoidance ability, but it has lost its path planning ability, so it is unable to complete the robot mobile task. Compared with the base-line methods, our proposed algorithm P-RL can maintain better navigation performance even when the pedestrian speed information is lost. From the graph, we can see that our algorithm maintains high robustness in terms of Success rate, Collision rate and Navigation Time. And through the reward curve, we can see that our algorithm maintains a higher learning efficiency than base-line methods, it can learn better navigation strategies faster than base-line methods.

We used 500 random navigation tests to evaluate the model, using the average of 500 tests data as the evaluation benchmark, as show in Table 1.

Analyzing Table 1, we can conclude that in the partially known environment where the pedestrian speed state is lost, the data of the baseline algorithm can not meet the standard, and the success rate is lower than 0.8, so the robot navigation task can not be completed well. We provide the P-RL algorithm has strong robustness, it still can maintain a success rate of 0.9 in this environment, and the collision rate and navigation time are better than the baseline algorithm. This shows that our algorithm has better performance in robot navigation tasks.

Table 1. Table of the average of 500 random navigation tests data.

Method	Success rate	Collision rate	Navigation time
CADRL	0.10	0.00	15.80
LSTM-RL	0.71	0.06	13.01
SARL	0.79	0.16	12.45
P-RL (Our)	**0.90**	**0.05**	**12.21**

6   Conclusion

In this work, we present the P-RL algorithm to dealing with the problem of crowd navigation. The algorithm replaces deep reinforcement learning the MDP model with the POMDP model, and introduces the LSTM neural network into the deep reinforcement learning algorithm. The LSTM neural network has the ability to process time series information to overcome the short-sighted problem of robot trajectory time, and enhances the robustness of the algorithm. And, add the attention network into the neural network, infer the relative importance of the adjacent frame data relative to its future state through the attention network, so as to focus on the key frame data and improve the learning efficiency. Experimental results show our algorithm is preferable to other algorithms in time overhead and navigation success rate in crowds environment.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social LSTM: human trajectory prediction in crowded spaces. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 961–971 (2016)
2. van den Berg, J.P., Lin, M.C., Manocha, D.: Reciprocal velocity obstacles for real-time multi-agent navigation. In: 2008 IEEE International Conference on Robotics and Automation, pp. 1928–1935 (2008)
3. Chen, C., Liu, Y., Kreiss, S., Alahi, A.: Crowd-robot interaction: crowd-aware robot navigation with attention-based deep reinforcement learning. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 6015–6022 (2019)
4. Chen, Y.F., Everett, M., Liu, M., How, J.P.: Socially aware motion planning with deep reinforcement learning. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1343–1350 (2017)
5. Chen, Y.F., Liu, M., Everett, M., How, J.P.: Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 285–292 (2017)
6. Ellis, D., Sommerlade, E., Reid, I.D.: Modelling pedestrian trajectory patterns with gaussian processes. In: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pp. 1229–1234 (2009)
7. Ferdowsi, A., Challita, U., Saad, W., Mandayam, N.B.: Robust deep reinforcement learning for security and safety in autonomous vehicle systems. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 307–312 (2018)

8. Helbing, D., Molnár, P.: Social force model for pedestrian dynamics. Phys. Rev. E, Stat. Phys. Plasmas, Fluids Related Interdisc. Top. **51**(5), 4282–4286 (1995)
9. Kayukawa, S., et al.: BBEEP: a sonic collision avoidance system for blind travellers and nearby pedestrians. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (2019)
10. Long, P., Fan, T., Liao, X., Liu, W., Zhang, H., Pan, J.: Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 6252–6259 (2018)
11. Snape, J., van den Berg, J.P., Guy, S.J., Manocha, D.: Smooth and collision-free navigation for multiple robots under differential-drive constraints. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4584–4589 (2010)
12. Sorokin, I., Seleznev, A., Pavlov, M., Fedorov, A., Ignateva, A.: Deep attention recurrent q-network. arXiv:abs/1512.01693 (2015)
13. Vemula, A., Muelling, K., Oh, J.: Social attention: modeling attention in human crowds. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 1–7 (2018)
14. Wang, J., Meng, M.Q.: Socially compliant path planning for robotic autonomous luggage trolley collection at airports. Sensors (Basel, Switzerland) **19**, 2759 (2019)
15. Zhu, Y., et al.: Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 3357–3364 (2017)

Workshop of Intelligent Systems and Control

Stationery Recognition System Using Dual Cameras

Kun Qian[1], Mingzhe He[1(✉)] , Shuyi Chen[1], Manru Li[1], and Shengjun Shi[2]

[1] State Grid Shandong Electric Power Material Supply Company,
Jinan 250000, Shandong, China
`mingzhe.he@qq.com`
[2] HRG (Shandong) Intelligent Equipment Research Institute,
No. 1268 Gongye 4 Road, Jinan 250000, Shandong, China

Abstract. In order to improve the recognition accuracy of stationery placed in two layers in the material box and overcome the influence of different depth of field on the recognition accuracy of stationery, one stationery recognition system using dual cameras is proposed in this work. In this mode, we employ two cameras, one has a shallow depth of field, which is mainly used to identify the stationery placed on the upper layer of the material box. The other camera has a deep depth of field, which is mainly used to identify the stationery placed on the lower layer of the material box. When the stationery recognition system works, the two cameras works in parallel, and according to the comparison accuracy of two cameras for the same stationery, the one with higher accuracy is selected as the final output of the system. One real test has been done to verify the performance of the proposed system, it shows that the proposed system is able to accurately identify stationery at different levels.

Keywords: Stationery recognition · Two layers · Dual cameras

1 Introduction

Image recognition technology is the ways to identify the image automatically according to the image color characteristics, texture features, shape features and spatial relationship features, which employs the computer vision, pattern recognition, machine learning and other technical methods [4,5]. The earliest image recognition technology can be traced back to the 1960 s [1,3], With the development of computer technology and artificial intelligence, image processing has gradually developed to target recognition, target recognition, fingerprint recognition, etc. The image recognition technology used has also evolved from the earliest template matching and prototype matching to deep learning and support vector machine methods [2].

In this paper, in order to improve the recognition accuracy of stationery placed in two layers in the material box and overcome the influence of different depth of field on the recognition accuracy of stationery, one stationery recognition system using dual cameras will be designed and investigated. In this mode, we employ two cameras, it is mainly used to identify the stationery placed on the upper layer of the material box. The other camera has a deep depth of field, it is mainly used to identify the stationery placed on the lower layer of the material box. When the stationery recognition system works, the two cameras works in parallel, and according to the comparison accuracy of two cameras for the same stationery, the one with higher accuracy is selected as the final output of the system. One real test has been done to verify the performance of the proposed system, it shows that the proposed system is able to accurately identify stationery at different levels.

2   Dual Camera-Based Stationery Recognition Method

In this section, we will design the dual camera-based stationery recognition method. The structure of the stationery recognition system using dual cameras is shown in Fig. 1. From the figure, we can see that the stationery recognition system includes two camera lens, which is able to overcome the influence of different depth of field on image recognition caused by storing two layers of stationery in the material box. The upper camera lens is used to identify stationery placed on the upper floor, and the lower camera lens is used to identify stationery placed on the lower floor.

The identification flow chart of the dual camera-based stationery recognition method is shown in Fig. 2. From the figure, we can see easily that the the identification flows of the upper and lower camera lens have two modes: photo mode and image mode. Firstly, when the stationery recognition system start to work, both the upper and lower camera lens are working in parallel, they carry out the following operations at the same time:

– Determine whether the camera is needed to take photos.
– If it needs the camera to take photos, the mode is photo mode, in this mode, the camera is used to take pictures of the stationery, which is used to match in the next steps.
– If it does not need the camera to take photos, the mode is image mode, in this mode, the camera gets the image template from the data base, which is built off line.
– Match the picture and the template of the stationery.

After the above operations are completed, the recognition system selects which lens should be selected as the final result of the system according to the comparison of the results of two camera lens.

3   Test

In this section, we will investigate one real test to show the effectiveness of the proposed method. In this section, the setting of the test will be introduced. And the results of the tests will be investigated.
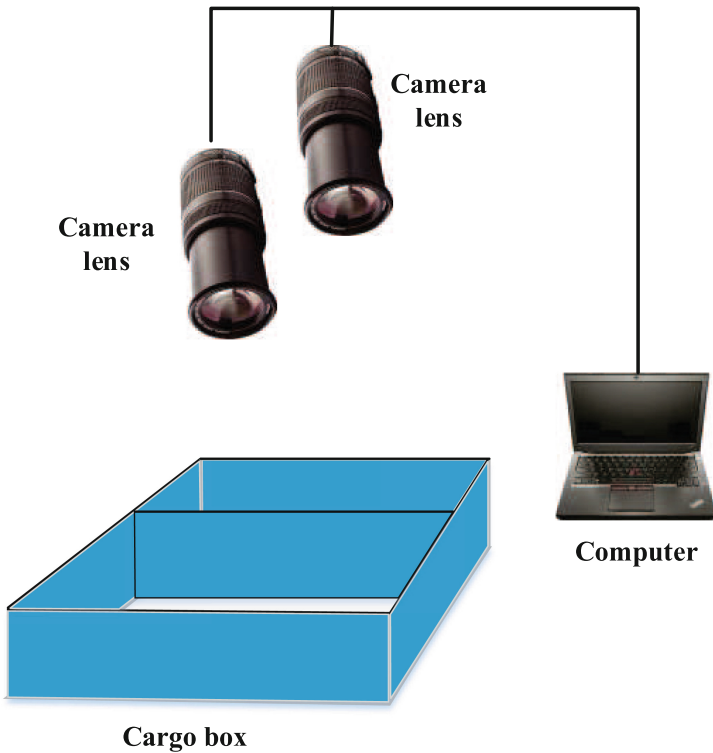
Fig. 1. The structure of the stationery recognition system using dual cameras.
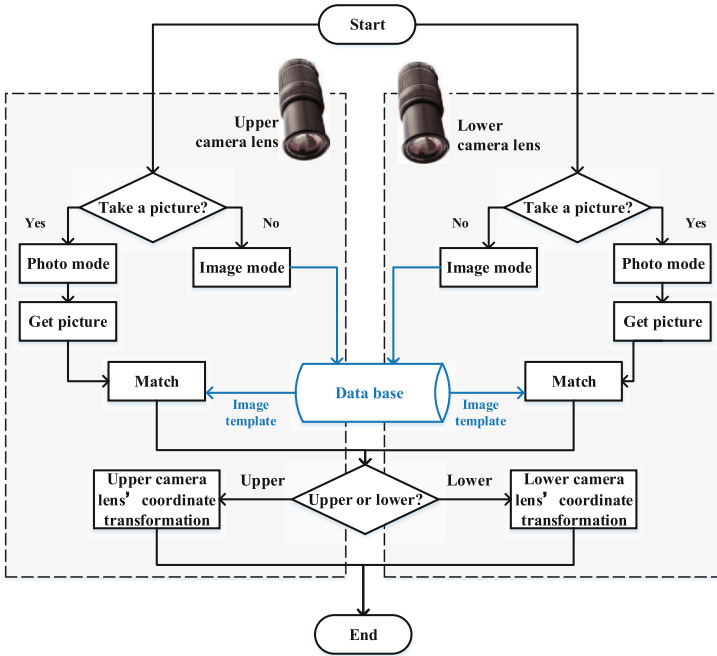
Fig. 2. The identification flow chart of the dual camera-based stationery recognition method.

3.1   The Setting of the Test

The real test has been done in the engineering training center of the Qilu University of Technology, Jinan, China. Figure 3 shows the real environment of the test. In this work, the test employs two industrial lenses: one is used as the upper lenses, the other one is used as the lower lenses, the two lenes can meet the requirements of different depth of field. Both the lenes are fixed on bracket. In this work, the HIKVISION MV-CE200-10GM and MV-CA050-11UM as the camera. And the computer is LAPTOP-OGCH3AJG, its CPU is Intel (R) Core(TM) i7-8550U CPU @ 1.80 GHz 2.00 GHz, and its RAM is 8.00 GB.
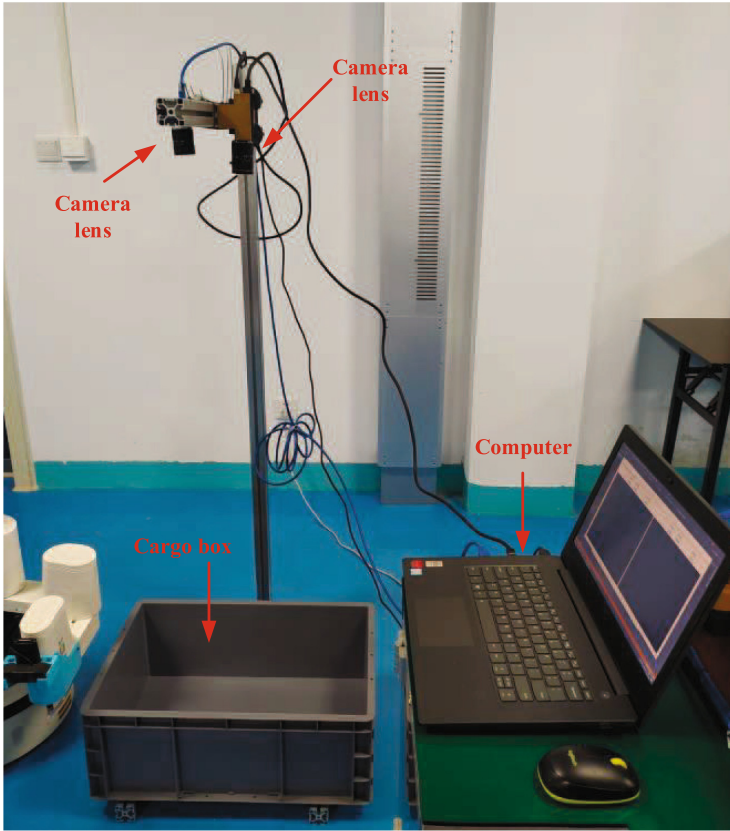
Fig. 3. The real environment of the test.

3.2 The Software of the System

In this work, we develop one software of the system, which is used to control the system to get the picture and identify the type of stationery. The Fig. 4 shows the main interface of software. From the figure, one can see easily that the main interface obviously includes the image display of two different lenses. The software of this system mainly includes camera debugging, template maintenance, network configuration and other functions. Before the system is enabled, we first need to complete the camera calibration using the software, which is shown in Fig. 5.

Fig. 4. The main interface of software.

Fig. 5. The software completes the calibration of the camera.

(a)
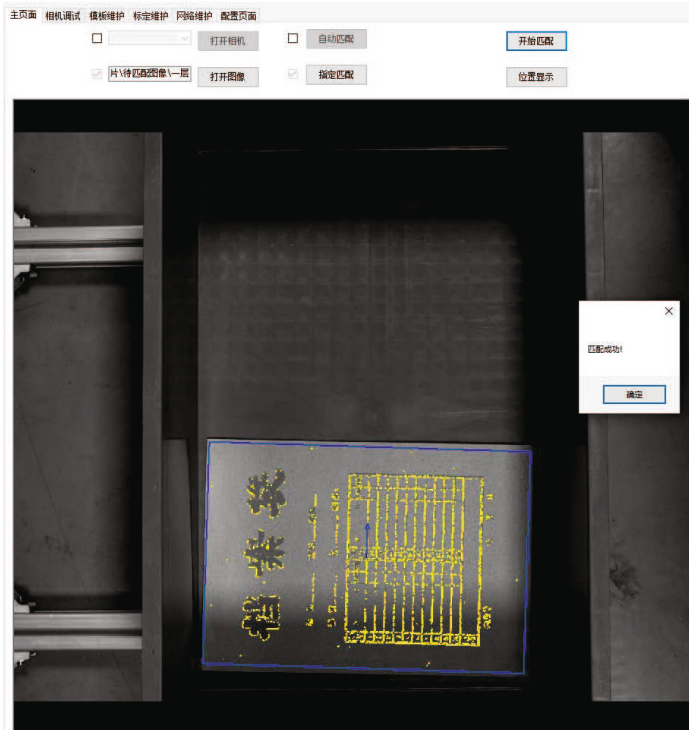
(b)

Fig. 6. The identification of the document bag at the lower level of the material box (a) the photo taken and (b) the identification result.

3.3 Stationery Identification

In this section, we will show the performance of the designed stationery identification system. In this section, we select two kinds of stationery: the file bag and the calculator. To the lower layer of material box, the identification of the document bag at the lower level of the material box is shown in Fig. 6. Here, Fig. 6 (a) is the photo taken and Fig. 6 (b) is the identification result. From the figure, we can see that the proposed system is useful to complete the identification of the document bag at the lower level of the material box (Fig. 7).

(a)
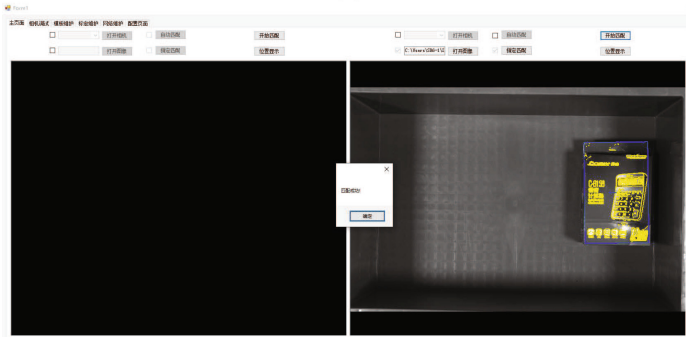
(b)

Fig. 7. The identification of the calculator at the upper level of the material box (a) the photo taken and (b) the identification result.

4 Conclusion

In this paper, in order to improve the recognition accuracy of stationery placed in two layers in the material box and overcome the influence of different depth of field on the recognition accuracy of stationery, one stationery recognition system using dual cameras will be designed and investigated. In this mode, we employ two cameras, it is mainly used to identify the stationery placed on the upper layer of the material box. The other camera has a deep depth of field, it is mainly used to identify the stationery placed on the lower layer of the material box. When the stationery recognition system works, the two cameras works in parallel, and according to the comparison accuracy of two cameras for the same stationery, the one with higher accuracy is selected as the final output of the system. One real test has been done to verify the performance of the proposed system, it shows that the proposed system is able to accurately identify stationery at different levels.

References

1. Andreopoulos, A., Tsotsos, J.K.: 50 years of object recognition: directions forward. Comput. Vis. Image Understand. **117**(8), 827–891 (2013)
2. Bi, P., Du, X.: Application of locally invariant robust PCA for underwater image recognition. IEEE Access **9**, 29470–29481 (2021)
3. Chen, Z., Ho, S.Y.: Computer vision for robust 3D aircraft recognition with fast library search. Pattern Recognit. **24**(5), 375–390 (1991)
4. Yao, Y., Wu, J., Lau, C., Wu, H., Jiang, F.: Reflection prediction of black silicon texture under the guidance of image recognition technology. IEEE J. Photovoltaics **11**(3), 600–605 (2021)
5. Zhou, H., Dai, Z.: Green urban garden landscape simulation platform based on high-resolution image recognition technology and GIS. Microprocess. Microsyst. **82**, 103893 (2021)

Lower Limb Posture Capture Using Quaternion Kalman Filter

Mingran Li[1] , Yuan Xu[1,4], Yanli Gao[2], Jidong Feng[1], and Guangchao Jin[3(✉)]

[1] School of Electrical Engineering, University of Jinan, Jinan 250022, China
[2] Qingdao Branch of Naval Aviation University, Qingdao 266000, China
[3] Jinan Central Hospital, Jinan 250000, Shandong, China
jgc3@163.com
[4] Shandong Beiming Medical Technology Co., Ltd, Jinan 250000, China

Abstract. Due to the accuracy requirements of the human body lower limb posture capture system, this paper proposes a quaternion Kalman filter-based human body lower limb posture capture method. Firstly, we employ the wearable inertial sensors to collect posture data of the lower limbs. Then, for the purpose of weakening the interference of the noise to the posture data of the sensors, the quaternion Kalman filter is designed. With the output of the quaternion Kalman filter, the lower limb joints' spatial position coordinates can be computed. The experimental results show that the proposed quaternion Kalman filter-based scheme can effectively reduce the attitude error, which make the attitude expression more intuitive and accurate.

Keywords: Quaternion · Kalman filter · Position calculation · Wearable inertial sensor

1 Introduction

In recent decades, due to the increasing aging of Chinese society and people's pursuit of high-quality life, patients with sports impairment have increasingly demanded effective rehabilitation training [1]. The inertial sensor posture capture system has the advantages of low cost, convenient operation and simple system composition, which is conducive to the use of civilians in rehabilitation training [4]. The advancement of wearable sensor technology provides an important breakthrough for clinicians and graduate schools engaged in rehabilitation medicine [3]. It integrates gyroscopes, accelerometers, and magnetometers. The accuracy will be interfered by drift errors, motion acceleration, and surrounding environmental magnetic fields, resulting in inaccurate attitude data output. How to ensure the accuracy of posture information has become a hot topic [8].

Kalman filter is a magic filter algorithm, it is a state estimation algorithm that combines prior experience and measurement update. Kalman filtering can be applied to any dynamic system with uncertain information to make basic

predictions for the next direction of the system. Once accompanied by various disturbances, the Kalman transform can always point out what actually happened. Recently, Kalman filtering has been implemented in many different waves [7].

The accuracy of using wearable sensors to obtain posture data may be affected by the environment, the joints tracked, and the type of exercise performed, resulting in insufficient accuracy. This article presents an algorithm for capturing the pose of the human body's lower limbs based on the quaternion Kalman filter. This method uses wearable inertial sensors to capture the posture data of the lower limbs, uses a quaternion Kalman filter as a local data fusion filter, and then uses the filtered posture quaternion to calculate the three-dimensional space position of the lower limb joints. After a lot of experimental verification, it can be determined that this scheme makes the acquired posture more accurate and intuitive.

The framework of this article is as follows: Sect. 2 designs the posture capture scheme of human lower limbs. Section 3 discusses the quaternion Kalman filter and the solution of lower limb joint positions. Section 4 verifies the quaternion Kalman filter through the results of semi-physical simulation. Section 5 summarizes this article.

2   Data Fusion Model

In this section, we will design the data fusion model used in this work. Since the quaternion has the advantages of no gimbal deadlock phenomenon, high efficiency, and convenient to use, we choose the quaternion as the attitude data of the inertial sensor. In order to collect posture information of the lower limbs, five wearable inertial sensors are used to place the abdomen, thigh and calf of the human body respectively, and then the posture information will be collected. The block diagram of the quaternion Kalman filter system is shown in Fig. 1. It includes 5 quaternion Kalman filter local filters, the filter outputs the posture quaternion $\mathbf{Q}$, and uses the output results to calculate the position of the lower limb joints to obtain accurate and intuitive lower limb posture.

Equation (1) represents the state equation of the filter in the fusion model, taking the attitude quaternion of each sensor as the state variable.

$$\mathbf{Q}(t)^l = \begin{bmatrix} 1 & -\frac{1}{2}\omega_x\,(t-1)\,T & -\frac{1}{2}\omega_y\,(t-1)\,T & -\frac{1}{2}\omega_z\,(t-1)\,T \\ \frac{1}{2}\omega_x\,(t-1)\,T & 1 & \frac{1}{2}\omega_z\,(t-1)\,T & -\frac{1}{2}\omega_y\,(t-1)\,T \\ \frac{1}{2}\omega_y\,(t-1)\,T & -\frac{1}{2}\omega_z\,(t-1)\,T & 1 & \frac{1}{2}\omega_x\,(t-1)\,T \\ \frac{1}{2}\omega_z\,(t-1)\,T & \frac{1}{2}\omega_y\,(t-1)\,T & -\frac{1}{2}\omega_x\,(t-1)\,T & 1 \end{bmatrix} \mathbf{Q}(t-1)^l + \mathbf{w}(t-1)^l,$$

(1)

where $\mathbf{Q}(\mathrm{t})^l = \begin{bmatrix} q_0 & q_1 & q_2 & q_3 \end{bmatrix}^T$ represents the state vector of the filter of the lth sensor at the time index t, $l \in (1, n)$, n=5. T represents the sampling period; $(\omega_x\,(t-1)\,, \omega_y\,(t-1)\,, \omega_z\,(t-1))$ represents the projection of the angular velocity from the n system to the b system in the b system at the time index t, that is output value of the gyroscope. $\mathbf{w}(t-1)^l \sim \mathcal{N}\left(0, \mathbf{G}^l\right)$ is the process noise.
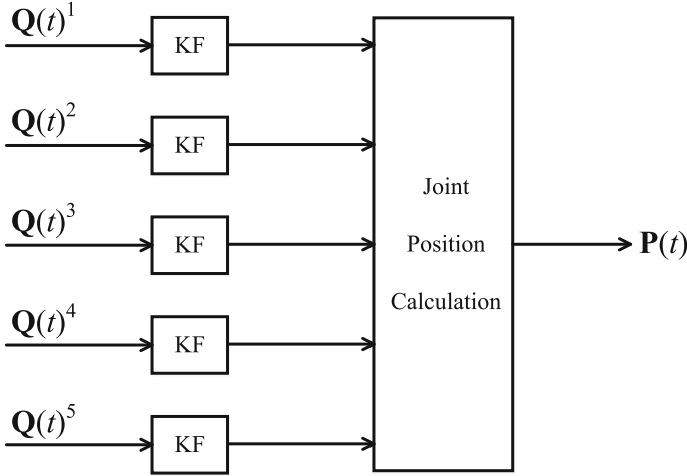
Fig. 1. Quaternion Kalman filter structure diagram.

The measurement equation used in this work can be written as:

$$\mathbf{Z}(t)^l = \mathbf{I}_{4\times4}\mathbf{Q}(t)^l + \mathbf{\Gamma}(t)^l, \tag{2}$$

where $\mathbf{Z}(\text{t})^l = \begin{bmatrix} q_0\ q_1\ q_2\ q_3 \end{bmatrix}^T$ represents the measured value of the lth sensor. $\mathbf{I}_{4\times4}$ represents the 4th order identity matrix. $\mathbf{\Gamma}(t)^l \sim \mathcal{N}\left(0, \mathbf{R}^l\right)$ is the measurement noise.

3   Quaternion Kalman Filter and Calculation of Lower Limb Joint Position

This section will design the quaternion Kalman filter based on model (1) and (2) as the filter of each sensor. First, we rewrite the model (1) (2) as follows:

$$\begin{cases} \mathbf{Q}(t)^l = \mathbf{F}(t-1)^l\mathbf{Q}(t-1)^l + \mathbf{w}(t-1)^l \\ \mathbf{Z}(t)^l = \mathbf{H}(t)^l\mathbf{Q}(t)^l + \mathbf{\Gamma}(t)^l \end{cases}, \tag{3}$$

In this model, the system parameter

$$\mathbf{F}(\text{t}-1) = \begin{bmatrix} 1 & -\frac{1}{2}\omega_x(t-1)T & -\frac{1}{2}\omega_y(t-1)T & -\frac{1}{2}\omega_z(t-1)T \\ \frac{1}{2}\omega_x(t-1)T & 1 & \frac{1}{2}\omega_z(t-1)T & -\frac{1}{2}\omega_y(t-1)T \\ \frac{1}{2}\omega_y(t-1)T & -\frac{1}{2}\omega_z(t-1)T & 1 & \frac{1}{2}\omega_x(t-1)T \\ \frac{1}{2}\omega_z(t-1)T & \frac{1}{2}\omega_y(t-1)T & -\frac{1}{2}\omega_x(t-1)T & 1 \end{bmatrix}$$

is used as the state transition matrix, and the attitude quaternion of the sensor is used as the State vector [5]. Choose attitude quaternion as the observation vector. Because the state quantity and observation vector are both attitude

quaternion, so the observation matrix is a 4th order identity matrix. $\mathbf{H}(t)^l = \mathbf{I}_{4\times 4}$. According to model (3), quaternion Kalman filter can be designed.

First, it is known that the attitude quaternion output by the sensor is used as the initial value $\mathbf{Q}(0)^l$ of the state estimator. Then, $\mathbf{Q}(t|t-1)^l$ can be estimated a priori on the time index t using Eq. (4) and Eq. (5).

$$\mathbf{Q}(t|t-1)^l = \mathbf{F}(t-1)^l \mathbf{Q}(t-1)^l, \tag{4}$$

$$\mathbf{P_Q}(t|t-1)^l = \mathbf{F}(t-1)^l \mathbf{P_Q}(t-1)^l \left(\mathbf{F}(t-1)^l\right)^T + \mathbf{G}^l, \tag{5}$$

where $\mathbf{Q}(t|t-1)^l$ represents the prior estimation of the state quantity at the time index of t, and $\mathbf{P_Q}(t|t-1)^l$ represents the predicted value of the covariance of the prior estimation error of the state quantity. $\mathbf{G}^l$ represents the process noise variance of the lth sensor. Then, Eqs. (6), (7) and (8)can be used to update the state quantity $\mathbf{Q}(t)^l$ on the time index t.

$$\mathbf{K}(t)^l = \mathbf{P_Q}(t|t-1)^l \left(\mathbf{H}(t)^l\right)^T \left[\mathbf{R}^l + \mathbf{H}(t)^l \mathbf{P_Q}(t|t-1)^l \left(\mathbf{H}(t)^l\right)^T\right]^{-1} \tag{6}$$

$$\mathbf{Q}(t)^l = \mathbf{Q}(t|t-1)^l + \mathbf{K}(t)^l \left[\mathbf{Z}(t)^l - \mathbf{H}(t)^l \mathbf{Q}(t|t-1)^l\right] \tag{7}$$

$$\mathbf{P_Q}(t)^l = \left[\mathbf{I} - \mathbf{K}(t)^l \mathbf{H}(t)^l\right] \mathbf{P_Q}(t|t-1)^l \tag{8}$$

where $\mathbf{K}(t)^l$ is expressed as the Kalman gain coefficient, which can fuse the measured value and the state quantity estimated a priori, and the result of the measured quantity and the state quantity estimated a priori can be weighed according to the magnitude of $\mathbf{K}(t)^l$. $\mathbf{R}^l$ represents the measurement noise variance of the lth sensor. $\mathbf{Q}(t)^l$ represents the posterior estimation of the state quantity at the time index t. $\mathbf{P_Q}(t)^l$ represents the updated value of the error covariance of the state quantity.

The filter outputs the attitude quaternion, which is expressed as the rotation from the navigation coordinate system to the carrier coordinate system, so the rotation matrix $\mathbf{C_n^b}$ can be expressed as:

$$\mathbf{C_n^b} = \begin{bmatrix} 2q_0^2+2q_1^2-1 & 2q_1q_2+2q_0q_3 & 2q_1q_3-2q_0q_2 \\ 2q_1q_2-2q_0q_3 & 2q_0^2+2q_2^2-1 & 2q_2q_3+2q_0q_1 \\ 2q_1q_3+2q_0q_2 & 2q_2q_3-2q_0q_1 & 2q_0^2+2q_3^2-1 \end{bmatrix}, \tag{9}$$

The coordinates of the knee joint and ankle joint are calculated based on the posture quaternion and limb size output by the quaternion Kalman filter and the position of the reference point. The detailed calculation method will be introduced below.

First, determine the reference system as the C system. This method uses the coordinate system of the sensor Xsens Dot where the abdomen is located as the reference coordinate system, that is, the C coordinate system. The joint coordinates calculated are all the position coordinates in this reference system.

Take the right leg joint as an example. This method requires that the direction of the right hip joint coordinate system is consistent with the direction of the reference coordinate system, and the coordinate system of the lower limb joints is required to be consistent with the coordinate system of the Xsens Dot sensor installed on the thigh and calf. Lp is the thigh length; Ld is the calf length. The direction of the sensor's own coordinate system is shown in Fig. 2, which conforms to the right-handed coordinate system [6]. The installation position of the sensor and the direction of the sensor and joint position coordinate system are shown in Fig. 3 [2].

Fig. 2. Sensor coordinate system.

Then the rotation matrix from the direction of the right knee joint coordinate system (k system) to the direction of the reference coordinate system (C system) can be expressed as:

$$\mathbf{C}_{k}^{C} = \mathbf{C}_{n}^{C}\mathbf{C}_{k}^{n} = \mathbf{C}_{n}^{C}(\mathbf{C}_{n}^{k})^{T}, \tag{10}$$

The rotation matrix from the direction of the right ankle joint coordinate system (a system) to the direction of the reference coordinate system (C system) can be expressed as:

$$\mathbf{C}_{a}^{C} = \mathbf{C}_{n}^{C}\mathbf{C}_{a}^{n} = \mathbf{C}_{n}^{C}(\mathbf{C}_{n}^{a})^{T}, \tag{11}$$
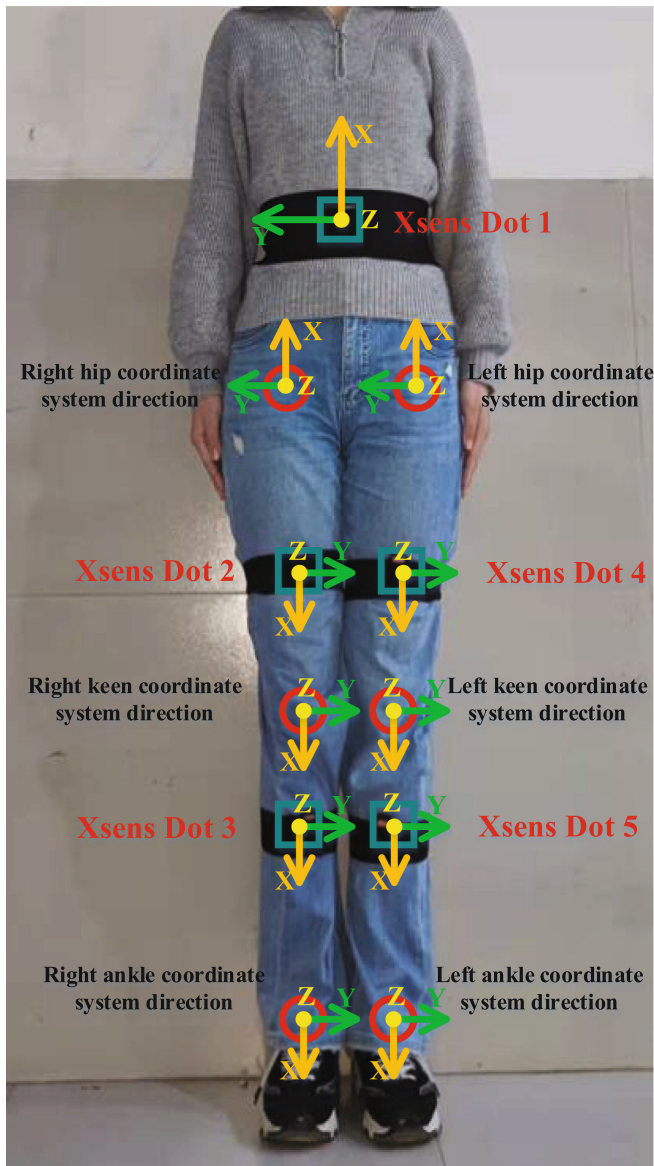
Fig. 3. The installation position of the sensor and the direction of the sensor and joint position coordinate system.

In the reference coordinate system (C coordinate system), through actual measurement, the three-dimensional space coordinates of the right hip joint position can be obtained:

$$\mathbf{P_h} = (x_h, y_h, z_h),\tag{12}$$

Since the X-axis direction of the knee joint position coordinate system coincides with the direction of the thigh, the right knee joint only has a component in the X-axis direction relative to the right hip joint point $\mathbf{P_h}$ [4]. Then in the reference coordinate system (C), the three-dimensional space coordinates of the right knee joint can be expressed as:

$$\mathbf{P_k} = \mathrm{Lp}\mathbf{C}_{k,1}^{C} + \mathbf{P_h}, \tag{13}$$

where $\mathbf{C}_{k,1}^{C}$ is the first column vector of the rotation matrix $\mathbf{C}_{k}^{C}$.

In the same way, in the reference coordinate system (C), the three-dimensional space coordinates of the right ankle joint can be expressed as:

$$\mathbf{P_a} = \mathrm{Ld}\mathbf{C}_{a,1}^{C} + \mathbf{P_k}. \tag{14}$$

where $\mathbf{C}_{a,1}^{C}$ is the first column vector of the rotation matrix $\mathbf{C}_{a}^{C}$.

The above are all steps to calculate the joint coordinates of the right leg, and the calculation method of the joint of the left leg is the same as above.

4   Experimental Test

4.1   Experimental Environment

Five wearable inertial sensors Xsens Dot were used in the experiment. The author placed the wearable inertial sensors at the reference point and the thigh and calf of the lower limbs, and tried to ensure that the sensors on the upper and lower legs were kept in the same straight line, and as far as possible to ensure that the sensor Xsens Dot at the reference point remained stationary. In the course of the experiment, with the assistance of a rehabilitation bicycle, the author simulates a patient with sports impairment to perform rehabilitation exercises for the lower limbs. The measured data environment and target personnel are shown in Fig. 4 and Fig. 5.

4.2   Performance Analysis of the Proposed Algorithm

For the convenience of observation, We compare the filtered pose quaternion-solved joint position results and the directly measured pose quaternion-solved joint position results with reference values.

Figure 6 and Fig. 7 are the distribution diagrams of the absolute error of the joint three-dimensional space position calculated by using the quaternion before and after filtering. The black point set represents the distribution of the absolute error of the joint three-dimensional space position calculated by the measured

Fig. 4. Experimental equipment and target personnel.
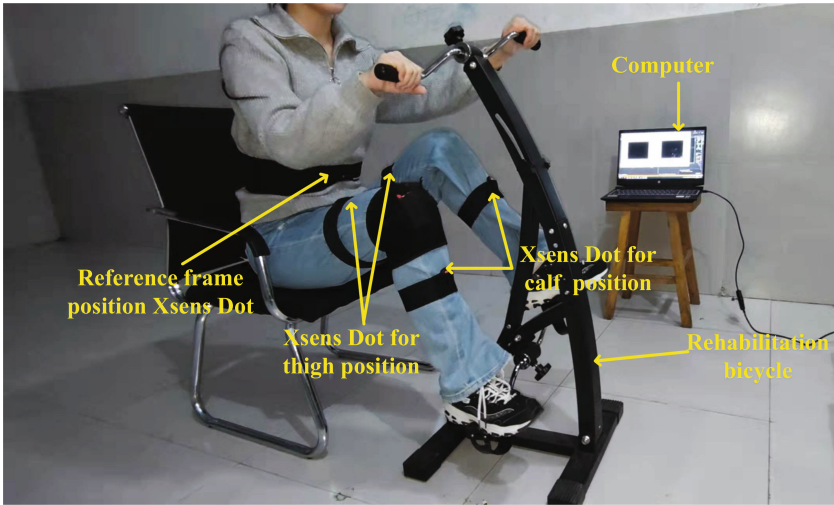
Fig. 5. Test environment.

value. The green point set represents the distribution of the absolute error of the joint three-dimensional space position calculated after filtering. It can be clearly seen that through filtering, the absolute error values of the X-axis, Y-axis, and Z-axis of the knee and ankle are significantly reduced. Therefore, it can be shown that the algorithm proposed in this paper has a good performance in reducing the absolute error.

Figure 8 and Fig. 9 are cumulative distribution function (CDF) plots of joint position errors before and after using the quaternion Kalman filtering method. As shown in the figure, the solid black line represents the CDF curve of the joint three-dimensional space position solved by the measured value. The green solid line represents the CDF curve of the joint three-dimensional space position calculated after filtering. It can be clearly seen from the figure that when y=0.9, the x value corresponding to the green solid line is smaller than the x value corresponding to the black solid line. That is to say, when the probability reaches 90 %, the joint position errors calculated by the filter value are significantly smaller than the joint position errors calculated by the measured value.

Table 1 shows the Root Mean Square Error (RMSE) before and after using the quaternion Kalman filter. It can be proved that the quaternion Kalman filter used in this experiment can effectively reduce the error of measuring joint position.
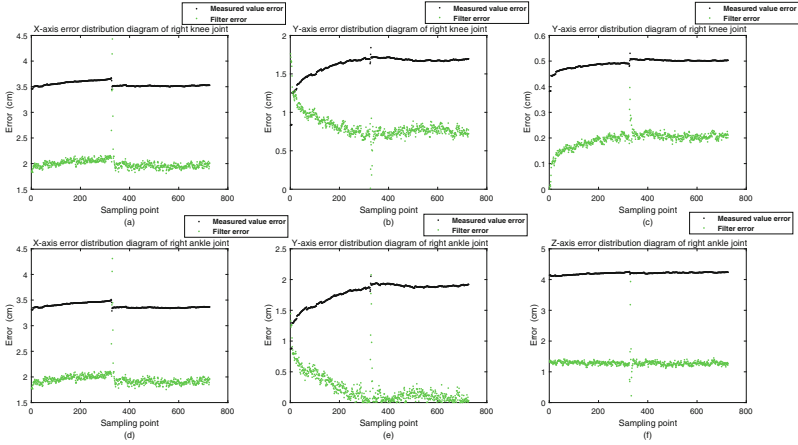
Fig. 6. Three dimensional position error distribution of right leg joint.
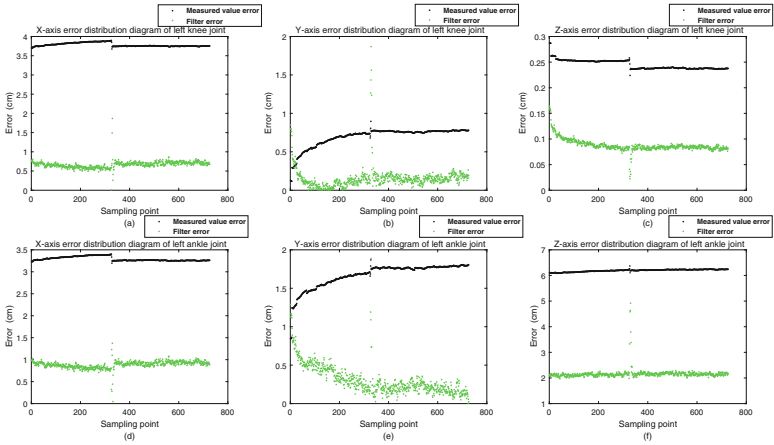
Fig. 7. Three dimensional position error distribution of left leg joint.
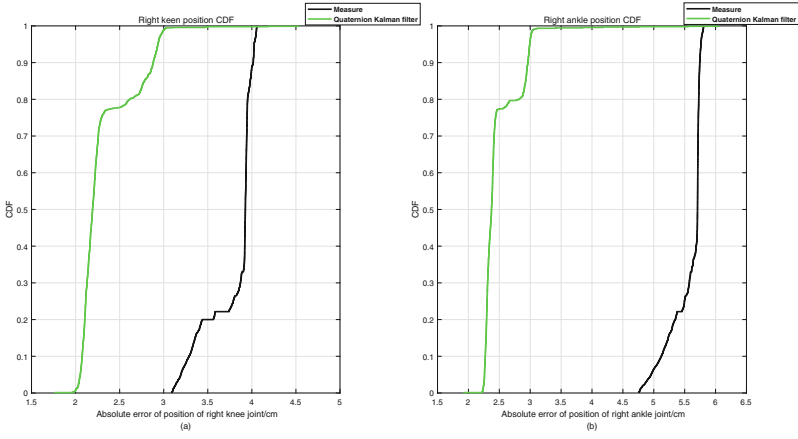
Fig. 8. CDF map of the three-dimensional space position of the right leg joint before and after filtering.
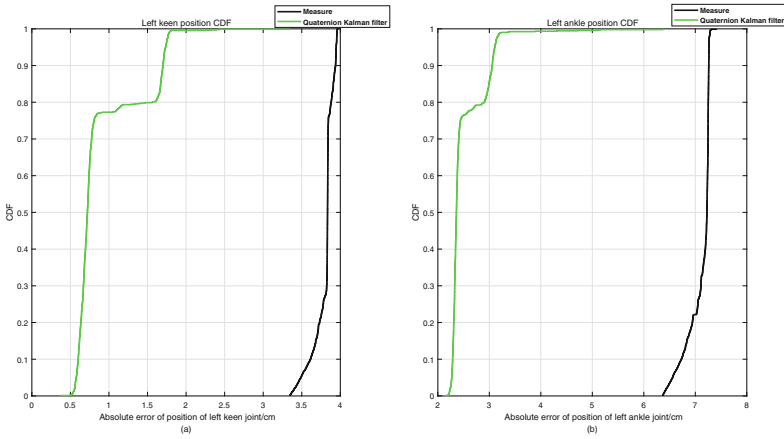
Fig. 9. CDF map of the three-dimensional space position of the left leg joint before and after filtering.

Table 1. Three-dimensional space position RMSE of joints before and after filtering.

Position	Aixs	Before filtering RMSE/cm	After filtering RMSE/cm
Right knee	X	3.4887	1.9565
	Y	1.4524	1.2612
	Z	0.45817	0.18342
Right ankle	X	3.3338	1.9025
	Y	1.5993	0.97097
	Z	4.1668	1.3026
Left knee	X	3.7244	0.7424
	Y	0.71415	0.6598
	Z	0.2622	0.12884
Left ankle	X	3.2437	0.95625
	Y	1.4911	0.90548
	Z	6.1471	2.1697

5   Conclusion

In this paper, quaternion Kalman filter is used to capture the lower limb posture. Firstly, using quaternion to express attitude can avoid the phenomenon of cardan deadlock, and the measurement noise can be reduced by filtering quaternion. Secondly, quaternion and limb length can be used to accurately calculate the three-dimensional position of the joint, so that the expression of human lower limb posture is more intuitive and easy to understand.

This method uses quaternion Kalman filters to filter the posture data output by the five sensors, and then the posture quaternion obtained before and after filtering was used to solve the three-dimensional space coordinate of the lower limb joint. Finally, the three dimensional position of the joint is compared with the reference value.

Through a large number of experiments in this article, it can be proved that through the quaternion Kalman filter and joint position calculation, more intuitive and accurate posture joint position results can be obtained.

References

1. Albán-Cadena, A., Villalba-Meneses, F., Pila-Varela, K.O., Moreno-Calvo, A., Almeida-Galárraga, D.: Wearable sensors in the diagnosis and study of parkinson's disease symptoms: a systematic review. J. Med. Eng. Technol. **45**(7), 1–14 (2021)

2. Cordillet, S., Bideau, N., Bideau, B., Nicolas, G.: Estimation of 3D knee joint angles during cycling using inertial sensors: accuracy of a novel sensor-to-segment calibration procedure based on pedaling motion. Sensors **19**(11), 2474 (2019)
3. Karina, L., Patrick, B., Hung, N., Christian, D.: Autonomous quality control of joint orientation measured with inertial sensors. Sensors **16**(7), 1037 (2016)
4. Li, S.A., Wang, Y.S., Xiang, R.P.: Human body motion capture system based on inertia sensing technology. Adv. Mater. Res. **722**, 454–458 (2013)
5. Pourtakdoust, S.H., Asl, H.G.: An adaptive unscented Kalman filter for quaternion-based orientation estimation in low-cost AHRS. Aircr. Eng. Aerosp. Technol. Int. J. **79**(5), 485–493 (2007)
6. Shen, S.Y., Fu, J., Xia, C.: Human upper limb motion capture system based on Xsens sensor. Elect. Mag. **06**, 16–18 (2019)
7. SzczSna, A., Pruszowski, P.: Model-based extended quaternion Kalman filter to inertial orientation tracking of arbitrary kinematic chains. SpringerPlus **5**(1), 1–13 (2016). https://doi.org/10.1186/s40064-016-3653-8
8. Xiang, Z.R., Zhi, J.Y., Xu, B.C., Li, J.: A review of motion capture technology and its applications. Appl. Res. Comput. **30**(8), 5 (2013)

Design of Infrared Spectrum Information Processing Algorithm for Fourier Infrared Spectrometer

Tuo Rui[1], Ren Wanjie[1], Hu Guoxing[1(✉)], Cai Chen[1], Lin Shuai[1], and Zhao Huan[2]

[1] Shandong Institute of Nonmetallic Materials, Jinan 250031, Shandong, China
sdjnhgx@163.com

[2] Army Armament Department Military Representative Office in Ji'nan, Jinan 250031, Shandong, China

Abstract. Fourier transform infrared spectrometer has a wide range of applications in many fields. In order to ensure the spectral quality of the spectrometer output, it is necessary to perform certain processing on the original spectrum. After having developed the Fourier transform infrared spectrometer, in this paper we design the infrared spectrum information processing algorithm. The basic transformations of the spectrum, such as spectral derivation, spectral normalization, centralization, and normalization, are realized. The methods of wavelet transform and S-G smoothing filtering are used to filter out the noise. By means of multivariate scattering correction method, the baseline shift and offset phenomenon of the infrared spectrum of the sample are corrected. Combining principal component analysis and Mahalanobis distance, a detection method of abnormal samples is proposed. Through the combination of multiple data processing algorithms, the processed spectra can play a better role in subsequent spectral analysis.

Keywords: Fourier transform infrared spectrometer · Spectral information processing · Filtering and denoising · Baseline correction

1 Introduction

Fourier transform infrared spectrometer can carry out qualitative and quantitative analysis of samples, and has been widely used in many fields such as medicine, chemical industry, geologic mining and so on [1, 2]. In the process of infrared spectrum signal acquisition, it may be affected by factors such as the state of the spectrometer, acquisition background, detection conditions, etc., resulting in interference in the measured spectrum [3, 4], such as noise interference. Since the background is collected every time while collecting a sample spectrum, the change of the background causes the spectrum to have a baseline drift phenomenon. Other factors such as abnormal sample interference and light scattering will also reduce the accuracy and stability of the model. Therefore, preprocessing the spectral data is a key step to ensure the output performance of the Fourier transform infrared spectrometer. In this paper, the processing of spectral

information mainly includes the basic transformation of the spectrum, the filtering of redundant noise interference and other irregular influencing factors, such as baseline drift caused by background interference during acquisition, noise interference of instruments and detection environments, and abnormal detection in spectra.

In order to remove the redundant noise interference of the spectrum, McClure et al. made a detailed study on the influence of the random noise of the spectrum on the model [5]. They confirmed that the random noise superimposed on the spectral signal will deteriorate the accuracy of the model. For the influence of baseline drift, baseline shift and uneven particle distribution on the spectrum, the commonly used solutions are the first derivative, the second derivative and the multivariate scattering correction [6]. The existence of abnormal sample data will affect the predictive ability of the model and cause deviations in the prediction. Commonly used methods for identifying abnormal samples include Mahalanobis distance method and principal component analysis, and the combination of partial least squares principal component score and Mahalanobis distance. The Mahalanobis distance method was used to identify the abnormal value of the leaf spectrum of Junzao [7]. During the sample spectrum collection process, the collected spectrum inevitably has interference due to the instrument, the sample itself or other reasons. Using the original spectrum directly will lead to poor model accuracy and instability. The sample spectrum information can be processed according to the research experience and the characteristics of the sample. This research will carry out the design of infrared spectrum information processing algorithm for our developed Fourier transform infrared spectrometer. The methods of spectral preprocessing include derivation, standard normal transformation, smoothing and filtering, multivariate scattering correction, etc. In actual processing, various methods will be combined in a certain order according to specific conditions.

2   Basic Transformation of Infrared Spectrum

2.1   Spectral Derivation

Spectral derivation is one of the commonly used preprocessing methods in infrared spectroscopy, which can eliminate baseline drift and improve spectral resolution. The direct difference method is used for the derivation of the spectrum. As a discrete spectrum derivation method, for the discrete spectrum x_i, $i = 1,\ldots, n$, the first order derivative and second derivative spectra at the wavelength i and the difference width g are calculated according to the following methods. The formula for the first derivative is

$$x_{i,1st} = \frac{x_i - x_{i+g}}{g} \tag{1}$$

And the second order derivative formula is

$$x_{i,2nd} = \frac{x_i + x_{i+2g} - 2x_{i+g}}{g^2} \tag{2}$$

Taking $Mn_3Al_2(SiO_4)_3$ material as an example, Fig. 1 shows the original spectrum and its first derivative, and Fig. 2 shows the original spectrum and its second derivative.
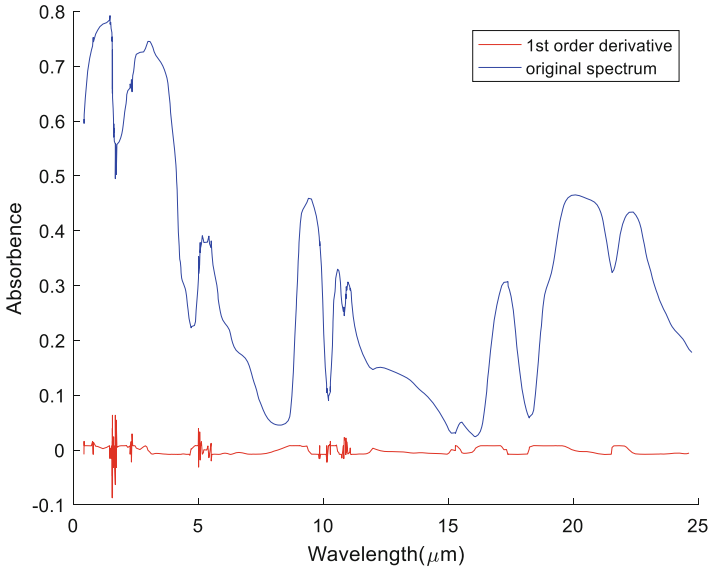
Fig. 1. Infrared spectrum of $Mn_3Al_2(SiO_4)_3$ and its first order derivative.
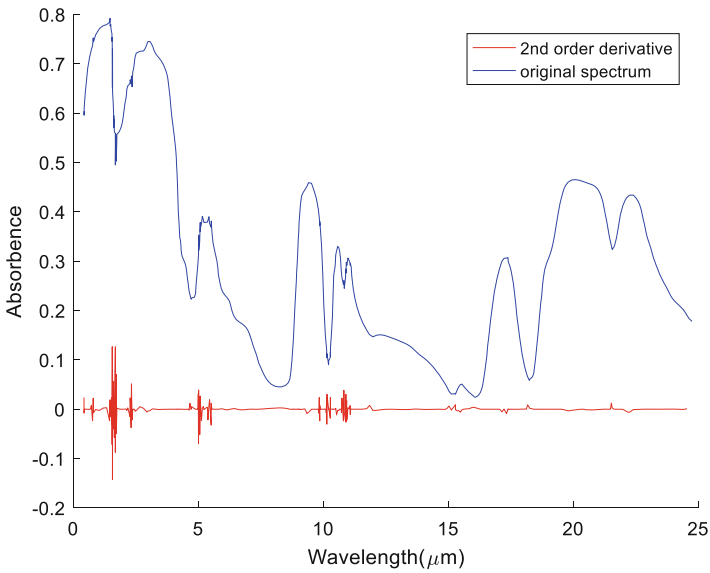
Fig. 2. Infrared spectrum of $Mn_3Al_2(SiO_4)_3$ and its second order derivative.

It can be seen that for the spectrum with high resolution and many wavelength sampling points, the derivative spectrum obtained by the direct difference method can meet the requirements. However, for the spectrum of sparse wavelength sampling points,

the derivative obtained by this method will have a large error. In this case, the Savitzky-Golay convolution derivation method can be used for calculation.

2.2 Normalization

When using infrared spectroscopy, it is necessary to correlate the characteristics of the spectrum with the properties or structural characteristics of the sample to be tested. Therefore, it is often necessary to use data enhancement algorithms to reduce or eliminate some redundant information. Commonly used algorithms include centralization, standardization and normalization. The main function of normalization is to normalize the ordinate of the spectrum, which is convenient for quantitative analysis of infrared spectrum. For absorbance spectra, the absorbance of the maximum absorption peak after normalization was normalized to 1 and the baseline was normalized to 0.

The specific calculation formula is shown in formula (3), where x is the absorbance corresponding to a certain wavelength, x_{min} is the minimum value of absorbance among all absorbance values in the spectrum, x_{max} is the maximum value of absorbance among all absorbance values in the spectrum, and x^* is the normalized absorbance value after processing, being between [0,1].

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3}$$

Figure 3 shows the result of normalizing the original infrared spectrum of $Mn_3Al_2(SiO_4)_3$ material.
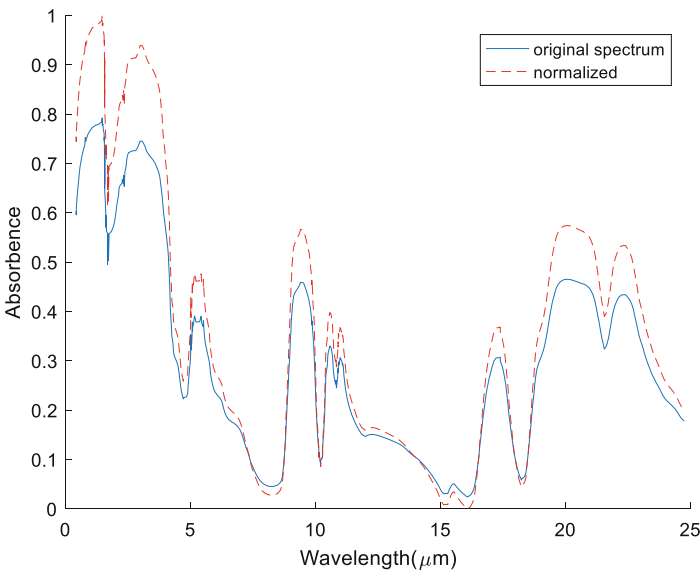
Fig. 3. Infrared spectrum and its normalization.

2.3   Centralization

Centralization, also known as zero-average, is the basic transformation of the infrared spectrum. It mainly completes the translation of the spectrum and moves it to the position with 0 as the center point. By subtracting the average value of all the data from each spectral data, the average value of the spectral data after centering is 0, and the variance is not limited. The centralization enables all spectral data to be distributed on both sides of the zero point, fully reflecting the change information, and effectively removing the impact of changes caused by objective factors such as temperature or human operation on the spectral data. The specific calculation formula is shown in formula (4), where x is the ordinate value corresponding to a certain wavelength, and μ is the average value of the ordinate corresponding to all wavelengths of the spectrum.

$$x^* = x - \mu \tag{4}$$

Still taking the $Mn_3Al_2(SiO_4)_3$ material in Fig. 1 as an example, after centering the original infrared spectrum, the spectrum is shown in Fig. 4.
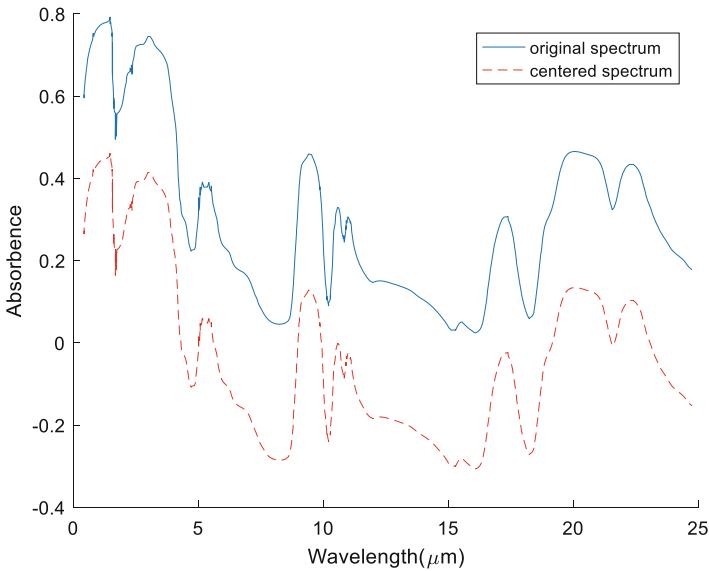
Fig. 4. Infrared spectrum and its centralization.

2.4   Standardization

As one of the basic transformations of infrared spectroscopy, standardization maps the data to a standard normal distribution with a mean of 0 and a standard deviation of 1. On the basis of data centralization, the data is divided by the standard deviation of all spectral data to make it satisfying the standard normal distribution.
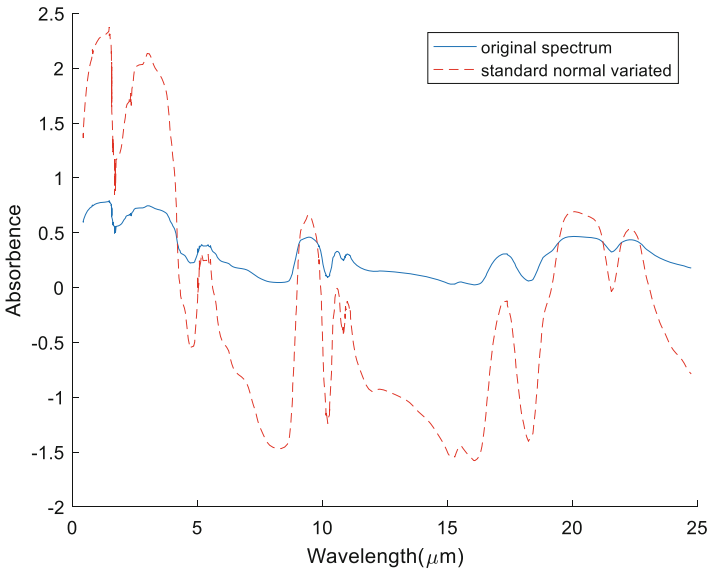
Fig. 5. Infrared spectrum and its normalization.

$$x^* = \frac{x - \mu}{\sigma} \tag{5}$$

The specific calculation of standardization is shown in formula (5), where x is the original spectral data, μ is the average value of all spectral data, σ is the standard deviation of all spectral data, x^* is the value after normalization, which obeys the standard normal distribution $x^* \sim N(0, 1)$. For the $Mn_3Al_2(SiO_4)_3$ material in Fig. 1, the normalized spectrum is shown in Fig. 5.

3 Filtering and Denoising of Infrared Spectrum

The output of infrared spectrometer not only contains useful information, but also super-imposes random errors, such as noise. The methods of noise filtering include Kalman filter, wavelet analysis, wavelet packet transform, smooth noise filtering and so on. In this paper, wavelet packet transform and S-G convolution smoothing are used to realize the filtering and denoising of infrared spectrum.

3.1 Wavelet Packet Filter Denoising

Wavelet Packet Transform (WPT) has higher accuracy and flexibility in signal analysis than wavelet transform, and has finer local analysis capabilities. Wavelet transform is mainly used for signal noise filtering, data compression and model transfer, while wavelet packet analysis is mainly used for signal noise removal [8]. As shown in Fig. 6,

the wavelet packet transform can not only decompose the low-frequency part of the signal, but also the high-frequency part. This decomposition has neither redundancy nor omissions, so it contains a lot of medium and high frequency information. Signals are able to perform better time-frequency localized analysis.
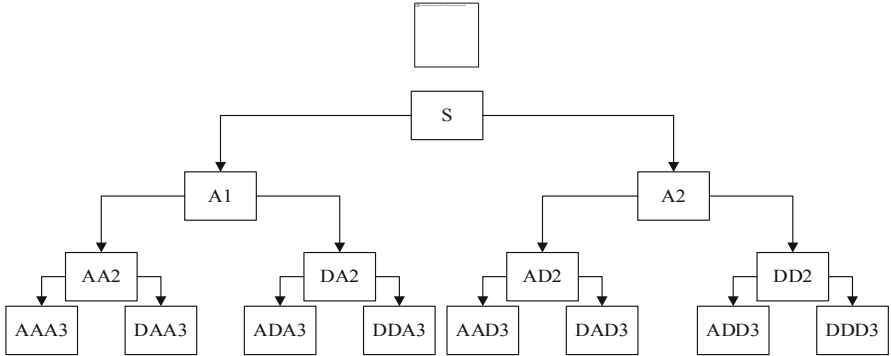
Fig. 6. Three-layer wavelet packet decomposition tree.

Wavelet packet transform can decompose spectral information into background information, component information and noise. The basic steps of wavelet packet threshold denoising are as follows [9]:

(1) Wavelet packet decomposition of the signal. Select the appropriate wavelet function and decomposition scale according to the signal, and calculate the optimal wavelet basis given the entropy criterion;

(2) Threshold quantization of wavelet packet decomposition coefficients. Choose the appropriate threshold rule according to experience, select the appropriate threshold, and process the decomposed wavelet packet coefficients;

(3) Wavelet packet reconstruction. The original spectral signal is reconstructed from the wavelet packet decomposition coefficients of the N-th layer and the processed coefficients.

For wavelet packet transformation, the selection of wavelet basis is very critical. Generally, a suitable wavelet basis function is selected from the four aspects of compactness, regularity, vanishing moment and symmetry. The commonly used wavelet basis functions are Daubechies wavelet system, SymletsA function system, Meyer wavelet, Coiflet wavelet system, Biorthogonal wavelet, and the commonly used wavelet functions in spectral denoising are db2, db4, sym6, boir2.4 [10]. After experimental comparison, this project selects a db4 wavelet for spectral denoising.

Once the signal undergoes wavelet packet transformation, the information is distributed in each frequency band. The effective spectral signal is usually concentrated in the low frequency band. On the larger wavelet packet coefficient, the noise energy is generally distributed on the entire coefficient axis, so it can be considered that the signal is generally concentrated in the amplitude value. The larger wavelet packet coefficients

and the noise are distributed on the smaller amplitude wavelet packet coefficients, so the threshold method can be used to extract useful signals. Because the threshold selection is too large, the details of the useful signal will be filtered out, and the threshold selection is too small, the denoising effect is not ideal, so it is necessary to select an appropriate threshold to quantize the wavelet packet decomposition coefficients. In this paper, the Sqtwolog length logarithm criterion is selected to set the threshold, and the specific calculation formula is as follows.

$$H = \sqrt{2 * \log(L(s))} \tag{6}$$

where H is the set threshold, and $L(s)$ is the length of the signal. Continuing the example of Fig. 1, Fig. 7 shows the actual effect of wavelet packet denoising of infrared spectra.
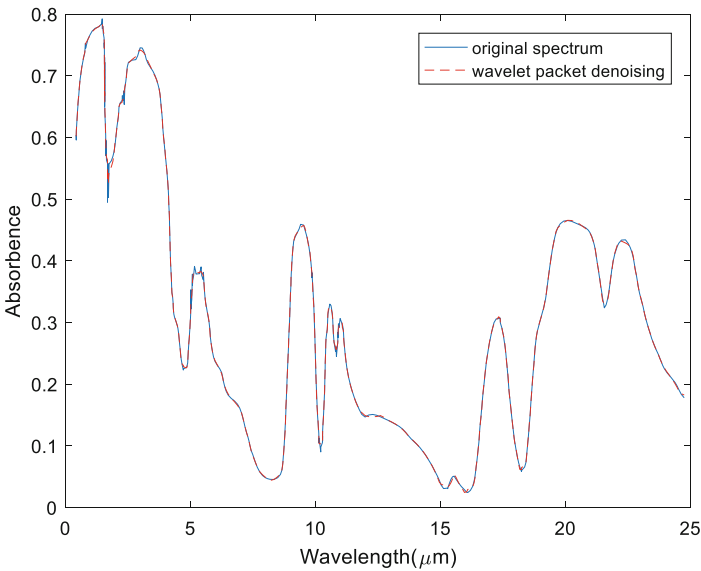
Fig. 7. Comparison before and after wavelet packet denoising of infrared spectrum.

3.2 Savitzky-Golay Convolution Smoothing

Savitzky-Golay convolution smoothing method, also called S-G smoothing, uses polynomials for data smoothing. S-G smoothing uses polynomials to perform polynomial least squares fitting on the data in the moving window, and its essence is a weighted average method. S-G smoothing can retain useful information in spectral signals, eliminate random noise, and make the curve smoother. It is a widely used denoising method at present [11].

A subset of the original spectral data is selected as the window instead of the entire spectrum. The width of the smoothing window is set to $2m + 1$, that is, the window

width $n = 2m + 1$. Assuming that the original data points within the window can be fitted with a k-1 polynomial, i.e.

$$y_i = a_0 + a_1 i + a_2 i^2 + \cdots + a_{k-1} i^{k-1} \tag{7}$$

where $i = (-m, -m + 1, \cdots, 0, 1, \cdots, m-1, m)$. Therefore, the above-mentioned polynomial can be obtained for each of the n original data points in the window, and n such polynomials constitute a k-element linear equation system, and the k fitting parameters a_j need to be solved. Generally, the selected filter window width n should be greater than or at least equal to k. When $n = k$, the fitting parameters can be solved by linear algebra; if $n > k$, the least squares method can be used to solve.

Juxtaposing the above polynomials, the following matrix operations can be obtained:

$$\begin{pmatrix} y_{-m} \\ y_{-m-1} \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 & -m & \cdots & (-m)^{k-1} \\ 1 & -m+1 & \cdots & (-m+1)^{k-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & m & \cdots & (m)^{k-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{k-1} \end{pmatrix} + \begin{pmatrix} e_{-m} \\ e_{-m-1} \\ \vdots \\ e_m \end{pmatrix} \tag{8}$$

It can be simplified to the following system of overdetermined equations

$$Y_{(2m+1) \times 1} = X_{(2m+1) \times k} \cdot A_{k \times 1} + E_{(2m+1) \times 1} \tag{9}$$

The calculation formula of the final solution of the filter value Y is as follows

$$\hat{Y} = XA = X(X^T X)^{-1} X^T Y = BY \tag{10}$$

Among them $B = X(X^T X)^{-1} X^T$ is the filter coefficient matrix, which is determined by and only by the X matrix, and the B matrix is a $(2m + 1) \times (2m + 1)$ matrix. According to the coefficient matrix, the S-G smooth fitting equation can be obtained.

Selecting the window width as 5 and the order of the fitting polynomial as 2, the result of S-G smoothing filtering on the infrared spectrum of $Mn_3Al_2(SiO_4)_3$ is shown in Fig. 8.

4 Multivariate Scattering Correction

Due to the influence of instrument background, sample particle size and other factors, baseline drift often occur in infrared analysis, and baseline correction can effectively eliminate these effects [12]. Methods such as peak-valley point leveling, offset deduction, differential processing, and baseline tilt can be used, and the most commonly used method is multivariate scattering correction.

Multiple Scattering Correction (MSC) was proposed by Martens et al. It is a commonly used method in spectral data preprocessing, mainly used to correct the shift and offset of the infrared spectral baseline of the sample. The resulting scattering effects improve the signal-to-noise ratio of the original absorbance spectrum. The method is based on the spectral array of a set of samples, and the basic idea is to effectively separate the absorption information of chemical substances from the scattered light signal

of the sample and its tilt shift, and divide its linear offset at the same time, so as to correct the baseline shift and shift of the spectrum. The spectral absorption information is not affected, so the signal-to-noise ratio of the spectrum is improved. The calculation method is as follows:

$$S_{i(MSC)} = \frac{(S_i - b_i)}{m_i} \tag{13}$$

During the spectrum acquisition process, the phenomenon of baseline drift will occur due to changes in the acquisition environment. Still taking $Mn_3Al_2(SiO_4)_3$ material as an example, the overall spectra of the spectrum repeated fifteen times are shown in Fig. 9, and it can be seen that the spectra has a baseline drift phenomenon. The spectra is processed by the method of multivariate scattering correction, and the result is shown in Fig. 10, which shows the effectively elimination of the baseline drift.
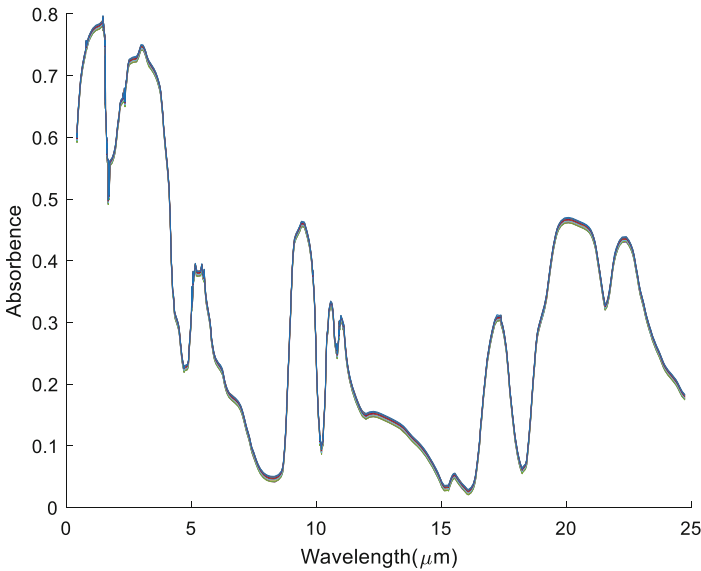
Fig. 9. Original multispectral images.

5   Abnormal Sample Removal

Abnormal samples will have a great impact on the infrared spectral model. The most widely used method is the Mahalanobis distance method [13]. In view of the characteristics of multi-variable spectral wavelength and easy overfitting, this paper combines principal component analysis and Mahalanobis distance to detect abnormal samples.

The principal components of each sample are first calculated, and the principal components of the sample replace the spectral data of the sample. The formula for
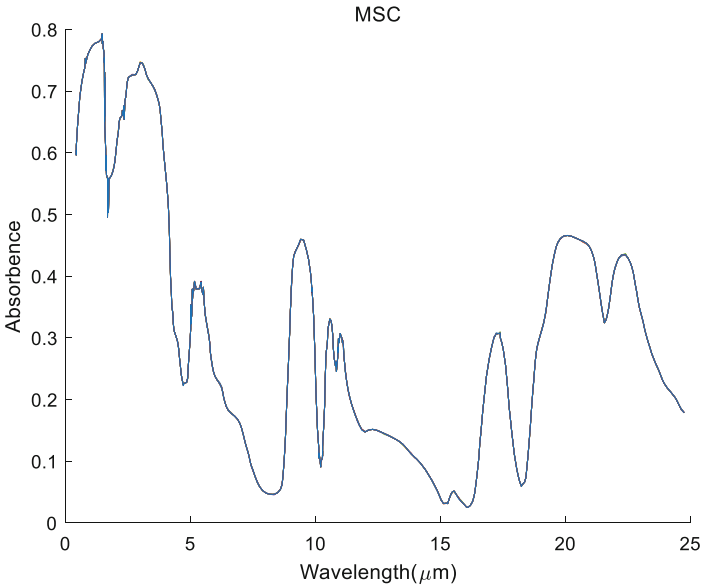
Fig. 10. Spectra after multivariate scattering correction.

calculating Mahalanobis distance is as follows:

$$D^2(\mathrm{i}) = (S_i - \overline{S})\mathbf{C}_S^{-1}(S_i - \overline{S})^T \tag{14}$$

where S is the sample data, $\mathbf{C}_S$ is the covariance of the sample data, and $\overline{S}$ is the mean of the data. The threshold is set to distinguish abnormal samples. The threshold setting formula is as follows:

$$D_\mathrm{t} = \overline{D} + e\delta_D \tag{15}$$

in which $\overline{D}$ is the mean of the Mahalanobis distance, δ_D is the standard deviation of the Mahalanobis distance, and e is a weight coefficient used to adjust the threshold for judging outliers. The larger the D_t value, the larger the value $D_i - D_t$, and the smaller the value, the less likely an anomaly is. When the average spectrum of a sample is very close to the sample, it is called the neighbor of the average sample.

Using the idea of discriminating spectral outliers combining Mahalanobis distance and principal component analysis, the 15 spectral samples shown in Fig. 9 are analyzed, and the sequence numbers of outliers are 6 and 7, and the abnormal samples are successfully eliminated.

6 Conclusion

Considering our developed Fourier transform infrared spectrometer, the spectral preprocessing algorithm is described in this paper. Firstly, the methods of first-order derivative, second-order derivative, centralization, normalization, and standardization of the

spectrum are discussed, and the spectrum can be transformed according to the detailed requirements. Using wavelet transform, S-G smoothing, principal component analysis-Malanobis distance method, multivariate scattering correction and other methods to process the spectrum, we can filter and denoise the spectrum, analyze the baseline shift and offset phenomenon of the infrared spectrum of the sample, as well as complete the detection of abnormal samples. As the first step of spectral analysis, the design of spectral preprocessing algorithm provides a solid foundation for the qualitative and quantitative analysis of subsequent spectra. Subsequent studies will also be carried out on spectral fitting, peak marking, and spectral feature extraction.

References

1. Dong, X., Guo, L., Wang, F., et al.: Quantitative analysis of microstructure of different coals based on Fourier transform infrared spectroscopy. Chin. Sci. Technol. Pap. **17**(1), 55–61 (2022)
2. Choi, K., Abbott, J., Park, B., Choi, C., et al.: Near-infrared diffuse reflectance for quantitative and qualitative measurement of soluble solids and firmness of delicious and Gala apples. Trans. ASAE **46**(6), 1721–1731 (2003)
3. Peirs, A., Schenk, A., Nicolai, B.: Effect of natural variability among appleson the accuracy of VIS-NIR calibration models for optimal harvest predictions. Postharvest Biol. Technol. **35**, 1–13 (2005)
4. Bessho, H., Kudo, K., Omori, J., et al.: A portable non-destructive quality meter for under standing fruit soluble solids in apple canopies. Acta Hort. **732**, 593–597 (2007)
5. Nicolai, B., Verlinden, B., Desmet, M., et al.: Time-resolved and continuous wave NIR reflectance spectroscopy to predict soluble solids content and firmness of pear. Postharvest Biol. Technol. **47**(1), 68–74 (2008)
6. Xu, G., Yuan, H., Lu, W.: Advances in modern near-infrared spectroscopy and its applications. Spectrosc. Spectral Anal. **02**, 134–142 (2000)
7. Hui, Y.: Study on Spectral Preprocessing Method and Moisture Detection Model of Junzao Leaves. Tarim University (2018)
8. Li, X., Zhang, Y., Liu, Z., et al.: Wavelet analysis and HHT transformation of blasting vibration signal. Explosion & Shock Impact **25**(6), 528–535 (2005)
9. Guo, X., Yang, H.: Wavelet packet denoising based on multi-threshold. In: Proceedings of the 27th China Conference on Control (2008)
10. Xu., X.: Application of enhanced wavelet transform in blasting vibration signal analysis. South China University of Technology (2013)
11. John, A., Sadasivan, J., Seelamantula, C.: Adaptive Savitzky-Golay filtering in non-gaussian noise. IEEE Trans. Signal Process. **69**, 5021–5036 (2021)
12. Zhong, J.: Study on spectral correction method and its application in soil detection. Jinan University (2018)
13. Zhang, L., Wang, W., Gu, Y., et al.: Principal component analysis of near-infrared spectroscopy-Markov distance clustering for authenticity identification of cigarettes. Spectrosc. Spectral Anal. **31**(05), 1254–1257 (2011)

An FES-Cycling Control System Based on Crank Angle

Tingting Wang[1,4] , Xuqun Pei[2,4(✉)], Min Liu[3,4], Chengqian Wang[1,4] ,
Mingxu Sun[1,4], and Han Zhang[1,4]

[1] School of Electrical Engineering, University of Jinan, Jinan 250022, China
202021100395@mail.ujn.edu.cn
[2] Jinan Central Hospital, Jinan 250000, China
[3] Shandong Provincial Third Hospital, Jinan 250000, China
[4] Shandong Betr Medical Technology Co. Ltd., Jinan 250022, China
13370582962@163.com

Abstract. Functional Electrical Stimulation (FES) cycling system is beneficial to the rehabilitation of patients with Spinal Cord Injury(SCI). In this paper, an FES cycling control system based on angular switching of electrical stimulation phases was developed using an indoor cycle, an electrical stimulator, a wireless inertial sensor and a PC. Then, this system was tested by healthy subjects. The device was able to send feedback data on the angle of the crank in real time to a PC, which processed the acquired data and sent serial data to the electrical stimulator to control the switching of the electrical stimulation phases. The electrical stimulator was configured in different set-up files by testing the muscle tolerance of the volunteers. Under the set experimental mode, at different crank angles, the target muscles are given appropriate stimulation pulses by the electrical stimulator. The electrical stimulation pattern: the quadriceps and biceps femoris of the volunteers were given alternate stimulation patterns. In the experiment, two healthy volunteers were test at 3 different cycling speed of 0.6 lap/sec, 1 lap/sec and 2 lap/sec. The test results show that the function and implementation of the controller are successful.

Keywords: Cycling · Lower limb · FES

1 Introduction

Spinal cord injury is the most serious complication of Trauma, often leading to severe dysfunction of the limbs below the injured segment. Specifically, spinal cord injuries can severely affect the ability to perform functional movements such as standing, walking or cycling. In the United States, approximately 291,000

people have spinal cord injuries [1]. This disease not only causes physical damage such as physical impairments of the patient's limbs, motor dysfunction, but also brings severe psychological pressure to the individual, which also has a great impact on its return to society. For patients with spinal cord injury, preventing muscle atrophy is a key factor, and it is necessary to train the affected side muscles. Studies have shown that FES is one of the most effective ways of muscle recovery [2]. FES technology uses electrical stimulation to activate paralyzed or paralyzed muscles with precise stimulation sequence and stimulation intensity [3], so that patients with spinal cord injury can restore certain motor functions.

Stimulating the leg muscles to achieve cyclic movements is much easier to achieve than using FES to achieve movements such as standing and walking. The FES cycling system is designed to achieve cyclic movements of the lower limbs by sequentially stimulating the patient's large muscle groups (including the quadriceps, gluteus, hamstrings and other muscles of the leg) and controlling the knee and ankle joints. By selecting the appropriate stimulation pattern, FES pedal exercise helps spinal cord injury patients to replenish muscle fatigue and increase rehabilitation training time [4,5]. And it can be used to prevent muscle atrophy during inactivity and aging, making it possible to repair muscle damage caused by various muscle diseases [6]. The system consists of four main components: the stimulator, the mechanical device, the controller and the signal acquisition equipment.

The effect of FES on cardiorespiratory function was proposed by Faghri PD et al. in 1984 [7]. This system was the prototype of the FES pedal cycle and has since attracted more attention from researchers. Various scholars have looked at the FES cycling in terms of improvements in physical function and have concluded that the FES cycling system has improved cardiopulmonary function, leg muscle recovery and bone density in patients [8–10]. In the field of FES cycling control systems, in 1997 Chen et al. Document [11] applied fuzzy control theory to the study of FES cycling system controllers for on-line real-time control of system parameters. In 2001, Gföhler et al. Document [12] looked at the effects of the geometry of the cycling not only in terms of the independent parameters used to generate moments, but also in terms of individual differences of paralysed patients. Kusgima et al. Document [13] proposed a FES cycling system for speed tracking control of human limbs, modelled Lagrangian dynamics and analysed its stability using Lyapunov's method. Cousin et al. Document [14,15] applied admittance control to the FES loop control system. There are also more constraints required for a control system that is too complex in realising a simple rehabilitation exercise process. This paper therefore proposes a simple control system for the FES cycling.

In this paper, in order to test the simple FES cycling control system, referring to T. Watanabe et al. proposed a FES control system that triggers electrical stimulation by the accelerometer signal attached to the crank [16] and Mingxu Sun et al. proposed the finite state machine [17]. Use the wireless inertial sensor to directly measure the crank acceleration to switch the stage of controlling the electric stimulator. By limiting the speed, two healthy volunteers were allowed

to perform the FES cycling control system proposed by the FES cycling exercise test. Finally, the state transition of the electric stimulator was analyzed.

2   Overview of the FES Cycling Control System

2.1   The Components of the FES Cycling Control System

The FES cycling system consists of an electrical stimulator, a cycle and a wireless inertial sensor, each of which is described below.

Stimulator. The FES cycling rehabilitation system can switch the working phase of the electrical stimulator through the crank angle. Electrical stimulation has the characteristics of expandable channels, programmable and portable. Electrical stimulator are divided into master and slave, which have the advantages of small size and light weight. Main control: $13\,cm * 10\,cm * 2\,cm$, slave electrical stimulator: $6\,cm * 5.5\,cm * 1.5\,cm$. It is powered by a rechargeable lithium battery, and the stimulation parameter $20\,Hz$. The biphasic pulse width is between 0–180 us, and the intensity is adjustable within the range of 0–$100\,mA$. These stimulating pulses are in direct contact with the target muscle through the electrode sheet, and the pulse is input to the muscle and the muscle is contracted to complete the target action. If the three channels of the master control cannot reach the required number of channels, the number of channels can be expanded by expanding the slave electrical stimulation.

Cycle. Use a fixed stationary cycle for testing. This cycle has a simple structure, so that the results can be better affected by external interference during the test process (for example, the resistance caused by the chain and the resistance caused by the friction between the wheel and the ground, etc.); in the static state, it reduces the accident.

Wireless Inertial Sensor. Since it is required to switch the next stage of the electrical stimulator according to the real-time angle, there are certain requirements for the receiving speed of the angle data. If the inertial sensor in the electric stimulator is used to transmit data directly, the bicycle movement is restricted by the wired condition of the serial port. Therefore, use the WIFI nine-axis sensor to collect the crankshaft angle data and transmit it to the control system using the User Datagram Protocol (UDP). UDP is a connectionless protocol, and data communication can be realized only by the application and the receiving end in a local area network. At the sending end, the speed of UDP transmission of data is only limited by the speed at which the application program generates data, the capacity of the computer, and the transmission bandwidth; at the receiving end, UDP puts each message segment in a queue, and the application reads it from the queue every time A message segment. According to the above characteristics of UDP, it meets the requirements of real-time communication. The
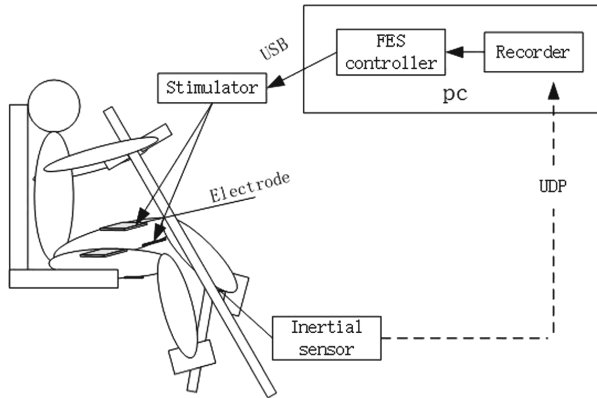
Fig. 1. The outline drawing of a cycle with FES control system. One of the electrodes is attached to the quadriceps and the other to the biceps.

outline drawing of a cycle with FES control system is shown in the Fig. 1. Fix the wireless inertial sensor on the center of the left crankshaft of the cycle with tape. The angle and angular velocity signals are measured by the inertial sensor, and the data of the inertial sensor is read to the PC through the UDP protocol. The timing of the electrical stimulator switching state is determined by the crank angle. Using the expandable electrical stimulator based on our previous research and development, the muscles are automatically electrically stimulated through the electrode pads at the appropriate time selected. The stimulus data is transmitted from the PC to the expandable electric stimulator through the serial port. The electric stimulator selects the state according to the received signal, thereby generating the stimulation pulse with the pulse frequency, pulse amplitude and pulse width suitable for the location. The output channel is composed of three channels of the master and one channel of the slave.

2.2 Experimental

Before the experiment, the volunteers' muscle tolerance was tested through the test mode of the electrical stimulator. Then configure the electrical stimulator with a profile suitable for volunteers. Also, before the experiment, the volunteers were told to subjectively control the riding speed of the FES bicycle at approximately 0.6 laps/sec, 1 lap/sec, and 2 laps/sec. Each FES riding exercise time exceeds 30 s, and the 30 s data with stable speed is taken during analysis.

Experimental Subjects. The subjects of this study included a 25-years-old healthy male and a 23-years-old healthy male.

Experimental Environment. We choose a wide and comfortable indoor environment as the test site. During the test, two healthy volunteers performed cycle

exercises of 0.6 laps/sec, 1 laps/sec, and 2 laps/sec, and recorded data such as the crank angle of the bicycle. Channel 1, channel 2, channel 3 and channel 4 of the electrical stimulation stimulate the right quadriceps, left quadriceps, right biceps femoris and left biceps femoris respectively. The physical map is shown in the Fig. 2.
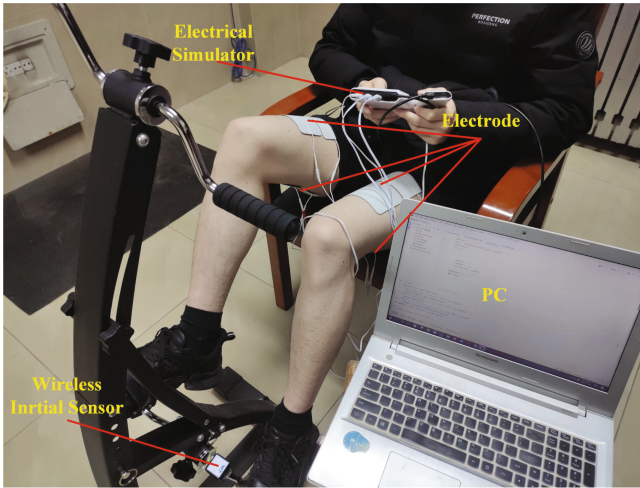
Fig. 2. FES cycling control system.

Experiment Methods. In this paper, we use the stimulation mode: quadriceps and biceps femoris stimulation alternately. The quadriceps is the thigh muscle of the human body, located in the front of the thigh muscle, with the function of extending the knee joint and bending the hip joint. The biceps femoris is one of the posterior muscles of the thigh and has the function of the knee joint. During the experiment, quadriceps and biceps femoris were alternately stimulated by electric stimulator to achieve the extension and flexion of the knee joint. The right quadriceps and left biceps femoris were stimulated and defined as phase 1, and the left quadriceps and right biceps femoris were stimulated as phase 2. Transitions of phase 1 and phase 2 are triggered by suitable methods. The structure of the left and right cranks on the horizontal plane is shown in Fig. 3.

The initial state is set as the left crank is on the top, the right crank is on the bottom, the left foot is pedaling forward as the movement direction, and the initial position of the electrical stimulator is phase 1. When the receiving crank Angle is greater than or equal to 130°, the electrical stimulator performs phase 2. When the receiving crank Angle is greater than or equal to 310°, the electrical stimulator performs phase 1. Phase 1 and phase 2 correspond to left pedal and right pedal respectively. Considering that in the case of convulsion of lower limbs, the method of triggering phase transition with a reading will cause instability of the controller.
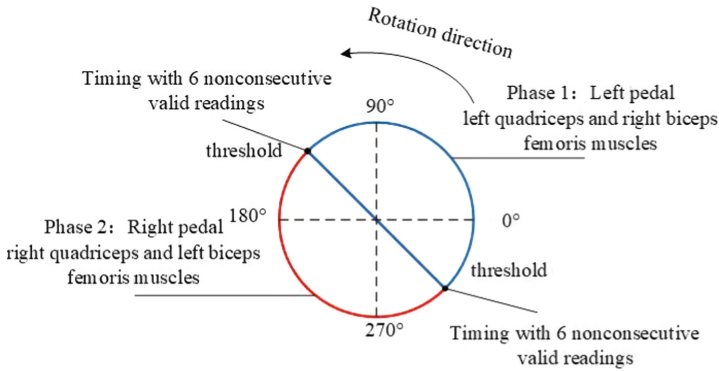
Fig. 3. Structure diagram of left and right cranks on horizontal plane. Blue line and red line symmetrical. (Color figure online)

The method of timing 6 discontinuous effective readings can better deal with the noise impact of upper limb convulsion [16]. Therefore, the triggering method in this paper is to start the effective count when one point reaches the threshold, automatically skip the no count when the number that does not reach the triggering condition in the count, and start the transition phase when the effective count reaches 6.

Implementation flow chart of FES controller is shown in Fig. 4. The initial state is set with the left crank on the top, the right crank on the bottom, and the forward cycling with the left foot as the direction of movement. The right quadriceps and the left biceps femoris muscles are stimulated first. When the received crank angle is between 310–130, the right quadriceps and left biceps femoris muscles (phase 1) are stimulated by the electrical stimulator and left pedal. When the received crank angle is between 130 and 310, the left quadriceps and right biceps femoris muscles (phase 2) are stimulated by the electrical stimulator and right pedal.

Results. During the volunteer FES cycle exercise, the relationship between the cycle exercise time and the crank angle at different speeds is shown in Fig. 5 and Fig. 6. It can be seen from the figure that within 30s, under different speed conditions, there are two phase transition points in a cycle.
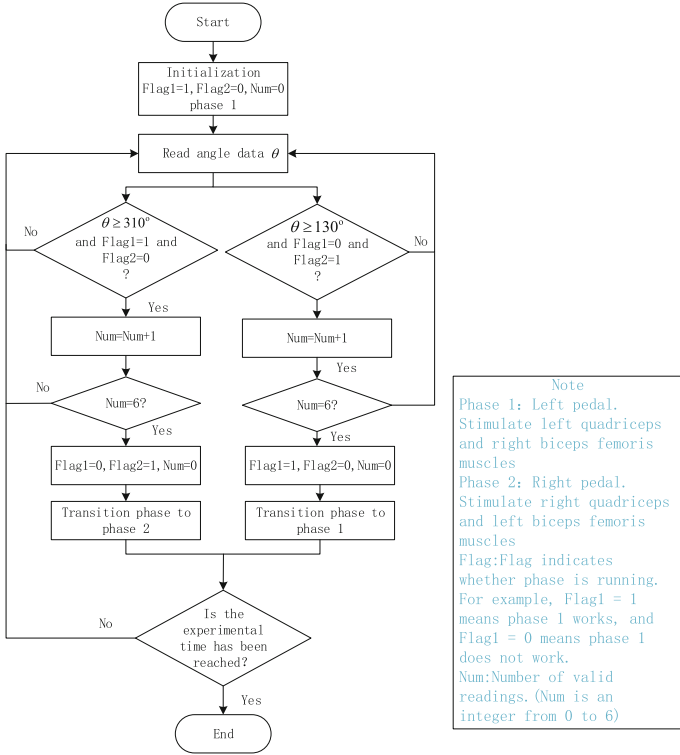
Fig. 4. Implementation flow chart of FES controller.

It can be concluded that the proposed FES cycling control system can transition points following the crank angle and stimulate the corresponding thigh muscles. In order to more intuitively observe the phase switching, we intercepted the data of two volunteers riding for two cycles in FES. It can be seen in Fig. 5 that the speed of volunteer 1 in the first two turns was not very stable. We used the data of the third and fourth cycles and the corresponding electrical stimulator channel pulses. Volunteer 2's speed was relatively stable. We used the data of the second and third cycles and the corresponding electrical stimulator channel pulses. Figure 7 and Fig. 8 show the controller phase transition of two volunteers during two cycles of FES cycle. In the figure, the transition time from the phase 1 to the phase 2 and from phase 2 to phase 1 corresponds to the transition between phase points. Moreover, in order to make the control system more stable, six discontinuous effective readings are used as the state transition condition. Therefore, the phase transition has a certain delay for the threshold point.
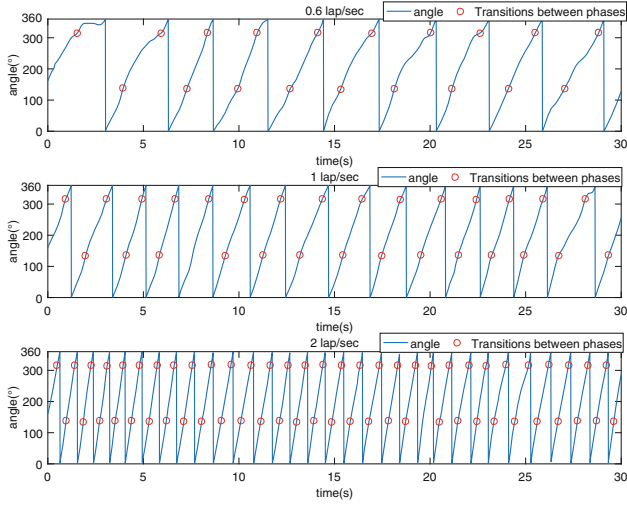
Fig. 5. Volunteer 1's FES cycling in 30 s. The first point is switched from phase 1 to phase 2, the second point is switched from phase 2 to phase 1 and so on.
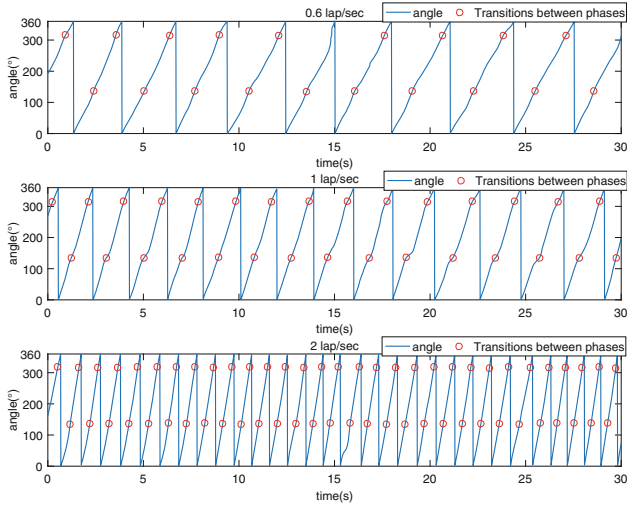
Fig. 6. Volunteer 2's FES cycling in 30 s. The first point is switched from phase 1 to phase 2, the second point is switched from phase 2 to phase 1 and so on.
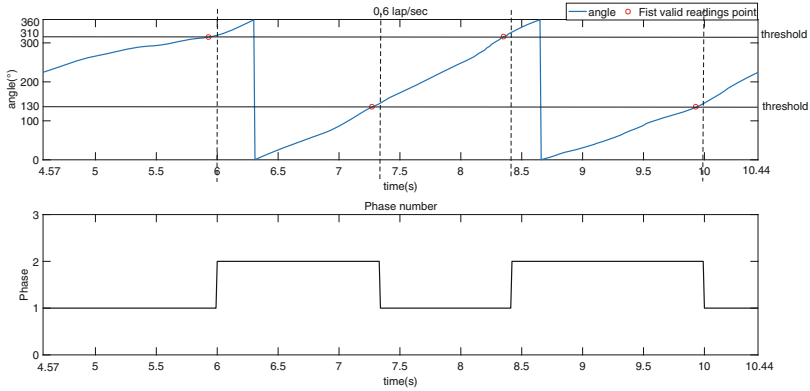
Fig. 7. Pulse waveform output from four channels of the electrical stimulator (take two cycles of the volunteer 1's FES cycling).
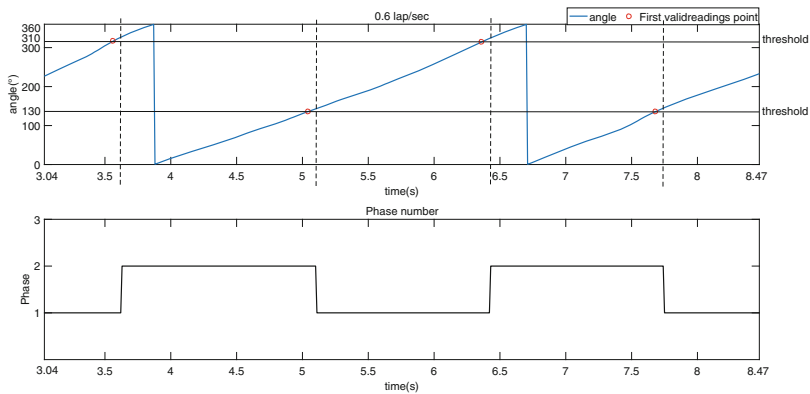

Fig. 8. Pulse waveform output from four channels of the electrical stimulator (take two cycles of the volunteer 2's FES cycling).

3   Conclusion

This paper first gives a brief introduction to the current research status and system composition of the FES cycling system, and then proposes a method of electrical stimulation mode applied to the FES cycling and conducts experiments. The experimental results show that the response of the FES cycling control system based on the electrical stimulation mode is timely and stable in operation. However, there will be some shortcomings in the system, for example, the testers in this experiment are healthy people, not tested on patients with spinal cord injury, so the system will have some shortcomings, and the next step will be to continue to improve the system with the actual situation of the patients.

References

1. Dunkelberger, N., Schearer, E.M., O'Malley, M.K.: A review of methods for achieving upper limb movement following spinal cord injury through hybrid muscle stimulation and robotic assistance. Exp. Neurol. **328**, 113274–113289 (2020)
2. Rky, T., Wang, X., Kwc, L., et al.: How to prepare a person with complete spinal cord injury to use surface electrodes for FES trike cycling. In: IEEE International Conference on Rehabilitation Robotics, pp. 801–805 (2017)
3. Mutton, D.L., Scremin, A., Barstow, T.J., et al.: Physiologic responses during functional electrical stimulation leg cycling and hybrid exercise in spinal cord injured subjects. Arch. Phys. Med. Rehabil. **78**(7), 712–718 (1997)
4. Ambrosini, E., Parati, M., Peri, E., et al.: Changes in leg cycling muscle synergies after training augmented by functional electrical stimulation in subacute stroke survivors: a pilot study. J. Neuroeng. Rehabil. **17**(1), 35–41 (2020)
5. Gelenitis, K., Foglyano, K., Lombardo, L., et al.: Selective neural stimulation methods improve cycling exercise performance after spinal cord injury: a case series. J Neuro Eng. Rehabil. **18**(1), 117–131 (2021)
6. Pette, D., Vrbov, G.: The contribution of neuromuscular stimulation in elucidating muscle plasticity revisited. Eur. J. Transl. Myol. **27**(1), 33–39 (2017)
7. Faghri, P.D., Glaser, R.M., Figoni, S.F.: Functional electrical stimulation leg cycle ergometer exercise: training effects on cardiorespiratory responses of spinal cord injured subjects at rest and during submaximal exercise. Arch. Phy. Med. Rehab. **73**(11), 1085–1093 (1992)
8. Barstow, T.J., Scremin, A., Mutton, D.L., et al.: Peak and kinetic cardiorespiratory responses during arm and leg exercise in patients with spinal cord injury. Spinal Cord **38**(6), 340–345 (2000)
9. Verellen, J., Vanlandewijck, Y., Andrews, B., et al.: Cardiorespiratory responses during arm ergometry, functional electrical stimulation cycling, and two hybrid exercise conditions in spinal cord injured. Disabil. Rehabil. Assistive Technol. **2**(2), 127–132 (2007)
10. Schutte, L.M., Rodgers, M.M.: Improving the efficacy of electrical stimulation-induced leg cycle ergometry: an analysis based on a dynamic musculoskeletal model. IEEE Trans. Rehabil. Eng. **1**(2), 109–125 (1993)
11. Chen, J.J., Yu, N.Y., Huang, D.G., et al.: Applying fuzzy logic to control cycling movement induced by functional electrical stimulation. IEEE Trans. Rehabil. Eng. **5**(2), 158–169 (2002)
12. Gföhler, M., Angeli, T., Eberharter, T., et al.: Test bed with force-measuring crank for static and dynamic investigations on cycling by means of functional electrical stimulation. IEEE Trans. Neural Syst. Rehabil. Eng. Publ. IEEE Eng. Med. Biol. Soc. **9**(2), 169–180 (2001)
13. Kushima, Y., Kawai, H., Murao, T., et al.: FES-assisted cycling with velocity tracking control for hemiparesis rehabilitation. In: IEEE International Conference on Advanced Intelligent Mechatronics (AIM). IEEE, vol. 11(9), pp. 645–650 (2017)
14. Cousin, C.A., Deptula, P., Rouse, C., Dixon, W.E.: Cycling with functional electrical stimulation and adaptive neural network admittance control. In: American Control Conference (ACC), vol. 28(1), pp. 1742–1747 (2019)
15. Cousin, C.A., Rouse, C.A., Dixon, W.E.: Split-crank functional electrical stimulation cycling: an adapting admitting rehabilitation robot. EEE Trans. Control Syst. Technol. **29**(5), 2153–2165 (2021)

16. Watanabe, T., Murakami, T., Handa, Y.: Preliminary tests of a prototype FES control system for cycling wheelchair rehabilitation. In: IEEE International Conference on Rehabilitation Robotics, pp. 1–6 (2013)
17. Sun, M., Christine, S., David, H., et al.: FES-UPP: a flexible functional electrical stimulation system to support upper limb functional activity practice. Front. Neurosci. **12**, 449 (2018)
18. Sun, M., Kenney, L., Howard, D., et al.: The design of an advanced functional electrical stimulation system (FES-UPP) for upper limb rehabilitation. In: Annual Conference of the International Functional Electrical Stimulation Society (2017)

A Novel Algorithm of Machine Learning: Fractional Gradient Boosting Decision Tree

Kangkai Gao and Yong Wang[(✉)]

Department of Automation, University of Science and Technology of China, Hefei 230026, China
yongwang@ustc.edu.cn

Abstract. The gradient boosting decision tree is a commonly used and effective ensemble machine learning method. In this paper, a factional gradient boosting decision tree scheme is first proposed with several loss functions discussed. This scheme implies that different algorithms can be established when adopting different fractional order gradient methods. It is then shown that with satisfactory convergence of fractional gradient descent algorithms, the proposed algorithms can train models with required accuracy in less number of iterations. Finally, several examples are provided to demonstrate the accuracy and efficiency of the algorithms.

Keywords: Fractional calculus · Gradient descent method · Ensemble method · GBDT

1 Introduction

Machine learning evolves from the broad field of artificial intelligence and plays an essential part in many areas. In the study of machine learning, ensemble methods refer to a kind of state-of-art learning approaches that combine multiple training learners. They have shown excellent performance in various machine learning applications and analytics competitions, e.g., Kaggle challenges. There are two primary techniques to associate the base learners: boosting [10,11,15] and bagging [1]. Gradient boosting is a classic ensemble technique for handling regression and classification problems. It combines the outputs of many base learners, typically decision trees, to produce a powerful learner in an additive form. Gradient boosting decision tree (GBDT) has been widely used recently mainly due to its high accuracy and fast training. Some works have been done to speed up the computation of GBDT under different parallel settings, e.g., XGBoost [3], LightGBM [12], and PLANET [14]. Some other works exploit benefits of GBDT for different machine learning applications. For instance, LambdaMART used GBDT to solve the ranking problem in [2].

In GBDT, each new tree is trained on the per-point residual defined as the negative gradient of loss function with regard to output of previous trees. Hence, GBDT can be regarded as an analogy of gradient descent method from parameter space to function space. A novel algorithm is proposed in this article by applying the fractional gradient descent method from this perspective.

The authors respectively proposed fractional gradient methods which can converge but with a low convergence rate in [5,7,16,20]. Nevertheless, it cannot be assured that the method will converge to the real global solution. The shortcoming has been partially overcome in [17]. After that, by approximating the exact fractional derivative, a new fractional gradient method was proposed by us [4,18]. Subsequently, three new viable solutions to the convergence problem were came up with in [19]. The algorithm proposed in [4,18] has been widely used in many applications, such as least mean square algorithm [6,21], system identification [8,9], recommender systems [13], etc. It is shown that the fractional gradient method is a promising algorithm. Additionally, the convergence speed should be taken into account. In [19], both theoretical analysis and simulation study indicate that all the designed methods can achieve the true convergence quickly. Therefore, the multiple learners boosted by fractional gradient may reach the required accuracy within less iterations and time costs.

Motivated by the previous discussions, a generalized fractional gradient boosting decision tree algorithm is presented in this article. It is shown by experiments with early stopping that the proposed algorithm trains the model quicker with the same tolerance compared to traditional gradient boosting decision tree. It also indicates the wide and potential applications of fractional gradient descent method.

The rest of this paper is organized as follows. A succinct review of fractional calculus, fractional gradient descent method, and ensemble methods is provided in Sect. 2. Then a detailed description of fractional gradient boosting decision tree algorithm is presented in Sect. 3, with several loss functions discussed. Section 4 exhibits several simulations to show the accuracy and effectiveness of the proposed algorithm. At last, Sect. 5 draws some conclusions.

Notation: Throughout the paper, the following notations are used. x and $\mathbf{x}$ denote the variable and vector respectively. $\Gamma(\alpha) = \int_0^\infty \mathrm{e}^{-t} t^{\alpha-1} \mathrm{d}t$ is the Gamma function which is an extension of the factorial function to real number arguments. $\arg\min_f L(f(\cdot))$ stands for argument of the minimum. The indicator function of a subset of a whole instance space is a function defined as $\mathbf{I}(\mathbf{x} \in R_{jm}) = \begin{cases} 0, \mathbf{x} \notin R_{jm}, \\ 1, \mathbf{x} \in R_{jm}. \end{cases}$ $E_x L(f(x))$ is the mathematical expectation of function $f(x)$ to x.

2   Preliminaries

A brief introduction of mathematical background of fractional order calculus and fractional gradient descent, and ensemble methods will be presented in this section.

2.1   Fractional Gradient Descent Method

Fractional order calculus is a generalization of integer order calculus. There are three widely used definitions for fractional order calculus, Riemann-Liouville definition, Caputo definition and Grünwald-Letnikov definition. In this article, only Riemann-Liouville and Caputo definitions are adopted.

The Riemann-Liouville derivative is expressed as

$$
{}^{\mathrm{RL}}_{c}\mathscr{D}^{\alpha}_{x}f(x) = \frac{1}{\Gamma(n-\alpha)}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\int_{c}^{x}\frac{f(\tau)}{(x-\tau)^{\alpha-n+1}}\mathrm{d}\tau, \tag{1}
$$

where $n-1 < \alpha < n, n \in \mathbb{Z}_{+}, c$ is the lower terminal.

The form of Caputo derivative is as follows

$$
{}^{\mathrm{C}}_{c}\mathscr{D}^{\alpha}_{x}f(x) = \frac{1}{\Gamma(n-\alpha)}\int_{c}^{x}\frac{f^{(n)}(\tau)}{(x-\tau)^{\alpha-n+1}}\mathrm{d}\tau. \tag{2}
$$

Furthermore, if function $f(x)$ can be expanded as a Taylor series, the fractional derivatives can be rewritten as

$$
{}^{\mathrm{RL}}_{c}\mathscr{D}^{\alpha}_{x}f(x) = \sum_{i=0}^{+\infty}\binom{\alpha}{i}\frac{f^{(i)}(x)}{\Gamma(i+1-\alpha)}(x-c)^{i-\alpha}, \tag{3}
$$

$$
{}^{\mathrm{C}}_{c}\mathscr{D}^{\alpha}_{x}f(x) = \sum_{i=n}^{+\infty}\binom{\alpha-n}{i-n}\frac{f^{(i)}(x)}{\Gamma(i+1-\alpha)}(x-c)^{i-\alpha}, \tag{4}
$$

or

$$
{}^{\mathrm{RL}}_{c}\mathscr{D}^{\alpha}_{x}f(x) = \sum_{i=0}^{+\infty}\frac{f^{(i)}(c)}{\Gamma(i+1-\alpha)}(x-c)^{i-\alpha}, \tag{5}
$$

$$
{}^{\mathrm{C}}_{c}\mathscr{D}^{\alpha}_{x}f(x) = \sum_{i=n}^{+\infty}\frac{f^{(i)}(c)}{\Gamma(i+1-\alpha)}(x-c)^{i-\alpha}, \tag{6}
$$

From the Formulas (3–6), we can note that the fractional derivative, both in the sense of Riemann-Liouville definition and Caputo definition, can be expanded to a series of integer order derivatives.

Gradient descent is one of the simplest of the frequently used numerical minimization methods. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point.

$$
x_{k+1} = x_k - \mu\nabla f(x_k), \tag{7}
$$

where x_k is the current point, x_{k+1} is the next point, μ is the learning rate and $\nabla f(x_k)$ is the first-order gradient at $x = x_k$.

There are three fractional order gradient descent methods introduced in [19]: fixed memory step, higher order truncation and variable fractional order. Then the update law can be rewritten as

$$
x_{k+1} = x_k - \mu h, \tag{8}
$$

where h is calculated by different equations which are listed in Eqs. (9–11).

$$h = \sum_{i=1}^{+\infty} \binom{\alpha - 1}{i - 1} \frac{f^{(i)}(x_k)}{\Gamma(i+1-\alpha)} (x_k - x_{k-K})^{i-\alpha}, \tag{9}$$

$$h = \frac{f^{(1)}(x_k)}{\Gamma(2-\alpha)} (|x_k - c| + \epsilon)^{1-\alpha}, \tag{10}$$

$$h = \sum_{i=1}^{+\infty} \binom{\alpha(x_k) - 1}{i - 1} \frac{f^{(i)}(x_k)}{\Gamma(i+1-\alpha(x))} (x_k - c)^{i-\alpha(x)}, \tag{11}$$

thereinto, $\alpha(x)$ in Eq. (11) can be designed as follows

$$\alpha(x) = \frac{1}{1 + \beta f(x)}, \tag{12}$$

$$\alpha(x) = \frac{2}{1 + e^{\beta f(x)}}, \tag{13}$$

$$\alpha(x) = 1 - \tanh(\beta f(x)). \tag{14}$$

2.2 Ensemble Methods

One major task of machine learning, pattern recognition and data mining is to construct good models from data sets. A formal formulation of learning process is as follows

$$\begin{aligned} f^* &= \arg\min_f E_{y,\mathbf{x}} L(y, f(\mathbf{x})) \\ &= \arg\min_f E_{\mathbf{x}}[E_y(L(y, f(\mathbf{x})))|\mathbf{x}], \end{aligned} \tag{15}$$

where $\mathbf{x}$ denotes the instance, y denotes the label, L is the loss function used in the problem, and f mapping $\mathbf{x}$ to y is the goal of the function estimation. Besides, if the label is categorical, the task is called classification and the learner is called classifier; if the label is numerical, the task is called regression and the learner is called fitted regression model [22].

The idea of ensemble learning is to build a prediction model by combining the strengths of a collection of simpler base models which can be decision tree, neural network, support vector machine or other kinds of machine learning algorithms. In GBDT, decision tree is adopted as the base learner. Ensemble learning can be broken down into two parts: developing a population of base learners from the training data and then combining them to form the composite predictor. The base learners can be generated in a parallel style, e.g., bagging, or in a sequential style, e.g., boosting.

3 Main Results

In this section, a fractional gradient boosting machine algorithm will be proposed with both regression and classification loss functions discussed.

3.1   Fractional Gradient Boosting Decision Tree

Decision trees partition the space of all joint predictor variable values into disjoint regions R_j, $j = 1, 2, \ldots, J$, as represented by the terminal nodes of the tree. Hence, the formula of a tree can be expressed as

$$T(\mathbf{x}; \Theta) = \sum_{j=1}^{J} \gamma_j \boldsymbol{I}(\mathbf{x} \in R_j), \tag{16}$$

with parameters $\Theta = \{R_j, \gamma_j\}_1^J$. There are several strategies to induce a tree, e.g., ID3, C4.5, CART. The boosted trees model takes an additive form

$$f_M(\mathbf{x}) = \sum_{m=1}^{M} T(\mathbf{x}; \Theta_m). \tag{17}$$

As mentioned before, the boosted trees are induced in a sequential way: the later trees will take the former trees into account. In this situation, we can try a greedy stagewise approach: assuming that the current model $f_{m-1}(\mathbf{x})$ is already attained, solve the Eq. (18) to grow the m-th tree at m iteration

$$\widehat{\Theta}_m = \underset{\Theta_m}{\arg\min} \sum_{i=1}^{N} L(y_i, f_{m-1}(\mathbf{x}_i) + T(\mathbf{x}_i; \Theta_m)). \tag{18}$$

Then how to build a tree at each iteration is described next. Since there is a finite data sample $\{y_i, \mathbf{x}_i\}_i^N$ and the loss function in using $f(\mathbf{x}_i)$ to predict label on the training data is

$$L(f) = \sum_{i=1}^{N} L(y_i, f(\mathbf{x}_i)), \tag{19}$$

to minimize the Eq. (19), a numerical optimization problem is obtained

$$\hat{\mathbf{f}} = \underset{\mathbf{f}}{\arg\min} \, L(\mathbf{f}), \tag{20}$$

where the parameters $\mathbf{f} \in \mathbb{R}^N$ are the values of the approximating function $f(\mathbf{x}_i)$ at each data point $\mathbf{x}_i$

$$\mathbf{f} = \{f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)\}. \tag{21}$$

Obviously, due to the addictive form, the equation $\mathbf{f} = \sum_{m=1}^{M} \mathbf{h}_m$, where $\mathbf{h}_m \in \mathbb{R}^N$ is the predicted value of the tree at m iteration, can be obtained. Taking the greedy stagewise approach, a natural and acceptable approach is inducing a tree to fit the negative gradient of loss function at each data points by applying gradient descent method.

$$g_{im} = \frac{\partial L(y_i, f_{m-1}(\mathbf{x}_i))}{\partial f_{m-1}(\mathbf{x}_i)}, \tag{22}$$

$$\widetilde{\Theta}_m = \underset{\Theta}{\arg\min} \sum_{i=1}^{N} (-g_{im} - T(\mathbf{x}_i; \Theta))^2, \tag{23}$$

where squared error is used to measure closeness.

GBDT can be viewed as an application of gradient descent method in function space, rather than parameter space. Hence, a modified fractional gradient boosting algorithm can be designed where fractional gradient descent is applied.

There are three fractional order gradient descent methods developed: fixed memory step, higher order truncation, and variable fractional order [19]. Since the commonly used loss function in classification problem is cross entropy and the series number of fractional gradient of cross entropy in Eq. (9) and (11) is infinite, we will only adopt higher order truncation to improve GBDT in this article. There is an emphasize of the update law expressed as

$$x_{k+1} = x_k - \mu \frac{f^{(1)}(x_k)}{\Gamma(2-\alpha)} (|x_k - c| + \epsilon)^{1-\alpha}, \tag{24}$$

where α is the fractional order, ϵ is a small nonnegative number.

Furthermore, with the initial instant varying, the equation can be re-expressed as

$$x_{k+1} = x_k - \mu \frac{f^{(1)}(x_k)}{\Gamma(2-\alpha)} (|x_k - x_{k-1}| + \epsilon)^{1-\alpha}, \tag{25}$$

where x_{k-1} is the predicted value in previous iterations.

The update law can be viewed as an optimized gradient descent method with the step length

$$\mu_k = \mu \frac{(|x_k - c| + \epsilon)^{1-\alpha}}{\Gamma(2-\alpha)}, \tag{26}$$

or

$$\mu_k = \mu \frac{(|x_k - x_{k-1}| + \epsilon)^{1-\alpha}}{\Gamma(2-\alpha)}, \tag{27}$$

varying according to the iteration.

Suppose the function f is convex and differentiable, and the gradient is Lipschitz continuous with constant $L > 0$, if the adaptive step size is less than $1/L$, the gradient descent will yield a solution f_k which satisfies

$$f(x_k) - f(x^*) \le \frac{\|x_0 - x^*\|_2^2}{2\mu_{\min} k}. \tag{28}$$

That means the convergence rate of the method is at least sublinear.

It is worth noticing that with the iterations increasing, the distance between x_k and x_{k-1} decreases. This causes the step length become larger than it should be in classical gradient descent method. By the virtue of growing step length, the learning rate is faster than the classical gradient descent in the eventual stages.

Naturally, since fractional order gradient descent method converges to extreme points with more efficiency compared to classical gradient one, we can use the value calculated by fractional gradient descent method to induce new decision trees. Then Eq. (22) is replaced by

$$g_{im} = \frac{\partial L(y_i, f_{m-1}(\mathbf{x}_i))}{\partial f_{m-1}(\mathbf{x}_i)} \frac{[|f_{m-1}(\mathbf{x}_i) - c| + \epsilon]^{1-\alpha}}{\Gamma(2-\alpha)}, \tag{29}$$

or

$$g_{im} = \frac{\partial L(y_i, f_{m-1}(\mathbf{x}_i))}{\partial f_{m-1}(\mathbf{x}_i)} \frac{[|f_{m-1}(\mathbf{x}_i) - f_{m-2}(\mathbf{x}_i)| + \epsilon]^{1-\alpha}}{\Gamma(2-\alpha)}. \tag{30}$$

A brief description of the process is demonstrated in Algorithm 1.

Algorithm 1. Fractional gradient boosting decision tree.

1: **Input:** Data sample $\{y_i, \mathbf{x}_i\}_1^N$

2: **Initialize:** $f_0(\mathbf{x}) = \arg\min_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$.

3: **for** $m = 1 \to M$ **do**

4:    Compute the fractional gradient as the working response

$$g_{im} = \frac{\partial L(y_i, f_{m-1}(\mathbf{x}_i))}{\partial f_{m-1}(\mathbf{x}_i)} \frac{[|f_{m-1}(\mathbf{x}_i) - c| + \epsilon]^{1-\alpha}}{\Gamma(2-\alpha)},$$

    or

$$g_{im} = \frac{\partial L(y_i, f_{m-1}(\mathbf{x}_i))}{\partial f_{m-1}(\mathbf{x}_i)} \frac{[|f_{m-1}(\mathbf{x}_i) - f_{m-2}(\mathbf{x}_i)| + \epsilon]^{1-\alpha}}{\Gamma(2-\alpha)},$$

    where $i = 1, \ldots, N$.

5:    Build a decision tree, $T(\mathbf{x}; \Theta) = \sum_{j=1}^J \gamma_j \boldsymbol{I}(\mathbf{x} \in R_j)$, through CART

$$\{R_{jm}\}_1^J = J - \text{terminal node } tree(\{-g_{im}, \mathbf{x}_i\}_1^N).$$

6:    Improve the fit by optimizing the coefficients for each of these terminal nodes

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{\mathbf{x}_i \in R_{jm}} L(y_i, f_{m-1}(\mathbf{x}_i) + \gamma).$$

7:    Update the approximation

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \sum_{j=1}^J \gamma_{jm} \boldsymbol{I}(\mathbf{x} \in R_{jm}).$$

8: **end for**

3.2   Loss Functions of Regression and Classification

Several popular loss criteria commonly used in regression and classification problems will be discussed in this part. In view of the only difference between Eq. (29) and (30) which is the initial point, we consider the fractional gradient with the same initial point only in this section.

Least-squares is the most commonly used regression loss function, generally known as

$$L(y_i, f(\mathbf{x}_i)) = \frac{[y_i - f(\mathbf{x}_i)]^2}{2}. \tag{31}$$

The corresponding fractional gradient is expressed as

$$g_{im} = -\frac{y_i - f_{m-1}(\mathbf{x}_i)}{\Gamma(2-\alpha)} [|f_{m-1}(\mathbf{x}_i) - c| + \epsilon]^{1-\alpha}. \tag{32}$$

The formula of least absolute deviation is shown below

$$L(y_i, f(\mathbf{x}_i)) = |y_i - f(\mathbf{x}_i)|, \tag{33}$$

and fractional gradient is displayed as

$$g_{im} = -\frac{\mathrm{sign}(y_i - f_{m-1}(\mathbf{x}_i))}{\Gamma(2-\alpha)}[|f_{m-1}(\mathbf{x}_i) - c| + \epsilon]^{1-\alpha}. \tag{34}$$

Huber loss is more robust to outliers in data than least square loss function and also differentiable at 0 compared to least absolute deviation. It is defined as

$$L(y_i, f(\mathbf{x}_i)) = \begin{cases} \frac{1}{2}[y_i - f(\mathbf{x}_i)]^2, & |y_i - f(\mathbf{x}_i)| \le \delta, \\ \delta[|y_i - f(\mathbf{x}_i)| - \frac{\delta}{2}]|y_i - f(\mathbf{x}_i)| > \delta. \end{cases} \tag{35}$$

Moreover, the fractional gradient goes like this

$$g_{im} = \begin{cases} -\dfrac{y_i - f_{m-1}(\mathbf{x}_i)}{\Gamma(2-\alpha)} \times \\ [|f_{m-1}(\mathbf{x}_i) - c| + \epsilon]^{1-\alpha}, & |y_i - f_{m-1}(\mathbf{x}_i)| \le \delta, \\ \\ -\dfrac{\delta \, \mathrm{sign}(y_i - f_{m-1}(\mathbf{x}_i))}{\Gamma(2-\alpha)} \times \\ [|f_{m-1}(\mathbf{x}_i) - c| + \epsilon]^{1-\alpha}, & |y_i - f_{m-1}(\mathbf{x}_i)| > \delta. \end{cases} \tag{36}$$

In addition, there are other common loss functions to choose: the quantile loss, likelihood loss, and so on.

With regard to classification problem, cross entropy is generally chosen as the loss function. The binomial deviance is shown as

$$L(y_i, f(\mathbf{x}_i)) = -[y_i \log p_i + (1 - y_i) \log(1 - p_i)], \tag{37}$$

where $p_i = \frac{1}{1 + e^{-f(\mathbf{x}_i)}}$, $y_i \in \{0, 1\}$. It can be naturally extended to the K-class multinomial deviance loss function

$$\begin{aligned} L(\{y_i, f_k(\mathbf{x}_i)\}_1^K) &= -\sum_{k=1}^{K} \mathbf{I}(y_i = k) \log p_k(\mathbf{x}_i) \\ &= -\sum_{k=1}^{K} \mathbf{I}(y_i = k) f_k(\mathbf{x}_i) + \log(\sum_{l=1}^{K} e^{f_l(\mathbf{x}_i)}), \end{aligned} \tag{38}$$

where $p_k(\mathbf{x}_i) = \frac{e^{f_k(\mathbf{x}_i)}}{\sum_{l=1}^{K} e^{f_l(\mathbf{x}_i)}}$, $y \in \{1, 2, \ldots, K\}$. The final estimates $\{f_k(\mathbf{x}_i)\}_1^K$ can be used to obtain corresponding probability estimates $\{p_k(\mathbf{x}_i)\}_1^K$. These in turn can be used for classification.

The fractional gradients of the two loss function can be expressed as

$$g_{im} = -\frac{y_i - p_{i,m-1}}{\Gamma(2-\alpha)}[|f_{m-1}(\mathbf{x}_i) - c| + \epsilon]^{1-\alpha}, \tag{39}$$

where $p_{i,m-1} = \frac{1}{1 + e^{-f_{m-1}(\mathbf{x}_i)}}$ for the binomial case, and

$$g_{kim} = -\frac{\mathbf{I}(y_i = k) - p_{k,m-1}(\mathbf{x}_i)}{\Gamma(2-\alpha)}[|f_{m-1}(\mathbf{x}_i) - c| + \epsilon]^{1-\alpha}, \tag{40}$$

where $p_{k,m-1}(\mathbf{x}_i) = \frac{e^{f_{k,m-1}(\mathbf{x}_i)}}{\sum_{l=1}^{K} e^{f_{l,m-1}(\mathbf{x}_i)}}$ for the multinomial case.

4   Experiments

Several examples are provided to illustrate the improved efficiency of the algorithms. As has been noted, fractional gradient boosting decision tree can handle both regression and classification problems, there are comparisons between the proposed and original algorithms of regression and classification datasets respectively. Since the proposed algorithm is designed to accelerate the process of training, we primarily focus on the iterations and time costs with the same tolerance. We will describe our experimental setup, datasets and the results of simulations, and analysis on the algorithms.

4.1   Regression

A major challenge in training models is how long is enough to obtain the model with desired accuracy. Early stopping is a simple, effective, and widely used approach to balance between underfitting and overfitting. That is when the improvement of the score in the last given-number iterations is less than the given tolerance $tol = 0.01$, the training procedure pauses.

For regression problem, there are three publicly available datasets adopted: boston housing price, diabetes and california housing price which is more complex than the first dataset. The basic information is summarized in Table 1.

Table 1. Datasets used in regression.

Name	Samples total	Dimensionality
Boston housing price	506	13
Diabetes	442	10
California housing price	20640	8

With the given tolerance, the averages of iterations, cost time and R^2 scores which are defined in (41) of 40 times training are listed in Table 3 with fixed initial point and Table 4 with changing initial point.

$$R^2(y, f(\mathbf{x})) = 1 - \frac{\sum_{i=1}^{N} [y_i - f(\mathbf{x}_i)]^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{41}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{N} [y_i - f(\mathbf{x}_i)]^2$.

Looking at Table 3, it is obvious that with narrow difference of 0.01, fractional gradient boosting decision tree can bring nearly 2× speed-up by selecting a proper fractional order α. Nevertheless, what stands out in the table is that the convergence rate will decrease with α increasing. If α is greater than 1, the convergence rate will be less than the original algorithm. Figure 1 shows the iterations-R^2 score curves of diabetes datasets using the algorithm with the same
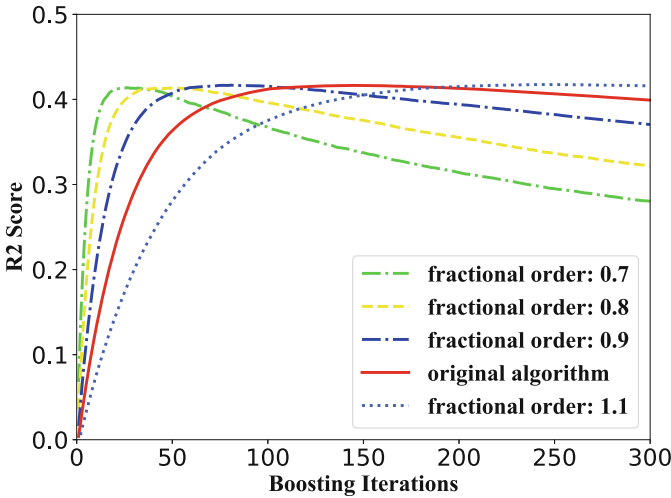
Fig. 1. Iterations-R^2 Score curve of diabetes.

initial point. The trend of the set of curves reveals the same result. Hence the range of α should be limited to $(0,1)$.

As to the algorithm with changing initial point, the results displayed in Table 4 indicate that it can also improve the convergence speed significantly. Furthermore, the range of α can extend to $(0,2)$. Figure 2 shows the iterations-R^2 score curves of california housing price with changing initial point.

It can be seen from these two tables that selecting appropriate parameters is an essential step of fractional gradient boosting decision tree. Otherwise, we may have a lengthy training process.

4.2   Classification

For classification problem, there are also three datasets adopted: iris, digits and a portion of forest cover type.

Table 2. Datasets used in classification.

Name	Samples total	Dimensionality
Iris	150	4
Digits	1797	64
Forest cover type	10000	54
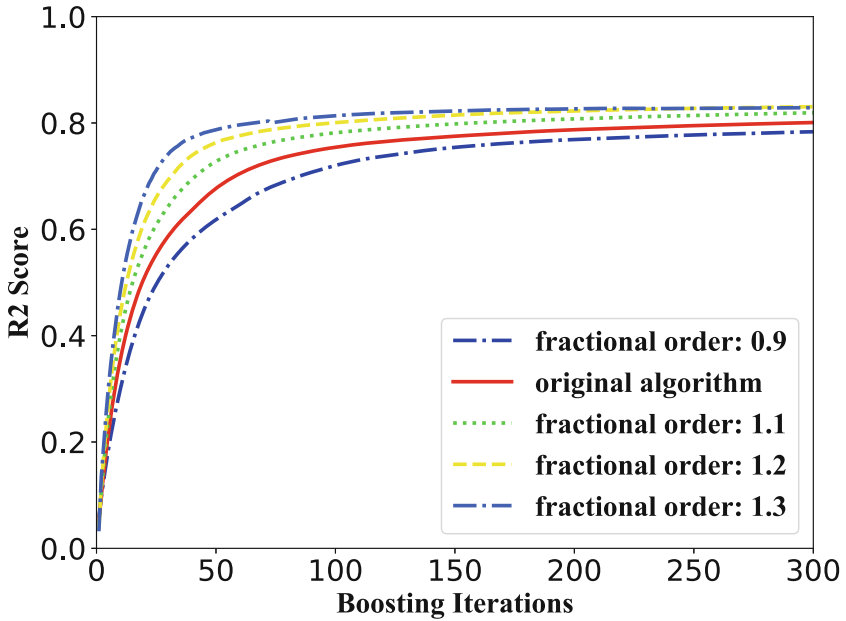
Fig. 2. Iterations-R^2 Score curve of california housing price.

Similarly, the averages of 40 times training are displayed in Table 6 with fixed initial point and Table 5 with changing initial point.

In Table 6, the results of iris and digits does not show evident improvement of fractional gradient compared to traditional algorithm. However, if we relax the restriction on accuracy score which is defined as accuracy$(y, f(\mathbf{x})) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}(f(\mathbf{x}_i) = y_i)$, there is also a nearly 2× speed-up of forest cover type.

Table 5 details the results of fractional gradient boosting decision tree with changing initial point. But the data is not that satisfactory. The algorithm does not achieve notable but still encouraging improvement which is highlighted in Table 5.

Though the results of classification are not exactly as expected, it still indicates that fractional gradient boosting algorithm can obtain models with the same accuracy score within approximate or even much less iterations compared to classical algorithm.

Table 3. Comparison on iterations, time cost, and test scores with different α and fixed initial point.

α	Boston housing price			Diabetes			California housing price		
	Iteration number	Cost time	R^2 score	Iteration number	Cost time	R^2 score	Iteration number	Cost time	R^2 score
0.5	**29.6000**	**0.0327**	**0.8424**	8.7000	0.0096	0.2524	54.3000	0.8056	0.7486
0.6	39.3500	0.0411	0.8515	10.9500	0.0111	0.3559	56.0000	0.8099	0.7441
0.7	48.6750	0.0506	0.8553	14.3750	0.0150	0.3917	**59.1250**	**0.7977**	**0.7411**
0.8	62.3000	0.0624	0.8535	**21.0250**	**0.0205**	**0.4038**	62.8000	0.8315	0.7373
0.9	88.5000	0.0864	0.8546	33.2250	0.0297	0.4045	67.6000	0.9014	0.7330
Origin	**117.3750**	**0.1128**	**0.8534**	**55.2000**	**0.0474**	**0.4031**	**72.7250**	**0.9146**	**0.7266**
1.1	165.0000	0.1598	0.8518	94.0250	0.0809	0.4032	79.4250	1.0536	0.7200
1.2	218.7750	0.2198	0.8485	154.3250	0.1420	0.4030	86.8000	1.2118	0.7093
1.3	310.5250	0.3390	0.8476	267.9000	0.2512	0.4021	96.6750	1.4491	0.6982
1.4	453.4500	0.5147	0.8445	488.0000	0.4961	0.4011	106.3000	1.8005	0.6795
1.5	667.6000	0.7105	0.8405	884.1750	0.8544	0.4003	113.1250	1.7205	0.6459

Table 4. Comparison on iterations, time cost, and test scores with different α and changing initial point

α	Boston housing price			Diabetes			California housing price		
	Iteration number	Cost time	R^2 score	Iteration number	Cost time	R^2 score	Iteration number	Cost time	R^2 score
0.5	267.3000	0.3323	0.7560	13.9250	0.0178	0.3592	72.7000	1.2214	0.2633
0.6	206.5000	0.2548	0.7833	24.2000	0.0303	0.3835	90.4750	1.4538	0.4106
0.7	147.9750	0.1680	0.8095	**30.8250**	**0.0318**	**0.3950**	100.4250	1.5616	0.5445
0.8	113.6000	0.1159	0.8250	36.8250	0.0354	0.4002	86.6750	1.2353	0.6296
0.9	105.5750	0.1182	0.8407	45.5000	0.0465	0.4037	80.4000	1.2634	0.6931
Origin	117.3750	0.1564	**0.8534**	**55.2000**	**0.0606**	**0.4031**	**72.7250**	**1.2455**	**0.7266**
1.1	99.6750	0.1141	0.8525	54.6250	0.0566	0.4044	61.2000	0.9091	0.7477
1.2	83.8250	0.1031	0.8508	60.0250	0.0670	0.3971	53.4750	0.8782	0.7674
1.3	**70.4000**	**0.0894**	**0.8510**	62.6750	0.0714	0.3966	44.3000	0.7607	0.7764
1.4	61.2250	0.0669	0.8273	60.3750	0.0589	0.4003	38.2750	0.5752	0.7739
1.5	46.4500	0.0522	0.7808	56.7000	0.0552	0.3504	**32.1250**	**0.4865**	**0.7601**

Table 5. Comparison on iterations, time cost, and test scores with different α and changing initial point

α	Iris			Digits			Forest cover type		
	Iteration number	Cost time	Accuracy scores	Iteration number	Cost time	Accuracy scores	Iteration number	Cost time	Accuracy scores
0.5	73.8750	0.3245	0.9483	202.9750	10.0915	0.9537	491.4500	57.6011	0.7906
0.6	71.4000	0.3246	0.9467	201.1000	10.4114	0.9543	468.2000	56.5454	0.7931
0.7	70.1000	0.3467	0.9458	190.2750	10.7328	0.9544	476.4000	61.7794	0.7999
0.8	67.1250	0.3240	0.9450	181.2000	10.2598	0.9543	435.9000	57.2084	0.8035
0.9	64.6250	0.3153	0.9458	**178.5500**	**9.9963**	**0.9536**	**415.5000**	**55.2197**	**0.8066**
Origin	**63.9000**	**0.3168**	**0.9450**	181.1750	9.9996	0.9537	454.6500	60.7997	**0.8131**
1.1	58.2500	0.2838	0.9467	21.0500	1.2051	0.7922	28.1500	3.9193	0.6998
1.2	**54.4500**	**0.2676**	**0.9458**	7.8750	0.4599	0.4778	7.7750	1.1326	0.4931
1.3	46.7500	0.2287	0.9183	6.5000	0.3722	0.1456	7.5500	1.0984	0.3156
1.4	20.4750	0.0998	0.5900	15.4750	0.8775	0.2952	8.4500	1.2138	0.2433
1.5	27.8500	0.1407	0.6558	12.3250	0.6984	0.1955	8.9000	1.2747	0.1734

Table 6. Comparison on iterations, time cost, and test scores with different α and fixed initial point

α	Iris			Digits			Forest cover type		
	Iteration number	Cost time	Accuracy scores	Iteration number	Cost time	Accuracy scores	Iteration number	Cost time	Accuracy scores
0.5	63.1250	0.3427	0.9450	133.7750	8.8068	0.9285	78.5000	12.1962	0.7355
0.6	63.3000	0.3023	0.9450	147.4500	8.7175	0.9410	104.1500	14.1998	0.7491
0.7	63.3750	0.3214	0.9450	156.8500	9.5226	0.9456	124.7750	18.2460	0.7622
0.8	62.6500	0.2983	0.9450	**166.4750**	**9.2084**	**0.9499**	192.2500	25.4559	0.7819
0.9	63.6750	0.2989	0.9433	181.8250	10.0452	0.9527	**252.4750**	**32.7774**	**0.7930**
Origin	**63.9000**	**0.2888**	**0.9450**	**181.1750**	**9.3711**	**0.9537**	**454.6500**	**57.9821**	**0.8131**
1.1	65.4000	0.3240	0.9442	244.8250	12.6987	0.8870	307.5000	40.6255	0.7933
1.2	65.1000	0.3180	0.9458	287.6500	14.2570	0.8136	96.9750	13.2535	0.6585
1.3	**63.9000**	**0.2944**	**0.9442**	51.5500	2.5292	0.4266	38.6250	5.6046	0.5369
1.4	59.7750	0.3068	0.9083	15.4000	0.9606	0.2727	25.9500	3.7102	0.4663
1.5	55.4000	0.2931	0.9000	10.1500	0.6157	0.2008	14.3500	2.0934	0.3289

5   Conclusions

In this paper, the fractional gradient boosting decision tree algorithm has been investigated. After introducing three proposed fractional gradient methods, this study applies higher order truncation to gradient boosting for the first time to develop fractional gradient boosting decision tree and then provides several examples to evaluate the effectiveness of the proposed algorithm. These experiments confirm that fractional gradient boosting decision tree can attain well-trained models with a significantly improved speed for regression compared with the traditional GBDT method. Furthermore precise fractional order ranges are obtained for regression and classification problem respectively through the results. In general, the paper demonstrate the promising applications of the fractional gradient descent method.

References

1. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
2. Burges, C.J.: From RankNet to LambdaRANK to LambdaMART: an overview. Learning **11**(23–581), 81 (2010)
3. 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)
4. Chen, Y., Gao, Q., Wei, Y., Wang, Y.: Study on fractional order gradient methods. Appl. Math. Comput. **314**, 310–321 (2017)
5. Chen, Y., Wei, Y., Liang, S., Wang, Y.: Indirect model reference adaptive control for a class of fractional order systems. Commun. Nonlinear Sci. Numer. Simul. **39**, 458–471 (2016)
6. Cheng, S., Wei, Y., Chen, Y., Li, Y., Wang, Y.: An innovative fractional order LMS based on variable initial value and gradient order. Signal Process. **133**, 260–269 (2017)
7. Cheng, S., Wei, Y., Chen, Y., Liang, S., Wang, Y.: A universal modified LMS algorithm with iteration order hybrid switching. ISA Trans. **67**, 67–75 (2017)

8. Cheng, S., Wei, Y., Sheng, D., Chen, Y., Wang, Y.: Identification for Hammerstein nonlinear ARMAX systems based on multi-innovation fractional order stochastic gradient. Signal Process. **142**, 1–10 (2018)
9. Cui, R., Wei, Y., Cheng, S., Wang, Y.: An innovative parameter estimation for fractional order systems with impulse noise. ISA Trans. **82**, 120–129 (2018)
10. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
11. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. **38**(4), 367–378 (2002)
12. Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, pp. 3146–3154 (2017)
13. Khan, Z.A., Chaudhary, N.I., Zubair, S.: Fractional stochastic gradient descent for recommender systems. Electron. Mark. **29**(2), 275–285 (2019)
14. Panda, B., Herbach, J.S., Basu, S., Bayardo, R.J.: PLANET: massively parallel learning of tree ensembles with MapReduce. In: Proceedings of the 35th International Conference on Very Large Data Bases, pp. 1426–1437 (2009)
15. Schapire, R.E.: A brief introduction to boosting. In: International Joint Conference on Artificial Intelligence, pp. 1401–1406 (1999)
16. Tan, Y., He, Z., Tian, B.: A novel generalization of modified LMS algorithm to fractional order. IEEE Signal Process. Lett. **22**(9), 1244–1248 (2015)
17. Tseng, C., Lee, S.: Designs of fractional derivative constrained 1-D and 2-D FIR filters in the complex domain. Signal Process. **95**, 111–125 (2014)
18. Wei, Y., Chen, Y., Cheng, S., Wang, Y.: A note on short memory principle of fractional calculus. Fractional Calculus Appl. Anal. **20**(6), 1382–1404 (2017)
19. Wei, Y., Kang, Y., Yin, W., Wang, Y.: Generalization of the gradient method with fractional order gradient direction. J. Franklin Inst. **357**, 2514–2532 (2020)
20. Wei, Y., Sun, Z., Hu, Y., Wang, Y.: On line parameter estimation based on gradient algorithm for fractional order systems. J. Control Dec. **2**(4), 219–232 (2015)
21. Yin, W., Wei, Y., Liu, T., Wang, Y.: A novel orthogonalized fractional order filtered-x normalized least mean squares algorithm for feedforward vibration rejection. Mech. Syst. Signal Process. **119**, 138–154 (2019)
22. Zhou, Z.H.: Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC, London (2012)

Apple Grading Model Based on Improved ResNet-50 Network

Lei Zhao[1], Qinjun Zhao[1(✉)], Tao shen[1,2], and Shuhui Bi[1]

[1] University of Jinan, Jinan 250000, China
cse_zhaoqj@ujn.edu.cn
[2] HIT Robot Group Co., Ltd, Harbin 150040, China

Abstract. In this paper, we study an apple grading model based on the convolutional neural network to classify Red Fuji apples according to features of size, color and external defects. Firstly, Red Fuji apple images are collected by professional equipment, and the RGB model of apple image is extracted and transformed into HSI model. Secondly, the segmentation between apple and background is realized by Otsu method in the S channel. Thirdly, the ResNet-50 network is improved by convolutional block attention module and LeakyReLU activation function. Finally, improved ResNet-50 network is applied to apple grading and compared with other mainstream convolutional neural networks. The experimental result shows that improved ResNet-50 network reaches the highest accuracy 95.1% in apple grading experiment, which is higher than AlexNet, VGG-16, GoogleNet, Mobilenet-V2 and the ResNet-50 network.

Keywords: Apple grading · ResNet network · Attention mechanism · LeakyReLU activate function

1 Introduction

Apple grading is an important part of the apple industry. In the apple growth, picking, transportation will be more or less rot, insect pests, crushing and other damage to the quality of apple. Apple shape, diameter and color will affect the sales, and then affect the profit, so apple grading is particularly important. The early apple grading method was manual sorting, which not only consumed a lot of manpower, but also slowly and inefficiently. Therefore, fast and accurate grading of apple is of great significance to the development of apple industry.

At present, neural network, machine learning and machine vision method has been widely used in fruit quality detection and classification due to its advantages in image processing [1]. Yuhui Ji [2] et al. applied the SVM model to the grading of apple based on features of defect and color, and the accuracy reached 91%. Payman Moallem [3] et al. detected Calyx region by K-means clustering method and applied the SVM model to apple grading, and the best accuracy reached 92.5%. Maoyong Nie [4] et al. applied Canny edge detection and SVM model based on particle swarm optimization, and the

accuracy reached 92%. Zheng Xu [5] et al. applied GA-SVM model to realize the grading of apple, and the accuracy reached 92.3%. Jinquan Li [6] et al. designed a shallow neural network through the Caffe framework to classify apple, and the accuracy reached 92%. To sum up, automatic classification has been widely used in apple grading, but there is still room for improvement in the accuracy of apple automation. In order to improve the accuracy of apple grading, this paper proposes improved ResNet-50 network to classify Red Fuji apples, and the experiment result proves that this network can effectively improve apple grading accuracy.

2   Overall Design of the Apple Grading Method

In this paper, the apple grading method mainly includes data acquisition, data expansion, image segmentation, improving ResNet-50 network, data set setting and setting comparison experiment to verify the accuracy of improved ResNet-50 network. Improved ResNet-50 network joins CBAM and LeakyReLU activation function. CBAM is Convolutional Block Attention Module [7, 8], which makes it easier for networks to capture significant information by focusing more on channel and spatial features. LeakyReLU activation function is used to solve the problem of neurons failing to learn to update (sudden death of neurons problem) [9]. In this paper, improved ResNet-50 network model is built on Pytorch framework, and its accuracy is tested through the Red Fuji apples data set. The apple grading system design is shown in Fig. 1. Apple grading system design.
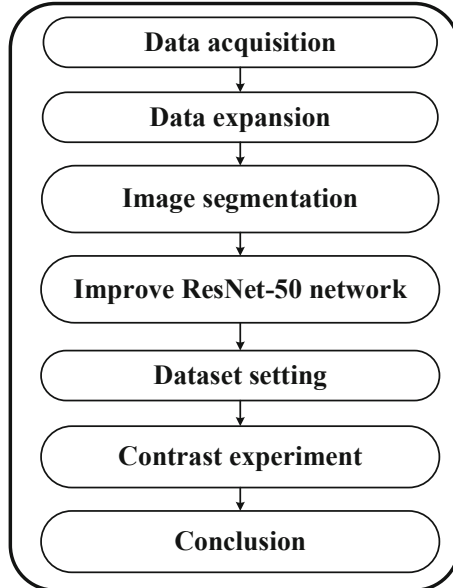
Fig. 1. Apple grading system design

2.1 Data Acquisition

The experimental material is Red Fuji apples, and apple images are collected by the apple image acquisition system, as shown in Fig. 2. Each apple is measured three times, and collected from the top, left side and right side respectively. A total of 1030 images of Red Fuji apple are collected, as shown in Fig. 3.

Fig. 2. Apple image acquisition system

Fig. 3. Red fuji apple image (left 1, 2, 3 sides; 4 top)

2.2 Data Expansion

Convolutional neural network requires a large amount of data for training and validating, so this paper applies the following methods to expand the data of apple images: (1) Vertical mirror flip, (2) Horizontal mirror flip, (3) Shrink or enlarge an image to a certain scale, (4) Using random rotating on image between $-60°-60°$, (5) Using random cut on image.

After the above picture expansion method, the data set is expanded to 6800 images. The data set after expansion is shown in Fig. 4.
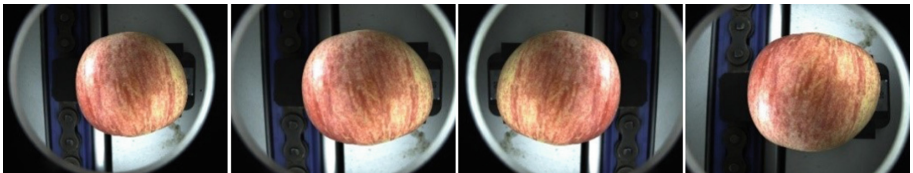
Fig. 4. Expanded red fuji apple data set (From left to right:1. Original picture 2. Expand at random scale 3. Horizontal mirror flipping after random scaling 4. Vertical mirror flipping after random scaling)

2.3 Image Segmentation

After data expansion, Red Fuji apple images of the data set is segmented to extract apple information. In this paper, the Otsu method is used for image segmentation, which is an algorithm to obtain the global threshold of the image proposed by the Japanese scholar Otsu [10, 11]. The algorithm principle is as follows:

Assuming that the image contains L gray levels, the threshold T divides the pixels of the image into two types B_1 (less than T) and B_2 (greater than T). The mean values of the two types of pixels are m_1 and m_2 respectively, and the mean values of the total pixels is m. The probability of pixel point to B_1 and B_2 are p_1 and p_2 respectively. When σ^2 reaches its maximum value, the gray level k value is the threshold T. The calculation formula of OTSU method is as follows:

$$p_1 = \sum_{i=0}^{k} p_i \tag{1}$$

$$p_2 = \sum_{i=k+1}^{L} p_i \tag{2}$$

$$p_1 m_1 + p_2 m_2 = m \tag{3}$$

$$p_1 + p_2 = 1 \tag{4}$$

$$\sigma^2 = p_1(m_1 - m)^2 + p_2(m_2 - m)^2. \tag{5}$$

The specific steps of apple image segmentation are as follows:

RGB model is extracted and transformed into HSI model by the following conversion formula.

$$H = \begin{cases} \theta & G \geq B \\ 360° - \theta & G < B \end{cases} \tag{6}$$

$$S = 1 - \frac{3min(R, G, B)}{R + G + B} \tag{7}$$

$$I = \frac{R + G + B}{3} \tag{8}$$

where

$$\theta = arccos \left\{ \frac{0.5[(R - G) + (R - B)]}{[(R - G)^2 + (R - B)(G - B)]^{\frac{1}{2}}} \right\}. \tag{9}$$

The HSI channel gray scale map and histogram information map are shown as Fig. 5.

After the experiment and comparison of histogram information in channels H, S and I, the difference between apple and background information in channel S is more obvious, so S channel is chosen for image segmentation. Firstly, 3×3 mean-filter is selected to process the S channel gray image, and then the segmentation between apple
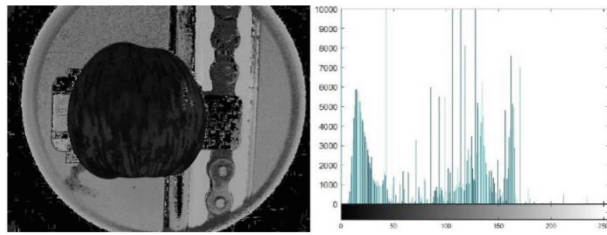
(a) H channel image (left) H channel histogram information (right)
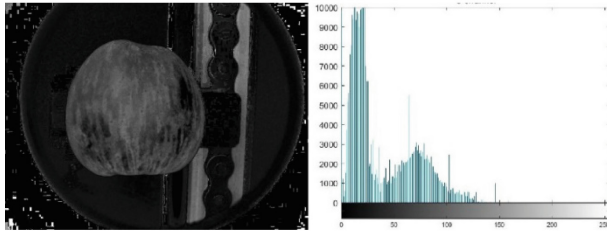
(b) S channel image (left) S channel histogram information (right)
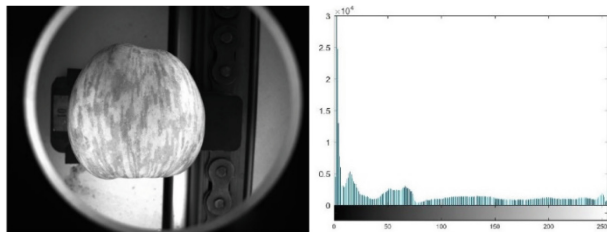
(c) I channel image (left) I channel histogram information (right)

Fig. 5. HSI channel grayscale image and histogram information

and background is realized by Otsu method. After this step, a small part of images have incomplete segmentation, so using the hole filling, expansion and corrosion operation to remove noise points, and making the white area as close as the shape of the original apple. Finally, the image is completely segmented, the middle area is filled with RGB color to obtain the final result. The image segmentation process is shown in Fig. 6.
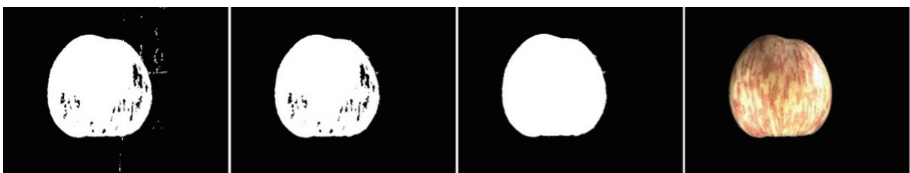
Fig. 6. Image segmentation process (From left to right:1. Image after OTSU method 2,3. Morphologic process 4. Image filling with RGB channel)

2.4   Improving ResNet-50 Network

Deep convolutional neural network has the problem of degradation, that is, with the increase of network depth, network accuracy becomes saturated or even decreases. ResNet Network is Residual Network, which adopts unique jump connection technology, that greatly alleviates the degradation problem of deep neural Network [12].

Convolutional Block Attention Module
Convolutional Block Attention Module (CBAM) is an attentional mechanism module, which is divided into Channel Attention Module and Spatial Attention Module. The structure of CBAM is shown in Fig. 7.
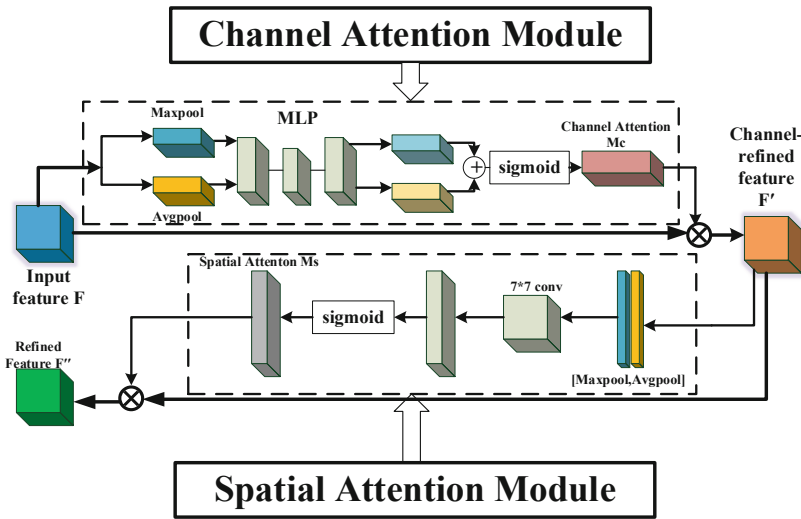
Fig. 7. CBAM structure

In the Channel Attention Module, the input feature is F, , which goes through global Max-pooling and global Avg-pooling based on image size respectively, and then enter MLP (Muti-Layer Perception). The MLP output two features are added based on elements. The combinative feature graph Mc is generated with sigmoid activation function. Mc and input feature F are multiplied based on elements to generate feature F', which is the output feature of the Channel Attention Module. The Mc calculation formula is as follows: (σ is the sigmoid activation function)

$$Mc = \sigma\{MLP[MaxPool(F)] + MLP[AvgPool(F)]\} \tag{10}$$

$$F' = Mc \otimes F. \tag{11}$$

In the Spatial Attention Module, the input is F', which goes through channel-based global Max-pooling and global Avg-pooling respectively, and then goes through a channel-based merge operation and a convolution layer to reduce the dimension to one channel. The feature graph Ms is generated with sigmoid activation function. Ms and input feature F' are multiplied based on elements to generate feature F'', which is the output feature of the Spatial Attention Module. Ms calculation formula is as follows: ($f^{7\times7}$ is the convolution layer of 7×7; σ is the sigmoid activation function)

$$Ms = \sigma\left\{f^{7\times7}\left[MaxPool(F'); AvgPool(F')\right]\right\} \tag{12}$$

$$F'' = Ms \otimes F'. \tag{13}$$

ResNet-50 Network and CBAM

In this paper, CBAM is added behind the block activation function. The output of the activation function after the first convolutional network in the block is set as F, and CBAM is added sequentially. The output is denoted as F''. The input F'' into the next convolution layer of the block until the optimization of the whole model is completed. By adding CBAM, the accuracy of the network can be effectively improved. Improved ResNet-50 network is shown in Fig. 8.
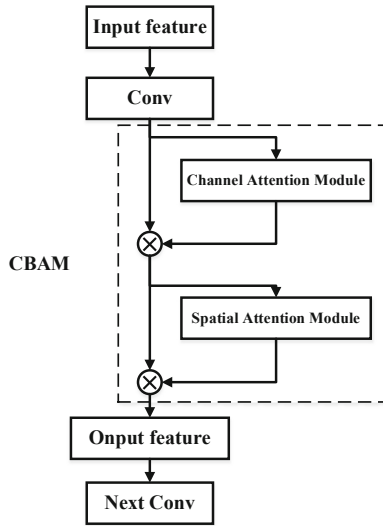
Fig. 8. Improved ResNet-50 network (Conv is convolution network)

Activation Function

In this paper, LeakyReLU activation function is used. Compared to the ReLU activation function, the LeakyReLU activation function has the following characteristics: When the input x is less than 0, there is still a very small gradient of the value $a * x$ output (a is a very small positive number, custom), thus avoiding the problem of neurons failing to update and learn. ReLU and LeakyReLU activation functions are shown in Fig. 9.
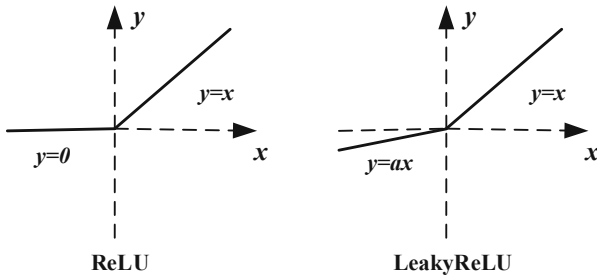
Fig. 9. Activation functions (left: ReLU; right LeakyRelu)

ReLU and LeakyReLU activation functions is calculated by the following formula: (a is a small positive number)

$$F(x) = \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases} \tag{14}$$

$$F(x) = \begin{cases} a * x & x \leq 0 \\ x & x > 0 \end{cases}. \tag{15}$$

3   Experiment and Conclusion

3.1   Data Set

In this paper, Red Fuji apple data set contains 3054 images. Firstly, according to our demand, we classify the overall data of apples into three parts: first-class, second-class and third-class. After being selected by several professionals according to the grading standard of Red Fuji apples as shown in Table 1, a total of 781 first-class apples, 1241 s-class apples and 1032 third-class apples are sorted out.

Then, the above data are randomly divided into training-set and verification-set in a ratio of 8:2. Training-set 2448, validation-set 606.The apple grading data set is shown in Table 2.

Table 1. Red fuji apple grading standard

Feature\Grade	First-class	Second-class	Third-class
Apple shape	Apple shape index ≥ 0.85	Apple shape index ≥ 0.80	Basic shape with this apple class
Apple color	Red colored area ≥ 75%	Red colored area ≥ 50%	Red colored area ≥ 25%
Diameter	≥ 80 mm	≥ 70 mm	≥ 65 mm
Diseases	No	No	No
Wound	No	No	No or slight
Chapped	No	No	No or slight
Sunburnt apple	No	No	No or slight

Table 2. Apple grading data set

	First-class	Second-class	Third-class	Total number
Training-set	625	997	826	2448
Validation-set	156	244	206	606
Total number	781	1241	1032	3054

3.2 Experiment Set

The apple grading experiment is carried out under the Ubuntu 18.04 system, and two GPU 2080Ti are used to accelerate the training of the program.

In order to validate the accuracy of improved ResNet-50 network in Red Fuji apples grading, five mainstream convolutional neural networks is selected for comparison, including, AlexNet, VGG-16, GoogleNet, Mobilenet-V2 and unimproved ResNet-50 network. We choose the highest accuracy rate and the average of the top five accuracy rate respectively as evaluation indexes to evaluate the accuracy of each network to apple

Table 3. Red fuji apples grading accuracy

Network\Aaccuracy	Highest-accuracy (%)	Average of the top five accuracy (%)
AlexNet	88.0	87.5
VGG-16	88.2	87.5
GoogleNet	91.8	90.7
Mobilenet-V2	87.2	86.8
ResNet-50	89.5	88.6
Improved ResNet-50	95.1	94.8

grading. Red Fuji apples grading accuracy is shown in Table 3. The accuracy comparison curve of each network is shown in Fig. 10.

The experimental result shows that the highest accuracy of improved ResNet-50 network is 95.1%, 3.3% higher than that of the second-rank GoogleNet network in Red Fuji apples grading, and the average of the top five accuracy of improved ResNet-50 network is 94.8%, 4.1% higher than that of the second-rank GoogleNet network. Improved ResNet-50 network achieves 95.1% accuracy in apple grading. To sum up, improved ResNet-50 network combined with combination of CBAM and LeakyReLU activation function has greater accuracy in apple grading.
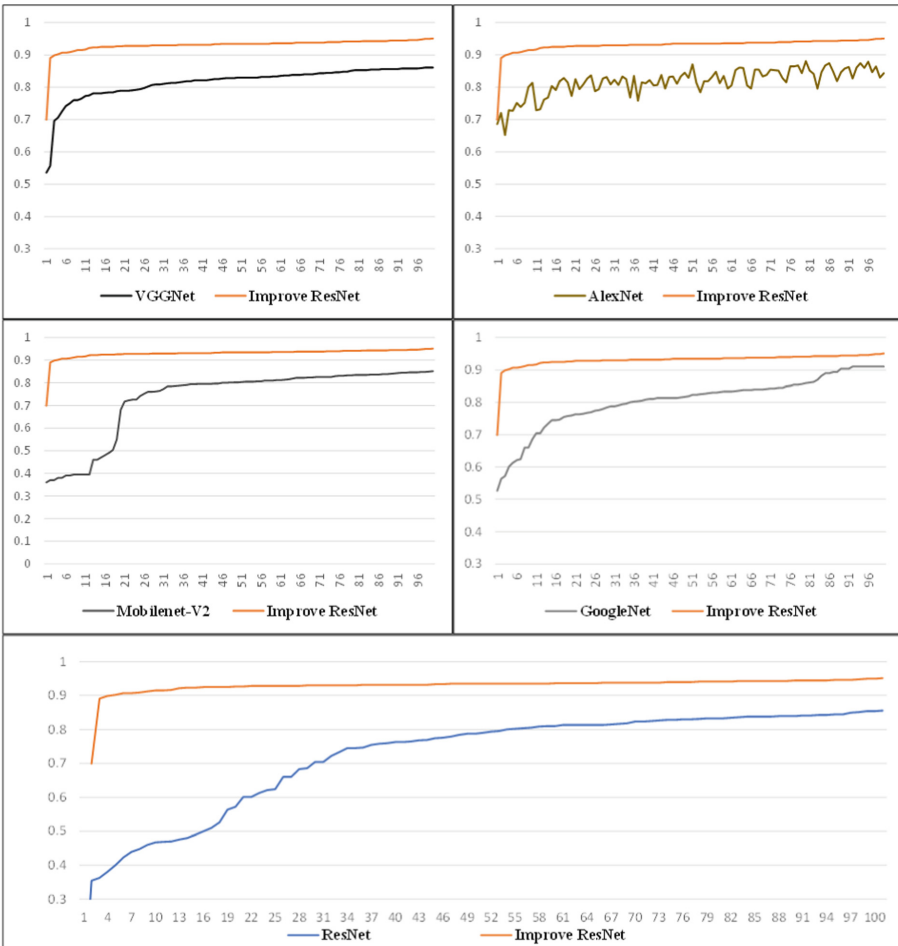
Fig. 10. Improved ResNet-50 network accuracy curve compared with VGGNet, AlexNet, Mobilenet-V2, GoogleNet, ResNet-50 (from left to right, from top to bottom).

3.3 Conclusion

This paper proposes an apple grading model based on improved ResNet-50 network by comparing the external features of the apple and draws the following conclusion: The addition of CBAM which contains a Channel Attention Module and a Spatial Attention Module, makes the network more aware of channel and spatial features and easier to extract key information. The addition of the LeakyReLU activation function alleviates the problem of neurons failing to update and learn. The above two improvements make ResNet-50 network more accurate for apple grading. In the apple grading experiment, the improved ResNet-50 network reaches high accuracy 95.1%, 5.6% improvement over the unimproved ResNet-50 network and 3.3% improvement over the second-rank network GoogleNet.

References

1. Roy, K., Chaudhuri, S.S., Pramanik, S.: Deep learning based real-time Industrial framework for rotten and fresh fruit detection using semantic segmentation. Microsyst. Technol. **27**(9), 3365–3375 (2021)
2. Ji, Y., Zhao, Q., Bi, S., et al.: Apple grading method based on features of color and defect. In: Proceedings of the 2018 37th Chinese Control Conference (CCC), pp. 5364–5368. IEEE (2018)
3. Moallem, P., Serajoddin, A., Pourghassem, H.: Computer vision-based apple grading for golden delicious apples based on surface features. Inf. Process. Aagric. **4**(1), 33–40 (2017)
4. Nie, M., Zhao, Q., Xu, Y., et al.: Machine vision-based apple external quality grading. In: Proceedings of the 2019 Chinese Control and Decision Conference (CCDC), pp. 5961–5966. IEEE (2019)
5. Xu, Z., Zhao, Q., Zhang, Y., Zhang, Y., Shen, T.: Apple grading method based on GA-SVM. In: Liu, Qi., Liu, X., Shen, T., Qiu, X. (eds.) CENet 2020. AISC, vol. 1274, pp. 79–89. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-8462-6_9
6. Li, J., Xie, S., Chen, Z., et al.: A shallow convolutional neural network for apple classification. IEEE Access **2020**(8), 111683–111692 (2020)
7. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
8. Ma, B., Wang, X., Zhang, H., Li, F., Dan, J.: CBAM-GAN: generative adversarial networks based on convolutional block attention module. In: Sun, X., Pan, Z., Bertino, E. (eds.) ICAIS 2019. LNCS, vol. 11632, pp. 227–236. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-24274-9_20
9. Dubey, A.K., Jain, V.: Comparative study of convolution neural network's relu and leaky-relu activation functions. In: Mishra, S., Sood, Y.R., Tomar, A. (eds.) Applications of Computing, Automation and Wireless Systems in Electrical Engineering. LNEE, vol. 553, pp. 873–880. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-6772-4_76

10. Shahabi, F., Poorahangaryan, F., Edalatpanah, S.A., et al.: A multilevel image thresholding approach based on crow search algorithm and otsu method. Int. J. Comput. Intell. Appl. **19**(02), 2050015 (2020)
11. Goh, T.Y., Basah, S.N., Yazid, H., et al.: Performance analysis of image thresholding: otsu technique. Measurement **2018**(114), 298–307 (2018)
12. McNeely-White, D., Beveridge, J.R., Draper, B.A.: Inception and ResNet features are (almost) equivalent. Cogn. Syst. Res. **2020**(59), 312–318 (2020)

Author Index